On Higher Order Elicitability and Some Limit Theorems on the Poisson and Wiener Space

Inauguraldissertation
der Philosophisch-naturwissenschaftlichen Fakultät
der Universität Bern

vorgelegt von

Tobias Fissler

von Deutschland

Leiterin der Arbeit:

Prof. Dr. Johanna Fasciati-Ziegel

Institut für mathematische Statistik und Versicherungslehre
der Universität Bern

On Higher Order Elicitability and Some Limit Theorems on the Poisson and Wiener Space

Inauguraldissertation
der Philosophisch-naturwissenschaftlichen Fakultät
der Universität Bern

vorgelegt von

Tobias Fissler

von Deutschland

Leiterin der Arbeit:

Prof. Dr. Johanna Fasciati-Ziegel

Institut für mathematische Statistik und Versicherungslehre
der Universität Bern

Von der Philosophisch-naturwissenschaftlichen Fakultät angenommen.

Bern, 10. April 2017

Der Dekan:

Prof. Dr. Gilberto Colangelo

Preface

This doctoral thesis comprises the scientific outcome of three and a half years of research as a doctoral student at the Institute of Mathematical Statistics and Actuarial Science of the University of Bern from September 2013 until March 2017. During that time, I had the opportunity to jointly work with my PhD supervisor Johanna F. Ziegel on the SNF-project 152609, having the title 'Understanding elicitability'. The joint research led to many new and fascinating insights into the topic, ultimately resulting in two peer-reviewed articles – one of them published in the *Annals of Statistics* – and further additional material. The first part of this thesis is devoted to these results, comprising the joint articles and the additional findings, with a main focus on *higher order elicitability*, which explains also the first part of the title of this thesis.

At the same time, I had the opportunity to do some joint research with Christoph Thäle from the Ruhr-University Bochum. We studied qualitative and quantitative limit theorems for Gaussian and Poisson functionals, using techniques from Malliavin Calculus in combination with Stein's method. With its rather probabilistic perspective, this work was not only a pleasant complement to the research on elicitability, but also built a connection to some topics of my Bachelor thesis at Heidelberg University, which was jointly supervised by Johanna F. Ziegel and Mark Podolskij. The fruitful outcome of this collaboration sublimated in two joint research articles, one of them is already published in *ALEA*, the other one is currently under review. They are the main content Part II of this thesis.

The two parts of this thesis are independent in their content, connected only by my own biographical ties, the fact that they both belong to the broad field of stochastics, and definitely by the fact that they are hosted in the same thesis. Since this thesis is cumulative in nature, with major parts consisting of original research papers – be it in their journal style, be it in their preprint version –, each of the articles has its own bibliography. All references appearing in the rest of the text can be found in a joint bibliography at the end of the thesis.

Acknowledgement

I am deeply grateful for receiving a lot of help, support, and advice while the various parts of this thesis came into being. First of all, I would like to express my great thankfulness for having the opportunity to write my PhD thesis with and under the supervision of Johanna F. Ziegel. It was always a big pleasure to collaborate with her, discuss about mathematics, or plan administrative things.

Whenever I encountered problems – be it related to mathematics, be it concerned to life as a PhD student in general with its ups and downs, particular merits, but also challenges – she could help me with her extremely helpful advice. Thank you for the time, encouragement, motivation, understanding, and help during the last three and a half years!

Besides Johanna, I would also like to warmly thank my other coauthors, Christoph Thäle, Tilmann Gneiting, and Mark Podolskij. It was a great joy and an inspiring experience to collaborate with such educated and dedicated scholars, but also friendly and nice persons. I could learn a lot in our various different projects. As a graduate student at Heidelberg University, I had the great experience of having many lectures with Mark on probability and statistics. As a motivating teacher, he could stimulate my curiosity for this field. The master thesis under his supervision has ultimately led to a joint paper, which will appear in *Bernoulli*.

Moreover, I enjoyed insightful discussions with a lot of colleagues from academia. Taking the substantial risk of giving an incomplete list, I want to say thank you to Carlo Acerbi, Fabio Bellini (also for the time to co-referee this thesis), Timo Dimitriadis, Paul Embrechts, Rafael Frongillo, Claudio Heinrich, Edgars Jakobsons, Marie Kratz, Giovanni Peccati, Matthias Schulte, and Ingo Steinwart.

During the recent years, I had the pleasure of being a member of the Institute of Mathematical Statistics and Actuarial Science (IMSV) at the University of Bern. With its friendly, supportive, and familiar atmosphere, it was a perfect environment for my doctoral studies. My thanks are due to all current and past members of the institute, especially the heads of the institute, Lutz Dümbgen and Ilya Molchanov. Moreover, I would like to thank all my current and past officemates, Lukas Martig, Anja Mühlemann, Eleo van der Lans, Gabriel Fischer, Kathrin Burkart, Janine Kuratli, and Mara Trübner, for good discussions and an enjoyable time. Moreover, I would like to gratefully acknowledge the financial support of the IMSV and the SNF.

Of course, my thanks go also to all my friends, both for supporting me and for distracting me from my mathematical work – and just for having a great time together! It is virtually impossible to mention all of them – however, I would like to mention my particular gratefulness for the linguistic comments on some parts of this thesis by my life-long friend Nils Hirsch (the remaining errors being exclusively in my own responsibility).

Sharing someone's life with a mathematician has possibly its merits, but definitely provides challenges, requiring a considerable amount of understanding – especially when the mathematician is about to finish his doctoral thesis! I am enormously thankful to my family and to my partner Isabel for taking these challenges, for their constant support, helpful understanding, and indispensable love!

Tobias Fissler Bern, March 2017

Contents

I. Higher Order Elicitability **7**

1. Introduction to Elicitability **9**
1.1. Forecasts over the times and an ancient approach towards forecast evaluation . . . 9
1.2. Forecast evaluation in modern times . . . 10
1.2.1. The ingredients of forecast comparison . . . 11
1.2.2. Consistency and elicitability . . . 13
1.3. The elicitation problem . . . 14
1.4. The relevance of elicitability in quantitative risk management . . . 16
1.5. Contributions of Part I of this thesis . . . 16

2. Some Basic Results and Preliminaries on Higher Order Elicitability **19**
2.1. Notation and definitions . . . 19
2.2. The prediction space setting – bringing elicitability to life . . . 21
2.2.1. Learning . . . 23
2.2.2. Regression . . . 23
2.2.3. The prediction space . . . 25
2.3. Transferring results from Gneiting (2011) to the higher-dimensional case . . . 28

3. Higher Order Elicitability and Osband's Principle **35**
3.1. Fissler and Ziegel (2016) . . . 36
3.2. Osband's Principle for identification functions . . . 67
3.3. Eliciting the divergence of a strictly consistent scoring function . . . 68

4. Scoring Functions Beyond Strict Consistency **71**
4.1. Order-Sensitivity . . . 73
4.1.1. Different notions of order-sensitivity . . . 73
4.1.2. The one-dimensional case . . . 76
4.1.3. Unique local minimum . . . 79
4.1.4. Self-calibration . . . 80
4.1.5. Componentwise order-sensitivity . . . 82
4.1.6. Metrical order-sensitivity . . . 85
4.1.7. Order-sensitivity on line segments . . . 98
4.1.8. Nested information sets . . . 104

4.2. Convexity of scoring functions 104
4.2.1. Motivation 107
4.2.2. Determining convex scoring functions for popular functionals 114
4.3. Equivariant functionals and order-preserving scoring functions . . . 127
4.3.1. Translation invariance 129
4.3.2. Homogeneity 134
4.4. Possible applications 135

5. Implications for Backtesting **137**

6. Discussion **147**
6.1. Reception in the literature 147
6.2. Outlook for possible future research projects 147
6.2.1. Vector-valued risk measures in the context of systemic risk 147
6.2.2. Testing the tail of the P&L-distribution 148
6.2.3. A sufficient condition for higher order elicitability 149

II. Limit Theorems on the Poisson and Wiener Space **151**

7. Introduction **153**
7.1. Chaos representation of Gaussian and Poisson functionals 153
7.2. Stein's method 155

8. A four moments theorem for Gamma limits on a Poisson chaos **159**

9. A new quantitative central limit theorem on the Wiener space with applications to Gaussian processes **193**

10. Discussion **219**

Part I.

Higher Order Elicitability

1. Introduction to Elicitability

From the cradle to the grave, human life is full of decisions. Due to the inherent nature of time, decisions have to be made today, but at the same time, they are supposed to account for unknown and uncertain future events. However, since these future events cannot be *known* today, the best thing to do is to base the decisions on *predictions* for these unknown and uncertain events. The call for and the usage of predictions for future events is literally ubiquitous and even dates back to ancient times.

1.1. Forecasts over the times and an ancient approach towards forecast evaluation

For ancient Greeks, the Delphic Oracle was one of the most respected authorities in the ancient Hellenistic world. It was consulted for decisions of private life such as marriage or business, but also for strategic political decisions concerning peace and war. In these days, dreams, divination, and revelation were considered as respected sources for forecasts, as was the course of nature resulting for example in the pre-science of astrology. During these days and at least till the 19^{th} century, the risk of famine was hanging over agricultural societies like the Sword of Damocles, giving evidence for the need of good means of weather forecasts. This found its expression mainly in terms of weather proverbs, incorporating the empirical knowledge and experience of ancestral generations. With the development of natural sciences, mathematics, and in particular statistics and probability theory – i.a. triggered by the need for guidance in gambling and betting –, the art of forecasting turned more and more into a discipline of science. Fields, dedicated to prediction making in specialized areas of nature, such as meteorology, or society, e.g. mathematical finance, or even futurology, evolved. Today, elaborated forecasts are present in a variety of different disciplines: Politics and government (e.g. forecasts for elections and votings), business (e.g. forecasts for demands), finance (e.g. forecasts of inflation, interest rates, exchange rates), defense and intelligence, the health sector, the energy market, agriculture, and everyday life.

At all times, due to previous experience or to pure logical reasoning in the presence of contradicting forecasts, there were people who were aware of the fact that predictions – whether made by divine oracles or by sophisticated scientific models – were not infallible. Assessing the quality of a forecast, they asked the two main questions:

(i) "How good is the forecast at hand in *absolute* terms?"

(ii) "How good is the forecast at hand in *relative* terms?"

Question (i) deals with *forecast validation*, whereas question (ii) is concerned with *forecast selection*, *forecast comparison*, or *forecast ranking*. A historic example for assessing the second question is the ancient king Croesus, king of Lydia. In the 6$^{\text{th}}$ century B.C., he sent messengers to the most famous oracles in the Hellenistic world. As Herodotus wrote:

> "His intent in sending was to test the knowledge of the oracles, so that, if they should be found to know the truth, he might send again and ask if he should take in hand an expedition against the Persians." (Herodotus and Godley, 1920, Book I. 46–48, pp. 53–55)

Herodotus also detailed on the way Croesus performed this ancient forecast comparison.

> "And when he sent to make trial of these shrines he gave the Lydians this charge: they were to keep count of the time from the day of their leaving Sardis, and on the hundredth day inquire of the oracles what Croesus, king of Lydia, son of Alyattes, was then doing; then they were to write down whatever were the oracular answers and bring them back to him. [...] Croesus then unfolded and surveyed all the writings." (*ibidem*)

Ultimately, the oracle at Delphi was the only one to report correctly what king Croesus was doing at that very day (namely, he was cooking a lamb-and-tortoise stew in a caldron of bronze), thus convincing him of her dominant predictive ability.

1.2. Forecast evaluation in modern times

Even though king Croesus' ultimate problem was not an incorrect forecast ranking, but rather a misinterpretation of a future oracular utterance,[1] his assessment of the problem of forecast ranking would definitely not satisfy the present state of the art, one of the major drawbacks being the fact that he based his judgement on one observation only. One of the suggestions of a 21$^{\text{st}}$–century–expert would certainly consist of, but would not be limited to, the proposal of using the full 100 days and repeating the very same experiment 100 times consecutively, then deeming that oracle to be the best which performed best *on average* over the 100 days.

Part I of this doctoral thesis is devoted to collecting and developing more advanced advice to king Croesus and his contemporary colleagues. Aware of the two

[1] When he was wondering whether to attack the Persians, he made an inquiry at the Delphic Oracle as to whether he should send an army against the Persians, and, if so, whether to take an ally. The oracle's answer was that "if he should send an army against the Persians he would destroy a great empire." (Herodotus and Godley, 1920, Book I. 53, p. 61) Misinterpreting this divination and having the Persian empire in mind to be destroyed by his attack, he sent his army against Persia, but ultimately destroyed his own 'great empire'.

crucial tasks of forecast evaluation – (i) forecast validation and (ii) forecast comparison – the thesis at hand is mostly concerned with the latter task, in particular concentrating on the issue of *elicitability*.

1.2.1. The ingredients of forecast comparison

We begin by collecting the ingredients of modern forecast comparison, some of them already present in Croesus' ancient oracle contest. To this end, we adopt a quite general decision-theoretic framework following Gneiting (2011); cf. (Savage, 1971; Osband, 1985; Lambert et al., 2008). For some number $n \geq 1$, one has

(a) observed *ex post* realizations $y_1, \dots, y_n$ of a time series $(Y_t)_{t\in\mathbb{N}}$, taking values in an *observation domain* O with a σ-algebra $\mathcal{O}$;

(b) a family $\mathcal{F}$ of probability distributions on $(\mathsf{O}, \mathcal{O})$, reflecting the potential distributions of Y_t;

(c) for $m \geq 1$ competing experts / forecasters, *ex ante* forecasts $x_1^{(i)}, \dots, x_n^{(i)}$, $i \in \{1, \dots, m\}$, taking values in an *action domain* A.

As outlined in Fissler and Ziegel (2016) (see Chapter 3), the observations y_t can be real-valued (GDP growth for one year, maximal temperature at one day), vector-valued (wind-speed, weight and height of persons), functional-valued (path of the exchange rate Euro–Swiss franc over one day), or also set-valued (area of rain on one day, area affected by a flood). Concerning the forecasts and the action domain, respectively, there is a dichotomy: On the one hand, predictions can take the form of *point forecasts*. In this situation, one is typically interested in a certain statistical property of the underlying distribution F_t of Y_t. Strictly speaking, this property can be expressed in terms of a *functional* $T\colon \mathcal{F} \to \mathsf{A}$ such as the mean, a certain quantile or a risk measure. In many cases, A coincides with O and is typically a subset of $\mathbb{R}$. But T can also be vector-valued (one is interested in the covariance matrix of a multivariate observation or in quantiles at different levels), but also set-valued (expectation of a random set, prediction region of a random variable at a certain level). On the other hand, forecasts can be *probabilistic*, taking into account the random nature of future events. In this case, the forecasts take the form of probability measures, probability distributions, or density functions. Gneiting (2011, p. 746) noted:

> "In many aspects of human activity, a major desire is to make forecasts for an uncertain future. Consequently, forecasts ought to be probabilistic in nature, taking the form of probability distributions over future quantities or events (Dawid, 1984; Gneiting, 2008). Still, many practical situations require single-valued point forecasts, for reasons of decision making, market mechanisms, reporting requirements, communication, or tradition, among others."

We remark that from a mathematical perspective, probabilistic forecasts can be seen as a particular instance of point forecasts where the functional T is infinite-dimensional and merely the identity map on $\mathcal{F}$.

Recalling that the observations $y_1, \dots, y_n$ are *ex post*, whereas the forecasts

$x_1^{(i)}, \dots, x_n^{(i)}$, $i \in \{1, \dots, m\}$, are *ex ante*, we emphasize that the forecasts do not need to be one step ahead forecasts, but can also be multistep ahead forecasts. Even more generally, the forecasts can be made with different time horizons (e.g. all forecasts are issued at the same time $t = 1$).

The last ingredient of the decision-theoretic framework is

(d) a loss function or *scoring function* $S\colon \mathsf{A} \times \mathsf{O} \to \mathbb{R}$. The scoring function is assumed to be negatively oriented, that is, if a forecaster reports the quantity $x \in \mathsf{A}$ and $y \in \mathsf{O}$ materializes, she is assigned the *penalty* $S(x, y) \in \mathbb{R}$.

Common examples for scoring rules are the absolute loss $S(x, y) = |x - y|$, the squared loss $S(x, y) = (x - y)^2$ (for $\mathsf{A} = \mathsf{O} = \mathbb{R}$), or the absolute percentage loss $S(x, y) = |(x - y)/y|$ (for $\mathsf{A} = \mathsf{O} = (0, \infty)$). In the literature, the function S has different names. In the case of probabilistic forecasts, it is commonly called *scoring rule* (Gneiting and Raftery, 2007; Gneiting, 2011). But from an abstract decision-theoretic point of view, the concepts coincide.

Having collected all the ingredients, modern forecast comparison is done in terms of the ranking of the *realized scores*

$$\bar{\mathbf{S}}_n^{(i)} = \frac{1}{n} \sum_{t=1}^{n} S(x_t^{(i)}, y_t), \qquad i \in \{1, \dots, m\}.$$

That is, a forecaster is deemed to be the better the lower her realized score is. We illustrate this practice with the following example.

Example 1.2.1. Suppose we have $m = 2$ forecaster, both issuing their predictions for tomorrow's temperature at noon for $n = 100$ subsequent days. After these 100 days, one evaluates the corresponding prediction-observation sequences $(x_t^{(i)}, y_t)_{t=1,\dots,100}$, $i = 1, 2$. Assume forecaster 1 is almost correct, but misspecified the correct values always by one degree. On the other hand, forecaster 2 is even better on the first 99 days and predicts exactly the correct temperatures.[2] However, on the the last day, he ultimately misspecified the real outcome and deviated by 50 degrees.

Who has performed better? The answer obviously depends on the choice of the scoring function: utilizing the absolute loss, forecaster 2 is clearly the better one. However, arguing that small errors are negligible, but larger misspecifications entail the risk of a severe damage to health, one could also argue in favour of using the squared loss assigning forecaster 1 a lower realized score.

Example 1.2.1 shows a *caveat*: the forecast ranking in terms of realized scores not only depends on the forecasts and the realizations (as it should definitely be the case), but also on the choice of the loss function. In order to avoid impure possibilities of manipulating the forecast ranking *ex post* with the data at hand,

[2] One can assume that the measurement and the forecasts for the temperature are always rounded to the closest integer.

it is necessary to specify a certain scoring function *before* the inspection of the data. *A fortiori*, for the sake of transparency and in order to encourage truthful forecasts, one ought to disclose the choice of the scoring function to the competing forecasters *ex ante*. But still, the optimal choice of the scoring function remains an open problem. One can think of two situations:

(i) A decision-maker might be aware of his actual economic costs of utilizing misspecified forecasts. In this case, the scoring function should reflect these economic costs.

(ii) The actual economic costs might be unclear and the scoring function might be just a tool for forecast ranking. However, the directive is given in terms of the functional $T\colon \mathcal{F} \to \mathsf{A}$ one is interested in.

1.2.2. Consistency and elicitability

For situation (i) described above, one should use the readily economically interpretable cost or scoring function. Therefore, the only concern is situation (ii). Assuming the forecasters are *homines oeconomici* and adopting the rationale of expected utility maximization, given a concrete scoring function S, the most sensible action consists in minimizing the *expected score* $\mathbb{E}_F[S(x,Y)]$ with respect to the forecast x, where Y follows the distribution F, thus issuing the *Bayes act* $\arg\min_{x\in\mathsf{A}} \mathbb{E}_F[S(x,Y)]$. Hence, a scoring function should be *incentive compatible* in that it encourages truthful and honest forecasts. In line with Murphy and Daan (1985) and Gneiting (2011), we call a scoring function *strictly $\mathcal{F}$-consistent* for the functional $T\colon \mathcal{F} \to \mathsf{A}$ if

$$T(F) = \arg\min_{x\in\mathsf{A}} \mathbb{E}_F[S(x,Y)]$$

for all distributions $F \in \mathcal{F}$. And following the terminology of Lambert et al. (2008) and Gneiting (2011), a functional $T\colon \mathcal{F} \to \mathsf{A}$ is called *elicitable* if it possesses a strictly $\mathcal{F}$-consistent scoring function. In the context of probabilistic forecasting, a strictly consistent scoring rule is called *strictly proper*. The necessity of utilizing strictly consistent scoring functions for meaningful forecast comparison is impressively demonstrated in terms of a simulation study in Gneiting (2011).

Further merits of elicitability, besides meaningful forecast comparison and ranking, are M-estimation (Huber, 1964; Huber and Ronchetti, 2009), and generalized regression such as quantile regression or expectile regression (Koenker, 2005; Newey and Powell, 1987); cf. Zwingmann and Holzmann (2016) for M-estimation as well as Bayer and Dimitriadis (2017) for a regression framework for the pair (Value at Risk, Expected Shortfall), respectively.

1.3. The elicitation problem

Having settled the basic definitions, one can formulate a threefold *elicitation problem* with respect to a fixed functional $T\colon \mathcal{F} \to \mathsf{A}$.

(i) Is T elicitable, i.e., is there a strictly $\mathcal{F}$-consistent scoring function for T?

(ii) What is the class of strictly $\mathcal{F}$-consistent scoring functions for T? What are handy sufficient and / or necessary conditions for a scoring function to be strictly $\mathcal{F}$-consistent for T?

(iii) Are there some particularly distinguished instances of strictly $\mathcal{F}$-consistent scoring functions?

Some comments are in order: Apparently, a certain hierarchy inheres in the three subquestions of the elicitation problem. One ought to try to answer the questions in the order presented above. But clearly, a question of inferior hierarchy may help to tackle a question of superior hierarchy: if, for example, one can find necessary conditions for the strict consistency of a scoring function which lead to a contradiction, this also gives a negative answer to question (i). A crucial fact which should be born in mind is that the questions of the elicitation problem are not only relative with respect to the functional, but also with respect to the domain $\mathcal{F}$ of the functional; see Fissler and Ziegel (2016, Lemma 2.5) and the discussion thereafter (Chapter 3 of this thesis). Finally, question (iii) is in its present form a vague statement which calls for further specification. This can be given in terms of a list of quality requirements for scoring functions beyond strict consistency. Some of them are certainly order-sensitivity, convexity, homogeneity or translation invariance of a scoring function.

Even though the denomination and the synopsis of the described problems under the term 'elicitation problem' are novel, there is a rich strand of literature in mathematical statistics and economics concerned with the threefold elicitation problem. Foremost, one should mention the pioneering work of Osband (1985), establishing a necessary condition for elicitability in terms of convex level sets of the functional, and a necessary representation of strictly consistent scoring functions, known as *Osband's principle* (Gneiting, 2011; Fissler and Ziegel, 2016). Whereas the necessity of convex level sets holds in a quite general framework, Lambert (2013) could specify sufficient conditions for elicitability for functionals taking values in a finite set, and Steinwart et al. (2014) showed sufficiency of convex level sets for real-valued functionals satisfying certain regularity conditions. Moments, ratios of moments, quantiles, and expectiles are in general elicitable, whereas other important functionals such as variance, Expected Shortfall or the mode functional are not (Savage, 1971; Gneiting, 2011; Heinrich, 2014; Osband, 1985; Weber, 2006). E.g., the squared loss is a strictly $\mathcal{F}$-consistent scoring function for the mean (with respect to the class $\mathcal{F}$ of distributions with finite second moments), and the absolute loss is the counterpart for the median (in this case with respect to the class $\mathcal{F}$ of distributions with unique medians and finite first moments).

Concerning subproblem (ii) of the elicitation problem, Savage (1971), Reichelstein and Osband (1984), Saerens (2000), and Banerjee et al. (2005) gave characterizations for strictly consistent scoring functions for the mean functional of a one-dimensional random variable in terms of *Bregman functions*. Strictly consistent scoring functions for quantiles have been characterized by Thomson (1979) and Saerens (2000); see Gneiting (2011) for a good review of the literature and a characterization of the class of strictly consistent scoring functions for expectiles. We refer to the recent paper Ehm et al. (2016) for an insightful complementary approach to the characterization of strictly consistent scoring functions for quantiles and expectiles in terms of Choquet type mixture representations. The case of vector-valued functionals apart from means of random vectors has been treated substantially less than the one-dimensional case (Osband, 1985; Banerjee et al., 2005; Lambert et al., 2008; Frongillo and Kash, 2015a,b).

Point (iii) of the elicitation problem in its different aspects was treated by several authors, again mainly concentrating on the case of real-valued functionals or even finite-valued functionals. The early work of Savage (1971) is concerned with special choices of strictly consistent scoring functions for the mean, in particular, resulting in translation invariance, homogeneity, or symmetry of the scoring function. The results of Patton (2011) point into a similar direction with a special emphasis on homogeneous scoring functions for mean forecasts. Finally, Nolde and Ziegel (2016) described the classes of strictly consistent and homogeneous scoring functions for the most common risk measures, such as Value at Risk (the quantile), expectiles, and the pair (Value at Risk, Expected Shortfall), being an example of a vector-valued functional. The issue of order-sensitivity of scoring functions has been treated mainly in the one-dimensional case (Lambert, 2013; Steinwart et al., 2014; Bellini and Bignozzi, 2015; Ehm et al., 2016), whereas Lambert et al. (2008) introduced a notion of order-sensitivity for the case of the action domain A being a subset of $\mathbb{R}^k$ as a kind of componentwise order-sensitivity. Friedman (1983) and Nau (1985) considered similar questions in the setting of probabilistic forecasts, coining the term of *effectiveness* of scoring rules which can be described as a kind of order-sensitivity in terms of a metric. Holzmann and Eulert (2014) proved a sort of order-sensitivity on a different level, considering *random* forecasts in the prediction space setting (Gneiting and Ranjan, 2013; Strähl and Ziegel, 2015), showing that ideal forecasts reward the more informed forecaster in the presence of nested information sets. Finally, Acerbi and Székely (2017) gave a new motivation for the usage of convex scoring functions (with convexity in the first argument of the scoring function), also describing the form of convex strictly consistent scoring functions for mean- and quantile-forecasts; cf. (Patton, 2011; Steinwart et al., 2014).

1.4. The relevance of elicitability in quantitative risk management

In recent years, there was a lively debate in academia, regulation, banks, and insurances about what risk measure to use in practice. One of the main requirements of a 'good' risk measure is comprised in a list of axioms given by Artzner et al. (1999), introducing the notion of *coherent risk measures*. Their axiomatic approach consists in requiring monotonicity, super-additivity, positive homogeneity, and translation invariance of a risk measure. Further consensually accepted requirements are robustness of risk measures and elicitability, the latter mainly for the purpose of backtesting historical data; see Emmer et al. (2015) for a good overview. The debate about the best choice of a risk measure has mainly concentrated on the dichotomy between Value at Risk (VaR) and Expected Shortfall (ES). In a nutshell, VaR is robust, and, as a quantile, it is elicitable (Cont et al., 2010; Emmer et al., 2015; Gneiting, 2011). However, VaR has been criticized for being 'blind' against losses beyond the the level α (Daníelsson et al., 2001) and for its lack of super-additivity (Acerbi, 2002). On the other hand, by definition, ES takes into account the losses beyond the level α, it is a coherent risk measure, but fails to be elicitable (Weber, 2006; Gneiting, 2011). Moreover, Daníelsson et al. (2001) pointed out that one needs more data to correctly estimate ES in comparison to VaR. During that debate, the role and relevance of elicitability was sometimes vague and often disputatious, however, recently has been clarified for being important for model selection, and not for model verification (Embrechts and Hofert, 2014; Acerbi and Székely, 2014; Emmer et al., 2015; Bellini and Bignozzi, 2015; Ziegel, 2016; Nolde and Ziegel, 2016; Davis, 2016; Acerbi and Székely, 2017).

1.5. Contributions of Part I of this thesis

The main contributions and achievements of Part I of this thesis affect all three aspects of the elicitation problem as well as a contribution to the discussion about the role of elicitability for backtesting in quantitative risk management. The main focus lies on a thorough study of higher order elicitability in the sense that the functionals under consideration are higher-dimensional, meaning vector-valued. The organization is as follows.

Chapter 2 settles the common notation for Part I of the thesis, and introduces the main rationale behind the usage of strictly consistent scoring functions also for the more realistic case of non-deterministic forecasts, thereby using and refining the prediction-space setting of Gneiting and Ranjan (2013) and Strähl and Ziegel (2015). Moreover, the chapter collects and transfers some basic results from the one-dimensional to the higher-dimensional setting.

The main content of Chapter 3 is the peer-reviewed and published paper Fissler and Ziegel (2016). It is primarily devoted to question (ii) of the elicitation problem in the higher-dimensional setting. A crucial tool and result is a refinement and generalization of Osband's principle (Osband, 1985), which gives a close connection between the gradient of the expected score and expectations of identification functions, exploiting first order conditions of the optimization problem linked to strict consistency. Using this and a second order version of Osband's principle, we are able to describe the class of all strictly consistent scoring functions for vectors of quantiles and / or expectiles and different levels. Making further use of the established technique, we can describe the class of strictly consistent scoring functions for a vector consisting of a spectral risk measure with finitely supported spectral measure and the corresponding quantiles. In particular, this shows that the pair (VaR, ES) at the same level $\alpha \in (0, 1)$ is jointly elicitable. This result is of high relevance in two aspects: First, it is the first result showing the elicitability of a functional which cannot be represented as a bijection of a functional consisting of elicitable components only. Second, it has added a piece to the mosaic of the debate about the best choice of a risk measure, opening the possibility of meaningful forecast comparison between competing joint (VaR, ES)-forecasts. The positive confirmation about the joint elicitability also gave rise to a more general principle of the joint elicitability of the 'divergence' of a strictly consistent scoring function discovered by Frongillo and Kash (2015b), which is described in Section 3.3. Section 3.2 presents a characterization of the class of (strict) identification functions for a fixed functional, utilizing an argument quite similar to Osband's principle.

Chapter 4 is concerned with the collection and investigation of further 'nice' properties of scoring functions besides and beyond strict consistency, apparently dealing with subproblem (iii) of the above introduced elicitation problem. The main topics are order-sensitivity and different variants of generalizing it to the multivariate case. In particular, we give characterizations of the different sorts of order-sensitivity for various popular functionals. The second goal is to study convexity and quasi-convexity of scoring functions in greater detail. We review and introduce various known as well as novel motivations of considering convex scoring functions. Similarly, we strive to establish the classes of (quasi-)convex scoring functions for many popular functionals. Finally, in Section 4.3, we inspect equivariant functionals covering the most important and popular cases of positively homogeneous functionals and translation equivariant functionals. Arguing that it is a desirable requirement of a scoring function that it preserves the ranking of competing forecasts under a transformation of the corresponding observations and forecasts, we study notions of *order-preserving* scoring functions. Again positively homogeneous and translation *in*variant scoring functions are the two main instances that should be born in mind. We collect some known results from the literature, but also present original findings concerning translation invariant scoring functions.

Chapter 5 consists of the article Fissler et al. (2016). Contributing to the debate about the choice of the best risk measure and commenting on the role of elicitability for backtesting in quantitative finance, we show how the finding of Fissler and Ziegel (2016) about the joint elicitability of (VaR, ES) can be used for *comparative backtests*. We introduce and explain this notion of *comparative backtests*, amounting to model *selection* rather than model *verification*, the latter being the goal of traditional backtests. Whereas for model verification can dispense with elicitability, it is vital for model selection. Remarkably, utilizing comparative backtests of Diebold-Mariano type, one can use a more conservative backtesting decision, amounting to a reversed onus of proof when establishing new internal models by banks, enticing financial institutions to improve the prediction performances of their internal models and being "beneficial to all stakeholders, including banks, regulators, and society at large" (Fissler et al., 2016, p. 60).

Part I concludes with a discussion in Chapter 6, reviewing the reception of the two articles Fissler and Ziegel (2016), Fissler et al. (2016) in the literature, and outlining possible future research projects related to Part I of this thesis.

2. Some Basic Results and Preliminaries on Higher Order Elicitability

Section 2.3 in this chapter gathers some basic observations about strictly consistent scoring functions for vector-valued functionals. It mainly generalizes results mentioned in Gneiting (2011) to the higher-dimensional case. Most of these generalizations are quite straight forward and coincide with the one-dimensional case – however, we present them for the sake of completeness. On the other hand, the novelty of this chapter is certainly the observations concerning the normalization conventions before Proposition 2.3.1, a generalized version of the revelation principle, Lemma 2.3.3 and the considerations about manipulating the second argument of a strictly consistent scoring function, which can be found after Proposition 2.3.4.

Before entering into the details of higher order elicitability, we strive to bring the rationale of elicitability to life. That is, we present the prediction space setting in Section 2.2, which is an adaptation of the concepts presented in Gneiting and Ranjan (2011) and Strähl and Ziegel (2015). It tries to explain how to use elicitability when comparing and ranking competing non-deterministic forecasts in general. *A fortiori*, it explains the 'conditioning argument' after which, in many situations, it is no loss of generality to consider forecasts as deterministic. It is accompanied by a brief excursion to the connection between elicitability and regression.

The whole notation in this part of the thesis is intended to be consistent, and formally, most notation is introduced in Fissler and Ziegel (2016), attached in Chapter 3. However, since we have decided to present these more elementary results before the paper mentioned above, it is necessary to fix some notation already here, thereby accepting the danger of redundancy and self-citation within the thesis to some extent.

2.1. Notation and definitions

We collect again the components of the decision-theoretic framework motivated in the Introduction and which can be found in Gneiting (2011). By O we denote the *observation domain* which is a measurable space equipped with a σ-algebra $\mathcal{O}$. In most applications, O will be a subset of the Euclidean space $\mathbb{R}^d$ equipped with the Borel σ-algebra. In those cases, we will deliberately identify a Borel probability measure on $(\mathsf{O}, \mathcal{O})$ with its distribution function. Generally, we denote by $\mathcal{F}$ a class of probability distributions in $(\mathsf{O}, \mathcal{O})$. Third, A stands for the *action domain*. For the presentation of the abstract setting, one can treat O and A as some generic sets equipped with σ-algebras. However, for most results of this thesis, we will assume that also $\mathsf{A} \subseteq \mathbb{R}^k$ for some $k \geq 1$ and that it is equipped with the corresponding Borel σ-algebra.

We say that a function $a\colon \mathsf{O} \to \mathbb{R}$ is $\mathcal{F}$-integrable if it is F-integrable for each $F \in \mathcal{F}$. A function $g\colon \mathsf{A} \times \mathsf{O} \to \mathbb{R}$ is $\mathcal{F}$-integrable if $g(x, \cdot)$ is $\mathcal{F}$-integrable for each $x \in \mathsf{A}$. If g is $\mathcal{F}$-integrable, we introduce the map

$$\bar{g}\colon \mathsf{A} \times \mathcal{F} \to \mathbb{R}, \quad (x, F) \mapsto \bar{g}(x, F) = \int g(x, y)\, \mathrm{d}F(y). \tag{2.1.1}$$

Consequently, for fixed $F \in \mathcal{F}$ we can consider the function $\bar{g}(\cdot, F)\colon \mathsf{A} \to \mathbb{R}$, $x \mapsto \bar{g}(x, F)$, and for fixed $x \in \mathsf{A}$ we can consider the (linear) functional $\bar{g}(x, \cdot)\colon \mathcal{F} \to \mathbb{R}$, $F \mapsto \bar{g}(x, F)$.

Definition 2.1.1 (Consistency)**.** A *scoring function* is a Borel-measurable, $\mathcal{F}$-integrable function $S\colon \mathsf{A} \times \mathsf{O} \to \mathbb{R}$. It is said to be *$\mathcal{F}$-consistent* for a functional $T\colon \mathcal{F} \to \mathsf{A}$ if

$$\bar{S}(T(F), F) \leq \bar{S}(x, F)$$

for all $F \in \mathcal{F}$ and for all $x \in \mathsf{A}$. Furthermore, S is *strictly $\mathcal{F}$-consistent* for T if it is $\mathcal{F}$-consistent for T and if

$$\bar{S}(T(F), F) = \bar{S}(x, F) \quad \Longrightarrow \quad x = T(F)$$

for all $F \in \mathcal{F}$ and for all $x \in \mathsf{A}$.

Definition 2.1.2 (Elicitability)**.** A functional $T\colon \mathcal{F} \to \mathsf{A}$ is called *elicitable*, if there exists a strictly $\mathcal{F}$-consistent scoring function for T.

Definition 2.1.3 (Identification function)**.** Let $\mathsf{A} \subseteq \mathbb{R}^k$. An *identification function* is a Borel-measurable, $\mathcal{F}$-integrable function $V\colon \mathsf{A} \times \mathsf{O} \to \mathbb{R}^k$. It is said to be an *$\mathcal{F}$-identification function* for a functional $T\colon \mathcal{F} \to \mathsf{A} \subseteq \mathbb{R}^k$ if

$$\bar{V}(T(F), F) = 0$$

for all $F \in \mathcal{F}$. Furthermore, V is a *strict $\mathcal{F}$-identification function* for T if

$$\bar{V}(x, F) = 0 \quad \Longleftrightarrow \quad x = T(F)$$

for all $F \in \mathcal{F}$ and for all $x \in \mathsf{A}$.

Definition 2.1.4 (Identifiability)**.** A functional $T\colon \mathcal{F} \to \mathsf{A} \subseteq \mathbb{R}^k$ is said to be *identifiable*, if there exists a strict $\mathcal{F}$-identification function for T.

Remark 2.1.5. Gneiting (2011) and some other parts of the literature consider functionals that are set-valued. That is, they consider $T\colon \mathcal{F} \to 2^{\mathsf{A}}$ where 2^{A} is the power set of A. For some functionals and choices the class $\mathcal{F}$, this is a natural assumption. Then, the notions introduced above can be generalized as follows. A scoring function $S\colon \mathsf{A}\times\mathsf{O}\to\mathbb{R}$ is $\mathcal{F}$-consistent for T if

$$\bar{S}(t,F) \le \bar{S}(x,F)$$

for all $F\in\mathcal{F}$, $t\in T(F)$ and $x\in\mathsf{A}$. The scoring function S is strictly $\mathcal{F}$-consistent for T if it is $\mathcal{F}$-consistent and if

$$\bar{S}(t,F) = \bar{S}(x,F) \quad\Longrightarrow\quad x\in T(F)$$

for all $F\in\mathcal{F}$ and for all $t\in T(F)$, $x\in\mathsf{A}$. The notions of a (strict) $\mathcal{F}$-identification functions can be generalized *mutatis mutandis*. It is important to remark that the two notions coincide whenever $T(F)$ is a singleton for all $F\in\mathcal{F}$.

Within this thesis, we consider the case of set-valued functionals only once explicitly, namely in Lemma 2.3.3 and the discussion thereafter. Nevertheless, many of the results of this thesis can be extended to set-valued functionals. However, to allow for a clear presentation, we confine ourselves to functionals with values in $\mathsf{A}\subseteq\mathbb{R}^k$ in most of the parts.

If we have a random variable Y on some probability space $(\Omega,\mathcal{A},\mathbb{P})$, mapping to O, we denote its distribution with $\mathcal{L}(Y)$. For some sub-σ-algebra $\mathcal{A}_0\subseteq\mathcal{A}$, we write $\mathcal{L}(Y|\mathcal{A}_0)$ for the (regular version of the) conditional distribution of Y given $\mathcal{A}_0$. If X is a random element on $(\Omega,\mathcal{A},\mathbb{P})$, we write $\sigma(X)\subseteq\mathcal{A}$ for the σ-algebra generated by the pre-images of X. As mentioned above, we often identify probability measures with their probability distribution function. We deliberately write $\mathcal{L}(Y|X)$ for the regular version of the conditional distribution $\mathcal{L}(Y|\sigma(X))$ as well as $\mathbb{E}[Y|X]$ for the conditional expectation $\mathbb{E}[Y|\sigma(X)]$. The notation $Y\sim F$ means that the random variable Y has distribution F, $\mathcal{L}(Y)=F$, and for two random variables $Y\sim Y'$ means that they have the same law, $\mathcal{L}(Y)=\mathcal{L}(Y')$.

2.2. The prediction space setting – bringing elicitability to life

After this theoretical introduction we demonstrate the rationale of elicitability. Assume we have a time series $(Y_t)_{t\in\mathbb{N}}$ taking values in an observation domain $\mathsf{O}\subseteq\mathbb{R}^d$ defined on some common underlying probability space $(\Omega,\mathcal{A},\mathbb{P})$. Assume further that we have m competing forecasters, each of them issuing a point forecast $X_t^{(i)}$, $i\in\{1,\dots,m\}$, for a certain k-dimensional property of the distribution of Y_t such as the mean or median for $k=1$ or a vector of different moments for $k>1$. More specifically, we have m sequences $(X_t^{(i)})_{t\in\mathbb{N}}$, $i\in\{1,\dots,m\}$, of random vectors taking values in an action domain $\mathsf{A}\subseteq\mathbb{R}^k$ and consider a functional T mapping from some generic class of distributions $\mathcal{F}$ on O to A. Consequently, we have a forecast-observation sequence

$$\big((X_t^{(1)},\dots,X_t^{(m)},Y_t)\big)_{t\in\mathbb{N}}\,. \tag{2.2.1}$$

Let $S\colon \mathsf{A}\times\mathsf{O}\to\mathbb{R}$ be a (strictly) $\mathcal{F}$-consistent scoring function for the functional T. Up to time n, it is common practice to compare and rank the forecast performances in terms in their realized scores

$$\bar{\mathbf{S}}_n^{(i)} = \frac{1}{n}\sum_{t=1}^{n} S(X_t^{(i)},Y_t).$$

Since the scores are negatively oriented, the forecaster with the lowest realized score is deemed to have the best forecast performance. But what is the rationale and the justification behind that practice? Assuming for a moment that the sequence at (2.2.1) is stationary and ergodic with $(X_t^{(1)},\dots,X_t^{(m)},Y_t)\sim(X^{(1)},\dots,X^{(m)},Y)$ and that $\mathbb{E}[|S(X^{(i)},Y)|]<\infty$ for all $i\in\{1,\dots,m\}$, then one has the convergence

$$\bar{\mathbf{S}}_n^{(i)} \longrightarrow \mathbb{E}[S(X^{(i)},Y)],\quad \text{as } n\to\infty$$

$\mathbb{P}$-almost surely and in L^p. Now suppose that the distribution $\mathcal{L}(Y)$ of Y is an element of $\mathcal{F}$ as well as for any sub-σ-algebra $\mathcal{A}_0\subseteq\mathcal{A}$ subsequently appearing, the regular version of the conditional distribution $\mathcal{L}(Y|\mathcal{A}_0)$ is $\mathbb{P}$-a.s. an element of $\mathcal{F}$. Then, by the tower property, one has for any $i\in\{1,\dots,m\}$

$$\begin{aligned}\mathbb{E}[S(X^{(i)},Y)] &= \mathbb{E}\big[\mathbb{E}[S(X^{(i)},Y)\,|\,\sigma(X^{(1)},\dots,X^{(m)})]\big]\\ &= \mathbb{E}\big[\bar{S}(X^{(i)},\mathcal{L}(Y\,|\,\sigma(X^{(1)},\dots,X^{(m)})))\big]\\ &\ge \mathbb{E}\big[\bar{S}(T(\mathcal{L}(Y\,|\,\sigma(X^{(1)},\dots,X^{(m)}))),\mathcal{L}(Y\,|\,\sigma(X^{(1)},\dots,X^{(m)})))\big],\end{aligned}$$

where we used the notation introduced at (2.1.1), the (strict) $\mathcal{F}$-consistency of S and the monotonicity of the conditional expectation. Hence, under the assumption of stationarity and ergodicity, a forecaster outperforms his colleagues – at least asymptotically – if she issues $\mathbb{P}$-a.s. the forecast

$$X^{(i)} = T(\mathcal{L}(Y\,|\,\sigma(X^{(1)},\dots,X^{(m)})))$$

and she outperforms a subset $\{j_1,\dots,j_\ell\}\subset\{1,\dots,m\}$ of her colleagues if $\mathbb{P}$-a.s.

$$X^{(i)} = T(\mathcal{L}(Y\,|\,\sigma(X^{(i)},X^{(j_1)},\dots,X^{(j_\ell)}))).$$

At first glance, this seems to be a very appealing reasoning. However, the assumption that the forecast observation sequence at (2.2.1) is stationary is quite simplistic and in most cases unrealistic. First, for some processes of observation from nature, assuming stationarity of $(Y_t)_{t\in\mathbb{N}}$ might be suitable due to the persistence of laws in nature over time. But if a forecaster has memory and can learn from previous observations of $(Y_t)_{t\in\mathbb{N}}$, a reasonable strategy for issuing the forecast X_t would be to compute some sample estimator of the functional based on $Y_1, \dots, Y_{t-1}$, e.g.

$$X_t = T(\hat{F}_{t-1}), \qquad \text{where } \hat{F}_{t-1} = \frac{1}{t-1}\sum_{j=1}^{t-1} \delta_{Y_j} \tag{2.2.2}$$

and δ_y is the Dirac measure in y. However, it is obvious that then $(X_t)_{t\in\mathbb{N}}$ is non-stationary. Second, in line with Davis (2016, p. 4)

> "…the typical stylised features found in financial price data [are] apparent non-stationarity and highly 'bursty' volatility."

So regarding elicitability, Davis (2016, p. 8) remarks:

> "What is not so clear is how to apply these results in a dynamic context such as risk management where the data is a sequence $Y_1, Y_2, \dots$ of random variables each having a different conditional distribution, …"

To cope with some non-stationarity, we introduce the prediction space setting for serial dependence which is due to Strähl and Ziegel (2015) and is a generalization of the one-period prediction space introduced in Gneiting and Ranjan (2013); see Subsection 2.2.3.

Nevertheless, we begin with two situations where stationarity and, *a fortiori*, the i.i.d. structure if the sequence at (2.2.1) are a common and accepted assumption, namely learning and regression.

2.2.1. Learning

If we have n observations of an i.i.d. (or, more generally, an ergodic) sequence $(Y_t)_{t\in\mathbb{N}}$ where Y_t has some unknown distribution $F \in \mathcal{F}$ and one wants to estimate a functional $T\colon \mathcal{F} \to \mathsf{A} \subseteq \mathbb{R}^k$, one can use any strictly $\mathcal{F}$-consistent scoring function S to do M-estimation in the sense that

$$\arg\min_{x\in\mathsf{A}} \frac{1}{n}\sum_{t=1}^{n} S(x, Y_t)$$

is a consistent estimator for

$$\arg\min_{x\in\mathsf{A}} \mathbb{E}[S(x,Y)] = \arg\min_{x\in\mathsf{A}} \bar{S}(x,F) = T(F)$$

under some regularity conditions detailed in Huber and Ronchetti (2009, Chapter 6).

2.2.2. Regression

The usual situation of regression is that we have i.i.d. data consisting of explanatory factors Z with values in $\mathbb{R}^\ell$ and an output variable Y with values in $\mathbb{R}^d$.

More specifically, we have n observations of the i.i.d. sequence $(Z_t, Y_t)_{t\in\mathbb{N}}$ where $(Z_t, Y_t) \sim (Z, Y)$. Given a (parametric) class of models $\mathcal{G}$ with measurable functions $g\colon \mathbb{R}^\ell \to \mathbb{R}^k$, $k = d$, as elements, one tries to fit the output Y with $g(Z)$ for $g \in \mathcal{G}$ in the sense that

$$Y = g(Z) + \varepsilon \tag{2.2.3}$$

under the assumption of the identity for the conditional mean

$$\mathbb{E}[Y|Z] = g^*(Z) \tag{2.2.4}$$

for some $g^* \in \mathcal{G}$ (at least approximately). Then, the fitting is usually done with a least squares approach

$$g^* = \arg\min_{g\in\mathcal{G}} \frac{1}{n}\sum_{t=1}^{n}(g(Z_t) - Y_t)^2\,, \tag{2.2.5}$$

thereby exploiting the fact that the squared loss is a strictly consistent scoring function for the mean. Now, one can generalize this classical approach on the one hand by computing the arg min at (2.2.5) with respect to another strictly consistent scoring function for the mean; see Gneiting (2011, Theorem 7) for possible candidates. On the other hand, one can generalize the assumption at (2.2.4): If

$$T(\mathcal{L}(Y|Z)) = g^*(Z)$$

for some $g^* \in \mathcal{G}$ (at least approximately) and some functional $T\colon \mathcal{F} \to \mathsf{A} \subseteq \mathbb{R}^k$, then one can compute g^* by

$$g^* = \arg\min_{g\in\mathcal{G}} \frac{1}{n}\sum_{t=1}^{n} S(g(Z_t), Y_t)\,,$$

where S is a strictly $\mathcal{F}$-consistent scoring function for T. Note that for the general situation, the dimensions k and d do not necessarily have to coincide. However, clearly for $k \neq d$, an identity in the spirit of (2.2.3) does not make sense any more.

To match this framework with the one of forecast comparison, assume that $\mathcal{G}$ is finite, that is $\mathcal{G} = \{g^{(1)}, \dots, g^{(m)}\}$. Then, we can define the 'forecasts' $X_t^{(i)} = g^{(i)}(Z_t)$ for $i \in \{1, \dots, m\}$. Notice, that the induced forecast-observation sequence corresponding to (2.2.1) is again i.i.d. and moreover $\sigma(X^{(1)}, \dots, X^{(m)}) \subseteq \sigma(Z)$. Due to the tower property, a model / forecaster $g^{(i)}$ outperforms the competing models / forecasters if

$$g^{(i)}(Z) = T(\mathcal{L}(Y|Z))\,.$$

More generally, it is also no problem to assume that the class $\mathcal{G}$ is not finite.

This way of bringing elicitability to life in the context of regression was mainly conducted in the field of quantile regression (Koenker, 2005) and expectile regression (Newey and Powell, 1987), but recently has been extended to the case of a joint regression framework for Value at Risk (VaR) and Expected Shortfall

(ES) (Bayer and Dimitriadis, 2017), which is based on the results of Fissler and Ziegel (2016) that the pair (VaR, ES) is jointly elicitable. We shall illuminate some aspects of regression using convex and strictly consistent scoring functions in Section 4.2. However, a detailed analysis of regression using strictly consistent scoring functions is beyond the scope of this thesis and is deferred to possible future research.

2.2.3. The prediction space

The prediction space setting for one period in Gneiting and Ranjan (2013) as well as its generalization to serial dependence in Strähl and Ziegel (2015) are tailored for probabilistic forecasts. However, it is straight forward to adopt the notion to point forecasts, which has been done by Ehm et al. (2016) for the one-period-setting. In the sequel, we give a definition for the setting of serial dependence for point forecasts, using the notation $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$.

Definition 2.2.1 (Prediction space for serial dependence for point forecasts)**.** Let $m, k, d \geq 1$ be integers. Consider a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ together with filtrations $(\mathcal{A}_t^{(1)})_{t\in\mathbb{N}_0}, \ldots, (\mathcal{A}_t^{(m)})_{t\in\mathbb{N}_0}$ with $\mathcal{A}_t^{(i)} \subseteq \mathcal{A}$ for all $i \in \{1, \ldots, m\}$, $t \in \mathbb{N}_0$. A *prediction space for serial dependence for point forecasts* is a collection of an $\mathbb{R}^d$-valued sequence $(Y_t)_{t\in\mathbb{N}}$ with the filtration $(\mathcal{T}_t)_{t\in\mathbb{N}_0}$ generated by $(Y_t)_{t\in\mathbb{N}}$, that is $\mathcal{T}_t = \sigma(Y_s,\, s \leq t)$ for $t \in \mathbb{N}$ and $\mathcal{T}_0 = \{\emptyset, \Omega\}$, and m sequences of $\mathbb{R}^k$-valued random variables $(X_t^{(i)})_{t\in\mathbb{N}}$, $i \in \{1, \ldots, m\}$, such that $X_t^{(i)}$ is measurable with respect to $\sigma(\mathcal{A}_{t-1}^{(i)}, \mathcal{T}_{t-1})$. Moreover, the prediction space contains a functional $T\colon \mathcal{F} \to \mathbb{R}^k$ where $\mathcal{F}$ is a class of probability distributions on $\mathbb{R}^d$ such that $\mathcal{L}(Y_t) \in \mathcal{F}$ for all $t \in \mathbb{N}$ and $\mathbb{P}$-a.s. $\mathcal{L}(Y_t \,|\, \sigma(\mathcal{A}_{t-1}^{(j_1)}, \ldots, \mathcal{A}_{t-1}^{(j_\ell)}, \mathcal{T}_{t-1})) \in \mathcal{F}$ for all $t \in \mathbb{N}$ and all subsets $\{j_1, \ldots, j_\ell\} \subseteq \{1, \ldots, m\}$.

Essentially, this definition can be understood in the following way: We have m competing forecasters. At time point $t-1$ they are given their personal (possibly exclusive) information $\mathcal{A}_{t-1}^{(i)}$ respectively, as well as the commonly available information consisting of the past observations, that is $\mathcal{T}_{t-1}$. Their point forecasts for time point t, that is $X_t^{(i)}$, must be based on these two sources of information only. For a detailed explanation of this definition in the setting of probabilistic forecasts, see Strähl and Ziegel (2015). Of course, the difference between point forecasts and probabilistic forecasts is the appearance of a certain fixed functional T in the case of point forecasts. In analogy to Strähl and Ziegel (2015), we introduce the following concepts.

Definition 2.2.2 (ideal)**.** Within the prediction space setting for serial dependence for point forecasts, a forecast $X_t^{(i)}$, $i \in \{1, \ldots, m\}$, is *ideal* for T with respect to $\mathcal{A}_{t-1}^{(i)}$ if

$$X_t^{(i)} = T(\mathcal{L}(Y_t \,|\, \sigma(\mathcal{A}_{t-1}^{(i)}, \mathcal{T}_{t-1}))) \qquad \mathbb{P}\text{-a.s.}$$

Definition 2.2.3 (cross-ideal)**.** Within the prediction space setting for serial dependence for point forecasts, a forecast $X_t^{(i)}$, $i \in \{1, \dots, m\}$, is *cross-ideal* for T with respect to $\mathcal{A}_{t-1}^{(j_1)}, \dots, \mathcal{A}_{t-1}^{(j_\ell)}$ where $i \in \{j_1, \dots, j_\ell\} \subseteq \{1, \dots, m\}$ if

$$X_t^{(i)} = T(\mathcal{L}(Y_t \,|\, \sigma(\mathcal{A}_{t-1}^{(j_1)}, \dots, \mathcal{A}_{t-1}^{(j_\ell)}, \mathcal{T}_{t-1}))) \qquad \mathbb{P}\text{-a.s.}$$

Moreover, in analogy to Nolde and Ziegel (2016, Definition 4) but with some slight differences, we introduce the notion of forecast dominance.

Definition 2.2.4 (Forecast dominance)**.** Let S be a strictly $\mathcal{F}$-consistent scoring function for a functional $T\colon \mathcal{F} \to \mathsf{A}$, let $\mathcal{H}$ be some σ-algebra, and $(\mathcal{H}_t)_{t\in\mathbb{N}_0}$ some filtration. Within the prediction space setting for serial dependence for point forecasts, a forecast $X_t^{(i)}$ S-dominates $X_t^{(j)}$ conditionally $\mathcal{H}$ if

$$\mathbb{E}[S(X_t^{(i)}, Y_t) - S(X_t^{(j)}, Y_t) \,|\, \mathcal{H}] \le 0 \qquad \mathbb{P}\text{-a.s.}$$

Moreover, the forecast sequence $(X_t^{(i)})_{t\in\mathbb{N}}$ S-dominates the sequence $(X_t^{(j)})_{t\in\mathbb{N}}$ on average conditionally on $(\mathcal{H}_t)_{t\in\mathbb{N}_0}$ if

$$\mathbb{E}[S(X_t^{(i)}, Y_t) - S(X_t^{(j)}, Y_t) \,|\, \mathcal{H}_{t-1}] \le 0 \qquad \mathbb{P}\text{-a.s. for all } t \in \mathbb{N}.$$

In particular, the forecast sequence $(X_t^{(i)})_{t\in\mathbb{N}}$ S-dominates the sequence $(X_t^{(j)})_{t\in\mathbb{N}}$ unconditionally on average if

$$\mathbb{E}[S(X_t^{(i)}, Y_t) - S(X_t^{(j)}, Y_t)] \le 0 \qquad \text{for all } t \in \mathbb{N}.$$

Remark 2.2.5. Clearly, the most interesting choices of the filtration is a filtration that makes one or even both predictions measurable. That is, in the prediction space setting, one could choose $\mathcal{H}_{t-1}$ such that it contains $\sigma(\mathcal{A}_{t-1}^{(i)}, \mathcal{T}_{t-1})$ or $\sigma(\mathcal{A}_{t-1}^{(j)}, \mathcal{T}_{t-1})$ or even $\sigma(\mathcal{A}_{t-1}^{(i)}, \mathcal{A}_{t-1}^{(j)}, \mathcal{T}_{t-1})$. This situation is in line with Nolde and Ziegel (2016, Definition 4). However, we also allow for smaller σ-algebras. In particular, choosing $\mathcal{H}_{t-1} = \{\emptyset, \Omega\}$ for all $t \in \mathbb{N}$, this definition also covers the unconditional case. Of course, by the tower property, the conditional S-dominance implies the unconditional S-dominance.

Lemma 2.2.6. *Let $S\colon \mathsf{A} \times \mathsf{O} \to \mathbb{R}$ be a strictly $\mathcal{F}$-consistent scoring function for a functional $T\colon \mathcal{F} \to \mathsf{A} \subseteq \mathbb{R}^k$. Within the prediction space setting for serial dependence for point forecasts, let the forecast $X_t^{(i)}$, $i \in \{1, \dots, m\}$, be cross-ideal for T with respect to $\mathcal{A}_{t-1}^{(j_1)}, \dots, \mathcal{A}_{t-1}^{(j_\ell)}$ where $i \in \{j_1, \dots, j_\ell\} \subseteq \{1, \dots, m\}$. Then $X_t^{(i)}$ S-dominates $X_t^{(j)}$ for any $j \in \{j_1, \dots, j_\ell\}$ conditionally on $\sigma(\mathcal{A}_{t-1}^{(j_1)}, \dots, \mathcal{A}_{t-1}^{(j_\ell)}, \mathcal{T}_{t-1})$ and unconditionally.*

Proof. Let $j \in \{j_1, \dots, j_\ell\}$. Then

$$\begin{aligned}
&\mathbb{E}\big[\mathbb{E}[S(X_t^{(j)}, Y_t) \,|\, \sigma(\mathcal{A}_{t-1}^{(j_i)}, \dots, \mathcal{A}_{t-1}^{(j_\ell)}, \mathcal{T}_{t-1})]\big] \\
&= \mathbb{E}\big[\bar{S}(X_t^{(j)}, \mathcal{L}(Y_t \,|\, \sigma(\mathcal{A}_{t-1}^{(j_i)}, \dots, \mathcal{A}_{t-1}^{(j_\ell)}, \mathcal{T}_{t-1})))\big] \\
&\ge \mathbb{E}\big[\bar{S}(T(\mathcal{L}(Y_t \,|\, \sigma(\mathcal{A}_{t-1}^{(i)}, \mathcal{T}_{t-1}))), \mathcal{L}(Y_t \,|\, \sigma(\mathcal{A}_{t-1}^{(j_i)}, \dots, \mathcal{A}_{t-1}^{(j_\ell)}, \mathcal{T}_{t-1})))\big] \\
&= \mathbb{E}\big[\bar{S}(X_t^{(i)}, \mathcal{L}(Y_t \,|\, \sigma(\mathcal{A}_{t-1}^{(j_i)}, \dots, \mathcal{A}_{t-1}^{(j_\ell)}, \mathcal{T}_{t-1})))\big] \,.
\end{aligned}$$

The unconditional case follows by the tower property. □

The notion of forecast dominance with respect to a certain fixed scoring function is due to Nolde and Ziegel (2016), whereas Ehm et al. (2016) consider the notion of forecast dominance with respect to the whole class of consistent scoring functions for some functional. Lemma 2.2.6 also means that the more information an ideal forecaster has the lower is her expected score. This recovers Corollary 2 in Holzmann and Eulert (2014) asserting that strictly consistent scoring functions are 'order-sensitive' with respect to nested information sets. The notion of order-sensitivity will be discussed in greater detail in Section 4.1 of this thesis.

It is possible to test for conditional and unconditional forecast dominance, thereby relaxing the assumption on stationarity.fortiori

Testing for (conditional) forecast dominance

Following the lines of Nolde and Ziegel (2016) and Giacomini and White (2006), it is theoretically possible to define asymptotic level α tests for the null hypothesis of conditional and unconditional S-forecast dominance. That is, one can test

$$H_0\colon\ \mathbb{E}[S(X_t^{(i)}, Y_t) - S(X_t^{(j)}, Y_t) \,|\, \mathcal{H}_{t-1}] \le 0 \qquad \mathbb{P}\text{-a.s. for all } t \in \mathbb{N}$$

for some filtration $(\mathcal{H}_t)_{t\in\mathbb{N}_0}$. It is important to notice that one can dispense with the stationarity assumption. However, one still has to impose mixing conditions on the sequences $(X_t^{(i)})_{t\in\mathbb{N}}$ and $(X_t^{(j)})_{t\in\mathbb{N}}$ (see the references for details). Moreover, the tests are also asymptotically consistent against corresponding alternatives.

For the unconditional case, meaning $\mathcal{H}_t = \{\emptyset, \Omega\}$ for all t, one can also test a broader null hypothesis. Assuming that the limit

$$\lambda := \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\big[S(X_t^{(i)}, Y_t) - S(X_t^{(j)}, Y_t)\big]$$

exists in $\mathbb{R}\cup\{\pm\infty\}$, one can define an asymptotic level α test for the null hypothesis

$$H_0\colon\ \lambda \le 0.$$

Of course, the null hypothesis holds under (conditional) S-dominance on average. For the proof, one still has to assume certain mixing conditions on the forecast sequences. Details for this unconditional approach can be found in Nolde and Ziegel (2016) and Giacomini and White (2006). Such unconditional tests involving rescaled versions of the test statistic

$$\frac{1}{n}\sum_{i=1}^{n} S(X_t^{(i)}, Y_t) - S(X_t^{(j)}, Y_t)$$

are so-called *Diebold-Mariano tests* (Diebold and Mariano, 1995). We propose the usage of such tests as *comparative backtests* in the setting of backtesting risk measures; see Fissler et al. (2016) which corresponds to Chapter 5 of this thesis.

The 'conditioning argument'

The essence of this section which should be borne in mind, also for the forthcoming sections and chapters of the first part of this thesis, is a certain conditioning argument. Whenever one intends to compare the expected scores of two competing forecasts one can first condition on a σ-algebra which makes the two forecast measurable. Then, one can treat them as if they were deterministic, and can evaluate the expected scores with respect to some conditional distribution of the observation. If one of the forecasters is ideal with respect to this information set, one can derive an inequality which carries over to the level of the expected scores by the monotonicity of the expectation. In some other cases such as order-sensitivity – see Section 4.1 – one can also derive an inequality on the level of the conditional expectations which carries over analogously. Following this rationale, one can usually work on the 'conditional level', meaning conditionally on a σ-algebra which makes all forecasts at hand measurable. And hence, we shall tacitly assume that the forecasts are deterministic and that the observations follow some deterministic distribution.

Nevertheless, we shall occasionally come back to the setting of the prediction space. Whenever we do that, we shall mention it explicitly.

2.3. Transferring results from Gneiting (2011) to the higher-dimensional case

Strictly consistent scoring functions for a given functional T are not unique. In particular, the following result generalizes directly from the one-dimensional case. Let $S\colon \mathsf{A} \times \mathsf{O} \to \mathbb{R}$ be a strictly $\mathcal{F}$-consistent scoring function for a functional $T\colon \mathcal{F} \to \mathsf{A}$. Then, for any $\lambda > 0$ and any $\mathcal{F}$-integrable function $a\colon \mathsf{O} \to \mathbb{R}$, the scoring function

$$\widetilde{S}(x, y) := \lambda S(x, y) + a(y) \tag{2.3.1}$$

is again strictly $\mathcal{F}$-consistent for T. In the sequel, we shall say that S and $\widetilde{S}$ are of equivalent form (or just equivalent) if there is some $\lambda > 0$ and some $\mathcal{F}$-integrable function $a\colon \mathsf{O} \to \mathbb{R}$ such that they are connected by equation (2.3.1). Nolde and Ziegel (2016) say that two scoring functions are equivalent if their difference is a function of the observation y only. Clearly, equivalence in their sense implies equivalence in our sense with $\lambda = 1$. Note that equivalence preserves (strict) consistency and also any kind of order-sensitivity and (quasi-)convexity.

Gneiting (2011, Theorem 2) showed that in the one-dimensional case under the assumption $S(x,y) \ge 0$, the class of consistent scoring functions is a convex cone. Generally, the assumption of scoring functions being nonnegative is natural if $\delta_y \in \mathcal{F}$ for all $y \in \mathsf{O}$ because for an $\mathcal{F}$-consistent scoring function S, the scoring function $\widetilde{S}(x,y) := S(x,y) - \bar{S}(T(\delta_y), \delta_y)$ is non-negative and it is of the form (2.3.1) if $y \mapsto \bar{S}(T(\delta_y), \delta_y)$ is $\mathcal{F}$-integrable. We can see that the normalization condition (S0) in Table 7 of Gneiting (2011) – which seems to appear quite a number of times in the literature, e.g. in Bellini and Bignozzi (2015), Ehm et al. (2016), or Davis (2016) – should be slightly adapted to $S(x,y) \ge 0$ and $S(x,y) = 0$ if $x = T(\delta_y)$, thereby replacing the condition that $S(x,y) = 0$ if $x = y$. In fact, this is quite important because otherwise, functionals like $T(F) = -\mathbb{E}_F[Y]$ would not be elicitable. So in general, assuming that $\mathcal{F}$ contains all point measures implies that one can assume that S is non-negative. However, not for all situations in this thesis, this assumption is satisfied, and in fact different normalization conventions are convenient for different situations. We generalize Gneiting (2011, Theorem 2) as follows, showing that the class of strictly $\mathcal{F}$-consistent scoring functions for T is a convex cone (not including zero). The proof follows easily using Fubini's theorem.

Proposition 2.3.1. *Let $T\colon \mathcal{F} \to \mathsf{A}$ be a functional and $(Z, \mathcal{Z})$ be a measurable space with a σ-finite measure ν where $\nu \ne 0$. Let $\{S_z\colon z \in Z\}$ be a family of strictly $\mathcal{F}$-consistent scoring functions $S_z\colon \mathsf{A} \times \mathsf{O} \to \mathbb{R}$ for T. If for all $x \in \mathsf{A}$ and for all $F \in \mathcal{F}$ the map*

$$Z \times \mathsf{O} \to \mathbb{R}, \qquad (z,y) \mapsto S_z(x,y)$$

is $\nu \otimes F$-integrable, then the scoring function

$$S\colon \mathsf{A} \times \mathsf{O} \to \mathbb{R}, \qquad (x,y) \mapsto S(x,y) = \int_Z S_z(x,y)\,\nu(\mathrm{d}z)$$

is strictly $\mathcal{F}$-consistent for T.

Many important statistical functionals are transformations of other statistical functionals, for example variance and first and second moment are related in this manner. The following *revelation principle*, which originates from Osband (1985, p. 8) and is also given in Gneiting (2011, Theorem 4), states that if two functionals are related by a bijection, then one of them is elicitable if and only if the other one is elicitable. The assertion also holds upon replacing 'elicitable' with 'identifiable'. We omit the proof which is straight forward.

Proposition 2.3.2 (Revelation principle)**.** *Let $g\colon \mathsf{A} \to \mathsf{A}'$ be a bijection with inverse g^{-1}, where $\mathsf{A}, \mathsf{A}' \subseteq \mathbb{R}^k$. Let $T\colon \mathcal{F} \to \mathsf{A}$ be a functional. Then the following two assertions hold.*

(i) The functional $T\colon \mathcal{F} \to \mathsf{A}$ is identifiable if and only if $T_g = g \circ T\colon \mathcal{F} \to \mathsf{A}'$ is identifiable. The function $V\colon \mathsf{A} \times \mathsf{O} \to \mathbb{R}^k$ is a strict $\mathcal{F}$-identification function for T if and only if

$$V_g\colon \mathsf{A}' \times \mathsf{O} \to \mathbb{R}^k, \qquad (x', y) \mapsto V_g(x', y) = V(g^{-1}(x'), y)$$

is a strict $\mathcal{F}$-identification function for T_g.

(ii) The functional $T\colon \mathcal{F} \to \mathsf{A}$ is elicitable if and only if $T_g = g \circ T\colon \mathcal{F} \to \mathsf{A}'$ is elicitable. The function $S\colon \mathsf{A} \times \mathsf{O} \to \mathbb{R}$ is a strictly $\mathcal{F}$-consistent scoring function for T if and only if

$$S_g\colon \mathsf{A}' \times \mathsf{O} \to \mathbb{R}, \qquad (x', y) \mapsto S_g(x', y) = S(g^{-1}(x'), y) \tag{2.3.2}$$

is a strictly $\mathcal{F}$-consistent scoring function for T_g.

One could possibly wonder what happens in the revelation principle if g is not necessarily a bijection. Clearly, then the scoring function S_g defined at (2.3.2) is not well-defined any more. However, one can give a meaningful interpretation to this question.

Lemma 2.3.3. *Let $T\colon \mathcal{F} \to 2^{\mathsf{A}}$ be a set-valued functional with a (strictly) $\mathcal{F}$-consistent scoring function $S\colon \mathsf{A} \times \mathsf{O} \to \mathbb{R}$. Let $g\colon \mathsf{A}' \to \mathsf{A}$ be some map. Then*

$$S_{g^{-1}}\colon \mathsf{A}' \times \mathsf{O} \to \mathbb{R}, \qquad (x', y) \mapsto S_{g^{-1}}(x', y) = S(g(x'), y) \tag{2.3.3}$$

is a (strictly) $\mathcal{F}$-consistent scoring function for $T_{g^{-1}} = g^{-1} \circ T\colon \mathcal{F} \to 2^{\mathsf{A}'}$.

Proof. Let $F \in \mathcal{F}$. If $T_{g^{-1}} = g^{-1}(T(F)) = \emptyset$, there is nothing to show. So assume that $t' \in T_{g^{-1}}(F)$ and let $x' \in \mathsf{A}'$. Then

$$\bar{S}_{g^{-1}}(x', F) - \bar{S}_{g^{-1}}(t', F) = \bar{S}(g(x'), F) - \bar{S}(g(t'), F) \geq 0 \tag{2.3.4}$$

due to the $\mathcal{F}$-consistency of S for T and due to the fact that $g(t') \in T(F)$. Now, assume that S is strictly $\mathcal{F}$-consistent for T and that we have an equality in (2.3.4). Then, necessarily $g(x') \in T(F)$ such that $x' \in T_{g^{-1}}(F)$. □

We remark that in this context, the sets A, A' can be arbitrary. If g is surjective and $T(F) \neq \emptyset$ for all $F \in \mathcal{F}$, then also $T_{g^{-1}}(F) \neq \emptyset$ for all $F \in \mathcal{F}$ and one can consequently define some selection $\widetilde{T}_{g^{-1}}\colon \mathcal{F} \to \mathsf{A}'$ of $T_{g^{-1}}$ meaning that $\widetilde{T}_{g^{-1}}(F) \in T_{g^{-1}}(F)$ for all $F \in \mathcal{F}$. Then $S_{g^{-1}}$ defined at (2.3.3) is an $\mathcal{F}$-consistent scoring function for $\widetilde{T}_{g^{-1}}$ which is only strictly $\mathcal{F}$-consistent if $T_{g^{-1}}(F)$ is a singleton for each $F \in \mathcal{F}$. A case of particular interest is that $\mathsf{A}' = \mathcal{F}$, $g = T$ and the selection of $T^{-1} \circ T$ being the identity on $\mathcal{F}$.

This consideration leads the way to a connection between the theories of comparing point forecasts and probabilistic forecasts. In the setting of probabilistic forecasts, one issues a forecast in form of a probability distribution. Using the language of point forecasts, one is interested in the identity functional on $\mathcal{F}$. Following Gneiting and Raftery (2007), we call scoring functions for probabilistic forecasts *scoring rules*. Formally, a scoring rule is a map $R\colon \mathcal{F} \times \mathsf{O} \to \mathbb{R}$ such that for each $G \in \mathcal{F}$, the map $R(G, \cdot)\colon \mathsf{O} \to \mathbb{R}$, $y \mapsto R(G, y)$ is $\mathcal{F}$-integrable. Using the notation introduced in Section 2.1, we say that a scoring rule is $\mathcal{F}$-proper if $\bar{R}(F, F) \le \bar{R}(G, F)$ for all $F, G \in \mathcal{F}$. A scoring rule is strictly $\mathcal{F}$-proper if it is $\mathcal{F}$-proper and $\bar{R}(F, F) = \bar{R}(G, F)$ implies that $F = G$ for all $F, G \in \mathcal{F}$. With Lemma 2.3.3 and the discussion thereafter, we can see that each $\mathcal{F}$-consistent scoring function for a functional $T\colon \mathcal{F} \to \mathsf{A}$ induces an $\mathcal{F}$-proper scoring rule. However, if we do not impose that the functional T is injective, we cannot conclude that it is a *strictly* $\mathcal{F}$-proper scoring rule, even though the scoring function we are starting with might be strictly $\mathcal{F}$-consistent. We recover Gneiting (2011, Theorem 3).

Proposition 2.3.4. *Let $T\colon \mathcal{F} \to \mathsf{A}$ be a functional with an $\mathcal{F}$-consistent scoring function $S\colon \mathsf{A} \times \mathsf{O} \to \mathbb{R}$. Then the scoring rule*

$$R\colon \mathcal{F} \times \mathsf{O} \to \mathbb{R}, \qquad (F, y) \mapsto R(F, y) = S(T(F), y)$$

is $\mathcal{F}$-proper.

Proof. Let $F, G \in \mathcal{F}$. Then

$$\bar{R}(G, F) = \bar{S}(T(G), F) \ge \bar{S}(T(F), F) = \bar{R}(F, F),$$

which yields the assertion. □

One way of interpreting the revelation principle, Proposition 2.3.2, is that one considers the *pushforward* $T_g = g \circ T$ of a functional $T\colon \mathcal{F} \to \mathsf{A}$ by applying the bijection $g\colon \mathsf{A} \to \mathsf{A}'$. We remark that for this result, it is not essential that A' is a subset of $\mathbb{R}^k$. On the level of the scoring functions, this amounts to a manipulation of the *first* argument of the scoring functions. If one takes this point of view as a starting point, it is also natural to ask what happens upon applying some measurable function $\phi\colon \mathsf{O} \to \mathsf{O}$ (not necessarily a bijection) to the *second* argument of a strictly $\mathcal{F}$-consistent scoring function for a functional $T\colon \mathcal{F} \to \mathsf{A}$. In fact, the resulting scoring function

$${}_\phi S\colon \mathsf{A} \times \mathsf{O} \to \mathbb{R}, \qquad (x, y) \mapsto {}_\phi S(x, y) = S(x, \phi(y))$$

is a strictly ${}_\phi\mathcal{F}$-consistent scoring function for the functional ${}_\phi T\colon {}_\phi\mathcal{F} \to \mathsf{A}$ where ${}_\phi T$ is the functional T applied to the pushforward measures. To give a more formal definition, assume that $\mathcal{F}$ is a class of probability measures P and let that ${}_\phi\mathcal{F} = \{P \in \mathcal{F}\colon P \circ \phi^{-1} \in \mathcal{F}\} \subseteq \mathcal{F}$. Then

$${}_\phi T\colon {}_\phi\mathcal{F} \to \mathsf{A}, \qquad P \mapsto {}_\phi T(P) = T(P \circ \phi^{-1})\,.$$

Now, we consider weighted scoring functions where the weight depends on the realized observation $y \in \mathsf{O}$ only, thus generalizing Gneiting (2011, Theorem 5).

Proposition 2.3.5. *Let $\mathsf{O} \subseteq \mathbb{R}^d$ and $w\colon \mathsf{O} \to [0,\infty)$ be a measurable weight function. Let μ be a σ-finite measure on $(\mathsf{O},\mathcal{O})$ and $\mathcal{F}$ be a class of probability distributions on $(\mathsf{O},\mathcal{O})$ which are dominated by μ, and which are such that $\bar{w}(F) \in (0,\infty)$ for all $F \in \mathcal{F}$. If $F \in \mathcal{F}$ has a density f with respect to μ, define $F^{(w)}$ as the probability distribution with μ-density*

$$f^{(w)}(y) = \frac{w(y)f(y)}{\bar{w}(F)}$$

for μ-almost-all $y \in \mathsf{O}$. Let $\mathcal{F}^{(w)} = \{F \in \mathcal{F}\colon F^{(w)} \in \mathcal{F}\}$. Fix a functional $T\colon \mathcal{F} \to \mathsf{A} \subseteq \mathbb{R}^k$, where $k \geq 1$, and define the functional

$$T^{(w)}\colon \mathcal{F}^{(w)} \to \mathsf{A}, \qquad F \mapsto T^{(w)}(F) = T(F^{(w)}).$$

Then, the following two assertions hold:

(i) If $V\colon \mathsf{A} \times \mathsf{O} \to \mathbb{R}^k$ is a (strict) $\mathcal{F}$-identification function for T, then

$$V^{(w)}\colon \mathsf{A} \times \mathsf{O} \to \mathbb{R}^k, \qquad (x,y) \mapsto V^{(w)}(x,y) = w(y)V(x,y)$$

is a (strict) $\mathcal{F}^{(w)}$-identification function for the functional $T^{(w)}$.

(ii) If $S\colon \mathsf{A} \times \mathsf{O} \to \mathbb{R}$ is a (strictly) $\mathcal{F}$-consistent scoring function for T, then

$$S^{(w)}\colon \mathsf{A} \times \mathsf{O} \to \mathbb{R}, \qquad (x,y) \mapsto S^{(w)}(x,y) = w(y)S(x,y)$$

is a (strictly) $\mathcal{F}^{(w)}$-consistent scoring function for the functional $T^{(w)}$.

Proof. Let $F \in \mathcal{F}^{(w)}$ and $x \in \mathsf{A}$. Then, we obtain the identity

$$\begin{aligned}
\bar{S}^{(w)}(x,F) &= \int_{\mathsf{O}} w(y)S(x,y)f(y)\,\mu(\mathrm{d}y) \\
&= \bar{w}(F)\Big[\int_{\mathsf{O}} S(x,y)f^{(w)}(y)\,\mu(\mathrm{d}y)\Big] \\
&= \bar{w}(F)\,\bar{S}(x,F^{(w)}),
\end{aligned}$$

and by an analogue calculation

$$\bar{V}^{(w)}(x,F) = \bar{w}(F)\,\bar{V}(x,F^{(w)}).$$

Now, the assertions follow upon recalling that $\bar{w}(F) \in (0,\infty)$. □

Remark 2.3.6. Interestingly, a combination of Lemma 2.3.3 and Proposition 2.3.5 retrieves Theorem 1 in Holzmann and Klar (2016). To see this, let $\mathcal{F}$ be a class of probability distributions on $\mathsf{O} \subseteq \mathbb{R}^d$ which are dominated by a σ-finite measure μ. Let $w\colon \mathsf{O} \to [0,\infty)$ be a measurable weight function such that $\bar{w}(F) > 0$ for all

$F \in \mathcal{F}$. Define the operator g_w on $\mathcal{F}$ via $g_w(F) = F^{(w)}$ where $F^{(w)}$ is defined like in Proposition 2.3.5. Set $\mathsf{A} := g_w(\mathcal{F}) \cap \mathcal{F}$ and $\mathsf{A}' := \mathcal{F}^{(w)} := g_w^{-1}(\mathsf{A}) \subseteq \mathcal{F}$, such that we are in the situation of Lemma 2.3.3 with $g = g_w|_{\mathsf{A}'} : \mathsf{A}' \to \mathsf{A}$. Overloading notation, we shall merely write g_w instead of $g_w|_{\mathsf{A}'}$. Let T be the identity on $\mathcal{F}$ and $S\colon \mathcal{F} \times \mathsf{O} \to \mathbb{R}$ be a strictly proper scoring rule. In the terminology of scoring functions, that means that S is a strictly $\mathcal{F}$-consistent scoring function for T. Then, Proposition 2.3.5 yields that

$$S^{(w)}\colon \mathsf{A} \times \mathsf{O} \to \mathbb{R}, \qquad (F, y) \mapsto S^{(w)}(F, y) = w(y)S(F, y)$$

is a strictly $\mathcal{F}$-consistent scoring function for $T^{(w)} = g_w\colon \mathcal{F}^{(w)} \to \mathsf{A}$. Finally, Lemma 2.3.3 asserts that

$$S^{(w)}_{g_w^{-1}}\colon \mathsf{A}' \times \mathsf{O} \to \mathbb{R}, \quad (F, y) \mapsto S^{(w)}_{g_w^{-1}}(F, y) = w(y)S(g_w(F), y) = w(y)S(F^{(w)}, y)$$

is a strictly $\mathcal{F}$-consistent scoring function for $T^{(w)}_{g_w^{-1}} = g_w^{-1} \circ g_w\colon \mathcal{F}^{(w)} \to 2^{\mathsf{A}'} = 2^{\mathcal{F}^{(w)}}$. In the terminology of Holzmann and Klar (2016) this means that $S^{(w)}_{g_w^{-1}}$ is proportionally locally proper. Indeed, For $F_1, F_2 \in \mathcal{F}^{(w)}$, the relation $F_2 \in T^{(w)}_{g_w^{-1}}(F_1)$ holds if and only if $g_w(F_1) = g_w(F_2)$. This in turn is equivalent to

$$w(y)\frac{f_1(y)}{\bar{w}(F_1)} = w(y)\frac{f_2(y)}{\bar{w}(F_2)}$$

for μ-almost-all $y \in \mathsf{O}$. But that means that $f_1 = \frac{\bar{w}(F_1)}{\bar{w}(F_2)} f_2$ on $\{w > 0\}$ μ-a.e.

Convexity of level sets is a necessary condition for elicitability. The result is classical in the literature and was first presented in Osband (1985, Proposition 2.5); see also Gneiting (2011, Theorem 6).

Proposition 2.3.7. *Let $T\colon \mathcal{F} \to \mathsf{A} \subseteq \mathbb{R}^k$ be an elicitable functional. Then for all $F_0, F_1 \in \mathcal{F}$ with $t := T(F_0) = T(F_1)$ and for all $\lambda \in (0,1)$ such that $F_\lambda := (1-\lambda)F_0 + \lambda F_1 \in \mathcal{F}$ it holds that $t = T(F_\lambda)$.*

Proof. Let $S\colon \mathsf{A} \times \mathsf{O} \to \mathbb{R}$ be a strictly $\mathcal{F}$-consistent scoring function for T. Let $F_0, F_1 \in \mathcal{F}$ with $t = T(F_0) = T(F_1)$ and $F_\lambda = (1-\lambda)F_0 + \lambda F_1 \in \mathcal{F}$ for some $\lambda \in (0,1)$. Let $x \in \mathsf{A}$, $x \neq t$. Then we have that

$$\begin{aligned}\bar{S}(t, F_\lambda) &= (1-\lambda)\bar{S}(t, F_0) + \lambda\bar{S}(t, F_1) \\ &< (1-\lambda)\bar{S}(x, F_0) + \lambda\bar{S}(x, F_1) = \bar{S}(x, F_\lambda).\end{aligned}$$

Since S is a strictly $\mathcal{F}$-consistent scoring function for T, we conclude that $t = T(F_\lambda)$. ∎

We remark that in a similar fashion one can see that convexity of the level sets of a functional is also necessary for its identifiability; see Lemma 4.1.13 for a slightly stronger result.

Steinwart et al. (2014) found that under some regularity conditions on the functional T such as its continuity and the fact that all distributions in $\mathcal{F}$ are dominated by some common measure, the convexity of the level sets is also a *sufficient* condition for the elicitability of T in the one-dimensional case $k = 1$. If one dispenses with the continuity of T, the convexity of the level sets fails to be a sufficient condition for elicitability, as shown for the mode-functional in Heinrich (2014). The proof of sufficiency in Steinwart et al. (2014) is done by a version of a separation theorem (actually, they show that the convexity of the level sets first implies the identifiability of the functional and then conclude that it is also elicitable). Hence, their proof cannot be generalized to the higher-dimensional case and it remains an open question, whether the sufficiency continues to hold for $k > 1$ under appropriate regularity conditions.

3. Higher Order Elicitability and Osband's Principle

This chapter consists of three sections. Section 3.1 is the main part and settles the most crucial results of this chapter, and, *a fortiori*, the most crucial results of Part I of this thesis. One particular achievement is Theorem 3.2 in Fissler and Ziegel (2016). Due to Gneiting (2011) we call it 'Osband's principle', since it originates from Osband's (1985) doctoral thesis, showing the intimate relation between strictly consistent scoring functions and identification functions. Our contribution is a generalization in that we formulated this relation on the level of expectations and not pointwise. Thus, we could relax the assumptions on the class $\mathcal{F}$ of distributions as well as smoothness assumptions substantially, making the result applicable to a larger class of functionals and scoring functions. Moreover, we established a kind of 'second order Osband's principle' in Fissler and Ziegel (2016, Corollary 3.3) exploiting the second order conditions related to the minimization problem of strictly consistent scoring functions. Osband's principle – also comprising the second order one – serves as a powerful tool in the derivation of the necessary form of strictly consistent scoring functions in a plenitude of different situations. One particular example for this and, at the same time, another major achievement is Fissler and Ziegel (2016, Theorem 5.2) where we could show that spectral risk measures with a finitely supported spectral measure are a component of an elicitable functional. Up to our knowledge, this is the first result of an elicitable functional having a non-elicitable component and which is not a bijection of another elicitable functional with elicitable components, such that the revelation principle does not apply. A consequence of particularly applied interest is Fissler and Ziegel (2016, Corollary 5.5) asserting that the pair (Value at Risk, Expected Shortfall), at the same level $\alpha \in (0, 1)$, is elicitable. The latter results gave rise to an observation made by Frongillo and Kash (2015b) that for a fixed elicitable functional T, the minimum of any strictly consistent scoring function – which defines another functional T' – is jointly elicitable with T. Section 3.3 is a brief summary of this observation. The chapter is complemented by Section 3.2 giving a characterization of the class of strict identification functions for a fixed functional exploiting a rationale very similar to Osband's principle.

3.1. Fissler and Ziegel (2016)

The content of this section is the article Fissler and Ziegel (2016) in the published version of the *Annals of Statistics* (`http://www.imstat.org/aos/`) which can be found on `http://projecteuclid.org/euclid.aos/1467894712`. It is followed by the online supplement of the article (`http://projecteuclid.org/euclid.aos/1467894712#supplemental`). Three preprint versions of this article can be found on `https://arxiv.org/abs/1503.08123`, and we refer to these versions by Fissler and Ziegel (2015).

The Annals of Statistics
2016, Vol. 44, No. 4, 1680–1707
DOI: 10.1214/16-AOS1439

HIGHER ORDER ELICITABILITY AND OSBAND'S PRINCIPLE[1]

BY TOBIAS FISSLER AND JOHANNA F. ZIEGEL

University of Bern

A statistical functional, such as the mean or the median, is called elicitable if there is a scoring function or loss function such that the correct forecast of the functional is the unique minimizer of the expected score. Such scoring functions are called strictly consistent for the functional. The elicitability of a functional opens the possibility to compare competing forecasts and to rank them in terms of their realized scores. In this paper, we explore the notion of elicitability for multi-dimensional functionals and give both necessary and sufficient conditions for strictly consistent scoring functions. We cover the case of functionals with elicitable components, but we also show that one-dimensional functionals that are not elicitable can be a component of a higher order elicitable functional. In the case of the variance, this is a known result. However, an important result of this paper is that spectral risk measures with a spectral measure with finite support are jointly elicitable if one adds the "correct" quantiles. A direct consequence of applied interest is that the pair (Value at Risk, Expected Shortfall) is jointly elicitable under mild conditions that are usually fulfilled in risk management applications.

1. Introduction. Point forecasts for uncertain future events are issued in a variety of different contexts such as business, government, risk-management or meteorology, and they are often used as the basis for strategic decisions. In all these situations, one has a random quantity Y with unknown distribution F. One is interested in a statistical property of F, that is a functional $T(F)$. Here, Y can be real-valued (GDP growth for next year), vector-valued (wind-speed, income from taxes for all cantons of Switzerland), functional-valued (path of the interchange rate Euro–Swiss franc over one day), or set-valued (area of rain tomorrow, area of influenza in a country). Likewise, also the functional T can have a variety of different sorts of values, among them the real- and vector-valued case (mean, vector of moments, covariance matrix, expectiles), the set-valued case (confidence regions) or also the functional-valued case (distribution functions). This article is concerned with the situation where Y is a d-dimensional random vector and T is a k-dimensional functional, thus also covering the real-valued case.

It is common to assess and compare competing point forecasts in terms of a loss function or *scoring function*. This is a function S such as the squared error or the

Received April 2015; revised January 2016.

[1] Supported by the SNF via Grant 152609.

MSC2010 subject classifications. Primary 62C99; secondary 91B06.

Key words and phrases. Consistency, decision theory, elicitability, Expected Shortfall, point forecasts, propriety, scoring functions, scoring rules, spectral risk measures, Value at Risk.

absolute error which is negatively oriented in the following sense: If the forecast $x \in \mathbb{R}^k$ is issued and the event $y \in \mathbb{R}^d$ materializes, the forecaster is *penalized* by the real value $S(x, y)$. In the presence of several different forecasters, one can compare their performances by ranking their realized scores. Hence, forecasters have an incentive to minimize their Bayes risk or *expected loss* $\mathbb{E}_F[S(x, Y)]$. Gneiting (2011) demonstrated impressively that scoring functions should be incentive compatible in that they should encourage the forecasters to issue truthful reports; see also Engelberg, Manski and Williams (2009), Murphy and Daan (1985). In other words, the choice of the scoring function S must be *consistent* with the choice of the functional T. We say a scoring function S is *strictly $\mathcal{F}$-consistent* for a functional T if $T(F)$ is the unique minimizer of the expected score $\mathbb{E}_F[S(x, Y)]$ for all $F \in \mathcal{F}$, where the class $\mathcal{F}$ of probability distributions is the domain of T. In some parts of the literature, strictly consistent scoring functions are called proper scoring rules. Our choice of terminology is in line with Gneiting (2011). Following Lambert, Pennock and Shoham (2008) and Gneiting (2011), we call a functional T with domain $\mathcal{F}$ *elicitable* if there exists a strictly $\mathcal{F}$-consistent scoring function for T.

The elicitability of a functional allows for regression, such as quantile regression and expectile regression [Koenker (2005), Newey and Powell (1987)] and for M-estimation [Huber (1964)]. Early work on elicitability is due to Osband (1985), Osband and Reichelstein (1985). More recent advances in the one-dimensional case, that is, $k = d = 1$ are due to Gneiting (2011), Lambert (2013), Steinwart et al. (2014) with the latter showing the intimate relation between elicitability and identifiability. Under mild conditions, many important functionals are elicitable such as moments, ratios of moments, quantiles and expectiles. However, there are also relevant functionals which are not elicitable such as variance, mode, or Expected Shortfall [Gneiting (2011), Heinrich (2014), Osband (1985), Weber (2006)].

With the so-called *revelation principle* Osband (1985) [see also Gneiting (2011), Theorem 4] was one of the first to show that a functional, albeit itself not being elicitable, can be a component of an elicitable vector-valued functional. The most prominent example in this direction is that the pair (mean, variance) is elicitable despite the fact that variance itself is not. However, it is crucial for the validity of the revelation principle that there is a bijection between the pair (mean, variance) and the first two moments. Until now, it appeared as an open problem if there are elicitable functionals with non-elicitable components other than those which can be connected to a functional with elicitable components via a bijection. Frongillo and Kash (2015) conjectured that this is generally not possible. We solve this open problem and can reject their conjecture: Corollary 5.5 shows that the pair (Value at Risk, Expected Shortfall) is elicitable, subject to mild regularity assumptions, improving a recent partial result of Acerbi and Székely (2014). To the best of our knowledge, we provide the first proof of this result in full generality. In fact, Corollary 5.4 demonstrates more generally that *spectral risk measures* with a spectral measure having finite support in $(0, 1]$ can be a component of an elicitable vector-valued functional. These results may lead to a new direction in the contemporary discussion about what risk measure is best in practice, and in particular about the importance of elicitability in risk measurement contexts [Acerbi and Székely (2014), Davis (2016), Embrechts and Hofert (2014), Emmer, Kratz and Tasche (2015)].

Complementing the question whether a functional is elicitable or not, it is interesting to determine the class of strictly consistent scoring functions for a functional, or at least to characterize necessary and sufficient conditions for the strict consistency of a scoring function. Most of the existing literature focuses on real-valued functionals meaning that $k = 1$. For the case $k > 1$, mainly linear functionals, that is, vectors of expectations of certain transformations, are classified where the only strictly consistent scoring functions are *Bregman functions* [Abernethy and Frongillo (2012), Banerjee, Guo and Wang (2005), Dawid and Sebastiani (1999), Osband and Reichelstein (1985), Savage (1971)]; for a general overview of the existing literature, we refer to Gneiting (2011). To the best of our knowledge, only Osband (1985), Lambert, Pennock and Shoham (2008) and Frongillo and Kash (2015) investigated more general cases of functionals, the latter also treating vectors of ratios of expectations as the first nonlinear functionals. In his doctoral thesis, Osband (1985) established a necessary representation for the first-order derivative of a strictly consistent scoring function with respect to the report x which connects it with identification functions. Following Gneiting (2011), we call results in the same flavor *Osband's principle*. Theorem 3.2 in this paper complements and generalizes Osband (1985), Theorem 2.1. Using our techniques, we retrieve the results mentioned above concerning the Bregman representation, however, under somewhat stronger regularity assumptions than the one in Frongillo and Kash (2015); see Proposition 4.4. On the other hand, we are able to treat a much broader class of functionals; see Proposition 4.2, Remark 4.5 and Theorem 5.2. In particular, we show that under mild richness assumptions on the class $\mathcal{F}$, any strictly $\mathcal{F}$-consistent scoring function for a vector of quantiles and/or expectiles is the sum of strictly $\mathcal{F}$-consistent one-dimensional scoring functions for each quantile/expectile; see Proposition 4.2.

The paper is organized as follows. In Section 2, we introduce notation and derive some basic results concerning the elicitability of k-dimensional functionals. Section 3 is concerned with *Osband's principle*, Theorem 3.2, and its immediate consequences. We investigate the situation where a functional is composed of elicitable components in Section 4, whereas Section 5 is dedicated to the elicitability of spectral risk measures. We end our article with a brief discussion; see Section 6. Most proofs are deferred to Section 7 and the supplementary material Fissler and Ziegel (2016).

2. Properties of higher order elicitability.

2.1. *Notation and definitions*. Following Gneiting (2011), we introduce a decision-theoretic framework for the evaluation of point forecasts. To this end, we introduce an *observation domain* $\mathsf{O} \subseteq \mathbb{R}^d$. We equip O with the Borel σ-algebra $\mathcal{O}$ using the induced topology of $\mathbb{R}^d$. We identify a Borel probability measure P on $(\mathsf{O}, \mathcal{O})$ with its cumulative distribution function (c.d.f.) $F_P : \mathsf{O} \to [0, 1]$ defined as $F_P(x) := P((-\infty, x] \cap \mathsf{O})$, where $(-\infty, x] = (-\infty, x_1] \times \cdots \times (-\infty, x_d]$ for $x = (x_1, \dots, x_d) \in \mathbb{R}^d$. Let $\mathcal{F}$ be a class of distribution functions on $(\mathsf{O}, \mathcal{O})$. Furthermore, for some integer $k \geq 1$, let $\mathsf{A} \subseteq \mathbb{R}^k$ be an *action domain*. To shorten notation, we usually write $F \in \mathcal{F}$ for a c.d.f. and also omit to mention the σ-algebra $\mathcal{O}$.

Let $T : \mathcal{F} \to \mathsf{A}$ be a functional. We introduce the notation $T(\mathcal{F}) := \{x \in \mathsf{A} : x = T(F) \text{ for some } F \in \mathcal{F}\}$. For a set $M \subseteq \mathbb{R}^k$, we will write $\operatorname{int}(M)$ for its interior with respect to $\mathbb{R}^k$, that is, $\operatorname{int}(M)$ is the biggest open set $U \subseteq \mathbb{R}^k$ such that $U \subseteq M$. The convex hull of M is defined as

$$\operatorname{conv}(M) := \left\{ \sum_{i=1}^n \lambda_i x_i \,\middle|\, n \in \mathbb{N}, x_1, \dots, x_n \in M, \lambda_1, \dots, \lambda_n > 0, \sum_{i=1}^n \lambda_i = 1 \right\}.$$

We say that a function $a : \mathsf{O} \to \mathbb{R}$ is $\mathcal{F}$-integrable if it is F-integrable for each $F \in \mathcal{F}$. A function $g : \mathsf{A} \times \mathsf{O} \to \mathbb{R}$ is $\mathcal{F}$-integrable if $g(x, \cdot)$ is $\mathcal{F}$-integrable for each $x \in \mathsf{A}$. If g is $\mathcal{F}$-integrable, we introduce the map

$$\bar{g} : \mathsf{A} \times \mathcal{F} \to \mathbb{R}, \qquad (x, F) \mapsto \bar{g}(x, F) = \int g(x, y) \, \mathrm{d}F(y).$$

Consequently, for fixed $F \in \mathcal{F}$ we can consider the function $\bar{g}(\cdot, F) : \mathsf{A} \to \mathbb{R}$, $x \mapsto \bar{g}(x, F)$, and for fixed $x \in \mathsf{A}$ we can consider the (linear) functional $\bar{g}(x, \cdot) : \mathcal{F} \to \mathbb{R}$, $F \mapsto \bar{g}(x, F)$.

If we fix $y \in \mathsf{O}$ and g is sufficiently smooth in its first argument, then for $m \in \{1, \dots, k\}$ we denote the mth partial derivative of the function $g(\cdot, y)$ with $\partial_m g(\cdot, y)$. More formally, we set

$$\partial_m g(\cdot, y) : \operatorname{int}(\mathsf{A}) \to \mathbb{R}, \qquad (x_1, \dots, x_k) \mapsto \frac{\partial}{\partial x_m} g(x_1, \dots, x_k, y).$$

We denote by $\nabla g(\cdot, y)$ the gradient of $g(\cdot, y)$ defined as $\nabla g(\cdot, y) := (\partial_1 g(\cdot, y), \dots, \partial_k g(\cdot, y))^\top$; and with $\nabla^2 g(\cdot, y) := (\partial_l \partial_m g(\cdot, y))_{l,m=1,\dots,k}$ the Hessian of $g(\cdot, y)$. *Mutatis mutandis*, we use the same notation for $\bar{g}(\cdot, F)$, $F \in \mathcal{F}$. We call a function on A differentiable if it is differentiable in $\operatorname{int}(\mathsf{A})$ and use the notation as given above. The restriction of a function f to some subset M of its domain is denoted by $f_{|M}$.

DEFINITION 2.1 (Consistency and elicitability). A *scoring function* is an $\mathcal{F}$-integrable function $S : \mathsf{A} \times \mathsf{O} \to \mathbb{R}$. It is said to be $\mathcal{F}$-*consistent* for a functional $T : \mathcal{F} \to \mathsf{A}$ if $\bar{S}(T(F), F) \leq \bar{S}(x, F)$ for all $F \in \mathcal{F}$ and for all $x \in \mathsf{A}$. Furthermore,

S is *strictly $\mathcal{F}$-consistent* for T if it is $\mathcal{F}$-consistent for T and if $\bar{S}(T(F), F) = \bar{S}(x, F)$ implies that $x = T(F)$ for all $F \in \mathcal{F}$ and for all $x \in \mathsf{A}$. Wherever it is convenient, we assume that $S(x, \cdot)$ is locally bounded for all $x \in \mathsf{A}$. A functional $T: \mathcal{F} \to \mathsf{A} \subseteq \mathbb{R}^k$ is called *k-elicitable*, if there exists a strictly $\mathcal{F}$-consistent scoring function for T.

DEFINITION 2.2 (Identification function). An *identification function* is an $\mathcal{F}$-integrable function $V: \mathsf{A} \times \mathsf{O} \to \mathbb{R}^k$. It is said to be an *$\mathcal{F}$-identification function* for a functional $T: \mathcal{F} \to \mathsf{A} \subseteq \mathbb{R}^k$ if $\bar{V}(T(F), F) = 0$ for all $F \in \mathcal{F}$. Furthermore, V is a *strict $\mathcal{F}$-identification function* for T if $\bar{V}(x, F) = 0$ holds if and only if $x = T(F)$ for all $F \in \mathcal{F}$ and for all $x \in \mathsf{A}$. Wherever it is convenient, we assume that $V(x, \cdot)$ is locally bounded for all $x \in \mathsf{A}$ and that $V(\cdot, y)$ is locally Lebesgue-integrable for all $y \in \mathsf{O}$. A functional $T: \mathcal{F} \to \mathsf{A} \subseteq \mathbb{R}^k$ is said to be *k-identifiable*, if there exists a strict $\mathcal{F}$-identification function for T.

If the dimension k is clear from the context, we say that a functional is elicitable (identifiable) instead of k-elicitable (k-identifiable).

REMARK 2.3. Depending on the class $\mathcal{F}$, some statistical functionals such as quantiles can be set-valued. In such situations, one can define $T: \mathcal{F} \to 2^{\mathsf{A}}$. Then a scoring function $S: \mathsf{A} \times \mathsf{O} \to \mathbb{R}$ is called (strictly) $\mathcal{F}$-consistent for T if $\bar{S}(t, F) \le \bar{S}(x, F)$ for all $x \in \mathsf{A}$, $F \in \mathcal{F}$ and $t \in T(F)$ [with equality implying $x \in T(F)$]. The definition of a (strict) $\mathcal{F}$-identification function for T can be generalized *mutatis mutandis*. Many of the results of this paper can be extended to the case of set-valued functionals—at the cost of a more involved notation and analysis. To allow for a clear presentation, we confine ourselves to functionals with values in $\mathbb{R}^k$ in this paper.

2.2. *Basic results*. The first lemma gives a useful equivalent characterization of strict consistency. Its proof is a direct consequence of the definition.

LEMMA 2.4. *A scoring function $S: \mathsf{A} \times \mathsf{O} \to \mathbb{R}$ is strictly $\mathcal{F}$-consistent for $T : \mathcal{F} \to \mathsf{A} \subseteq \mathbb{R}^k$ if and only if the function*

$$\psi: D \to \mathbb{R}, \qquad s \mapsto \bar{S}(t + sv, F)$$

has a global unique minimum at $s = 0$ for all $F \in \mathcal{F}$, $t = T(F)$ and $v \in \mathbb{S}^{k-1}$ where $D = \{s \in \mathbb{R}: t + sv \in \mathsf{A}\}$.

The following result follows directly from the definition of consistency (Definition 2.1). However, it is crucial to understand many of the results of this paper.

LEMMA 2.5. *Let $T: \mathcal{F} \to \mathsf{A} \subseteq \mathbb{R}^k$ be a functional with a strictly $\mathcal{F}$-consistent scoring function $S: \mathsf{A} \times \mathsf{O} \to \mathbb{R}$. Then the following two assertions hold*:

(i) *Let* $\mathcal{F}' \subseteq \mathcal{F}$ *and* $T_{|\mathcal{F}'}$ *be the restriction of* T *to* $\mathcal{F}'$. *Then* S *is also a strictly* $\mathcal{F}'$*-consistent scoring function for* $T_{|\mathcal{F}'}$.

(ii) *Let* $\mathsf{A}' \subseteq \mathsf{A}$ *such that* $T(\mathcal{F}) \subseteq \mathsf{A}'$ *and* $S_{|\mathsf{A}'\times\mathsf{O}}$ *be the restriction of* S *to* $\mathsf{A}' \times \mathsf{O}$. *Then* $S_{|\mathsf{A}'\times\mathsf{O}}$ *is also a strictly* $\mathcal{F}$*-consistent scoring function for* T.

The main results of this paper consist of necessary and sufficient conditions for the strict $\mathcal{F}$-consistency of a scoring function S for some functional T. What are the consequences of Lemma 2.5 for such conditions? Assume that we start with a functional $T' : \mathcal{F}' \to \mathsf{A}' \subseteq \mathbb{R}^k$ and deduce some *necessary* conditions for a scoring function $S' : \mathsf{A}' \times \mathsf{O} \to \mathbb{R}$ to be strictly $\mathcal{F}'$-consistent for T'. Then Lemma 2.5(i) implies that these conditions continue to be necessary conditions for the strict $\mathcal{F}$-consistency of S' for $T : \mathcal{F} \to \mathsf{A}'$ where $\mathcal{F}' \subseteq \mathcal{F}$, and T is some extension of T' such that $T(\mathcal{F}) \subseteq \mathsf{A}'$. On the other hand, Lemma 2.5(ii) implies that the necessary conditions for the strict $\mathcal{F}'$-consistency of a scoring function $S' : \mathsf{A}' \times \mathsf{O} \to \mathbb{R}$ continue to be necessary conditions for the strict $\mathcal{F}'$-consistency of $S : \mathsf{A} \times \mathsf{O} \to \mathbb{R}$ for T', where $\mathsf{A}' \subseteq \mathsf{A}$ and S is some extension of S'.

Summarizing, given a functional $T : \mathcal{F} \to \mathsf{A}$, a collection of necessary conditions for the strict $\mathcal{F}$-consistency of scoring functions for T is the more restrictive the smaller the class $\mathcal{F}$ and the smaller the set A is [provided that $T(\mathcal{F}) \subseteq \mathsf{A}$, of course]. Hence, in the forthcoming results concerning necessary conditions, it is no loss of generality to just mention which distributions must necessarily be in the class $\mathcal{F}$ to guarantee the validity of the results. Furthermore, it is no loss of generality to make the assumption that T is surjective, so $\mathsf{A} = T(\mathcal{F})$.

Some of the subsequent results also provide *sufficient* conditions for the strict $\mathcal{F}$-consistency of a scoring function $S : \mathsf{A} \times \mathsf{O} \to \mathbb{R}$ for a functional $T : \mathcal{F} \to \mathsf{A}$. Those results are the stronger the bigger the class $\mathcal{F}$ and the bigger the set A is. For the notion of elicitability, this means that the assertion that a functional $T : \mathcal{F} \to \mathsf{A}$ is elicitable is also the stronger the bigger the class $\mathcal{F}$ and the bigger the set A is. To demonstrate this reasoning, observe that if the functional $T : \mathcal{F} \to \mathsf{A}$ is degenerate in the sense that it is constant, so $T \equiv t$ for some $t \in \mathsf{A}$ (which covers the particular case that $\mathcal{F}$ contains only one element), then T is automatically elicitable with a strictly $\mathcal{F}$-consistent scoring function $S : \mathsf{A} \times \mathsf{O} \to \mathbb{R}$, defined as $S(x, y) := \|x - t\|$.

As a last result in this section, we present the intuitive observation that a vector of elicitable functionals itself is elicitable.

LEMMA 2.6. *Let* $k_1, \dots, k_l \geq 1$ *and let* $T_m : \mathcal{F} \to \mathsf{A}_m \subseteq \mathbb{R}^{k_m}$ *be a* k_m*-elicitable functional*, $m \in \{1, \dots, l\}$. *Then the functional* $T = (T_1, \dots, T_l)$: $\mathcal{F} \to \mathsf{A}$ *is* k*-elicitable where* $k = k_1 + \cdots + k_l$ *and* $\mathsf{A} = \mathsf{A}_1 \times \cdots \times \mathsf{A}_l \subseteq \mathbb{R}^k$.

PROOF. For $m \in \{1, \dots, l\}$, let $S_m : \mathsf{A}_m \times \mathsf{O} \to \mathbb{R}$ be a strictly $\mathcal{F}$-consistent scoring function for T_m. Let $\lambda_1, \dots, \lambda_l > 0$ be positive real numbers. Then

$$S(x_1, \dots, x_l, y) := \sum_{m=1}^{l} \lambda_m S_m(x_m, y) \tag{2.1}$$

is a strictly $\mathcal{F}$-consistent scoring function for T. $\square$

A particularly simple and relevant case of Lemma 2.6 is the situation $k_1 = \cdots = k_l = 1$ such that $k = l$. It is an interesting question whether the scoring functions of the form (2.1) are the only strictly $\mathcal{F}$-consistent scoring functions for T, which amounts to the question of *separability* of scoring rules that was posed by Frongillo and Kash (2015). The answer is generally negative. As mentioned in the Introduction, it is known that all Bregman functions elicit T, if the components of T are all expectations of transformations of Y [Abernethy and Frongillo (2012), Banerjee, Guo and Wang (2005), Dawid and Sebastiani (1999), Osband and Reichelstein (1985), Savage (1971)] or ratios of expectations with the same denominator [Frongillo and Kash (2015)]; see also Proposition 4.4. However, for other situations, such as a combination of different quantiles and/or expectiles, the answer is positive; see Proposition 4.2. These results rely on "Osband's principle" which gives necessary conditions for scoring functions to be strictly $\mathcal{F}$-consistent for a given functional T; see Section 3.

There are more involved functionals that are k-elicitable than combinations of k 1-elicitable components. An immediate example that is the pair (expectation, variance) which can be obtained through the revelation principle from the 2-elicitable pair (expectation, second moment). In Section 5, we show that the concept of k-elicitability is also not restricted to functionals that can be obtained by combining Lemma 2.6 and the revelation principle. It is shown in Weber (2006), Example 3.4 and Gneiting (2011), Theorem 11, that the coherent risk measure Expected Shortfall at level α, $\alpha \in (0, 1)$, does not have convex level sets and is therefore not elicitable. In contrast, we show in Corollary 5.5 that the pair (Value at Risk$_\alpha$, Expected Shortfall$_\alpha$) is 2-elicitable relative to the class of distributions on $\mathbb{R}$ with finite first moment and unique α-quantiles. This refutes Proposition 2.3 of Osband (1985); see Remark 5.3 for a discussion.

3. Osband's principle.

In this section, we give necessary conditions for the strict $\mathcal{F}$-consistency of a scoring function S for a functional $T : \mathcal{F} \to \mathsf{A}$. In the light of Lemma 2.5 and the discussion thereafter, we have to impose some richness conditions on the class $\mathcal{F}$ as well as on the "variability" of the functional T. To this end, we establish a link between strictly $\mathcal{F}$-consistent scoring functions and strict $\mathcal{F}$-identification functions. We illustrate the idea in the one-dimensional case. Let $\mathcal{F}$ be a class of distribution functions on $\mathbb{R}$, $T : \mathcal{F} \to \mathbb{R}$ a functional and $S : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ a strictly $\mathcal{F}$-consistent scoring function for T. Furthermore, let $V : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be an oriented strict $\mathcal{F}$-identification function for T. Then, under certain regularity conditions, there is a nonnegative function $h : \mathbb{R} \to \mathbb{R}$ such that

$$\frac{\mathrm{d}}{\mathrm{d}x} S(x, y) = h(x) V(x, y). \tag{3.1}$$

If we naïvely swap differentiation and expectation and h does not vanish, the form (3.1) plus the identification property of V are sufficient for the first order condition on $\bar{S}(\cdot, F)$, $F \in \mathcal{F}$, to be satisfied and the orientation of V (see Remark 4.1) as well as the fact that h is positive are sufficient for $\bar{S}(\cdot, F)$ to satisfy the second-order condition for strict $\mathcal{F}$-consistency. So the really interesting part is to show that the form given in (3.1) is necessary for the strict $\mathcal{F}$-consistency of a scoring function for T.

The idea of this characterization originates from Osband (1985). He gives a characterization including $\mathbb{R}^k$-valued functionals, but for his proof he assumes that $\mathcal{F}$ contains all distributions with finite support. This is not a problem *per se*, but in the light of Lemma 2.5 and the discussion thereafter it is desirable to weaken this assumption. In particular, the results in Section 5 on spectral risk measures cannot be derived if $\mathcal{F}$ has to contain all distributions with finite support. Relying on a functional space extension of the Kuhn–Tucker theorem, Osband (1985) conjectures that his characterization continues to hold if $\mathcal{F}$ consists only of absolutely continuous distributions, but we do not believe that his approach is feasible in this case. In Steinwart et al. (2014), Theorem 5, there is a rigorous statement of Osband's principle for the one-dimensional functionals where the distributions in $\mathcal{F}$ must be absolutely continuous with respect to some finite measure. We shall give a proof in the setting of an $\mathbb{R}^k$-valued functional that does not have to specify the kinds of distributions in $\mathcal{F}$, but only uses the following (minimal) collection of regularity assumptions. To this end, we apply a similar technique as in the proof of Osband (1985), Lemma 2.2, which is based on a finite-dimensional argument.

Let $\mathcal{F}$ be a class of distribution functions on $\mathsf{O} \subseteq \mathbb{R}^d$. Fix a functional $T: \mathcal{F} \to \mathsf{A} \subseteq \mathbb{R}^k$, an identification function $V: \mathsf{A} \times \mathsf{O} \to \mathbb{R}^k$ and a scoring function $S: \mathsf{A} \times \mathsf{O} \to \mathbb{R}$.

ASSUMPTION (V1). Let $\mathcal{F}$ be a convex class of distributions functions on $\mathsf{O} \subseteq \mathbb{R}^d$ and assume that for every $x \in \operatorname{int}(\mathsf{A})$ there are $F_1, \dots, F_{k+1} \in \mathcal{F}$ such that

$$0 \in \operatorname{int}\bigl(\operatorname{conv}\bigl(\bigl\{\bar{V}(x, F_1), \dots, \bar{V}(x, F_{k+1})\bigr\}\bigr)\bigr).$$

REMARK 3.1. Assumption (V1) implies that for every $x \in \operatorname{int}(\mathsf{A})$ there are $F_1, \dots, F_k \in \mathcal{F}$ such that the vectors $\bar{V}(x, F_1), \dots, \bar{V}(x, F_k)$ are linearly independent.

Assumption (V1) ensures that the class $\mathcal{F}$ is "rich" enough meaning that the functional T varies sufficiently in order to derive a necessary form of the scoring function S in Theorem 3.2. Assumptions like (V1) are classical in the literature. For the case of k-elicitability, Osband (1985) assumes that $0 \in \operatorname{int}(\operatorname{conv}(\{V(x, y): y \in \mathsf{O}\}))$. Steinwart et al. (2014), Definition 8 and Lambert (2013) treat the case $k = 1$ and work under the assumption that the functional is *strictly locally nonconstant* which implies assumption (V1) if the functional is identifiable.

ASSUMPTION (V2). For every $F \in \mathcal{F}$, the function $\bar{V}(\cdot, F)\colon \mathsf{A} \to \mathbb{R}^k$, $x \mapsto \bar{V}(x, F)$, is continuous.

ASSUMPTION (V3). For every $F \in \mathcal{F}$, the function $\bar{V}(\cdot, F)$ is continuously differentiable.

If the function $x \mapsto V(x, y)$, $y \in \mathsf{O}$, is continuous (continuously differentiable), assumption (V2) [assumption (V3)] is satisfied, and it is equivalent to (V2) [(V3)] if $\mathcal{F}$ contains all measures with finite support. However, (V2) and (V3) are much weaker requirements if we move away from distributions with finite support. To illustrate this fact, let $k, d = 1$ and $V(x, y) = \mathbb{1}\{y \le x\} - \alpha$, $\alpha \in (0, 1)$, which is a strict $\mathcal{F}$-identification function for the α-quantile. Of course, $V(\cdot, y)$ is not continuous. But if $\mathcal{F}$ contains only probability distributions F that have a continuous derivative $f = F'$, then $\bar{V}(x, F) = F(x) - \alpha$ and $(\mathrm{d}/\mathrm{d}x)\bar{V}(x, F) = f(x)$ and V satisfies (V2) and (V3). The following assumptions (S1) and (S2) are similar conditions as (V2) and (V3) but for scoring functions instead of identification functions.

ASSUMPTION (S1). For every $F \in \mathcal{F}$, the function $\bar{S}(\cdot, F)\colon \mathsf{A} \to \mathbb{R}$, $x \mapsto \bar{S}(x, F)$, is continuously differentiable.

ASSUMPTION (S2). For every $F \in \mathcal{F}$, the function $\bar{S}(\cdot, F)$ is continuously differentiable and the gradient is locally Lipschitz continuous. Furthermore, $\bar{S}(\cdot, F)$ is twice continuously differentiable at $t = T(F) \in \mathrm{int}(\mathsf{A})$.

Note that assumption (S2) implies that the gradient of $\bar{S}(\cdot, F)$ is (totally) differentiable for almost all $x \in \mathsf{A}$ by Rademacher's theorem, which in turn indicates that the Hessian of $\bar{S}(\cdot, F)$ exists for almost all $x \in \mathsf{A}$ and is symmetric by Schwarz's theorem; see Grauert and Fischer (1978), page 57.

THEOREM 3.2 (Osband's principle). *Let $T\colon \mathcal{F} \to \mathsf{A} \subseteq \mathbb{R}^k$ be a surjective, elicitable and identifiable functional with a strict $\mathcal{F}$-identification function $V\colon \mathsf{A} \times \mathsf{O} \to \mathbb{R}^k$ and a strictly $\mathcal{F}$-consistent scoring function $S\colon \mathsf{A} \times \mathsf{O} \to \mathbb{R}$. If the assumptions (V1) and (S1) hold, then there exists a matrix-valued function $h\colon \mathrm{int}(\mathsf{A}) \to \mathbb{R}^{k\times k}$ such that for $l \in \{1, \dots, k\}$*

$$\partial_l \bar{S}(x, F) = \sum_{m=1}^{k} h_{lm}(x)\bar{V}_m(x, F) \tag{3.2}$$

for all $x \in \mathrm{int}(\mathsf{A})$ and $F \in \mathcal{F}$. If in addition, assumption (V2) holds, then h is continuous. Under the additional assumptions (V3) and (S2), the function h is locally Lipschitz continuous.

Under the conditions of Theorem 3.2, equation (3.2) gives a characterization of the partial derivatives of the expected score. If we impose more smoothness assumptions on the expected score, we are also able to give a characterization of the second-order derivatives of the expected score.

COROLLARY 3.3. *For a surjective, elicitable and identifiable functional* $T\colon \mathcal{F} \to \mathsf{A} \subseteq \mathbb{R}^k$ *with a strict* $\mathcal{F}$*-identification function* $V\colon \mathsf{A} \times \mathsf{O} \to \mathbb{R}^k$ *and a strictly* $\mathcal{F}$*-consistent scoring function* $S\colon \mathsf{A} \times \mathsf{O} \to \mathbb{R}$ *that satisfy assumptions* (V1), (V3) *and* (S2), *we have the following identities for the second-order derivatives*:

$$\begin{aligned}\partial_m \partial_l \bar{S}(x, F) &= \sum_{i=1}^{k} \partial_m h_{li}(x)\bar{V}_i(x, F) + h_{li}(x)\partial_m \bar{V}_i(x, F) \\ &= \sum_{i=1}^{k} \partial_l h_{mi}(x)\bar{V}_i(x, F) + h_{mi}(x)\partial_l \bar{V}_i(x, F) = \partial_l \partial_m \bar{S}(x, F),\end{aligned} \tag{3.3}$$

for all $l, m \in \{1, \dots, k\}$, *for all* $F \in \mathcal{F}$ *and almost all* $x \in \mathrm{int}(\mathsf{A})$, *where* h *is the matrix-valued function appearing at* (3.2). *In particular*, (3.3) *holds for* $x = T(F) \in \mathrm{int}(\mathsf{A})$.

Theorem 3.2 and Corollary 3.3 establish necessary conditions for strictly $\mathcal{F}$-consistent scoring functions on the level of the expected scores. If the class $\mathcal{F}$ is rich enough and the scoring and identification function smooth enough in the following sense, we can also deduce a necessary condition for S which holds pointwise.

ASSUMPTION (F1). For every $y \in \mathsf{O}$, there exists a sequence $(F_n)_{n\in\mathbb{N}}$ of distributions $F_n \in \mathcal{F}$ that converges weakly to the Dirac-measure δ_y such that the support of F_n is contained in a compact set K for all n.

ASSUMPTION (VS1). Suppose that the complement of the set

$$C := \{(x, y) \in \mathsf{A} \times \mathsf{O} \mid V(x, \cdot) \text{ and } S(x, \cdot) \text{ are continuous at the point } y\}$$

has $(k + d)$-dimensional Lebesgue measure zero.

PROPOSITION 3.4. *Assume that* $\mathrm{int}(\mathsf{A}) \subseteq \mathbb{R}^k$ *is a star domain and let* $T\colon \mathcal{F} \to \mathsf{A}$ *be a surjective, elicitable and identifiable functional with a strict* $\mathcal{F}$*-identification function* $V\colon \mathsf{A} \times \mathsf{O} \to \mathbb{R}^k$ *and a strictly* $\mathcal{F}$*-consistent scoring function* $S\colon \mathsf{A} \times \mathsf{O} \to \mathbb{R}$. *Suppose that assumptions* (V1), (V2), (S1), (F1) *and* (VS1) *hold. Let* h *be the matrix valued function appearing at* (3.2). *Then the scoring function* S *is necessarily of the form*

$$\begin{aligned}S(x, y) = \sum_{r=1}^{k}\sum_{m=1}^{k} \int_{z_r}^{x_r} & h_{rm}(x_1, \dots, x_{r-1}, v, z_{r+1}, \dots, z_k) \\ & \times V_m(x_1, \dots, x_{r-1}, v, z_{r+1}, \dots, z_k, y)\, \mathrm{d}v + a(y)\end{aligned} \tag{3.4}$$

for almost all $(x, y) \in \mathsf{A} \times \mathsf{O}$, *for some star point* $z = (z_1, \dots, z_k) \in \operatorname{int}(\mathsf{A})$ *and some* $\mathcal{F}$*-integrable function* $a: \mathsf{O} \to \mathbb{R}$. *On the level of the expected score* $\bar{S}(x, F)$, *equation* (3.4) *holds for all* $x \in \operatorname{int}(A)$, $F \in \mathcal{F}$.

While Theorem 3.2, Corollary 3.3 and Proposition 3.4 only establish necessary conditions for strictly $\mathcal{F}$-consistent scoring functions for some functional T, often they guide a way how to construct strictly $\mathcal{F}$-consistent scoring functions starting with a strict $\mathcal{F}$-identification function V for T.

For the one-dimensional case, one can use the fact that, subject to some mild regularity conditions, if V is a strict $\mathcal{F}$-identification function, then either V or $-V$ is oriented; see Remark 4.1. Supposing that V is oriented, we can choose any strictly positive function $h: \mathsf{A} \to \mathbb{R}$ to get the derivative of a strictly $\mathcal{F}$-consistent scoring function. Then integration yields the desired strictly $\mathcal{F}$-consistent scoring function.

Establishing *sufficient* conditions for scoring functions to be strictly $\mathcal{F}$-consistent for T is generally more involved in the case $k > 1$. First of all, working under assumption (S2), the symmetry of the Hessian $\nabla^2 \bar{S}(x, F)$ imposes strong necessary conditions on the functions h_{lm}; see, for example, Proposition 4.2 which treats the case where all components of the functional $T = (T_1, \dots, T_k)$ are elicitable and identifiable. The example of spectral risk measures is treated in Section 5. Second, (3.2) and (3.3) are necessary conditions for $\bar{S}(x, F)$ having a *local* minimum in $x = T(F)$, $F \in \mathcal{F}$. Even if we additionally suppose that the Hessian $\nabla^2 \bar{S}(x, F)$ is strictly positive definite at $x = T(F)$, this is a sufficient condition only for a *local* minimum at $x = T(F)$, but does not provide any information concerning a *global* minimum. Consequently, even if the functions h_{lm} satisfy (3.3), one must verify the strict consistency of the scoring function on a case by case basis. This can often be done by showing that the one-dimensional functions $\mathbb{R} \to \mathbb{R}$, $s \mapsto \bar{S}(t + sv, F)$, with $t = T(F)$, have a global minimum in $s = 0$ for all $v \in \mathbb{S}^{k-1}$ and for all $F \in \mathcal{F}$.

4. Functionals with elicitable components. Suppose that the functional $T = (T_1, \dots, T_k): \mathcal{F} \to \mathsf{A} \subseteq \mathbb{R}^k$ consists of 1-elicitable components T_m. As prototypical examples of such 1-elicitable components, we consider the functionals given in Table 1 where we implicitly assume that $\mathsf{O} \subseteq \mathbb{R}$ if a quantile or an expectile are a part of T. If V_m are strict $\mathcal{F}$-identification functions for T_m then $V: \mathsf{A} \times \mathbb{R} \to \mathbb{R}^k$ with

$$V(x_1, \dots, x_k, y) = \big(V_1(x_1, y), \dots, V_k(x_k, y)\big)^\top \tag{4.1}$$

is a strict $\mathcal{F}$-identification function for T. Under (V3), the partial derivatives of $\bar{V}(x, F)$, $x \in \mathsf{A}$ and $F \in \mathcal{F}$ exist, and if the class $\mathcal{F}$ is sufficiently rich T (or some subset of its components) often fulfills the following assumption.

TABLE 1
Strict identification functions for $k = 1$; see Gneiting (2011), Table 9

Functional	**Strict identification function**
Ratio $\mathbb{E}_F[p(Y)]/\mathbb{E}_F[q(Y)]$	$V(x, y) = xq(y) - p(y)$
α-Quantile	$V(x, y) = \mathbb{1}\{y \le x\} - \alpha$
τ-Expectile	$V(x, y) = 2\|\mathbb{1}\{y \le x\} - \tau\|(x - y)$

ASSUMPTION (V4). Let assumption (V3) hold. For all $r \in \{1, \ldots, k\}$ and for all $t \in \mathrm{int}(\mathsf{A}) \cap T(\mathcal{F})$ there are $F_1, F_2 \in T^{-1}(\{t\})$ such that

$$\partial_l \bar{V}_l(t, F_1) = \partial_l \bar{V}_l(t, F_2) \qquad \forall l \in \{1, \ldots, k\} \setminus \{r\}, \qquad \partial_r \bar{V}_r(t, F_1) \ne \partial_r \bar{V}_r(t, F_2).$$

The following proposition gives a characterization of the class of strictly $\mathcal{F}$-consistent scoring functions under (V4). In particular, the result covers vectors of different quantiles and/or different expectiles (with the exception of the 1/2-expectile), thus answering a question raised in Gneiting and Raftery (2007), page 370.

One relevant exception when (V4) is not satisfied is when T is a ratio of expectations with the same denominator, that is, $q_m = q$ for all m. We treat this case in Proposition 4.4 below.

REMARK 4.1. Steinwart et al. (2014) introduced the notion of an *oriented* strict $\mathcal{F}$-identification function for the case $k = 1$ and $d = 1$. They say that $V: \mathsf{A} \times \mathsf{O} \to \mathbb{R}$ is an oriented strict $\mathcal{F}$-identification function for the functional $T: \mathcal{F} \to \mathsf{A}$ if V is a strict $\mathcal{F}$-identification function for T and, moreover, $\bar{V}(x, F) > 0$ if and only if $x > T(F)$ for all $F \in \mathcal{F}$ and for all $x \in \mathsf{A}$.

PROPOSITION 4.2. *Let $T_m: \mathcal{F} \to \mathsf{A}_m \subseteq \mathbb{R}$ be 1-elicitable and 1-identifiable functionals with oriented strict $\mathcal{F}$-identification functions $V_m: \mathsf{A}_m \times \mathsf{O} \to \mathbb{R}$ for $m \in \{1, \ldots, k\}$. Define $T = (T_1, \ldots, T_k)$ with identification function V as at (4.1) and a strictly $\mathcal{F}$-consistent scoring function $S: \mathsf{A} \times \mathsf{O} \to \mathbb{R}$ with $\mathsf{A} := T(\mathcal{F}) \subseteq \mathsf{A}_1 \times \cdots \times \mathsf{A}_k$. Suppose that* $\mathrm{int}(\mathsf{A})$ *is a star domain, and assumptions (V1), (V3), (V4), (S2) hold. Define $\mathsf{A}'_m := \{x_m: \exists (z_1, \ldots, z_k) \in \mathrm{int}(A), z_m = x_m\}$.*

(i) *Let $h : \mathrm{int}(\mathsf{A}) \to \mathbb{R}^{k \times k}$ be the function given at (3.2). Then there are functions $g_m: \mathsf{A}'_m \to \mathbb{R}$, $g_m > 0$, such that $h_{mm}(x_1, \ldots, x_k) = g_m(x_m)$ for all $m \in \{1, \ldots, k\}$ and $(x_1, \ldots, x_k) \in \mathrm{int}(\mathsf{A})$ and*

$$h_{rl}(x) = 0 \tag{4.2}$$

for all $r, l \in \{1, \ldots, k\}$, $l \ne r$, and for all $x \in \mathrm{int}(\mathsf{A})$.

(ii) *Assume that (F1) and (VS1) hold. Then S is strictly $\mathcal{F}$-consistent for T if and only if it is of the form*

$$S(x_1, \dots, x_k, y) = \sum_{m=1}^{k} S_m(x_m, y), \tag{4.3}$$

for almost all $(x, y) \in \mathsf{A} \times \mathsf{O}$, where $S_m : \mathsf{A}_m \times \mathsf{O} \to \mathbb{R}$, $m \in \{1, \dots, k\}$, are strictly $\mathcal{F}$-consistent scoring functions for T_m.

REMARK 4.3. Lambert, Pennock and Shoham (2008), Theorem 5, show that a scoring function is accuracy-rewarding if and only if it is the sum of strictly consistent scoring functions for each component. Their assumptions are different from Proposition 4.2(ii). For example, they assume that all distributions in $\mathcal{F}$ have finite support, and that scoring functions are twice continuously differentiable. Therefore, despite the same form of the characterization in (4.3), neither result implies the other. However, the key components of the proofs of both results is to show that the cross-derivatives of the expected scoring functions are zero. This implies then a decomposition as in (4.3). The converse is trivial in both cases.

If T is a ratio of expectations with the same denominator, it is well known that the class of strictly $\mathcal{F}$-consistent scoring functions is bigger than the one given in Proposition 4.2(ii).

PROPOSITION 4.4. *Let $T : \mathcal{F} \to \mathsf{A} \subseteq \mathbb{R}^k$ be a ratio of expectations with the same denominator, that is, $T(F) = \mathbb{E}_F[p(Y)]/\mathbb{E}_F[q(Y)]$ for some $\mathcal{F}$-integrable functions $p : \mathsf{O} \to \mathbb{R}^k$, $q : \mathsf{O} \to \mathbb{R}$. Assume that $\bar{q}(F) > 0$ for all $F \in \mathcal{F}$ and let $V : \mathsf{A} \times \mathsf{O} \to \mathbb{R}^k$, $V(x, y) = q(y)x - p(y)$. Let $S : \mathsf{A} \times \mathsf{O} \to \mathbb{R}$ be a strictly $\mathcal{F}$-consistent scoring function for T and $h : \mathrm{int}(\mathsf{A}) \to \mathbb{R}^{k \times k}$ be the function given at (3.2). Suppose that T is surjective, and assumptions (V1), (V3), (S2) hold.*

(i) *It holds that*

$$\partial_l h_{rm}(x) = \partial_r h_{lm}(x), \qquad h_{rl}(x) = h_{lr}(x) \tag{4.4}$$

for all $r, l, m \in \{1, \dots, k\}$, $l \neq r$, where the first identity holds for almost all $x \in \mathrm{int}(\mathsf{A})$ and the second identity for all $x \in \mathrm{int}(\mathsf{A})$. Moreover, the matrix $(h_{rl}(x))_{r,l=1,\dots,k}$ is positive definite for all $x \in \mathrm{int}(A)$.

(ii) *Let $\mathrm{int}(\mathsf{A})$ be a star domain and assume that (F1) and (VS1) hold. Then S is strictly $\mathcal{F}$-consistent for T if and only if it is of the form*

$$S(x, y) = -\phi(x)q(y) + \sum_{m=1}^{k} \big(q(y)x_m - p_m(y)\big)\partial_m\phi(x) + a(y), \tag{4.5}$$

with

$$\phi(x) = \sum_{r=1}^{k} \int_{z_r}^{x_r} \int_{z_r}^{v} h_{rr}(x_1, \dots, x_{r-1}, w, z_{r+1}, \dots, z_k)\, \mathrm{d}w\, \mathrm{d}v, \tag{4.6}$$

for almost all $(x, y) \in \mathsf{A} \times \mathsf{O}$, *where* $(z_1, \dots, z_k) \in \mathrm{int}(\mathsf{A})$ *is some star point and* $a: \mathsf{O} \to \mathbb{R}$ *is* $\mathcal{F}$*-integrable. Moreover,* ϕ *has Hessian* h *and is strictly convex.*

Part (ii) of this proposition recovers results of Abernethy and Frongillo (2012), Banerjee, Guo and Wang (2005), Osband and Reichelstein (1985) if $q \equiv 1$, which show that all consistent scoring functions for vectors of expectations are so-called *Bregman functions*, that is, functions of the form (4.5) with $q \equiv 1$ and a convex function ϕ. Frongillo and Kash (2015), Theorem 13, also treat the case of more general functions q.

REMARK 4.5. One might wonder about necessary conditions on the matrix-valued function h in the flavor of Propositions 4.2(i) and 4.4(i) if the k components of the functional T can be regrouped into (a) a new functional $T_1': \mathcal{F} \to \mathsf{A}_1' \subset \mathbb{R}^{k_1'}$ with an oriented strict $\mathcal{F}$-identification function $V_1': \mathsf{A}_1' \times \mathsf{O} \to \mathbb{R}^{k_1'}$ which satisfies assumption (V4), and (b) several, say l, new functionals $T_m': \mathcal{F} \to \mathsf{A}_{k_m'}' \subseteq \mathbb{R}^{k_m'}$, $m \in \{2, \dots, l+1\}$ which are ratios of expectations with the same denominator, and $k_1' + \cdots + k_{l+1}' = k$. We can apply the propositions to obtain necessary conditions for each of the $(k_m' \times k_m')$-valued functions h_m', $m \in \{1, \dots, l+1\}$. Applying Lemma 2.6, we get a possible choice for a strictly $\mathcal{F}$-consistent scoring function S for T. On the level of the $k \times k$-valued function h associated to S this means that h is a block diagonal matrix of the form $\mathrm{diag}(h_1', \dots, h_{l+1}')$. But what about the necessity of this form? Indeed, if we assume that the blocks in (b) have maximal size (or equivalently that l is minimal) then one can verify that h must be necessarily of the block diagonal form described above.

5. Spectral risk measures. Risk measures are a common tool to measure the risk of a financial position Y. A risk measure is usually defined as a mapping ρ from some space of random variables, for example, L^∞, to the real line. Arguably, the most common risk measure in practice is *Value at Risk* at level α (VaR_α) which is the generalized α-quantile $F^{-1}(\alpha)$, that is,

$$\mathrm{VaR}_\alpha(Y) := F^{-1}(\alpha) := \inf\{x \in \mathbb{R}: F(x) \geq \alpha\},$$

where F is the distribution function of Y. An important alternative to VaR_α is *Expected Shortfall* at level α (ES_α) (also known under the names Conditional Value at Risk or Average Value at Risk). It is defined as

$$\mathrm{ES}_\alpha(Y) := \frac{1}{\alpha} \int_0^\alpha \mathrm{VaR}_u(Y)\,\mathrm{d}u, \qquad \alpha \in (0, 1], \tag{5.1}$$

and $\mathrm{ES}_0(Y) = \operatorname{ess\,inf} Y$. Since the influential paper of Artzner et al. (1999) introducing coherent risk measures, there has been a lively debate about which risk measure is best in practice, one of the requirements under discussion being the *coherence* of a risk measure. We call a functional ρ *coherent* if it is monotone, meaning that $Y \le X$ a.s. implies that $\rho(Y) \le \rho(X)$; it is super-additive in the sense that $\rho(X+Y) \ge \rho(X) + \rho(Y)$; it is positively homogeneous which means that $\rho(\lambda Y) = \lambda\rho(Y)$ for all $\lambda \ge 0$; and it is translation invariant which amounts to $\rho(Y+a) = \rho(Y) + a$ for all $a \in \mathbb{R}$. In the literature on risk measures, there are different sign conventions which co-exist. In this paper, a positive value of Y denotes a profit. Moreover, the position Y is considered the more risky the smaller $\rho(Y)$ is. Strictly speaking, we have chosen to work with utility functions instead of risk measures as, for example, in Delbaen (2012). The risk measure ρ is called *comonotonically additive* if $\rho(X+Y) = \rho(X) + \rho(Y)$ for comonotone random variables X and Y. Coherent and comonotonically additive risk measures are also called *spectral* risk measures [Acerbi (2002)]. All risk measures of practical interest are law-invariant, that is, if two random variables X and Y have the same law F, then $\rho(X) = \rho(Y)$. As we are only concerned with law-invariant risk measures in this paper, we will abuse notation and write $\rho(F) := \rho(X)$, if X has distribution F.

One of the main criticisms on VaR_α is its failure to fulfill the super-additivity property in general [Acerbi (2002)]. Furthermore, it fails to take the size of losses beyond the level α into account [Daníelsson et al. (2001)]. In both of these aspects, ES_α is a better alternative as it is coherent and comonotonically additive, that is, a spectral risk measure. However, with respect to robustness, some authors argue that VaR_α should be preferred over ES_α [Cont, Deguest and Scandolo (2010), Kou, Peng and Heyde (2013)], whereas others argue that the classical statistical notions of robustness are not necessarily appropriate in a risk measurement context [Krätschmer, Schied and Zähle (2012, 2015, 2014)]. Finally, ES_α fails to be 1-elicitable [Gneiting (2011), Weber (2006)], whereas VaR_α is 1-elicitable for most classes of distributions $\mathcal{F}$ of practical relevance. In fact, except for the expectation, all spectral risk measures fail to be 1-elicitable [Ziegel (2014)]; further recent results on elicitable risk measures include [Kou and Peng (2014), Wang and Ziegel (2015)] showing that distortion risk measures are rarely elicitable and [Bellini and Bignozzi (2015), Delbaen et al. (2016), Weber (2006)] demonstrating that convex risk measures are only elicitable if they are shortfall risk measures.

We show in Theorem 5.2 (see also Corollaries 5.4 and 5.5) that spectral risk measures having a spectral measure with finite support can be a component of a k-elicitable functional. In particular, the pair $(\mathrm{VaR}_\alpha, \mathrm{ES}_\alpha)\colon \mathcal{F} \to \mathbb{R}^2$ is 2-elicitable for any $\alpha \in (0,1)$ subject to mild conditions on the class $\mathcal{F}$. We remark that our results substantially generalize the result of Acerbi and Székely (2014) as detailed below.

DEFINITION 5.1 (Spectral risk measures). Let μ be a probability measure on $[0,1]$ (called *spectral measure*) and let $\mathcal{F}$ be a class of distribution functions on $\mathbb{R}$ with finite first moments. Then the *spectral risk measure* associated to μ is the functional $\nu_\mu : \mathcal{F} \to \mathbb{R}$ defined as

$$\nu_\mu(F) := \int_{[0,1]} \mathrm{ES}_\alpha(F)\mu(\mathrm{d}\alpha).$$

Jouini, Schachermayer and Touzi (2006), Kusuoka (2001) have shown that law-invariant coherent and comonotonically additive risk measures are exactly the spectral risk measures in the sense of Definition 5.1 for distributions with compact support. If $\mu = \delta_\alpha$ for some $\alpha \in [0,1]$, then $\nu_\mu(F) = \mathrm{ES}_\alpha(F)$. In particular, $\nu_{\delta_1}(F) = \int y \,\mathrm{d}F(y)$ is the expectation of F.

In the following theorem, we show that spectral risk measures whose spectral measure μ has finite support in $(0,1)$ are k-elicitable for some k. The key to finding the form of the strictly consistent scoring functions at (5.2) is the observation that spectral risk measures jointly with the correct quantiles are identifiable with identification function given at (5.4). It is possible to extend the result to spectral measures with finite support in $(0,1]$; see Corollary 5.4.

THEOREM 5.2. *Let $\mathcal{F}$ be a class of distribution functions on $\mathbb{R}$ with finite first moments. Let $\nu_\mu : \mathcal{F} \to \mathbb{R}$ be a spectral risk measure where μ is given by*

$$\mu = \sum_{m=1}^{k-1} p_m \delta_{q_m},$$

with $p_m \in (0,1]$, $\sum_{m=1}^{k-1} p_m = 1$, $q_m \in (0,1)$ and the q_m's are pairwise distinct. Define the functional $T = (T_1, \dots, T_k) : \mathcal{F} \to \mathbb{R}^k$, where $T_m(F) := F^{-1}(q_m)$, $m \in \{1, \dots, k-1\}$, and $T_k(F) := \nu_\mu(F)$. Then the following assertions are true:

(i) *If the distributions in $\mathcal{F}$ have unique q_m-quantiles, $m \in \{1, \dots, k-1\}$, then the functional T is k-elicitable with respect to $\mathcal{F}$.*

(ii) *Let $\mathsf{A} \supseteq T(\mathcal{F})$ be convex and set $\mathsf{A}'_r := \{x_r : \exists (z_1, \dots, z_k) \in \mathsf{A}, x_r = z_r\}$, $r \in \{1, \dots, k\}$. Define the scoring function $S : \mathsf{A} \times \mathbb{R} \to \mathbb{R}$ by*

$$\begin{aligned} S(x,y) = {} & \sum_{r=1}^{k-1} \big(\mathbb{1}\{y \le x_r\} - q_r\big) G_r(x_r) - \mathbb{1}\{y \le x_r\} G_r(y) \\ & + G_k(x_k)\left(x_k + \sum_{m=1}^{k-1} \frac{p_m}{q_m}\big(\mathbb{1}\{y \le x_m\}(x_m - y) - q_m x_m\big)\right) \\ & - \mathcal{G}_k(x_k) + a(y), \end{aligned} \tag{5.2}$$

where $a : \mathbb{R} \to \mathbb{R}$ is $\mathcal{F}$-integrable, $G_r : \mathsf{A}'_r \to \mathbb{R}$, $r \in \{1, \dots, k\}$, $\mathcal{G}_k : \mathsf{A}'_k \to \mathbb{R}$ with $\mathcal{G}'_k = G_k$ and for all $r \in \{1, \dots, k\}$ and all $x_r \in \mathsf{A}'_r$ the functions $\mathbb{1}_{(\infty, x_r]} G_r$ are $\mathcal{F}$-integrable.

If $\mathcal{G}_k$ is convex and for all $r \in \{1, \dots, k-1\}$ and $x_k \in \mathsf{A}'_k$, the function

$$\mathsf{A}'_{r,x_k} \to \mathbb{R}, \qquad x_r \mapsto x_r \frac{p_r}{q_r} G_k(x_k) + G_r(x_r) \tag{5.3}$$

with $\mathsf{A}'_{r,x_k} := \{x_r : \exists(z_1, \dots, z_k) \in \mathsf{A}, x_r = z_r, x_k = z_k\}$ *is increasing, then* S *is* $\mathcal{F}$*-consistent for* T*. If additionally the distributions in* $\mathcal{F}$ *have unique* q_m*-quantiles,* $m \in \{1, \dots, k-1\}$, $\mathcal{G}_k$ *is strictly convex and the functions given at* (5.3) *are strictly increasing, then* S *is strictly* $\mathcal{F}$*-consistent for* T.

(iii) *Assume the elements of* $\mathcal{F}$ *have unique* q_m*-quantiles,* $m \in \{1, \dots, k-1\}$ *and continuous densities. Define the function* $V : \mathsf{A} \times \mathbb{R} \to \mathbb{R}^k$ *with components*

$$\begin{aligned} V_m(x_1, \dots, x_k, y) &= \mathbb{1}\{y \le x_m\} - q_m, \qquad m \in \{1, \dots, k-1\}, \\ V_k(x_1, \dots, x_k, y) &= x_k - \sum_{m=1}^{k-1} \frac{p_m}{q_m} y \mathbb{1}\{y \le x_m\}. \end{aligned} \tag{5.4}$$

Then V *is a strict* $\mathcal{F}$*-identification function for* T *satisfying assumption* (V3).

If additionally the interior of $\mathsf{A} := T(\mathcal{F}) \subseteq \mathbb{R}^k$ *is a star domain,* (V1) *and* (F1) *hold, and* $(V_1, \dots, V_{k-1})$ *satisfies* (V4)*, then every strictly* $\mathcal{F}$*-consistent scoring function* $S : \mathsf{A} \times \mathbb{R} \to \mathbb{R}$ *for* T *satisfying* (S2), (VS1) *is necessarily of the form given at* (5.2) *almost everywhere. Additionally,* $\mathcal{G}_k$ *must be strictly convex and the functions at* (5.3) *must be strictly increasing.*

REMARK 5.3. According to Theorem 5.2, the pair $(\mathrm{VaR}_\alpha(F), \mathrm{ES}_\alpha(F))$, and more generally $(F^{-1}(q_1), \dots, F^{-1}(q_{k-1}), \nu_\mu(F))$, admits only nonseparable strictly consistent scoring functions. This result gives an example demonstrating that Osband (1985), Proposition 2.3, cannot be correct as it states that any strictly consistent scoring function for a functional with a quantile as a component must be separable in the sense that it must be the sum of a strictly consistent scoring function for the quantile and a strictly consistent scoring function for the rest of the functional.

Using Theorem 5.2 and the revelation principle, we can now state one of the main results of this paper.

COROLLARY 5.4. *Let* $\mathcal{F}$ *be a class of distribution functions on* $\mathbb{R}$ *with finite first moments and unique quantiles. Let* $\nu_\mu : \mathcal{F} \to \mathbb{R}$ *be a spectral risk measure. If the support of* μ *is finite with* L *elements and contained in* $(0, 1]$*, then* ν_μ *is a component of a* k*-elicitable functional where*:

(i) $k = 1$, *if* μ *is concentrated at* 1 *meaning* $\mu(\{1\}) = 1$;
(ii) $k = 1 + L$, *if* $\mu(\{1\}) < 1$.

In the special case of $T = (\mathrm{VaR}_\alpha, \mathrm{ES}_\alpha)$, the maximal sensible action domain is $\mathsf{A}_0 := \{x \in \mathbb{R}^2 : x_1 \ge x_2\}$ as we always have $\mathrm{ES}_\alpha(F) \le \mathrm{VaR}_\alpha(F)$. For this action domain, the characterization of consistent scoring functions of Theorem 5.2 simplifies as follows.

COROLLARY 5.5. *Let $\alpha \in (0,1)$. Let $\mathcal{F}$ be a class of distribution functions on $\mathbb{R}$ with finite first moments and unique α-quantiles. Let $\mathsf{A}_0 = \{x \in \mathbb{R}^2 : x_1 \geq x_2\}$. A scoring function $S: \mathsf{A}_0 \times \mathbb{R} \to \mathbb{R}$ of the form*

$$\begin{aligned} S(x_1, x_2, y) = {} & \big(\mathbb{1}\{y \leq x_1\} - \alpha\big) G_1(x_1) - \mathbb{1}\{y \leq x_1\} G_1(y) \\ & + G_2(x_2)\bigg(x_2 - x_1 + \frac{1}{\alpha}\mathbb{1}\{y \leq x_1\}(x_1 - y)\bigg) \\ & - \mathcal{G}_2(x_2) + a(y), \end{aligned} \tag{5.5}$$

where $G_1, G_2, \mathcal{G}_2, a: \mathbb{R} \to \mathbb{R}$, $\mathcal{G}_2' = G_2$, a is $\mathcal{F}$-integrable and $\mathbb{1}_{(-\infty, x_1]} G_1$ is $\mathcal{F}$-integrable for all $x_1 \in \mathbb{R}$, is $\mathcal{F}$-consistent for $T = (\mathrm{VaR}_\alpha, \mathrm{ES}_\alpha)$ if G_1 is increasing and $\mathcal{G}_2$ is increasing and convex. If $\mathcal{G}_2$ is strictly increasing and strictly convex, then S is strictly $\mathcal{F}$-consistent for T.

Under the conditions of Theorem 5.2(iii) *all strictly $\mathcal{F}$-consistent scoring functions for T are of the form* (5.5) *almost everywhere.*

Acerbi and Székely (2014) also give an example of a scoring function for the pair $T = (\mathrm{VaR}_\alpha, \mathrm{ES}_\alpha): \mathcal{F} \to \mathsf{A} \subseteq \mathbb{R}^2$. They use a different sign convention for VaR_α and ES_α than we do in this paper. Using our sign convention, their proposed scoring function $S^W: \mathsf{A} \times \mathbb{R} \to \mathbb{R}$ reads

$$\begin{aligned} S^W(x_1, x_2, y) = {} & \alpha\big(x_2^2/2 + W x_1^2/2 - x_1 x_2\big) \\ & + \mathbb{1}\{y \leq x_1\}\big(-x_2(y - x_1) + W\big(y^2 - x_1^2\big)/2\big), \end{aligned} \tag{5.6}$$

where $W \in \mathbb{R}$. The authors claim that S^W is a strictly $\mathcal{F}$-consistent scoring function for $T = (\mathrm{VaR}_\alpha, \mathrm{ES}_\alpha)$ provided that

$$\mathrm{ES}_\alpha(F) > W \mathrm{VaR}_\alpha(F) \tag{5.7}$$

for all $F \in \mathcal{F}$. This means that they consider a strictly smaller action domain than A_0 in Corollary 5.5. They assume that the distributions in $\mathcal{F}$ have continuous densities, unique α quantiles, and that $F(x) \in (0,1)$ implies $f(x) > 0$ for all $F \in F$ with density f. Furthermore, in order to ensure that $\bar{S}^W(\cdot, F)$ is finite, one needs to impose the assumption that $\int_{-\infty}^x y^2 \, \mathrm{d}F(y)$ is finite for all $x \in \mathbb{R}$ and $F \in \mathcal{F}$. This is slightly less than requiring finite second moments. As a matter of fact, they only show that $\nabla \bar{S}^W(t_1, t_2, F) = 0$ for $F \in \mathcal{F}$ and $(t_1, t_2) = T(F)$ and that $\nabla^2 \bar{S}^W(t_1, t_2, F)$ is positive definite. This only shows that $\bar{S}^W(x, F)$ has a *local* minimum at $x = T(F)$ but does not provide a proof concerning a *global* minimum; see also the discussion after Corollary 3.3. However, we can use Theorem 5.2(ii) to verify their claims with $G_1(x_1) = -(W/2)x_1^2$, $\mathcal{G}_2(x_2) = (\alpha/2)x_2^2$ and $a = 0$. Hence, $\mathcal{G}_2$ is strictly convex, and the function $x_1 \mapsto x_1 G_2(x_2)/\alpha + G_1(x_1)$ is strictly increasing in x_1 if and only if $x_2 > W x_1$ as at (5.7).

The scoring function S^W has one property which is potentially relevant in applications. If x_1, x_2 and y are expressed in the same units of measurement, then $S^W(x_1, x_2, y)$ is a quantity with these units squared. If one insists that we should only add quantities with the same units, then the necessary condition that $x_1 \mapsto x_1 G_2(x_2)/\alpha + G_1(x_1)$ is strictly increasing enforces a condition of the type (5.7). The action domain is restricted for S^W and the choice of W may not be obvious in practice. Similarly, for the maximal action domain A_0, an open question of practical interest is the choice of the functions G_1 and $\mathcal{G}_2$ in (5.5). We would like to remark that S remains strictly consistent upon choosing $G_1 = 0$ and $\mathcal{G}_2$ strictly increasing and strictly convex.

6. Discussion. We have investigated necessary and sufficient conditions for the elicitability of k-dimensional functionals of d-dimensio nal distributions. In order to derive necessary conditions, we have adapted Osband's principle for the case where the class $\mathcal{F}$ of distributions does not necessarily contain distributions with finite support. This comes at the cost of certain smoothness assumptions on the expected scores $\bar{S}(\cdot, F)$. For particular situations, for example, when characterizing the class of strictly $\mathcal{F}$-consistent scoring functions for ratios of expectations, it is possible to weaken the smoothness assumptions; see Frongillo and Kash (2015). While moving away from distributions with finite support is not a great gain in the case of linear functionals or ratios of expectations, it comes in handy when considering spectral risk measures. Value at Risk, VaR_α, being defined as the smallest α-quantile, is generally not elicitable for distributions where the α-quantile is not unique. Therefore, we believe that it is also not possible to show joint elicitability of $(\mathrm{VaR}_\alpha, \mathrm{ES}_\alpha)$ for classes $\mathcal{F}$ of distributions with nonunique α-quantiles. However, we can give consistent scoring functions which become strictly consistent as soon as the elements of $\mathcal{F}$ have unique quantiles. Fortunately, the classes $\mathcal{F}$ of distributions that are relevant in risk management usually consist of absolutely continuous distributions having unique quantiles.

Emmer, Kratz and Tasche (2015) have remarked that ES_α is *conditionally elicitable*. Slightly generalizing their definition, a functional $T_k : \mathcal{F} \to \mathsf{A}_k \subseteq \mathbb{R}$ is called *conditionally elicitable of order* k, $k \geq 1$, if there are $k - 1$ elicitable functionals $T_m : \mathcal{F} \to \mathsf{A}_m \subseteq \mathbb{R}$, $m \in \{1, \dots, k - 1\}$, such that T_k is elicitable restricted to the class $\mathcal{F}_{x_1,\dots,x_{k-1}} := \{F \in \mathcal{F} : T_1(F) = x_1, \dots, T_{k-1}(F) = x_{k-1}\}$ for any $(x_1, \dots, x_{k-1}) \in \mathsf{A}_1 \times \cdots \times \mathsf{A}_{k-1}$. *Mutatis mutandis*, one can define a notion of *conditional identifiability* by replacing the term "elicitable" with "identifiable" in the above definition. It is not difficult to check that any conditionally identifiable functional T_k of order k is a component of an identifiable functional $T = (T_1, \dots, T_k)$. Spectral risk measures ν_μ with spectral measure μ with finite support in $(0, 1)$ provide an example of a conditionally elicitable functional of order $L + 1$, where L is the cardinality of the support of μ; see Theorem 5.2. However, we would like to stress that it is generally an open question whether any conditionally elicitable and identifiable functional T_k of order $k \geq 2$ is always a component of a k-elicitable functional.

Slightly modifying Lambert, Pennock and Shoham (2008), Definition 11, one could define the *elicitability order* of a real-valued functional T as the smallest number k such that the functional is a component of a k-elicitable functional. It is clear that the elicitability order of the variance is two, and we have shown that the same is true for ES_α for reasonably large classes $\mathcal{F}$. For spectral risk measures ν_μ, the elicitability order is at most $L+1$, where L is the cardinality of the support; see Corollary 5.4.

In the one-dimensional case, Steinwart et al. (2014) have shown that convex level sets in the sense of Osband (1985), Proposition 2.5 [see also Gneiting (2011), Theorem 6] is a sufficient condition for the elicitability of a functional T under continuity assumptions on T. Without such continuity assumptions, the converse of Osband (1985), Proposition 2.5, is generally false; see Heinrich (2014) for the example of the mode functional. It is an open (and potentially difficult) question under which conditions a converse of Osband (1985), Proposition 2.5, is true for higher order elicitability.

7. Proofs.

PROOF OF THEOREM 3.2. Let $x \in \mathrm{int}(\mathsf{A})$. The identifiability property of V plus the first-order condition stemming from the strict $\mathcal{F}$-consistency of S yields the relation $\bar{V}(x,F)=0 \Longrightarrow \nabla\bar{S}(x,F)=0$ for all $F\in\mathcal{F}$. Let $l\in\{1,\dots,k\}$. To show (3.2), consider the composed functional

$$\bar{B}(x,\cdot)\colon \mathcal{F}\to\mathbb{R}^{k+1}, \qquad F\mapsto \big(\partial_l\bar{S}(x,F),\bar{V}(x,F)\big).$$

By construction, we know that

$$\bar{V}(x,F)=0 \quad\Longleftrightarrow\quad \bar{B}(x,F)=0 \tag{7.1}$$

for all $F\in\mathcal{F}$. Assumption (V1) implies that there are $F_1,\dots,F_{k+1}\in\mathcal{F}$ such that the matrix $\mathbb{V}=\mathrm{mat}(\bar{V}(x,F_1),\dots,\bar{V}(x,F_{k+1}))\in\mathbb{R}^{k\times(k+1)}$ has maximal rank, meaning $\mathrm{rank}(\mathbb{V})=k$. If $\mathrm{rank}(\mathbb{V})<k$, then the space $\mathrm{span}\{\bar{V}(x,F_1),\dots,\bar{V}(x,F_{k+1})\}$ would be a linear subspace such that the interior of $\mathrm{conv}(\{\bar{V}(x,F_1),\dots,\bar{V}(x,F_{k+1})\})$ would be empty. Let $G\in\mathcal{F}$. Then still $0\in\mathrm{int}(\mathrm{conv}(\{\bar{V}(x,G),\bar{V}(x,F_1),\dots,\bar{V}(x,F_{k+1})\}))$, so $\mathrm{rank}(\mathbb{V}_G)=k$ where $\mathbb{V}_G=\mathrm{mat}(\bar{V}(x,G),\bar{V}(x,F_1),\dots,\bar{V}(x,F_{k+1}))\in\mathbb{R}^{k\times(k+2)}$. Define the matrix

$$\mathbb{B}_G=\begin{pmatrix}\partial_l\bar{S}(x,G) & \partial_l\bar{S}(x,F_1) & \cdots & \partial_l\bar{S}(x,F_{k+1})\\ & & \mathbb{V}_G & \end{pmatrix}\in\mathbb{R}^{(k+1)\times(k+2)}.$$

We use (7.1) to show that $\ker(\mathbb{B}_G)=\ker(\mathbb{V}_G)$. First, observe that the relation $\ker(\mathbb{B}_G)\subseteq\ker(\mathbb{V}_G)$ is clear by construction. To show the other inclusion, let $\theta\in\ker(\mathbb{V}_G)$ be an element of the simplex. Then (7.1) and the convexity of $\mathcal{F}$ yields that $\theta\in\ker(\mathbb{B}_G)$. By linearity, the inclusion holds also for all $\theta\in\ker(\mathbb{V}_G)$ with nonnegative components. Finally, let $\theta\in\ker(\mathbb{V}_G)$ be arbitrary. Assumption (V1) implies that there is $\theta^*\in\ker(\mathbb{V}_G)$ with strictly positive components.

Hence, there is an $\varepsilon > 0$ such that $\theta^* + \varepsilon\theta$ has nonnegative components. Since $\mathbb{V}_G(\theta^* + \varepsilon\theta) = \mathbb{V}_G\theta^* + \varepsilon\mathbb{V}_G\theta = 0$, we know that $\theta^* + \varepsilon\theta \in \ker(\mathbb{B}_G)$. Again using linearity and the fact that $\theta^* \in \ker(\mathbb{B}_G)$, we obtain that $\theta \in \ker(\mathbb{B}_G)$.

With the rank-nullity theorem, this gives $\operatorname{rank}(\mathbb{B}_G) = \operatorname{rank}(\mathbb{V}_G) = k$. Hence, there is a unique vector $(h_{l1}(x), \dots, h_{lk}(x)) \in \mathbb{R}^k$ such that one has $\partial_l \bar{S}(x, G) = \sum_{m=1}^k h_{lm}(x)\bar{V}_m(x, G)$. Since $G \in \mathcal{F}$ was arbitrary, the assertion at (3.2) follows.

The second part of the claim can be seen as follows. For $x \in \operatorname{int}(\mathsf{A})$, pick $F_1, \dots, F_k \in \mathcal{F}$ such that $\bar{V}(x, F_1), \dots, \bar{V}(x, F_k)$ are linearly independent and let $\mathbb{V}(z)$ be the matrix with columns $\bar{V}(z, F_i)$, $i \in \{1, \dots, k\}$ for $z \in \operatorname{int}(\mathsf{A})$. Due to assumption (V2) or (V3), $\mathbb{V}(z)$ has full rank in some neighborhood U of x. Let $r \in \{1, \dots, k\}$ and let e_r be the rth standard unit vector of $\mathbb{R}^k$. We define $\lambda(z) := \mathbb{V}(z)^{-1}e_r$ for $z \in U$. Taking the inverse of a matrix is a continuously differentiable operation, so it is in particular locally Lipschitz continuous. Therefore, the vector λ inherits the regularity properties of $\bar{V}(z, F_i)$, that is, under (V2) λ is continuous, and under (V3) λ is locally Lipschitz continuous. Therefore, these properties carry over to h because for $l \in \{1, \dots, k\}$, $z \in U$

$$h_{lr}(z) = \sum_{i=1}^k \lambda_i(z) \sum_{m=1}^k h_{lm}(z)\bar{V}_m(z, F_i) = \sum_{i=1}^k \lambda_i(z)\partial_l \bar{S}_m(z, F_i)$$

using the assumptions on S. □

PROOF OF PROPOSITIONS 4.2 AND 4.4. We show parts (i) of the two propositions simultaneously. We have that $\partial_l \bar{V}_r(x, F) = 0$ for all $l, r \in \{1, \dots, k\}$, $l \neq r$, and $x \in \operatorname{int}(\mathsf{A})$, $F \in \mathcal{F}$. Equation (3.3) evaluated at $x = t = T(F)$ yields

$$h_{rl}(t)\partial_l \bar{V}_l(t, F) = h_{lr}(t)\partial_r \bar{V}_r(t, F). \tag{7.2}$$

If (V4) holds, then (7.2) implies that $h_{rl}(t) = 0$ for $r \neq l$, hence we obtain (4.2) with the surjectivity of T. On the other hand, if $V_r(x, y) = q(y)x_m - p_m(y)$, (7.2) implies that $h_{rl}(t) = h_{lr}(t)$, whence the second part of (4.4) is shown, again using the surjectivity of T. In both cases, (3.3) is equivalent to

$$\sum_{m=1}^k \big(\partial_l h_{rm}(x) - \partial_r h_{lm}(x)\big)\bar{V}_m(x, F) = 0. \tag{7.3}$$

Using assumption (V1), there are $F_1, \dots, F_k \in \mathcal{F}$ such that the vectors $\bar{V}(x, F_1), \dots, \bar{V}(x, F_k)$ are linearly independent. This yields that $\partial_l h_{rm}(x) = \partial_r h_{lm}(x)$ for almost all $x \in \operatorname{int}(\mathsf{A})$. For Proposition 4.2, we can conclude that $\partial_l h_{rr}(x) = \partial_r h_{lr}(x) = 0$ for $r \neq l$ for almost all $x \in \operatorname{int}(\mathsf{A})$. Consequently, invoking that A is connected, the functions h_{mm} only depend on x_m and we can write $h_{mm}(x) = g_m(x_m)$ for some function $g_m : \mathsf{A}'_m \to \mathbb{R}$. By Lemma 2.4(i), for $v \in \mathbb{S}^{k-1}$, $t =$

$T(F) \in \mathrm{int}(\mathsf{A})$, the function $s \mapsto \bar{S}(t + sv, F)$ has a global unique minimum at $s = 0$, hence

$$v^\top \nabla \bar{S}(t + sv, F) = \sum_{m=1}^{k} g_m(t_m + sv_m)\bar{V}_m(t_m + sv_m, F)v_m$$

vanishes for $s = 0$, is negative for $s < 0$ and positive for $s > 0$, where s is in some neighborhood of zero. Choosing v as the lth standard basis vector of $\mathbb{R}^k$ we obtain that $g_l > 0$ exploiting the orientation of V_l and the surjectivity of T.

For Proposition 4.4(i) to show the assertion about the definiteness of h, observe that for $v \in \mathbb{S}^{k-1}$, $t = T(F) \in \mathrm{int}(\mathsf{A})$ we have $\bar{V}(t + sv, F) = \bar{q}(F)sv$ where $\bar{q}(F) > 0$. Hence, $v^\top \nabla \bar{S}(t + sv, F) = \bar{q}(F)sv^\top h(t + sv)v$, which implies the claim using again the surjectivity of T.

For part (ii) of Proposition 4.2, the sufficiency is immediate; see the proof of Lemma 2.6. For necessity, we apply Proposition 3.4 and part (i) such that

$$S(x, y) = \sum_{m=1}^{k} \int_{z_m}^{x_m} g_m(v)V_m(v, y)\,\mathrm{d}v + a(y),$$

for almost all $(x, y) \in \mathsf{A} \times \mathsf{O}$, where $z \in \mathrm{int}(\mathsf{A})$ is a star point of $\mathrm{int}(\mathsf{A})$ and a is an $\mathcal{F}$-integrable function. Let $t = T(F)$ and $x_m \neq t_m$. The strict consistency of S implies that $\bar{S}(t, F) < \bar{S}(t_1, \dots, t_{m-1}, x_m, t_{m+1}, \dots, t_m)$. This means $\bar{S}_m(t_m, F) < \bar{S}_m(x_m, F)$ with $S_m(x_m, y) := \int_{z_m}^{x_m} g_m(v)V_m(v, y)\,\mathrm{d}v + (1/k)a(y)$.

For part (ii) of Proposition 4.4, observe that due to part (i) h is the Hessian of ϕ, and thus, ϕ is strictly convex. For the sufficiency of the form (4.5), let $x \neq t = T(F)$ for some $F \in \mathcal{F}$. Then

$$\bar{S}(x, F) - \bar{S}(t, F) = \bar{q}(F)\big(\phi(t) - \phi(x) + \langle \nabla\phi(x), x - t\rangle\big) > 0$$

due to the strict convexity of ϕ and $\bar{q}(F) > 0$. For the necessity of the form (4.5), apply Proposition 3.4 and use partial integration. $\square$

PROOF OF THEOREM 5.2. (i) The second part of Theorem 5.2(ii) implies the k-elicitability of T.

(ii) Let $S\colon \mathsf{A} \times \mathbb{R} \to \mathbb{R}$ be of the form (5.2), $\mathcal{G}_k$ be convex and the functions at (5.3) be increasing. Let $F \in \mathcal{F}$, $x = (x_1, \dots, x_k) \in \mathsf{A}$ and set $t = (t_1, \dots, t_k) = T(F)$, $w = \min(x_k, t_k)$. Then we obtain

$$\begin{aligned} S(x, y) = {} & -\mathcal{G}_k(x_k) + G_k(w)(x_k - y) + a(y) \\ & + \sum_{r=1}^{k-1} \big(\mathbb{1}\{y \le x_r\} - q_r\big)\left(G_r(x_r) + \frac{p_r}{q_r} G_k(w)(x_r - y)\right) \\ & - \mathbb{1}\{y \le x_r\}G_r(y) + \big(G_k(x_k) - G_k(w)\big) \\ & \times \left(x_k + \sum_{m=1}^{k-1} \frac{p_m}{q_m}\big(\mathbb{1}\{y \le x_m\}(x_m - y) - q_m x_m\big)\right). \end{aligned}$$

This implies that $\bar{S}(x, F) - \bar{S}(t, F) = R_1 + R_2$ with

$$R_1 = \sum_{r=1}^{k-1} \big(F(x_r) - q_r\big)\bigg(G_r(x_r) + \frac{p_r}{q_r} G_k(w) x_r\bigg) - \int_{t_r}^{x_r} \bigg(G_r(y) + \frac{p_r}{q_r} G_k(w) y\bigg) \,\mathrm{d}F(y),$$

$$R_2 = \big(G_k(x_k) - G_k(w)\big)\Bigg(x_k + \sum_{m=1}^{k-1} \frac{p_m}{q_m}\bigg(\int_{-\infty}^{x_m} (x_m - y)\,\mathrm{d}F(y) - q_m x_m\bigg)\Bigg) - \mathcal{G}_k(x_k) + \mathcal{G}_k(t_k) + G_k(w)(x_k - t_k).$$

We denote the rth summand of R_1 by ξ_r and suppose that $t_r < x_r$. Due to the assumptions, the term $G_r(y) + (p_r/q_r)G_k(w)y$ is increasing in $y \in [t_r, x_r]$ which implies that $\xi_r \geq (F(x_r) - q_r)(G_r(x_r) + (p_r/q_r)G_k(w)x_r) - (F(x_r) - F(t_r))(G_r(x_r) + (p_r/q_r)G_k(w)x_r) = 0$. Analogously, one can show that $\xi_r \geq 0$ if $x_r < t_r$. If F has a unique q_r-quantile and the term $G_r(y) + (p_r/q_r)G_k(w)y$ is strictly increasing in y, then we even get $\xi_r > 0$ if $x_r \neq t_r$.

Now consider the term R_2. Splitting the integrals from ∞ to x_m into integrals from $-\infty$ to t_m and from t_m to x_m and partially integrating the latter, we obtain

$$\begin{aligned}
R_2 &= \big(G_k(x_k) - G_k(w)\big) \\
&\quad \times \Bigg(x_k + \sum_{m=1}^{k-1} p_m\bigg(t_m - x_m - \frac{1}{q_m}\int_{-\infty}^{t_m} y\,\mathrm{d}F(y) + \frac{1}{q_m}\int_{t_m}^{x_m} F(y)\,\mathrm{d}y\bigg)\Bigg) \\
&\quad - \mathcal{G}_k(x_k) + \mathcal{G}_k(t_k) + G_k(w)(x_k - t_k) \\
&= \big(G_k(x_k) - G_k(w)\big)\Bigg(x_k - t_k + \sum_{m=1}^{k-1} p_m\bigg(t_m - x_m + \frac{1}{q_m}\int_{t_m}^{x_m} F(y)\,\mathrm{d}y\bigg)\Bigg) \\
&\quad - \mathcal{G}_k(x_k) + \mathcal{G}_k(t_k) + G_k(w)(x_k - t_k) \\
&\geq \big(G_k(x_k) - G_k(w)\big)(x_k - t_k) - \mathcal{G}_k(x_k) + \mathcal{G}_k(t_k) + G_k(w)(x_k - t_k) \\
&= \mathcal{G}_k(t_k) - \mathcal{G}_k(x_k) - G_k(x_k)(t_k - x_k) \geq 0.
\end{aligned}$$

The first inequality is due to the fact that (i) G_k is increasing and (ii) for $x_m \neq t_m$ we have $(1/q_m)\int_{t_m}^{x_m} F(y)\,\mathrm{d}y \geq x_m - t_m$ with strict inequality if F has a unique q_m-quantile. The last inequality is due to the fact that $\mathcal{G}_k$ is convex. The inequality is strict if $x_k \neq t_k$ and if $\mathcal{G}_k$ is strictly convex.

(iii) If f denotes the density of F, it holds that

$$\mathrm{ES}_\alpha(F) = \frac{1}{\alpha}\int_{-\infty}^{F^{-1}(\alpha)} y f(y)\,\mathrm{d}y, \qquad \alpha \in (0, 1]. \tag{7.4}$$

We first show the assertions concerning V given at (5.4). Let $F \in \mathcal{F}$ with density $f = F'$ and let $t = T(F)$. Then we have for $m \in \{1, \dots, k-1\}$, $x \in \mathsf{A}$, that

$\bar{V}_m(x, F) = F(x_m) - q_m$ which is zero if and only if $x_m = t_m$. On the other hand, using the identity at (7.4)

$$\bar{V}_k(t_1, \dots, t_{k-1}, x_k, F) = x_k - \sum_{m=1}^{k-1} \frac{p_m}{q_m} \int_{-\infty}^{t_m} y f(y)\, \mathrm{d}y = x_k - t_k.$$

Hence, it follows that V is a strict $\mathcal{F}$-identification function for T. Moreover, V satisfies assumption (V3), and we have for $m \in \{1, \dots, k-1\}$, $l \in \{1, \dots, k\}$ and $x \in \mathrm{int}(\mathsf{A})$ that $\partial_l \bar{V}_m(x, F) = 0$ if $l \neq m$ and $\partial_m \bar{V}_m(x, F) = f(x_m)$, $\partial_m \bar{V}_k(x, F) = -(p_m/q_m) x_m f(x_m)$ and $\partial_k \bar{V}_k(x, F) = 1$.

From now on, we assume that $t = T(F) \in \mathrm{int}(\mathsf{A})$. Let S be a strictly $\mathcal{F}$-consistent scoring function for T satisfying (S2). Then we can apply Theorem 3.2 and Corollary 3.3 to get that there are locally Lipschitz continuous functions $h_{lm} : \mathrm{int}(\mathsf{A}) \to \mathbb{R}$ such that (3.2) and (3.3) hold. If we evaluate (3.3) for $l = k$, $m \in \{1, \dots, k-1\}$ at the point $x = t$, we get

$$h_{km}(t) \partial_m \bar{V}_m(t, F) + h_{kk}(t) \partial_m \bar{V}_k(t, F) = h_{mk}(t) \partial_k \bar{V}_k(t, F),$$

which takes the form $h_{km}(t) f(t_m) - h_{kk}(t)(p_m/q_m) t_m f(t_m) = h_{mk}(t)$. Invoking assumption (V4) for $(V_1, \dots, V_{k-1})$, we get that necessarily $h_{mk}(t) = 0$ and $h_{km}(t) = (p_m/q_m) t_m h_{kk}(t)$. So with the surjectivity of T, we get for $x \in \mathrm{int}(\mathsf{A})$ that

$$h_{mk}(x) = 0, \qquad h_{km}(x) = \frac{p_m}{q_m} x_m h_{kk}(x) \qquad \text{for all } m \in \{1, \dots, k-1\}. \tag{7.5}$$

Now, we can evaluate (3.3) for $m, l \in \{1, \dots, k-1\}$, $m \neq l$, at $x = t$ and use the first part of (7.5) to get that $h_{ml}(t) f(t_l) = h_{lm}(t) f(t_m)$. Using again the same argument, we get for $x \in \mathrm{int}(\mathsf{A})$ that

$$h_{ml}(x) = 0 \qquad \text{for all } m, l \in \{1, \dots, k-1\}, l \neq m. \tag{7.6}$$

At this stage, we can evaluate (3.3) for $l \in \{1, \dots, k-1\}$, $m \in \{1, \dots, k\}$, $m \neq l$, for some $x \in \mathrm{int}(\mathsf{A})$. Using (7.5) and (7.6), we obtain

$$\sum_{i=1}^{k} \big(\partial_l h_{mi}(x) - \partial_m h_{li}(x)\big) \bar{V}_i(x_i, F) = 0.$$

Invoking assumption (V1) and using (7.5) and (7.6), we can conclude that for almost all $x \in \mathsf{A}$,

$$\partial_l h_{mm}(x) = 0 \qquad \text{for all } l \in \{1, \dots, k-1\}, m \in \{1, \dots, k\}, l \neq m \tag{7.7}$$

and

$$\partial_k h_{ll}(x) = \frac{p_l}{q_l} h_{kk}(x) \qquad \text{for all } l \in \{1, \dots, k-1\}. \tag{7.8}$$

Equation (7.7) for $m = k$ shows that there is a locally Lipschitz continuous function $g_k : \mathsf{A}'_k \to \mathbb{R}$ such that for all $(x_1, \dots, x_k) \in \mathrm{int}(\mathsf{A})$, we have $h_{kk}(x_1, \dots, x_k) =$

$g_k(x_k)$. Equation (7.8) together with (7.7) gives that for $l \in \{1, \dots, k-1\}$, and $(x_1, \dots, x_k) \in \operatorname{int}(\mathsf{A})$, we obtain $h_{ll}(x_1, \dots, x_k) = (p_l/q_l)G_k(x_k) + g_l(x_l)$, where $g_l\colon \mathsf{A}'_l \to \mathbb{R}$ is locally Lipschitz continuous and $G_k\colon \mathsf{A}'_k \to \mathbb{R}$ is such that $G'_k = g_k$.

Knowing the form of the matrix-valued function h, we can apply Proposition 3.4. Let $z \in \operatorname{int}(\mathsf{A})$ be some star point. Then there is some $\mathcal{F}$-integrable function $b\colon \mathbb{R} \to \mathbb{R}$ such that

$$
\begin{aligned}
S(x, y) = {} & \sum_{r=1}^{k-1} \int_{z_r}^{x_r} \left(\frac{p_r}{q_r} G_k(z_k) + g_r(v)\right)\big(\mathbb{1}\{y \le v\} - q_r\big)\,\mathrm{d}v \\
& + \big(G_k(x_k) - G_k(z_k)\big) \sum_{m=1}^{k-1} \frac{p_m}{q_m}\big(x_m\big(\mathbb{1}\{y \le x_m\} - q_m\big) - y\mathbb{1}\{y \le x_m\}\big) \\
& + G_k(x_k)x_k - \mathcal{G}_k(x_k) + b(y),
\end{aligned} \tag{7.9}
$$

for almost all (x, y) where $\mathcal{G}_k\colon \mathsf{A}'_k \to \mathbb{R}$ is such that $\mathcal{G}'_k = G_k$. One can check by a straightforward computation that the representation of S at (7.9) is equivalent to the one at (5.2) upon choosing a suitable $\mathcal{F}$-integrable function $a\colon \mathbb{R} \to \mathbb{R}$.

It remains to show that $\mathcal{G}_k$ is strictly convex and that the functions given at (5.3) are strictly increasing. To this end, we use Lemma 2.4. Let $D = \{s \in \mathbb{R}\colon t + sv \in \operatorname{int}(\mathsf{A})\}$, and let $v = (v_1, \dots, v_k) \in \mathbb{S}^{k-1}$ and without loss of generality assume $v_k \ge 0$. We define $\psi\colon D \to \mathbb{R}$ by $\psi(s) := \bar{S}(t + sv, F)$, that is,

$$
\begin{aligned}
\psi(s) = {} & \sum_{r=1}^{k-1} \int_{z_r}^{\bar{s}_r} \left(\frac{p_r}{q_r} G_k(z_k) + g_r(v)\right)\big(F(v) - q_r\big)\,\mathrm{d}v \\
& + \big(G_k(\bar{s}_k) - G_k(z_k)\big) \sum_{m=1}^{k-1} \frac{p_m}{q_m}\left(\bar{s}_m\big(F(\bar{s}_m) - q_m\big) - \int_{-\infty}^{\bar{s}_m} y f(y)\,\mathrm{d}y\right) \\
& + \bar{s}_k G_k(\bar{s}_k) - \mathcal{G}_k(\bar{s}_k) + \bar{b}(F),
\end{aligned}
$$

where we use the notation $\bar{s} = t + sv$. The function ψ has a minimum at $s = 0$. Hence, there is $\varepsilon > 0$ such that $\psi'(s) < 0$ for $s \in (-\varepsilon, 0)$ and $\psi'(s) > 0$ for $s \in (0, \varepsilon)$. If $v_k = 0$, then

$$
\psi'(s) = \sum_{r=1}^{k-1} \big(F(\bar{s}_r) - q_r\big) v_r \left(g_r(\bar{s}_r) + \frac{p_r}{q_r} G_k(\bar{s}_k)\right).
$$

Choosing v as the rth standard basis vector of $\mathbb{R}^k$ for $r \in \{1, \dots, k-1\}$, we obtain that $g_r(\bar{s}_r) + (p_r/q_r)G_k(\bar{s}_k) > 0$. Exploiting the surjectivity of T we can deduce that the functions at (5.3) are strictly increasing. On the other hand, if v is the kth standard basis vector, we obtain that $\psi'(s) = g_k(\bar{s}_k)s$. Again using the surjectivity of T, we get that $g_k > 0$ which shows the strict convexity of $\mathcal{G}_k$. □

PROOF OF COROLLARY 5.5. The sufficiency follows directly from Theorem 5.2. We will show that G_2 is necessarily bounded below. Suppose the contrary.

For the action domain A_0, we have $\mathsf{A}'_{1,x_2} = [x_2, \infty)$, therefore, for $x_2 \le x_1 < x_1'$ (5.3) yields $-\infty < G_1(x_1) - G_1(x_1') \le (1/\alpha)G_2(x_2)(x_1' - x_1)$. Letting $x_2 \to -\infty$, one obtains a contradiction. Let $C_2 = \lim_{x_2 \to -\infty} G_2(x_2) > -\infty$. Then, by (5.3), we obtain that $G_1(x_1) + (C_2/\alpha)x_1$ is increasing in $x_1 \in \mathbb{R}$. We can write S at (5.5) as

$$\begin{aligned} S(x_1, x_2, y) = {} & \big(G_2(x_2) - C_2\big)\bigg(\frac{1}{\alpha}\mathbb{1}\{y \le x_1\}(x_1 - y) - (x_1 - x_2)\bigg) \\ & + \big(\mathbb{1}\{y \le x_1\} - \alpha\big)\bigg(G_1(x_1) + \frac{C_2}{\alpha}x_1\bigg) \\ & - \mathbb{1}\{y \le x_1\}\bigg(G_1(y) + \frac{C_2}{\alpha}y\bigg) \\ & - \big(\mathcal{G}_2(x_2) - C_2x_2\big) + a(y). \end{aligned}$$

The last expression is again of the form at (5.5) with increasing functions $\tilde{G}_1(x_1) = G_1(x_1) + (C_2/\alpha)x_1$ and $\tilde{G}_2(x_2) = G_2(x_2) - C_2 \ge 0$. $\square$

Acknowledgements. We would like to thank Carlo Acerbi, Rafael Frongillo and Tilmann Gneiting for fruitful discussions which inspired some of the results of this paper. We are grateful for the valuable comments of two anonymous referees which significantly improved the paper.

SUPPLEMENTARY MATERIAL

Supplement to "Higher order elicitability and Osband's principle" (DOI: 10.1214/16-AOS1439SUPP; .pdf). The proofs of Proposition 3.4 and Corollary 5.4 are deferred to this supplement.

REFERENCES

ABERNETHY, J. D. and FRONGILLO, R. M. (2012). A characterization of scoring rules for linear properties. In *Proceedings of the* 25*th Annual Conference on Learning Theory. Journal of Machine Learning Research*: *Workshop and Conference Proceedings*, *Vol*. 23 (S. Mannor, N. Srebro and R. C. Williamson, eds.) 27.1–27.13.

ACERBI, C. (2002). Spectral measures of risk: A coherent representation of subjective risk aversion. *J. Bank. Financ.* **26** 1505–1518.

ACERBI, C. and SZÉKELY, B. (2014). Backtesting expected shortfall. *Risk Mag.* **27** 76–81.

ARTZNER, P., DELBAEN, F., EBER, J.-M. and HEATH, D. (1999). Coherent measures of risk. *Math. Finance* **9** 203–228. MR1850791

BANERJEE, A., GUO, X. and WANG, H. (2005). On the optimality of conditional expectation as a Bregman predictor. *IEEE Trans. Inform. Theory* **51** 2664–2669. MR2246384

BELLINI, F. and BIGNOZZI, V. (2015). On elicitable risk measures. *Quant. Finance* **15** 725–733. MR3334566

CONT, R., DEGUEST, R. and SCANDOLO, G. (2010). Robustness and sensitivity analysis of risk measurement procedures. *Quant. Finance* **10** 593–606. MR2676786

DANÍELSSON, J., EMBRECHTS, P., GOODHART, C., KEATING, C., MUENNICH, F., RENAULT, O. and SHIN, H. S. (2001). An academic response to Basel II. Special Paper No. 130.

DAVIS, M. H. A. (2016). Verification of internal risk measure estimates. *Stat. Risk Model.* To appear.

DAWID, A. P. and SEBASTIANI, P. (1999). Coherent dispersion criteria for optimal experimental design. *Ann. Statist.* **27** 65–81. MR1701101

DELBAEN, F. (2012). *Monetary Utility Functions*. Osaka Univ. Press, Osaka.

DELBAEN, F., BELLINI, F., BIGNOZZI, V. and ZIEGEL, J. F. (2016). Risk measures with the CxLS property. *Finance Stoch.* **20** 433–453. MR3479327

EMBRECHTS, P. and HOFERT, M. (2014). Statistics and quantitative risk management for banking and insurance. *Ann. Rev. Stat. Appl.* **1**.

EMMER, S., KRATZ, M. and TASCHE, D. (2015). What is the best risk measure in practice? A comparison of standard measures. *J. Risk* **18** 31–60.

ENGELBERG, J., MANSKI, C. F. and WILLIAMS, J. (2009). Comparing the point predictions and subjective probability distributions of professional forecasters. *J. Bus. Econom. Statist.* **27** 30–41. MR2484982

FISSLER, T. and ZIEGEL, J. F. (2016). Supplement to "Higher order elicitability and Osband's principle." DOI:10.1214/16-AOS1439SUPP.

FRONGILLO, R. and KASH, I. (2015). Vector-valued property elicitation. *J. Mach. Learn. Res. Workshop Conf. Proc.* **40** 1–18.

GNEITING, T. (2011). Making and evaluating point forecasts. *J. Amer. Statist. Assoc.* **106** 746–762. MR2847988

GNEITING, T. and RAFTERY, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc.* **102** 359–378. MR2345548

GRAUERT, H. and FISCHER, W. (1978). *Differential- und Integralrechnung. II: Differentialrechnung in Mehreren Veränderlichen Differentialgleichungen*, corrected ed. *Heidelberger Taschenbücher* [*Heidelberg Paperbacks*] **36**. Springer, Berlin. MR0514400

HEINRICH, C. (2014). The mode functional is not elicitable. *Biometrika* **101** 245–251. MR3180670

HUBER, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* **35** 73–101. MR0161415

JOUINI, E., SCHACHERMAYER, W. and TOUZI, N. (2006). Law invariant risk measures have the Fatou property. In *Advances in Mathematical Economics. Vol.* 9. *Adv. Math. Econ.* **9** 49–71. Springer, Tokyo. MR2277714

KOENKER, R. (2005). *Quantile Regression*. *Econometric Society Monographs* **38**. Cambridge Univ. Press, Cambridge. MR2268657

KOU, S. and PENG, X. (2014). On the measurement of economic tail risk. Preprint. Available at http://arxiv.org/pdf/1401.4787v2.pdf.

KOU, S., PENG, X. and HEYDE, C. C. (2013). External risk measures and Basel accords. *Math. Oper. Res.* **38** 393–417. MR3092538

KRÄTSCHMER, V., SCHIED, A. and ZÄHLE, H. (2012). Qualitative and infinitesimal robustness of tail-dependent statistical functionals. *J. Multivariate Anal.* **103** 35–47. MR2823706

KRÄTSCHMER, V., SCHIED, A. and ZÄHLE, H. (2015). Quasi-Hadamard differentiability of general risk functionals and its applications. *Stat. Risk Model.* **32** 25–47. MR3331780

KRÄTSCHMER, V., SCHIED, A. and ZÄHLE, H. (2014). Comparative and qualitative robustness for law-invariant risk measures. *Finance Stoch.* **18** 271–295. MR3177407

KUSUOKA, S. (2001). On law invariant coherent risk measures. In *Advances in Mathematical Economics, Vol.* 3. *Adv. Math. Econ.* **3** 83–95. Springer, Tokyo. MR1886557

LAMBERT, N. (2013). Elicitation and evaluation of statistical functionals. Preprint. Available at http://web.stanford.edu/~nlambert/papers/elicitation.pdf.

LAMBERT, N., PENNOCK, D. M. and SHOHAM, Y. (2008). Eliciting properties of probability distributions. In *Proceedings of the* 9*th ACM Conference on Electronic Commerce* 129–138. ACM, Chicago, IL.

MURPHY, A. H. and DAAN, H. (1985). Forecast evaluation. In *Probability, Statistics and Decision Making in the Atmospheric Sciences* (A. H. Murphy and R. W. Katz, eds.) 379–437. Westview Press, Boulder, CO.

NEWEY, W. K. and POWELL, J. L. (1987). Asymmetric least squares estimation and testing. *Econometrica* **55** 819–847. MR0906565

OSBAND, K. H. (1985). Providing incentives for better cost forecasting. Ph.D. thesis, Univ. California, Berkeley.

OSBAND, K. and REICHELSTEIN, S. (1985). Information-eliciting compensation schemes. *J. Public Econ*. **27** 107–115.

SAVAGE, L. J. (1971). Elicitation of personal probabilities and expectations. *J. Amer. Statist. Assoc*. **66** 783–801. MR0331571

STEINWART, I., PASIN, C., WILLIAMSON, R. and ZHANG, S. (2014). Elicitation and identification of properties. In *Proceedings of the* 27*th Conference on Learning Theory. Journal of Machine Learning Research*: *Workshop and Conference Proceedings*, *Vol*. 35 (M. F. Balcan and C. Szepesvari, eds.) 482–526.

WANG, R. and ZIEGEL, J. F. (2015). Elicitable distortion risk measures: A concise proof. *Statist. Probab. Lett*. **100** 172–175. MR3324090

WEBER, S. (2006). Distribution-invariant risk measures, information, and dynamic consistency. *Math. Finance* **16** 419–441. MR2212272

ZIEGEL, J. F. (2014). Coherence and elicitability. *Math. Finance*. To appear.

INSTITUTE OF MATHEMATICAL STATISTICS
AND ACTUARIAL SCIENCE
UNIVERSITY OF BERN
SIDLERSTRASSE 5, CH-3012 BERN
SWITZERLAND
E-MAIL: tobias.fissler@stat.unibe.ch
johanna.ziegel@stat.unibe.ch

Submitted to the Annals of Statistics
arXiv: 1503.08123

SUPPLEMENT TO "HIGHER ORDER ELICITABILITY AND OSBAND'S PRINCIPLE"

BY TOBIAS FISSLER* AND JOHANNA F. ZIEGEL*

University of Bern

This is a supplement to Fissler and Ziegel (2016) and contains the proofs of Proposition 3.4 and Corollary 5.4.

The notation and references in this supplementary material are relative to the original paper Fissler and Ziegel (2016).

Proof of Proposition 3.4. Let $x \in \mathrm{int}(\mathsf{A})$, $F \in \mathcal{F}$ and let $z \in \mathrm{int}(\mathsf{A})$ be some star point. Using a telescoping argument we obtain

$$\begin{aligned}\bar{S}(x,F) - \bar{S}(z,F) &= \bar{S}(x_1,\dots,x_k,F) - \bar{S}(x_1,\dots,x_{k-1},z_k,F) \\ &\quad + \bar{S}(x_1,\dots,x_{k-1},F) - \bar{S}(x_1,\dots,x_{k-2},z_{k-1},z_k,F) \\ &\quad + \dots \\ &\quad + \bar{S}(x_1,z_2,\dots,z_k,F) - \bar{S}(z_1,\dots,z_k,F) \\ &= \sum_{r=1}^{k} \int_{z_r}^{x_r} \partial_r \bar{S}(x_1,\dots,x_{r-1},v,z_{r+1},\dots,z_k,F)\,\mathrm{d}v.\end{aligned}$$

Invoking the identity at (3.2) yields (3.4) for the expected scores with $\bar{a}(F) = \bar{S}(z,F)$. We denote the right hand side of (3.4) minus $a(y)$ by $I(x,y)$, hence $\bar{I}(x,F) = \bar{S}(x,F) - \bar{S}(z,F)$.

For almost all $y \in \mathsf{O}$, the set $\{x \in \mathbb{R}^k \mid (x,y) \in C^c\} =: A_y$ has k-dimensional Lebesgue measure zero, where C^c is the complement of the set C defined in assumption (VS1). Let $y \in \mathsf{O}$ be such that A_y has measure zero. Then we obtain that for almost all x the sets $\{x_i \in \mathbb{R} \mid (x,y) \in A_y\} =: N_i$ have one-dimensional Lebesgue-measure zero for all $i \in \{1,\dots,k\}$. Therefore, $S(x,\cdot)$ and $I(x,\cdot)$ are continuous in y for almost all x.

Let $(F_n)_{n\in\mathbb{N}}$ be a sequence as in assumption (F1), that is, $(F_n)_{n\in\mathbb{N}}$ converges weakly to δ_y and the support of all F_n is contained in some compact

*Supported by the SNF via grant 152609.
MSC 2010 subject classifications: Primary 62C99; secondary 91B06
Keywords and phrases: Consistency, decision theory, elicitability, Expected Shortfall, point forecasts, propriety, scoring functions, scoring rules, spectral risk measures, Value at Risk

set K. Let φ be a function on O which is locally bounded and continuous at y. By the dominated convergence theorem and the continuous mapping theorem we get that then $\int_{\mathsf{O}} \varphi \, \mathrm{d}F_n \to \varphi(y)$.

By this argument (recalling that $S(x,\cdot)$, $V(x,\cdot)$ are assumed to be locally bounded), if $S(x,\cdot)$ and $I(x,\cdot)$ are continuous at y, then $\bar{S}(x,F_n) - \bar{I}(x,F_n) \to S(x,y) - I(x,y)$. We have shown that $\bar{S}(x,F_n) - \bar{I}(x,F_n)$ does not depend on x, hence the same is true for the limit. Therefore, we can define $a(y) = S(x,y) - I(x,y)$ for almost all y. The function a is $\mathcal{F}$-integrable, since S and I are $\mathcal{F}$-integrable. □

Proof of Corollary 5.4. For the first part of the claim, note that if $\mu(\{1\}) = 1$, then ν_μ coincides with the expectation and is thus 1-elicitable. If $\mu(\{1\}) = 0$, the assertion of the corollary is a direct consequence of Theorem 5.2 (i). If $\lambda := \mu(\{1\}) \in (0,1)$, then we can write

$$\mu = \sum_{m=1}^{k-2} p_m \delta_{q_m} + \lambda \delta_1,$$

where $p_m \in (0,1)$, $\sum_{m=1}^{k-2} p_m = 1 - \lambda$, $q_m \in (0,1)$ and the q_m's are pairwise distinct. Define the probability measure

$$\tilde{\mu} := \sum_{m=1}^{k-2} \frac{p_m}{1-\lambda} \delta_{q_m}.$$

Using Theorem 5.2 (i), the functional $(T'_1, \dots, T'_{k-1}) \colon \mathcal{F} \to \mathbb{R}^{k-1}$ is $(k-1)$-elicitable where $T'_m(F) := F^{-1}(q_m)$, $m \in \{1, \dots, k-2\}$, and $T'_{k-1}(F) = \nu_{\tilde{\mu}}(F)$. Using Lemma 2.6 we can deduce that the functional $(T'_1, \dots, T'_{k-1}, \nu_{\delta_1}) \colon \mathcal{F} \to \mathbb{R}^k$ is k-elicitable. Note that

$$\nu_\mu = (1-\lambda)\nu_{\tilde{\mu}} + \lambda \nu_{\delta_1}.$$

Hence, we can apply the revelation principle to deduce that the functional $T = (T_1, \dots, T_k) : \mathcal{F} \to \mathbb{R}^k$ is k-elicitable where $T_m = T'_m$, $m \in \{1, \dots, k-2\}$, $T_{k-1} = \nu_{\delta_1}$ and $T_k = \nu_\mu$. □

References.

FISSLER, T. and ZIEGEL, J. F. (2016). Higher order elicitability and Osband's principle. *Ann. Statist.*

INSTITUTE OF MATHEMATICAL STATISTICS
AND ACTUARIAL SCIENCE
UNIVERSITY OF BERN
SIDLERSTRASSE 5, CH-3012 BERN
SWITZERLAND
E-MAIL: tobias.fissler@stat.unibe.ch
johanna.ziegel@stat.unibe.ch

3.2. Osband's Principle for identification functions

As we have seen in Theorem 3.2 of Fissler and Ziegel (2016), Osband's principle gives a necessary connection between the gradient of a strictly consistent scoring function and a strict identification function. One can use the same reasoning of the proof to characterize the class of strict identification functions.

Let $V\colon \mathsf{A}\times\mathsf{O}\to\mathbb{R}^k$ be a strict $\mathcal{F}$-identification function for a functional $T\colon \mathcal{F}\to \mathsf{A}\subseteq\mathbb{R}^k$. Let $h\colon \mathsf{A}\to\mathbb{R}^{k\times k}$ be matrix valued function. Then the function

$$hV\colon \mathsf{A}\times\mathsf{O}\to\mathbb{R}^k, \qquad (x,y)\mapsto h(x)V(x,y).$$

is an $\mathcal{F}$-identification function for T. If $\det(h)\neq 0$, it is even a strict $\mathcal{F}$-identification function for T. We claim that the class of strict $\mathcal{F}$-identification functions for T is given by

$$\mathcal{V}=\{hV\colon \mathsf{A}\times\mathsf{O}\to\mathbb{R}^k,\ \det(h)\neq 0\}.$$

To prove the necessity, one can use the same argumentation as in Osband's principle.

Proposition 3.2.1. *Let $T\colon \mathcal{F}\to\mathsf{A}\subseteq\mathbb{R}^k$ be a surjective functional with a strict $\mathcal{F}$-identification function $V\colon \mathsf{A}\times\mathsf{O}\to\mathbb{R}^k$. Then the following two assertions hold.*

(i) If $h\colon \mathsf{A}\to\mathbb{R}^{k\times k}$ is a matrix-valued function with $\det(h)\neq 0$, then $hV\colon \mathsf{A}\times\mathsf{O}\to\mathbb{R}^k$, $(x,y)\mapsto h(x)V(x,y)$, is also a strict $\mathcal{F}$-identification function for T.

(ii) Let $V'\colon \mathsf{A}\times\mathsf{O}\to\mathbb{R}^k$ be a strict $\mathcal{F}$-identification function for T such that both V and V' satisfy Assumption (V1) in Fissler and Ziegel (2016). Then there is a matrix-valued function $h\colon \operatorname{int}(\mathsf{A})\to\mathbb{R}^{k\times k}$ with $\det(h)\neq 0$, such that

$$\bar{V}'(x,F)=h(x)\bar{V}(x,F)$$

for all $x\in\operatorname{int}(\mathsf{A})$ and for all $F\in\mathcal{F}$.

Proof. Assertion (i) is straight forward. Concerning (ii), the proof of the existence of h follows along the lines of the proof of Osband's principle, which is Theorem 3.2 in Fissler and Ziegel (2016) upon replacing $\nabla\bar{S}(x,F)$ by $\bar{V}'(x,F)$. To show the invertibility of $h(x)$, one can interchange the roles of V and V'. Hence, one obtains the existence of a matrix-valued function $g\colon \operatorname{int}(\mathsf{A})\to\mathbb{R}^{k\times k}$ such that

$$\bar{V}(x,F)=g(x)\bar{V}'(x,F)$$

for all $x\in\operatorname{int}(\mathsf{A})$ and for all $F\in\mathcal{F}$. Apparently, using both identities and exploiting Assumption (V1), one gets that $g(x)=(h(x))^{-1}$. □

Remark 3.2.2. If the two identification functions V and V' in Proposition 3.2.1 (ii) both satisfy assumption (V2) in Fissler and Ziegel (2016) then h is also continuous. Therefore, since the determinant is also a continuous map, either $\det(h)>0$ or $\det(h)<0$.

Remark 3.2.3. Patton (2015) considered the case where the class $\mathcal{F}$ is such that two *a priori* different functionals coincide. In particular, he considers the case where $\mathcal{F}$ contains only symmetric distributions such that the mean and the median coincide (of course under the integrability condition that the mean exists). It is immediate that any convex-combination of a strictly consistent scoring function for the mean and for the median elicits this functional. Consequently, one can use both $V_{\mathrm{mean}}(x,y) = x-y$ and $V_{\mathrm{median}}(x,y) = \mathrm{sgn}(x-y)$ as identification functions. However, there is no $h\colon \mathbb{R}\to\mathbb{R}$ such that $V_{\mathrm{mean}}(x,y) = h(x)V_{\mathrm{median}}(x,y)$.

What seems as a contradiction can be explained by the fact that assuming the convexity of $\mathcal{F}$ is crucial for the proof of Osband's principle. However, the convex combination of two symmetric distributions is generally not symmetric (provided that the two distributions have a different center).[1]

3.3. Eliciting the divergence of a strictly consistent scoring function

There is an interesting parallel between the pair (mean, variance) = $(\mathbb{E}, \mathrm{Var})$ and the pair $(\mathrm{VaR}_\alpha, \mathrm{ES}_\alpha)$, $\alpha \in (0,1)$. Using the revelation principle in the first case and invoking Corollary 5.5 in Fissler and Ziegel (2016) both functionals are 2-elicitable. However, there is a stronger connection. Consider the scoring functions

$$S_{\mathbb{E}}(x,y) = (x-y)^2, \qquad S_{\mathrm{VaR}_\alpha}(x,y) = \frac{1}{\alpha}\mathbb{1}\{y\le x\}(x-y) - x.$$

Under mild regularity conditions on $\mathcal{F}$, they are strictly $\mathcal{F}$-consistent scoring functions for $\mathbb{E}$ and VaR_α, respectively. Moreover, for a distribution $F\in\mathcal{F}$, a direct computation yields

$$\min_{x\in\mathbb{R}} \bar S_{\mathbb{E}}(x,F) = \mathrm{Var}(F), \qquad \min_{x\in\mathbb{R}} \bar S_{\mathrm{VaR}_\alpha}(x,F) = -\,\mathrm{ES}_\alpha(F).$$

And moreover, a corresponding relation is also true for the spectral risk measures with a finitely supported spectral measure considered in Fissler and Ziegel (2016, Theorem 5.2). That is, for some pairwise disjoint $\alpha_1,\dots,\alpha_k\in(0,1)$ and $p_1,\dots,p_k\in(0,1)$ with $\sum_{i=1}^k p_i = 1$ one has

$$\min_{x\in\mathbb{R}^k} \sum_{i=1}^k p_i\, \bar S_{\mathrm{VaR}_{\alpha_i}}(x_i,F) = -\sum_{i=1}^k p_i\,\mathrm{ES}_{\alpha_i}(F) = -\nu_\mu(F),$$

using the notation of Definition 5.1 in Fissler and Ziegel (2016) with $\mu = \sum_{i=1}^k p_i\delta_{\alpha_i}$. While writing the paper Fissler and Ziegel (2016), we were not aware of these connections. However, shortly after finishing a preprint of this article, Frongillo and Kash (2015b) realized this connection and could prove that this is not just a coincidence. They showed the following theorem.

[1] And if all distributions in $\mathcal{F}$ had the same center, then Assumption (V1) would be violated.

Theorem 3.3.1 (Frongillo and Kash (2015b)). *Under suitable regularity assumptions, the following holds. Let $T\colon \mathcal{F} \to \mathsf{A} \subseteq \mathbb{R}^k$ be a k-elicitable functional with a strictly consistent scoring function $S\colon \mathsf{A} \times \mathsf{O} \to \mathbb{R}$. Define the functional*

$$T'\colon \mathcal{F} \to \mathbb{R}, \qquad F \mapsto \min_{x \in \mathsf{A}} \bar{S}(x, F) = \bar{S}(T(F), F).$$

Then the functional $(T, T')\colon \mathcal{F} \to \mathbb{R}^{k+1}$ is $(k+1)$-elicitable.

Invoking the revelation principle (Proposition 2.3.2), it directly follows that, under the conditions of Theorem 3.3.1, the pair $(T, -T')$ is also $(k+1)$-elicitable.

Remark 3.3.2. Since every (strictly) consistent scoring function S induces a proper scoring rule $R(G, y) = S(T(G), y)$ by Proposition 2.3.4, one can consider the term $\bar{S}(T(F), F) = \bar{R}(F, F)$ as the divergence of the scoring rule R; see Gneiting and Raftery (2007).

With their analysis, Frongillo and Kash (2015b) were also able to solve the question posed in the discussion of Fissler and Ziegel (2016) about the *elicitability order* of a spectral risk measure with a finitely supported spectral measure. According to our definition

> "the *elicitability order* of a real-valued functional T as the smallest number k such that the functional is a component of a k-elicitable functional. It is clear that the elicitability order of the variance is two, and we have shown that the same is true for ES_α for reasonably large classes $\mathcal{F}$. For spectral risk measures ν_μ, the elicitability order is at most $L+1$, where L is the cardinality of the support [of μ]" (Fissler and Ziegel, 2016, p. 1699)

Frongillo and Kash (2015b, Definition 6) define the *elicitation complexity* of a functional T as the smallest number k such that there is a function f and a k-elicitable functional T' such that $T = f \circ T'$. So it is clear that the elicitation complexity is less than or equal to the elicitability order.[2] However, they prove for the relevant case $\mu(\{1\}) < 1$ that the elicitation complexity of the spectral risk measure ν_μ is $L + 1$. This proves also that the elicitability order of ν_μ is exactly $L + 1$.

[2]For the elicitability order, one only considers projections as candidates for the map f in the definition of the elicitation complexity.

4. Scoring Functions Beyond Strict Consistency

If a strictly consistent scoring function S for a functional T is intrinsically meaningful in the sense that it correctly specifies the costs of an incorrect forecast, it is no problem to rank competing forecasts in terms of their realized scores, no matter if one of them was able to report the 'true' functional value or not. On the other hand, if a strictly consistent scoring function was tailored to elicit a functional T and consequently is just a tool, things look different. Ehm et al. (2016, p. 506) write:

> "As there is no obvious reason for a consistent scoring function to be preferred over any other, this raises the question which one of the many alternatives to use."

To give guidance which (strictly consistent) scoring functions to choose, recall that the strict consistency of S only justifies a comparison of two competing forecasts if one of them reports the true functional value. If both of them are misspecified, it is *per se* not possible to draw a conclusion which forecast is 'closer' to the true functional value by comparing the realized scores. To this end, some notions of *order-sensitivity* are desirable. According to Lambert (2013) we say that a scoring function S is $\mathcal{F}$-order-sensitive for a one-dimensional functional $T\colon \mathcal{F} \to \mathsf{A} \subseteq \mathbb{R}$ if for any $F \in \mathcal{F}$ and any $x, z \in \mathsf{A}$ such that either $z \le x \le T(F)$ or $z \ge x \ge T(F)$, then $\bar{S}(x,F) \le \bar{S}(z,F)$. That means if a forecast lies *between* the true functional value and some other forecast, then issuing the forecast between should yield a smaller expected score than issuing the forecast further away. In particular, order-sensitivity implies consistency. *Vice versa*, under weak regularity conditions on the functional, strict consistency also implies order-sensitivity if the functional is real-valued. However, this notion does not allow for a comparison of two misspecified forecasts lying on different sides of the true functional value. This shortcoming could be overcome by introducing a kind of *metrical* order-sensitivity imposing that $d(x, T(F)) \le d(z, T(F))$ implies that $\bar{S}(x,F) \le \bar{S}(z,F)$ where d is some metric on A. This notion of metrical order-sensitivity has a long history in the context of probabilistic forecasting and dates back to Friedman (1983) and Nau (1985). They call a scoring rule $R\colon \mathcal{F} \times \mathsf{O} \to \mathbb{R}$ 'effective' relative to some probability metric d on $\mathcal{F}$ if for all $F, G, H \in \mathcal{F}$ it holds that $\bar{R}(G,F) \le \bar{R}(H,F)$ if and only if $d(G,F) \le d(H,F)$.

Let us consider the other side of the coin and draw attention to the analytic properties of the expected score $x \mapsto \bar{S}(x,F)$, $x \in \mathsf{A}$, for some scoring function S

and some distribution $F \in \mathcal{F}$. The (strict) consistency of S for some functional T is then equivalent to the fact that the expected score has a (unique) global minimum at $x = T(F)$. The order-sensitivity strives for assessing monotonicity properties of the expected score. Whilst this notion is straight forward for the one-dimensional case due to the canonical order on $\mathbb{R}$, one faces the burden of choice when generalizing this notion to the higher-dimensional setting due to the very fact that $\mathbb{R}^k$ has no total order. In Section 4.1, we introduce three generalizations of order-sensitivity in the higher-dimensional setting that appear natural to us: metrical order-sensitivity, componentwise order-sensitivity and order-sensitivity on line segments. Furthermore, we discuss their connections and strive to give conditions when such scoring functions exist and of what form they are. Besides this discussion, we give connections to the notion of *oriented* identification functions. Moreover, passing to the level of the prediction space setting (see Subsection 2.2.3) we recall a different notion of order-sensitivity in the case of two ideal forecasters when their information sets are nested; see Holzmann and Eulert (2014). As a technical result, we show that under weak regularity assumptions on T, the expected score of a strictly consistent scoring function has a unique local minimum – which, of course, coincides with the global minimum at $x = T(F)$. Accompanied with a result on self-calibration (which is a kind of continuity property of the inverse of the expected score), these two findings are of interest on their own in the context of learning.

Besides different notions of monotonicity, the expected score can have other appealing properties such as convexity or quasi-convexity. Apparently, the convexity of the expected score is a desirable property in the context of learning and regression, leading to convex optimization problems; see Subsection 4.2.1. However, the motivation is not limited to these two fields. In the context of forecast comparison, convex scoring functions incentivize to *maximize sharpness, subject to calibration*, thereby generalizing a paradigm raised by Gneiting et al. (2007) in the framework of probabilistic forecasts to point forecasts. Finally, (quasi-)convex scoring functions have an insurance-like interpretation and give incentives for a collaboration between forecasters. An exposition of these diverse motivations and possible applications as well as classifications of (quasi-)convex strictly consistent scoring functions for some popular functionals are given in Section 4.2.

Many interesting functionals, e.g. coherent risk measures, are positively homogeneous. Then, simultaneously changing the unit of the observation and the forecast should not affect the ranking of competing forecasts. However, as shown in Patton (2011), this requires using positively homogenous scoring functions as well. If a functional is translation-equivariant – again standard examples are given by coherent risk measure – then the translation invariance of the corresponding scoring function is again a natural requirement. In Section 4.3 we give a formal concept of generalized equivariance of functional, nesting the practically important cases of homogeneity and translation-equivariance. We discuss that it is a natural requirement on a 'good' scoring function to be compatible with the equivariance of the functional. This leads to the notion of being *order-preserving*.

Section 4.4 briefly classifies possible applications for the three preceding notions of order-sensitivity, (quasi-)convexity, and equivariance / order-preservingness.

4.1. Order-Sensitivity

4.1.1. Different notions of order-sensitivity

As already discussed, the idea of order-sensitivity is that a forecast lying between the true functional value and some other forecast is also assigned an expected score lying between the two other expected scores. If the action domain is one dimensional, there are only two cases to consider: both forecasts are on the left hand side of the functional value or on the right hand side. However, if $\mathsf{A} \subseteq \mathbb{R}^k$ for $k \geq 2$, the notion of 'lying between' is ambiguous. Two obvious interpretations for the multidimensional case are the componentwise interpretation and the interpretation that one forecast is the convex combination of the true functional value and the other forecast. In other words, the latter means that the two forecasts and the true functional value are lying on the same line segment and on this line segment, the two forecasts are on the same side of the functional. This amounts to the two formal definitions where we fix some generic functional $T\colon \mathcal{F} \to \mathsf{A} \subseteq \mathbb{R}^k$ and a scoring function $S\colon \mathsf{A} \times \mathsf{O} \to \mathbb{R}$.

Definition 4.1.1 (Componentwise order-sensitivity)**.** A scoring function S is componentwise $\mathcal{F}$-order-sensitive for T if for all $F \in \mathcal{F}$, $t = T(F)$ and for all $x, z \in \mathsf{A}$ we have that

$$z \leq x \leq t \text{ or } z \geq x \geq t \quad \Longrightarrow \quad \bar{S}(x, F) \leq \bar{S}(z, F), \tag{4.1.1}$$

where the inequalities on the left hand side are meant componentwise. If additionally

$$\big(z \leq x \leq t \text{ or } z \geq x \geq t\big) \text{ and } x \neq z \quad \Longrightarrow \quad \bar{S}(x, F) < \bar{S}(z, F), \tag{4.1.2}$$

we call S strictly componentwise $\mathcal{F}$-order-sensitive for T.

Remark 4.1.2. In economic terms, a componentwise order-sensitive scoring function rewards *Pareto improvements*[1] in the sense that improving the prediction performance in one component without deteriorating the prediction ability in the other components results in a lower expected score.

[1]The definition of the *Pareto principle* according to Scott and Marshall (2009): "A principle of welfare economics derived from the writings of Vilfredo Pareto, which states that a legitimate welfare improvement occurs when a particular change makes at least one person better off, without making any other person worse off. A market exchange which affects nobody adversely is considered to be a 'Pareto-improvement' since it leaves one or more persons better off. 'Pareto optimality' is said to exist when the distribution of economic welfare cannot be improved for one individual without reducing that of another."

Definition 4.1.3 (Order-sensitivity on line segments)**.** Let $\|\cdot\|$ be the Euclidean norm on $\mathbb{R}^k$. A scoring function S is $\mathcal{F}$-order-sensitive on line segments for a functional T if for all $F \in \mathcal{F}$, $t = T(F)$, and for all $v \in \mathbb{S}^{k-1} = \{x \in \mathbb{R}^k : \|x\| = 1\}$ the map

$$\psi\colon D = \{s \in [0,\infty)\colon t + sv \in \mathsf{A}\} \to \mathbb{R}, \quad s \mapsto \bar{S}(t + sv, F) \tag{4.1.3}$$

is increasing. If the map ψ is strictly increasing, we call S strictly $\mathcal{F}$-order-sensitive on line segments for T.

Remark 4.1.4. The definition of order-sensitivity on line segments does not depend on the particular choice of the norm.

We have already discussed that these two notions of order-sensitivity do not allow for a comparison of any two misspecified forecasts, no matter where they are relative to the true functional value. An intuitive requirement could be 'the closer to the true functional value the smaller the expected score', thus calling for the notion of a metric.

Definition 4.1.5 (Metrical order-sensitivity)**.** A scoring function S is metrically $\mathcal{F}$-order-sensitive for T relative to some metric d on A if for all $F \in \mathcal{F}$, $t = T(F)$ and for all $x, z \in \mathsf{A}$ we have that

$$d(x,t) \le d(z,t) \implies \bar{S}(x,F) \le \bar{S}(z,F). \tag{4.1.4}$$

If additionally the inequalities in (4.1.4) are strict, we say that S is strictly metrically $\mathcal{F}$-order-sensitive for T relative to d.

Since, for a fixed functional T and some fixed distribution F, we always have a fixed reference point $T(F)$ and we have the induced vector-space structure of $\mathbb{R}^k$ on A, we shall tacitly assume throughout the thesis that the metric is induced by a norm. Recalling the fact that all norms are equivalent on $\mathbb{R}^k$, this means that we also work with the usual (Euclidean) topology on A and, in particular, the notion of continuity is the same. Sometimes, when we need that the norm $\|x\|$, $x \in \mathsf{A}$, is sensitive with respect to changes in any of the components of x, we will assume that the norm corresponds to an ℓ^p-norm $\|\cdot\|_p$ where $\|x\|_p := (\sum_{i=1}^k |x_i|^p)^{1/p}$, where usually $p \in [1,\infty)$ and sometimes we will consider $p = \infty$ separately (then, $\|x\|_\infty = \sup_{i=1,\dots,k} |x_i|$). This has the advantage of a tighter presentation of our analysis and results, and should also cover most cases of applied interest on a finite-dimensional action domain. However, for a more general action domain, and especially for the infinite-dimensional case of forecasting functions, or probabilisitic forecasting, this assumption would be somehow restrictive and one should allow for general metrics. But this is beyond the scope of this thesis and is relegated to future work. Note that we will often not mention the norm / metric explicitly if the choice of the norm / metric is clear from the context or does not matter.

Remark 4.1.6. It is valuable to mention that, similarly to (strict) consistency, all three notions of (strict) order-sensitivity are preserved by the equivalence relation on scoring functions.

The three different notions of order-sensitivity have some inspiration from the literature. The notion of componentwise order-sensitivity corresponds almost literally to the notion of *accuracy-rewarding* introduced by Lambert et al. (2008). And apparently, metrically order-sensitivity scoring functions have their counterparts in the field of probabilistic forecasting in *effective scoring rules* introduced by Friedman (1983) and further investigated by Nau (1985). Actually, the latter paper – and in particular Proposition 3 therein which we will present in our notation in the sequel – have also given the inspiration for the notion of order-sensitivity on line segments.

Proposition 4.1.7 (Proposition 3 in Nau (1985))**.** *Let $R\colon \mathcal{F}\times\mathsf{O}\to\mathbb{R}$ be a scoring rule and let $F,G\in\mathcal{F}$, $F\neq G$ and $H=(1-\lambda)F+\lambda G$ for some $\lambda\in[0,1)$. Then, if R is proper, one has the inequality*

$$\bar{R}(H,F)\leq\bar{R}(G,F),$$

with a strict inequality if R is strictly proper.

It is obvious that any of the three notions of (strict) order-sensitivity implies (strict) consistency. The next lemma formally states this result and gives some logical implications concerning the different notions of order-sensitivity. The proof is standard and therefore omitted.

Lemma 4.1.8. *Let $T\colon\mathcal{F}\to\mathsf{A}\subseteq\mathbb{R}^k$ be a functional and $S\colon\mathsf{A}\times\mathsf{O}\to\mathbb{R}$ a scoring function. Then the following implications hold.*

(i) *If S is (strictly) metrically $\mathcal{F}$-order-sensitive for T relative to the ℓ^p-norm for $p\in[1,\infty)$, then S is (strictly) componentwise $\mathcal{F}$-order-sensitive for T.*

(i') *If S is (strictly) metrically $\mathcal{F}$-order-sensitive for T relative to the ℓ^∞-norm, then S is componentwise $\mathcal{F}$-order-sensitive for T.*

(i") *If S is (strictly) metrically $\mathcal{F}$-order-sensitive for T relative to the ℓ^∞-norm, then S is (strictly) $\mathcal{F}$-consistent for T.*

(ii) *If S is (strictly) componentwise $\mathcal{F}$-order-sensitive for T, then S is (strictly) $\mathcal{F}$-order-sensitive on line segments for T.*

(iii) *If S is (strictly) $\mathcal{F}$-order-sensitive on line segments for T, then S is (strictly) $\mathcal{F}$-consistent for T.*

The purpose of points (i') and (i") in Lemma 4.1.8 is to specify point (i) for the case of the ℓ^∞-norm. Essentially, the only difference is that metrical order-sensitivity with respect to the ℓ^∞-norm – whether strict or not – does not necessarily imply strict componentwise order-sensitivity, but merely implies componentwise order-sensitivity.

4.1.2. The one-dimensional case

Since the notions of componentwise order-sensitivity and order-sensitivity on line segments coincide for $k = 1$, we will loosely refer to both notions as order-sensitivity in the one-dimensional case, in contrast to metrical order-sensitivity.

Steinwart et al. (2014, Theorem 5 and Corollary 9) showed that under the regularity conditions that the functional $T\colon \mathcal{F} \to \mathsf{A} \subseteq \mathbb{R}$ is continuous and 'strictly locally non-constant' – a property similar to assumption (V1) in Fissler and Ziegel (2016) – the elicitability of T is equivalent to (i) T having convex level sets; (ii) the existence of an *oriented* strict identification function for T; and (iii) the existence of an order-sensitive scoring function for T. Their setting is the following: As a class $\mathcal{F}$ they consider measures which are absolutely continuous with respect to some finite reference measure; they consider the topology which is induced by the total variation distance on $\mathcal{F}$; and they use the usual Euclidean distance on A. They say that a strict $\mathcal{F}$-identification function for T is oriented if

$$\bar{V}(x,F)\begin{cases} <0, & x<T(F)\\ =0, & x=T(F)\\ >0, & x>T(F)\end{cases} \tag{4.1.5}$$

for any $F \in \mathcal{F}$ and $x \in \mathsf{A}$; see Subsection 4.1.7 for a more detailed discussion of oriented identification functions and their connections to order-sensitivity. Then, they use Osband's principle and a construction in the spirit of Proposition 3.4 in Fissler and Ziegel (2016) to show the existence of an $\mathcal{F}$-order-sensitive scoring function for T. Indeed, under the conditions of this proposition, one has the following representation for any strictly $\mathcal{F}$-consistent scoring function

$$\bar{S}(x,F) = \int_z^x h(v)\bar{V}(v,F)\,\mathrm{d}v + \bar{a}(F)$$

for all $x \in \mathsf{A}$, $F \in \mathcal{F}$, some $z \in \mathsf{A}$, some $\mathcal{F}$-integrable function a and a continuous function $h\colon \operatorname{int}(\mathsf{A}) \to \mathbb{R}$. If V is oriented, then the strict $\mathcal{F}$-consistency of S and the surjectivity of T imply that $h > 0$ almost everywhere. This in turn implies the strict $\mathcal{F}$-order-sensitivity of S.

However, using an approach in the spirit of Proposition 4.1.7 one can omit to use the machinery of Osband's principle and hence dispense with a lot of regularity assumptions on T and $\mathcal{F}$. What is essential to assume is the mixture-continuity of the functional T.

Definition 4.1.9 (Mixture-continuity)**.** Let $\mathcal{F}$ be convex. A functional $T\colon \mathcal{F} \to \mathsf{A} \subseteq \mathbb{R}^k$ is called *mixture-continuous* if for all $F, G \in \mathcal{F}$ the map

$$[0,1] \to \mathbb{R}, \quad \lambda \mapsto T((1-\lambda)F + \lambda G)$$

is continuous.

For the purposes of this thesis, it suffices to consider only the (induced) Euclidean topology on A. However, especially if one works with an infinite dimensional action domain, the choice of the topology plays a more important role.

Proposition 4.1.10. *Let $\mathcal{F}$ be convex and $T\colon \mathcal{F} \to \mathsf{A} \subseteq \mathbb{R}$ be a surjective and mixture-continuous functional. Then any (strictly) $\mathcal{F}$-consistent scoring function for T is (strictly) $\mathcal{F}$-order-sensitive for T.*

The proof uses an argument which is quite similar to the proof of Nau (1985, Proposition 3); see also Lambert (2013, Proposition 2) and Bellini and Bignozzi (2015, Proposition 3.4).

Proof. Let $F \in \mathcal{F}$ and $t = T(F)$. Without loss of generality, we assume that $t < x < z$ for some $x, z \in \mathsf{A}$. Then there exists some $G \in \mathcal{F}$ such that $T(G) = z$. Using the mixture-continuity of T there is a $\lambda \in (0,1)$ such that for $H = (1-\lambda)F + \lambda G$ one has $T(H) = x$. Due to the $\mathcal{F}$-consistency of S one obtains that

$$(1-\lambda)\bar{S}(x,F) + \lambda\bar{S}(x,G) = \bar{S}(x,H) \le \bar{S}(z,H) = (1-\lambda)\bar{S}(z,F) + \lambda S(z,G).$$

This is equivalent to

$$\frac{\lambda}{1-\lambda}\big(\bar{S}(x,G) - \bar{S}(z,G)\big) \le \bar{S}(z,F) - \bar{S}(x,F).$$

If S is $\mathcal{F}$-consistent for T, the left hand side is non-negative which yields the claim. If moreover S is strictly $\mathcal{F}$-consistent, the left hand side is strictly positive and all inequalities in the definition of order-sensitivity become strict. $\square$

It is quite appealing that one does not have to specify any topology on $\mathcal{F}$ to define mixture-continuity because it suffices to work with the induced Euclidean topology on $[0,1]$ and on $\mathsf{A} \subseteq \mathbb{R}^k$. Moreover, this criterion is often quite handsome to check. For example, it is straight forward to see that the ratio of expectations is mixture-continuous. Moreover, by the implicit function theorem it is possible to verify the mixture-continuity of quantiles and expectiles directly under appropriate regularity conditions (e.g., in the case of quantiles, all distributions in $\mathcal{F}$ should be C^1 with non-vanishing derivatives). However, we pursue a different approach and generalize Bellini and Bignozzi (2015, Proposition 3.4c)). The main generalization of our version is that our result is valid for distributions that do not have compact support (however, the image of the functional must be bounded), and moreover, we directly give a higher-dimensional version of it.

Proposition 4.1.11. *Let $T\colon \mathcal{F} \to \mathbb{R}^k$ be an elicitable functional with a strictly $\mathcal{F}$-consistent scoring function $S\colon \mathbb{R}^k \times \mathbb{R}^d \to \mathbb{R}$ such that $\bar{S}(\cdot, F)$ is continuous for all $F \in \mathcal{F}$. Then T is mixture-continuous on any $\mathcal{F}_0 \subseteq \mathcal{F}$ such that $\mathcal{F}_0$ is convex and the image $T(\mathcal{F}_0)$ is bounded.*

Proof. Let $\mathcal{F}_0 \subseteq \mathcal{F}$ be convex such that $T(\mathcal{F}_0) \subseteq [-C, C]^k$ for some $C > 0$. Let $F, G \in \mathcal{F}_0$. Define $h_{F,G} \colon [-C, C]^k \times [0,1] \to \mathbb{R}$ via

$$h_{F,G}(x, \lambda) = \bar{S}(x, (1-\lambda)F + \lambda G) = (1-\lambda)\bar{S}(x, F) + \lambda \bar{S}(x, G).$$

Then $h_{F,G}$ is jointly continuous, and due to the strict consistency

$$T((1-\lambda)F + \lambda G) = \operatorname*{arg\,min}_{x \in [-C,C]^k} h_{F,G}(x, \lambda).$$

By virtue of the Berge Maximum Theorem (Aliprantis and Border, 2006, Theorem 17.31 and Lemma 17.6), the function

$$\operatorname*{arg\,min}_{x \in [-C,C]^k} h_{F,G}(x, \lambda)$$

is continuous in λ. □

Remark 4.1.12. Similarly to the original proof of Bellini and Bignozzi (2015) a sufficient criterion for the continuity of $\bar{S}(\cdot, F)$ for any $\mathcal{F}$ is that for all $y \in \mathbb{R}^d$ $S(x, y)$ is quasi-convex and continuous in x.

Recall that, under appropriate regularity conditions on $\mathcal{F}$, the asymmetric piecewise linear loss $S_\alpha(x, y) = (\mathbb{1}\{y \le x\} - \alpha)(x - y)$ and the asymmetric piecewise quadratic loss $S_\tau(x, y) = |\mathbb{1}\{y \le x\} - \tau|(x - y)^2$ are strictly consistent scoring functions for the α-quantile and the τ-expectile, respectively, and both, S_α as well as S_τ, are continuous in their first argument and convex. Hence, Proposition 4.1.11 yields that both quantiles and expectiles are mixture-continuous.

Bellini and Bignozzi (2015) showed that the weak continuity of a functional T implies its mixture-continuity. Therefore, the assumption in Steinwart et al. (2014) of continuity of T with respect to the total variation distance clearly implies that T is mixture-continuous. Consequently, one can also derive the order-sensitivity in their framework directly with Proposition 4.1.10.

Ehm et al. (2016, Subsection 2.4) showed order-sensitivity of any (regular) strictly consistent scoring function for the α-quantile and τ-expectile by a completely different method, namely by a mixture representation of any (regular) strictly consistent scoring function in the spirit of Proposition 2.3.1.

Lambert (2013) showed that it is a harder requirement to have order-sensitivity if $T(\mathcal{F})$ is discrete. Then both approaches, the one invoking Osband's principle and the other one using mixture-continuity, do not work because then the interior of the image of T is empty and moreover mixture-continuity implies that the functional is constant (such that only trivial cases can be considered). Furthermore, it is proven in Lambert (2013) that for a functional T with a discrete image, all strictly consistent scoring functions are order-sensitive if and only if there is one order-sensitive scoring function for T. In particular, there are functionals admitting strictly consistent scoring functions that are not order-sensitive, one such example being the mode functional.[2]

4.1.3. Unique local minimum

Before investigating the different notions of componentwise order-sensitivity, metrical order-sensitivity and order-sensitivity on line segments in the higher-dimensional case, we first pose the question of what is possible to deduce concerning monotonicity properties of the expected score of a strictly consistent scoring function in the framework of Proposition 4.1.10, that is, if one only assumes the mixture-continuity of a functional $T\colon \mathcal{F} \to \mathsf{A}$.

To this end, it is essential to obtain a deeper understanding of the paths $\gamma\colon [0,1] \to \mathsf{A} \subseteq \mathbb{R}^k$, $\gamma(\lambda) = T(\lambda F + (1-\lambda)G)$ for $F, G \in \mathcal{F}$. If T is elicitable, it necessarily has convex level sets by Proposition 2.3.7. This in turn implies that the level sets of γ can only be closed intervals (including the case of singletons and the empty set). This rules out loops and some other possible pathologies of γ. Furthermore, under the assumption that T is identifiable, one can even show that the path γ is either injective or constant. In particular, this implies that T has convex level sets.

Lemma 4.1.13. *Let $\mathcal{F}$ be convex and $T\colon \mathcal{F} \to \mathsf{A} \subseteq \mathbb{R}^k$ be identifiable with a strict $\mathcal{F}$-identification function $V\colon \mathsf{A} \times \mathsf{O} \to \mathbb{R}^k$. Then for any $F, G \in \mathcal{F}$, the path $\gamma\colon [0,1] \to \mathsf{A}$, $\gamma(\lambda) = T(\lambda F + (1-\lambda)G)$ is either constant or injective.*

Proof. Let $F, G \in \mathcal{F}$ such that $t = T(F) = T(G)$. Define γ as in the lemma. Then for any $\lambda \in [0,1]$, one has $\bar{V}(t, \lambda F + (1-\lambda)G) = \lambda \bar{V}(t, F) + (1-\lambda)\bar{V}(t, G) = 0$. Since V is a strict $\mathcal{F}$-identification function for T, $t = \gamma(\lambda)$ for all $\lambda \in [0,1]$.

Now let $T(F) \neq T(G)$ and let $0 \le \lambda < \lambda' \le 1$. Since V is a strict $\mathcal{F}$-identification function, $\bar{V}(T(F), G) \neq 0$ (and symmetrically $\bar{V}(T(G), F) \neq 0$.) Assume that $\gamma(\lambda) = \gamma(\lambda')$. Define

$$H_\lambda = \lambda F + (1-\lambda)G, \qquad H_{\lambda'} = \lambda' F + (1-\lambda')G\,.$$

Consequently, there are $\mu, \mu' \in \mathbb{R}$ such that $F = \mu H_\lambda + (1-\mu)H_{\lambda'}$ and $G = \mu' H_\lambda + (1-\mu')H_{\lambda'}$. Then

$$\bar{V}(\gamma(\lambda), F) = \mu \bar{V}(\gamma(\lambda), H_\lambda) + (1-\mu)\bar{V}(\gamma(\lambda), H_{\lambda'}) = 0.$$

And similarly $\bar{V}(\gamma(\lambda), G) = 0$. Consequently, $T(F) = \gamma(\lambda) = T(G)$, which is a contradiction to the assumption that $T(F) \neq T(G)$. Hence, $\gamma(\lambda) \neq \gamma(\lambda')$. □

[2] Note that due to Proposition 1 in Heinrich (2014), the mode functional is elicitable relative to the class of probability measure $\mathcal{F}$ containing unimodal discrete measures. Moreover, interpreting the mode functional as a set-valued functional, it is elicitable in the sense of Remark 2.1.5. A strictly $\mathcal{F}$-consistent scoring function is given by $S(x,y) = \mathbb{1}\{x \neq y\}$. The main result of Heinrich (2014) is that the mode functional is *not* elicitable relative to the class $\mathcal{F}$ of unimodal probability measures with Lebesgue densities.

Proposition 4.1.14. *Let $\mathcal{F}$ be convex and $T\colon \mathcal{F} \to \mathsf{A} \subseteq \mathbb{R}^k$ be mixture-continuous and surjective. Let $S\colon \mathsf{A} \times \mathsf{O} \to \mathbb{R}$ be strictly $\mathcal{F}$-consistent for T. Then for each $F \in \mathcal{F}$, $t = T(F)$ and each $x \in \mathsf{A}$, $x \neq t$ there is a continuous path $\gamma\colon [0,1] \to \mathsf{A}$ such that $\gamma(0) = x$, $\gamma(1) = t$ and the function $[0,1] \ni \lambda \mapsto \bar{S}(\gamma(\lambda), F)$ is decreasing. Additionally, for $\lambda < \lambda'$ such that $\gamma(\lambda) \neq \gamma(\lambda')$ it holds that $\bar{S}(\gamma(\lambda), F) > \bar{S}(\gamma(\lambda'), F)$.*

Proof. Let $F \in \mathcal{F}$, $t = T(F)$ and $x \neq t$. Then there is some $G \in \mathcal{F}$ with $x = T(G)$. Define the map

$$\gamma\colon [0,1] \to \mathsf{A}, \quad \lambda \mapsto T(\lambda F + (1-\lambda)G).$$

Clearly, $\gamma(0) = x$ and $\gamma(1) = t$. Due to the mixture-continuity of T, the path γ is also continuous. The rest follows along the lines of the proof of Proposition 4.1.10. □

Remark 4.1.15. (i) Proposition 4.1.14 remains valid if S is only $\mathcal{F}$-consistent. Then, we merely have that the function $[0,1] \ni \lambda \mapsto \bar{S}(\gamma(\lambda), F)$ is decreasing, so the last inequality in Proposition 4.1.14 is not necessarily strict.

(ii) If one assumes in Proposition 4.1.14 that T is also identifiable, one can use the injectivity of γ implied by Lemma 4.1.13 to see that the function $[0,1] \ni \lambda \mapsto \bar{S}(\gamma(\lambda), F)$ is strictly decreasing.

Under certain (weak) regularity conditions, the expected scores of a strictly consistent scoring function has only one local minimum.

Proposition 4.1.16. *Let $\mathcal{F}$ be convex and $T\colon \mathcal{F} \to \mathsf{A} \subseteq \mathbb{R}^k$ be mixture-continuous and surjective. If $S\colon \mathsf{A} \times \mathsf{O} \to \mathbb{R}$ is strictly $\mathcal{F}$-consistent for T, then for all $F \in \mathcal{F}$ the expected score $\bar{S}(\cdot, F)\colon \mathsf{A} \to \mathbb{R}$ has only one local minimum which is at $t = T(F)$.*

Proof. Let $F \in \mathcal{F}$ with $t = T(F)$. Due to the strict $\mathcal{F}$-consistency of S, the expected score $\bar{S}(\cdot, F)$ has a local minimum at t. Assume there is another local minimum at some $x \neq t$. Then there is a distribution $G \in \mathcal{F}$ with $x = T(G)$. Consider the path

$$\gamma\colon [0,1] \to \mathsf{A}, \quad \lambda \mapsto T(\lambda F + (1-\lambda)G).$$

Due to Proposition 4.1.14 the function $\lambda \mapsto \bar{S}(\gamma(\lambda), F)$ is decreasing and strictly decreasing when we move on the image of the path from x to t. Hence $\bar{S}(\cdot, F)$ cannot have a local minimum at $x = \gamma(0)$. □

4.1.4. Self-calibration

With Proposition 4.1.14 it is possible to prove that, under mild regularity conditions, strictly consistent scoring functions fulfill a *local* property called ‘self-calibration’.

Definition 4.1.17 (Self-calibration)**.** A scoring function $S\colon \mathsf{A} \times \mathsf{O} \to \mathbb{R}$ is called $\mathcal{F}$-self-calibrated for a functional $T\colon \mathcal{F} \to \mathsf{A} \subseteq \mathbb{R}^k$ with respect to a metric d on A if for all $\varepsilon > 0$ and for all $F \in \mathcal{F}$ there is a $\delta = \delta(\varepsilon, F) > 0$ such that for all $x \in \mathsf{A}$ and $t = T(F)$

$$\bar{S}(x, F) - \bar{S}(t, F) < \delta \quad \Longrightarrow \quad d(t, x) < \varepsilon.$$

In some sense, self-calibration can be considered as the continuity of the inverse of the expected score $\bar{S}(\cdot, F)$ at the global minimum $x = T(F)$. This property ensures that convergence of the expected score to its global minimum implies convergence of the forecast to the true functional value. Therefore, this property is particularly advantageous in the context of learning and M-estimation.[3] The notion of self-calibration was introduced by Steinwart (2007) in the context of machine learning. In a preprint of Steinwart et al. (2014)[4], the authors translate this concept to the setting of scoring functions and give a definition which corresponds to Definition 4.1.17. We cite their explanation, using our notation:

> "For self-calibrated S, every δ-approximate minimizer of $\bar{S}(\cdot, F)$, approximates the desired property $T(F)$ with precision not worse than ε. [...] In some sense order sensitivity is a global and qualitative notion while self-calibration is a local and quantitative notion."

It is relatively straight forward that self-calibration implies strict consistency.

Lemma 4.1.18. *Let $S\colon \mathsf{A} \times \mathsf{O} \to \mathbb{R}$ be an $\mathcal{F}$-self-calibrated scoring function for some functional $T\colon \mathcal{F} \to \mathsf{A} \subseteq \mathbb{R}^k$ with respect to some metric d. Then S is strictly $\mathcal{F}$-consistent for T.*

Proof. Let S be $\mathcal{F}$-self-calibrated for T with respect to d. Let $F \in \mathcal{F}$, $t = T(F)$ and $x \in \mathsf{A}$ with $x \neq t$. Then for $\varepsilon := d(t, x)/2 > 0$ there is a $\delta > 0$ such that

$$\bar{S}(x, F) - \bar{S}(t, F) \geq \delta > 0.$$

□

In the preprint of Steinwart et al. (2014) it is shown for $k = 1$ that order-sensitivity implies self-calibration. The next Proposition shows that the kind of

[3] Let us briefly sketch the rationale. Assume that we have an i.i.d. sequence $(Y_t)_{t \in \mathbb{N}}$ of observations with some distribution $F \in \mathcal{F}$. Let $S\colon \mathsf{A} \times \mathsf{O} \to \mathbb{R}$ be strictly $\mathcal{F}$-consistent and $\mathcal{F}$-self-calibrated for a functional $T\colon \mathcal{F} \to \mathsf{A}$ with respect to a metric d. Let $\hat{F}_n$ be the empirical distribution based on the first n observations $Y_1, \dots, Y_n$. Then, by the strong law of large numbers $\bar{S}(x, \hat{F}_n)$ converges to $\bar{S}(x, F)$ almost surely for all $x \in \mathsf{A}$. Now assume that this convergence is *uniform* in x, that is $\sup_{x \in \mathsf{A}} |\bar{S}(x, \hat{F}_n) - \bar{S}(x, F)| \to 0$ almost surely. Define $\hat{x}_n := \arg\min_{x \in \mathsf{A}} \bar{S}(x, \hat{F}_n)$, the empirical M-estimator. Then one can show that $d(x_n, t) \to 0$ almost surely where $t = T(F)$. Indeed, let $\varepsilon > 0$. Due to the self-calibration, there is a $\delta > 0$ such that $\bar{S}(x, F) - \bar{S}(t, F) < \delta$ implies that $d(x, t) < \varepsilon$. Let $n_0 \in \mathbb{N}$ such that almost surely $|\bar{S}(x, \hat{F}_n) - \bar{S}(x, F)| < \delta/2$ for all $x \in \mathsf{A}$, and in particular for $x = \hat{x}_n$. Consequently, $\min_{x \in \mathsf{A}} \bar{S}(x, \hat{F}_n) = \bar{S}(\hat{x}_n, \hat{F}_n) \leq \bar{S}(t, F) + \delta/2$. Then, the triangle inequality yields $\bar{S}(\hat{x}_n, F) - \bar{S}(t, F) < \delta$, which concludes the argument.

[4] At the moment of writing this thesis, this preprint was still online available at http://users.cecs.anu.edu.au/~williams/papers/P196.pdf.

order-sensitivity given by Proposition 4.1.14 also implies self-calibration for $k > 1$ when the metric is induced by a norm (recall that all norms are equivalent on $\mathbb{R}^k$ such that is suffices to consider the Euclidean norm in this case).

Proposition 4.1.19. *Let $\mathcal{F}$ be convex and $T\colon \mathcal{F} \to \mathsf{A} \subseteq \mathbb{R}^k$ a surjective and mixture-continuous functional. If $S\colon \mathsf{A}\times\mathsf{O} \to \mathbb{R}$ is strictly $\mathcal{F}$-consistent for T and $\bar{S}(\cdot, F)\colon \mathsf{A} \to \mathbb{R}$ is continuous for all $F \in \mathcal{F}$, then S is $\mathcal{F}$-self-calibrated for T.*

Proof. Let $F \in \mathcal{F}$, $t = T(F)$ and $\varepsilon > 0$. Define

$$\delta := \min\{\bar{S}(x, F) - \bar{S}(t, F)\colon x \in \mathsf{A},\ \|x - t\| = \varepsilon\}.$$

Due to the continuity of $\bar{S}(\cdot, F)$, the minimum is well-defined and, as a consequence of the strict $\mathcal{F}$-consistency of S for T, δ is positive. Now we show the implication

$$\|x - t\| \ge \varepsilon \quad \Longrightarrow \quad \bar{S}(x, F) - \bar{S}(t, F) \ge \delta$$

for all $x \in \mathsf{A}$.

Let $x \in \mathsf{A}$. If $\|x-t\| = \varepsilon$, we have, by the definition of δ, that $\bar{S}(x, F)-\bar{S}(t, F) \ge \delta$. Assume that $\|x - t\| > \varepsilon$. Then there is a distribution $G \in \mathcal{F}$ with $T(G) = x$. Due to Proposition 4.1.14 there is a continuous path $\gamma\colon [0, 1] \to \mathsf{A}$ such that $\gamma(0) = x$, $\gamma(1) = t$ and such that $\bar{S}(\gamma(\lambda), F)$ is decreasing in λ. Moreover, if $\lambda < \lambda'$ such that $\gamma(\lambda) \ne \gamma(\lambda')$ it holds that $\bar{S}(\gamma(\lambda), F) > \bar{S}(\gamma(\lambda'), F)$. Due to the continuity of γ there is some $x' \in \gamma([0, 1])$ with $\|x' - t\| = \varepsilon$. Then we obtain

$$\bar{S}(x, F) - \bar{S}(t, F) > \bar{S}(x', F) - \bar{S}(t, F) \ge \delta.$$

This concludes the proof. □

4.1.5. Componentwise order-sensitivity

As mentioned further above, the notion of componentwise order-sensitivity can be traced back to Lambert et al. (2008). In their Theorem 5, they claim that whenever a functional has a componentwise order-sensitive scoring function, the components of the functional must necessarily be elicitable. Moreover, they assert that any componentwise order-sensitive scoring function is the sum of strictly consistent scoring functions for the components. However, they use quite restrictive regularity assumptions, e.g. they assume that Ω is finite and that the scoring function itself is twice continuously differentiable. We give a proof of both assertions in a more general setting.

Lemma 4.1.20. *Let $T = (T_1, \dots, T_k)\colon \mathcal{F} \to \mathsf{A} \subseteq \mathbb{R}^k$ be a k-dimensional functional with components $T_m\colon \mathcal{F} \to \mathsf{A}_m \subseteq \mathbb{R}$ where $\mathsf{A} = \mathsf{A}_1 \times \cdots \times \mathsf{A}_k$. If there is a strictly componentwise $\mathcal{F}$-order-sensitive scoring function $S\colon \mathsf{A}\times\mathsf{O} \to \mathbb{R}$ for T, then the components T_m, $m \in \{1, \dots, k\}$, are elicitable.*

Proof. Fix $m \in \{1,\dots,k\}$. Let $F \in \mathcal{F}$ and $x, z \in \mathsf{A}$ such that $T_m(F) = x_m$, $x_i = z_i$ for all $i \neq m$ and $x_m \neq z_m$. Due to the strict componentwise $\mathcal{F}$-order-sensitivity of S this implies that

$$\bar{S}(x, F) < \bar{S}(z, F).$$

This in turn means that for any $z = (z_1, \dots, z_k) \in \mathsf{A}$ the map

$$\begin{aligned} &S_{m,z} \colon \mathsf{A}_m \times \mathsf{O} \to \mathbb{R}, \\ &(x_m, y) \mapsto S_{m,z}(x_m, y) := S(z_1, \dots, z_{m-1}, x_m, z_{m+1}, \dots, z_k, y) \end{aligned} \tag{4.1.6}$$

is a strictly $\mathcal{F}$-consistent scoring function for T_m. □

Lemma 4.1.21. *Assume that $\mathcal{F}$ is convex and let $T_m \colon \mathcal{F} \to \mathsf{A}_m \subseteq \mathbb{R}$, $m \in \{1,\dots,k\}$, be mixture-continuous functionals with strictly $\mathcal{F}$-consistent scoring functions $S_m \colon \mathsf{A}_m \times \mathsf{O} \to \mathbb{R}$. Then, the scoring function $S \colon \mathsf{A}_1 \times \cdots \times \mathsf{A}_k \times \mathsf{O} \to \mathbb{R}$ defined by*

$$S(x_1, \dots, x_k, y) = \sum_{m=1}^{k} S_m(x_m, y) \tag{4.1.7}$$

is strictly componentwise $\mathcal{F}$-order-sensitive for $T = (T_1, \dots, T_k)$.

Proof. Due to Proposition 4.1.10 the scoring functions S_m, $m \in \{1,\dots,k\}$, are strictly $\mathcal{F}$-order-sensitive for T_m. Thanks to its additive form, S is strictly componentwise $\mathcal{F}$-order-sensitive. □

Now, we shall establish the reverse direction of Lemma 4.1.21 in the sense that any strictly componentwise order-sensitive scoring function must necessarily be of the additive form given at (4.1.7). In Fissler and Ziegel (2016, Section 4), we established a dichotomy for functionals with elicitable components: In most relevant cases, the functional (the corresponding strict identification function, respectively) satisfies Assumption (V4) therein (e.g., when the functional is a vector of different quantiles and / or different expectiles with the exception of the 1/2-expectile), or it is a vector of ratios of expectations with the same denominator, or it is a combination of both situations. Under some regularity conditions, Propositions 4.2 and 4.4 in Fissler and Ziegel (2016) characterize the form of strictly consistent scoring functions for the first two situations, whereas Remark 4.5 is concerned with the third situation. For this latter situation, any strictly consistent scoring function must be necessarily additive for the respective blocks of the functional. And for the first situation, Fissler and Ziegel (2016, Proposition 4.2) yields the additive form of S automatically. It remains to consider the case of Proposition 4.4, that is, a vector of ratios of expectations with the same denominator.

Proposition 4.1.22. *Let $T \colon \mathcal{F} \to \mathsf{A} \subseteq \mathbb{R}^k$ be a ratio of expectations with the same denominator, that is, $T(F) = \mathbb{E}_F[p(Y)]/\mathbb{E}_F[q(Y)]$ for some $\mathcal{F}$-integrable functions $p \colon \mathsf{O} \to \mathbb{R}^k$, $q \colon \mathsf{O} \to \mathbb{R}$ such that $\mathbb{E}_F[q(Y)] > 0$ for all $F \in \mathcal{F}$. Assume that T is surjective, and that* $\mathrm{int}(\mathsf{A})$ *is a star domain. Moreover, consider the standard* *strict $\mathcal{F}$-identification function $V\colon \mathsf{A}\times\mathsf{O}\to\mathbb{R}^k$, $V(x,y)=q(y)x-p(y)$ and some strictly $\mathcal{F}$-consistent scoring function $S\colon \mathsf{A}\times\mathsf{O}\to\mathbb{R}$ such that the Assumptions (V1), (S2), (F1), and (VS1) in Fissler and Ziegel (2016) hold. If S is strictly componentwise $\mathcal{F}$-order-sensitive for T, then S is of the form*

$$S(x_1,\dots,x_k,y)=\sum_{m=1}^{k}S_m(x_m,y), \tag{4.1.8}$$

for almost all $(x,y)\in\mathsf{A}\times\mathsf{O}$, where $S_m\colon \mathsf{A}_m\times\mathsf{O}\to\mathbb{R}$, $m\in\{1,\dots,k\}$, are strictly $\mathcal{F}$-consistent scoring functions for $T_m\colon\mathcal{F}\to\mathsf{A}_m$, $\mathsf{A}_m:=T_m(\mathcal{F})\subseteq\mathbb{R}$, and $T_m(F)=\mathbb{E}_F[p_m(Y)]/\mathbb{E}_F[q(Y)]$.

Proof. First, note that due to the fact that for fixed $y\in\mathsf{O}$, $V(x,y)$ is a polynomial in x, Assumption (V3) in Fissler and Ziegel (2016) is automatically satisfied. Let $h\colon \operatorname{int}(\mathsf{A})\to\mathbb{R}^{k\times k}$ be the matrix-valued function given in Osband's principle; see Fissler and Ziegel (2016, Theorem 3.2). By Proposition 4.4 (i) of Fissler and Ziegel (2016) we have that

$$\partial_l h_{rm}(x)=\partial_r h_{lm}(x), \qquad\qquad h_{rl}(x)=h_{lr}(x) \tag{4.1.9}$$

for all $r,l,m\in\{1,\dots,k\}$, $l\neq r$, where the first identity holds for almost all $x\in\operatorname{int}(\mathsf{A})$ and the second identity for all $x\in\operatorname{int}(\mathsf{A})$. Moreover, the matrix $\big(h_{rl}(x)\big)_{l,r=1,\dots,k}$ is positive definite for all $x\in\operatorname{int}(\mathsf{A})$. If we show that $h_{lr}=0$ for $l\neq r$, we can use the first part of (4.1.9) and deduce that for all $m\in\{1,\dots,k\}$ there are positive functions $g_m\colon \mathsf{A}'_m\to\mathbb{R}$, where $\mathsf{A}'_m=\{x_m\in\mathbb{R}\colon \exists(z_1,\dots,z_k)\in\operatorname{int}(\mathsf{A})$ and $z_m=x_m\}$, such that

$$h_{mm}(x_1,\dots,x_k)=g_m(x_m)$$

for all $(x_1,\dots,x_k)\in\operatorname{int}(\mathsf{A})$. Then, we can conclude like in the proof of Fissler and Ziegel (2016, Proposition 4.2(ii)).

Fix $l,r\in\{1,\dots,k\}$ with $l\neq r$ and $F\in\mathcal{F}$ such that $T(F)\in\operatorname{int}(\mathsf{A})$. Due to the strict $\mathcal{F}$-consistency of $S_{l,z}$ defined at (4.1.6) we have that

$$0=\frac{\mathrm{d}}{\mathrm{d}x_l}\bar S_{l,z}(x_l,F)=\partial\bar S_{l,z}(x_l,F)=\partial_l\bar S(z_1,\dots,z_{l-1},x_l,z_{l+1},\dots,z_k,F) \tag{4.1.10}$$

whenever $x_l=T_l(F)$ and for all $z\in\operatorname{int}(\mathsf{A})$. This means the map $\operatorname{int}(\mathsf{A})\ni z\mapsto \partial\bar S_{l,z}(T_l(F),F)$ is constantly 0. Hence, for all $x\in\operatorname{int}(\mathsf{A})$

$$\partial_r\partial_l\bar S(x,F)=0$$

whenever $x_l=T_l(F)$. Using the special form of V and Corollary 3.3 in Fissler and Ziegel (2016), we have for $x=t=T(F)$ that

$$0=\partial_r\partial_l\bar S(t,F)=h_{lr}(t)\partial_r\bar V_r(t,F)=h_{lr}(t)\bar q(F)$$

and by assumption $\bar q(F)>0$. Using the surjectivity of T we obtain that $h_{lr}(t)=0$ for all $t\in\operatorname{int}(\mathsf{A})$, which ends the proof. $\square$

Remark 4.1.23. It is no loss of generality to assume that $\bar{q}(F) > 0$ for all $F \in \mathcal{F}$ in Proposition 4.1.22. In order to ensure that T is well-defined, necessarily $\bar{q}(F) \neq 0$ for all $F \in \mathcal{F}$. However, Assumption (V1) implies that $\mathcal{F}$ is convex. So if there are $F_1, F_2 \in \mathcal{F}$ such that $\bar{q}(F_1) < 0$ and $\bar{q}(F_2) > 0$ then there is a convex combination G of F_1 and F_2 such that $\bar{q}(G) = 0$. Consequently, either $\bar{q}(F) > 0$ for all $F \in \mathcal{F}$ or $\bar{q}(F) < 0$ for all $F \in \mathcal{F}$, and by possibly changing the sign of p one can assume that the first case holds.

Remark 4.1.24. One might wonder if the necessary characterization of strictly componentwise order-sensitive scoring functions as the sum of strictly consistent scoring functions for each component carries over to more general functionals with elicitable components, apart from vectors of ratios of expectations with the same denominator or functionals satisfying Assumption (V4) in Fissler and Ziegel (2016). The answer is not completely clear. In the preprint version (Fissler and Ziegel, 2015) we have introduced Assumption (V5). That is, under assumption (V3) in Fissler and Ziegel (2016), for all $F \in \mathcal{F}$ there is a constant $c_F \neq 0$ such that for all $r \in \{1, \dots, k\}$ and for all $x \in \operatorname{int}(\mathsf{A})$ it holds that

$$\partial_r \bar{V}_r(x, F) = c_F.$$

In principle, one can extend the result of Proposition 4.1.22 to the situation, when this Assumption (V5) is satisfied (see also Proposition 4.1 in Fissler and Ziegel (2015)). However, the only functionals with elicitable components satisfying Assumption (V5), that came to our minds, are ratios of expectations with the same denominator. That is why we decided to give the results directly in terms of the latter functionals in the published version (Fissler and Ziegel, 2016) and why we also restricted to this situation in Proposition 4.1.22.

We have already seen in Remark 4.1.2 that the notion of componentwise order-sensitivity has a very appealing interpretation in the sense that it rewards Pareto improvements of the predictions. In some sense, the results of Lemma 4.1.20 and Proposition 4.1.22 give a very clear understanding of that concept including its limitations to the case of functionals only consisting of elicitable components.

4.1.6. Metrical order-sensitivity

We start with an equivalent formulation of metrical order-sensitivity that is easier to check in practice than the mere definition. Recall the convention that we tacitly assume that the metric d on A is induced by a norm, and hence, the induced topology corresponds to the Euclidean topology on A. In particular, then, the notion of mixture-continuity is not ambiguous.

Lemma 4.1.25. *Let $\mathcal{F}$ be convex and $T\colon \mathcal{F} \to \mathsf{A} \subseteq \mathbb{R}^k$ be mixture-continuous and surjective. Let $S\colon \mathsf{A} \times \mathsf{O} \to \mathbb{R}$ be (strictly) $\mathcal{F}$-consistent for T. Then S is* *(strictly) metrically $\mathcal{F}$-order-sensitive for T relative to d if and only if for all $F \in \mathcal{F}$, $t = T(F)$ and $x, z \in \mathsf{A}$ we have the implication*

$$d(x,t) = d(z,t) \implies \bar{S}(x,F) = \bar{S}(z,F). \tag{4.1.11}$$

Proof. Let S be metrically $\mathcal{F}$-order sensitive for T relative to d. Let $F \in \mathcal{F}$, $t = T(F)$, $x, z \in \mathsf{A}$ such that $d(x,t) = d(z,t)$. Then we have both $\bar{S}(x,F) \le \bar{S}(z,F)$ and $\bar{S}(z,F) \le \bar{S}(x,F)$.

Assume that (4.1.11) holds and S is (strictly) $\mathcal{F}$-consistent. Let $F \in \mathcal{F}$ with $t = T(F)$ and $x, z \in \mathsf{A}$. W.l.o.g. assume that $d(x,t) \le d(z,t)$. If $d(x,t) = d(z,t)$, (4.1.11) implies that $\bar{S}(x,F) = \bar{S}(z,F)$ and there is nothing to show. If $d(x,t) < d(z,t)$, we can apply Proposition 4.1.14. There is a continuous path $\gamma\colon [0,1] \to \mathsf{A}$ such that $\gamma(0) = z$ and $\gamma(1) = t$, and the function $[0,1] \ni \lambda \mapsto \bar{S}(\gamma(\lambda), F)$ is decreasing. Due to continuity there is a $\lambda' \in [0,1]$ such that $d(\gamma(\lambda'), t) = d(x,t)$. Due to (4.1.11) it holds that $\bar{S}(x,F) = \bar{S}(\gamma(\lambda'),F) \le \bar{S}(z,F)$. If S is strictly $\mathcal{F}$-consistent then the latter inequality is strict. $\square$

Lemma 4.1.25 directly leads the way to the assertion that for a real-valued functional T there can be at most one strictly metrically order-sensitive scoring function, up to equivalence, of course. To show this, we have to put ourselves into the setting of Osband's principle.

Proposition 4.1.26. *Let $T\colon \mathcal{F} \to \mathsf{A} \subseteq \mathbb{R}$ be a surjective, elicitable and identifiable functional with an oriented strict $\mathcal{F}$-identification function $V\colon \mathsf{A} \times \mathsf{O} \to \mathbb{R}$. If $\mathrm{int}(\mathsf{A}) \neq \emptyset$ is convex and $S, S^*\colon \mathsf{A} \times \mathsf{O} \to \mathbb{R}$ are two strictly metrically $\mathcal{F}$-order-sensitive scoring functions for T such that the Assumptions (V1), (V2), (S1), (F1) and (VS1) from Fissler and Ziegel (2016) (with respect to both scoring functions) hold, then S and S^* are equivalent almost everywhere.*

Proof. We apply Osband's principle, that is, Fissler and Ziegel (2016, Theorem 3.2) to S. Consequently, there is a function $h\colon \mathrm{int}(\mathsf{A}) \to \mathbb{R}$ such that

$$\frac{\mathrm{d}}{\mathrm{d}x}\bar{S}(x,F) = h(x)\bar{V}(x,F) \tag{4.1.12}$$

for all $F \in \mathcal{F}$ and $x \in \mathrm{int}(\mathsf{A})$. Due to the strict $\mathcal{F}$-consistency of S and the orientation of V, $h \ge 0$. We show that actually $h > 0$. Applying Lemma 4.1.25, one has that

$$\bar{S}(T(F)+x,F) = \bar{S}(T(F)-x,F) \tag{4.1.13}$$

for all $F \in \mathcal{F}$, $x \in \mathbb{R}$ such that $T(F)+x, T(F)-x \in \mathrm{int}(\mathsf{A})$. Hence, also the derivative with respect to x of the left-hand side of (4.1.13) must coincide with the derivative on the right-hand side. This yields, using (4.1.12),

$$h(T(F)+x)\bar{V}(T(F)+x,F) = -h(T(F)-x)\bar{V}(T(F)-x,F) \tag{4.1.14}$$

for all $F \in \mathcal{F}$, $x \in \mathbb{R}$ such that $T(F)+x, T(F)-x \in \mathrm{int}(\mathsf{A})$. Assume $h(z) = 0$ for some $z \in \mathrm{int}(\mathsf{A})$. Then, invoking the surjectivity of T and the convexity of $\mathrm{int}(\mathsf{A})$, for all $z' \in \mathrm{int}(\mathsf{A}) \setminus \{z\}$ there exists an $F \in \mathcal{F}$ and $x \in \mathbb{R} \setminus \{0\}$ such that $z = T(F)+x$ and $z' = T(F)-x$. Since V is a *strict* $\mathcal{F}$-orientation function for T, both $\bar{V}(T(F)+x, F) \neq 0$ and $\bar{V}(T(F)-x, F) \neq 0$. Hence, (4.1.14) implies that $h(z') = 0$. This implies that h identically vanishes on $\mathrm{int}(\mathsf{A})$ which contradicts the strict $\mathcal{F}$-consistency of S.

That means $V^*(x,y) := h(x)V(x,y)$ is an oriented strict $\mathcal{F}$-identification function for T. Applying Osband's principle to S^*, one gets a function $h^*\colon \mathrm{int}(\mathsf{A}) \to \mathbb{R}$ such that

$$\frac{\mathrm{d}}{\mathrm{d}x}\bar{S}^*(x,F) = h^*(x)\bar{V}^*(x,F)$$

for all $F \in \mathcal{F}$, $x \in \mathbb{R}$ such that $T(F)+x, T(F)-x \in \mathrm{int}(\mathsf{A})$. Due to the analogue of (4.1.13) for S^* and (4.1.14), one obtains

$$\begin{aligned} h^*(T(F)+x)\bar{V}^*(T(F)+x,F) &= -h^*(T(F)-x)\bar{V}^*(T(F)-x,F) \\ &= h^*(T(F)-x)\bar{V}^*(T(F)+x,F). \end{aligned}$$

for all $F \in \mathcal{F}$, $x \in \mathbb{R}$ with $T(F)+x, T(F)-x \in \mathrm{int}(\mathsf{A})$. By a similar reasoning as above, one can deduce that h^* must be constant and positive. Now, the claim follows by Proposition 3.4 in Fissler and Ziegel (2016). □

Now, we shall show a similar version of Proposition 4.1.26 for the higher-dimensional setting. However, it is a bit more limited: We show that two scoring functions that are strictly metrically order-sensitive for the same functional and are additively separable in the form of (4.1.8) must be necessarily equivalent. For most practically relevant cases – namely the case when the metric is induced by an ℓ^p-norm with $p \in [1,\infty)$ and when the functional consists of blocks satisfying Assumption (V4) or that are ratios of expectations with the same denominator – Lemma 4.1.8, Proposition 4.1.22 and Fissler and Ziegel (2016, Proposition 4.2) yield that any metrically order-sensitive scoring function is additively separable. Hence, for these situations, metrically order-sensitive scoring functions are unique, up to equivalence.

Proposition 4.1.27. *Let $S\colon \mathsf{A} \times \mathsf{O} \to \mathbb{R}$ be a strictly metrically $\mathcal{F}$-order-sensitive scoring function for a surjective functional $T = (T_1, \dots, T_k)\colon \mathcal{F} \to \mathsf{A} \subseteq \mathbb{R}^k$ of the form*

$$S(x_1,\dots,x_k,y) = \sum_{m=1}^{k} S_m(x_m,y)$$

for all $(x,y) \in \mathsf{A} \times \mathsf{O}$ where $S_m\colon \mathsf{A}_m \times \mathsf{O} \to \mathbb{R}$, $m \in \{1,\dots,k\}$, $\mathsf{A}_m = \{x_m \in \mathbb{R} \colon \exists (z_1,\dots,z_k) \in \mathsf{A}$ and $z_m = x_m\}$, are strictly $\mathcal{F}$-consistent scoring functions for T_m. Assume that $\mathrm{int}(\mathsf{A}) \neq \emptyset$. Then, the following assertions hold:

(i) The scoring functions S_m, $m \in \{1,\dots,k\}$, are strictly metrically $\mathcal{F}$-order-sensitive for T_m.

(ii) *Let $\lambda_1, \dots, \lambda_k > 0$ and define the scoring function $S^*\colon \mathsf{A} \times \mathsf{O} \to \mathbb{R}$ via*

$$S^*(x_1, \dots, x_k) = \sum_{m=1}^{k} \lambda_m S_m(x_m, y).$$

Then S^* is strictly metrically $\mathcal{F}$-order-sensitive (with respect to the same metric as S) if and only if $\lambda_1 = \dots = \lambda_k$.

Proof. (i) Recall that we assume that the metric is induced by a norm. Due to the positive homogeneity of norms, all norms on $\mathbb{R}$ are of the form $c|\cdot|$ for some $c > 0$.

Let $m \in \{1, \dots, k\}$, $F \in \mathcal{F}$ with $t = T(F) \in \operatorname{int}(\mathsf{A})$. Let $\mu \in \mathbb{R}$ and $x, z \in \mathsf{A}$ with $x_i = z_i = t_i$ for all $i \neq m$ and with $x_m = t_m + \mu$ and $z_m = t_m - \mu$, such that $d(x,t) = d(z,t)$. Due to Lemma 4.1.25 and due to the particular additive form of S, we have

$$\begin{aligned} 0 = \bar{S}(x,F) - \bar{S}(z,F) &= \bar{S}_m(x_m,F) - \bar{S}_m(z_m,F) \\ &= \bar{S}_m(t_m+\mu,F) - \bar{S}_m(t_m-\mu,F). \end{aligned}$$

Again with Lemma 4.1.25 one obtains the assertion.

(ii) The only interesting direction is to assume that S^* is strictly metrically $\mathcal{F}$-order-sensitive (with respect to the same metric d as S). We will show that $\lambda_1 = \lambda_m$ for all $m \in \{2, \dots, k\}$. But for notational convenience, we give the proof only for $m = 2$.

Let $F \in \mathcal{F}$, $t = T(F) \in \operatorname{int}(\mathsf{A})$, $x, z \in \mathsf{A}$ with $d(x,t) = d(z,t) > 0$ and $x_i = z_i = t_i$ for all $i \in \{3, \dots, k\}$. Moreover, let $x_1 \neq z_1 = t_1$. Due to Lemma 4.1.25 we have that

$$\bar{S}(x,F) - \bar{S}(z,F) = \bar{S}^*(x,F) - \bar{S}^*(z,F) = 0.$$

Moreover, due to the assumptions on x, z and the fact that the scoring functions S_m are strictly $\mathcal{F}$-consistent

$$\begin{aligned} 0 = \bar{S}(x,F) - \bar{S}(z,F) &= \sum_{i=1}^{k} \bar{S}_i(x_i,F) - \bar{S}_i(z_i,F) \\ &= \bar{S}_1(x_1,F) - \bar{S}_1(z_1,F) + \bar{S}_2(x_2,F) - \bar{S}_2(z_2,F). \end{aligned}$$

Setting $\varepsilon := \bar{S}_1(x_1,F) - \bar{S}_1(z_1,F) > 0$, one obtains with the same calculation

$$\begin{aligned} 0 &= \bar{S}^*(x,F) - \bar{S}^*(z,F) \\ &= \lambda_1\big(\bar{S}_1(x_1,F) - \bar{S}_1(z_1,F)\big) + \lambda_2\big(\bar{S}_2(x_2,F) - \bar{S}_2(z_2,F)\big) \\ &= \varepsilon(\lambda_1 - \lambda_2). \end{aligned}$$

□

Corollary 4.1.28. *Let $T = (T_1, \dots, T_k)\colon \mathcal{F} \to \mathsf{A} \subseteq \mathbb{R}^k$ be a surjective, elicitable and identifiable functional with a strict $\mathcal{F}$-identification function $V = (V_1, \dots, V_k)$: $\mathsf{A} \times \mathsf{O} \to \mathbb{R}^k$ such that each component V_m, $m \in \{1, \dots, k\}$, is an oriented strict $\mathcal{F}$-identification function for T. Define $\mathsf{A}_m = \{x_m \in \mathbb{R} \colon \exists (z_1, \dots, z_k) \in \mathsf{A}$ and $z_m = x_m\}$ for $m \in \{1, \dots, k\}$ and assume that $\operatorname{int}(\mathsf{A}) \neq \emptyset$ as well as $\operatorname{int}(\mathsf{A}_m)$ is convex and non-empty for all $m \in \{1, \dots, k\}$. If $S, S^* \colon \mathsf{A} \times \mathsf{O} \to \mathbb{R}$ are two strictly metrically $\mathcal{F}$-order-sensitive scoring functions for T such that Assumptions (V1), (V2), (S1), (F1) and (VS1) from Fissler and Ziegel (2016) (with respect to both scoring functions) hold, and they are of the form*

$$S(x_1, \dots, x_k, y) = \sum_{m=1}^{k} S_m(x_m, y), \qquad S^*(x_1, \dots, x_k, y) = \sum_{m=1}^{k} S^*_m(x_m, y),$$

for almost all $(x, y) \in \mathsf{A} \times \mathsf{O}$, where $S_m, S^*_m \colon \mathsf{A}_m \times \mathsf{O} \to \mathbb{R}$, $m \in \{1, \dots, k\}$, are strictly $\mathcal{F}$-consistent scoring functions for $T_m \colon \mathcal{F} \to \mathsf{A}_m$, then S and S^* are equivalent almost everywhere.

Proof. Due to Proposition 4.1.27(i) S_m and S^*_m are strictly metrically $\mathcal{F}$-order-sensitive for T_m. Invoking Proposition 4.1.26, S_m and S^*_m are equivalent almost everywhere. So there are positive constants $\lambda_1, \dots, \lambda_k > 0$ and $\mathcal{F}$-integrable functions $a_1, \dots, a_k \colon \mathsf{O} \to \mathbb{R}$ such that

$$S^*(x_1, \dots, x_k, y) = \sum_{m=1}^{k} \lambda_m S_m(x_m, y) + a_m(y)$$

for almost all $(x, y) \in \mathsf{A} \times \mathsf{O}$. Set $a(y) = \sum_{m=1}^{k} a_m(y)$. Then $S(x, y) - a(y)$ is also strictly metrically $\mathcal{F}$-order-sensitive for T. With respect to Proposition 4.1.27(ii), it holds that $\lambda_1 = \dots = \lambda_k$. So the claim follows. □

Next, we use the derived theoretical results to examine when some popular functionals admit strictly metrically order-sensitive scoring functions, and if so, of what form they are.

Ratios of expectations with the same denominator

We start with the one-dimensional characterization.

Lemma 4.1.29. *Let $\mathcal{F}$ be convex and $p, q \colon \mathsf{O} \to \mathbb{R}$ two $\mathcal{F}$-integrable functions such that $\bar{q}(F) > 0$ for all $F \in \mathcal{F}$. Define $T \colon \mathcal{F} \to \mathsf{A} \subseteq \mathbb{R}$, $T(F) = \bar{p}(F)/\bar{q}(F)$ and assume that T is surjective as well as $\operatorname{int}(\mathsf{A}) \neq \emptyset$. Then the following two assertions are true:*

(i) Any scoring function which is equivalent to

$$S \colon \mathsf{A} \times \mathsf{O} \to \mathbb{R}, \qquad S(x, y) = \frac{1}{2} q(y) x^2 - p(y) x \tag{4.1.15}$$

is strictly metrically $\mathcal{F}$-order-sensitive for T.

(ii) *If $\mathcal{F}$ is such that Assumptions (V1), (F1) in Fissler and Ziegel (2016) are satisfied with $V(x,y) = q(y)x - p(y)$, then any scoring function $S^*\colon \mathsf{A} \times \mathsf{O} \to \mathbb{R}$, which is strictly metrically $\mathcal{F}$-order-sensitive and satisfies Assumptions (S1) and (VS1), is equivalent to S defined at (4.1.15) almost everywhere.*

Proof. (i) We can apply Lemma 4.1.25. Let $F \in \mathcal{F}$. Then

$$\mathbb{R} \ni x \mapsto \bar{S}(T(F)+x, F) = \frac{1}{2}\bar{q}(F)x^2 - \frac{1}{2}\frac{\bar{p}(F)^2}{\bar{q}(F)}$$

is an even function in x. Moreover, equivalence of scoring functions preserves (strict) metrical order-sensitivity.

(ii) The convexity of A is implied by the mixture-continuity of T and the convexity of $\mathcal{F}$. Then, the claim follows with Proposition 4.1.26.

□

Now, we turn to the multivariate characterization.

Proposition 4.1.30. *Let $\mathcal{F}$ be convex and $p\colon \mathsf{O} \to \mathbb{R}^k$, $q\colon \mathsf{O} \to \mathbb{R}$ two $\mathcal{F}$-integrable functions such that $\bar{q}(F) > 0$ for all $F \in \mathcal{F}$. Define $T\colon \mathcal{F} \to \mathsf{A} \subseteq \mathbb{R}^k$, $T(F) = \bar{p}(F)/\bar{q}(F)$ and assume that T is surjective as well as $\operatorname{int}(\mathsf{A}) \neq \emptyset$. Then, the following assertions are true:*

(i) *Any scoring function which is equivalent to*

$$S\colon \mathsf{A} \times \mathsf{O} \to \mathbb{R}, \qquad S(x_1,\dots,x_k,y) = \sum_{m=1}^{k} \frac{1}{2}q(y)x_m^2 - p_m(y)x_m \qquad (4.1.16)$$

is strictly metrically $\mathcal{F}$-order-sensitive for T with respect to the ℓ^2-norm.

(ii) *If $\mathcal{F}$ is such that Assumptions (V1), (F1) in Fissler and Ziegel (2016) are satisfied with $V(x,y) = q(y)x - p(y)$, then any scoring function $S^*\colon \mathsf{A} \times \mathsf{O} \to \mathbb{R}$, which is strictly metrically $\mathcal{F}$-order-sensitive with respect to the ℓ^2-norm and satisfies Assumptions (S1) and (VS1), is equivalent to S defined at (4.1.16) almost everywhere.*

(iii) *If $\mathcal{F}$ is such that Assumptions (V1), (F1) in Fissler and Ziegel (2016) are satisfied with $V(x,y) = q(y)x - p(y)$, then there is no scoring function $S^*\colon \mathsf{A} \times \mathsf{O} \to \mathbb{R}$ satisfying Assumptions (S1) and (VS1), which is strictly metrically $\mathcal{F}$-order-sensitive with respect to an ℓ^p-norm with $p \in [1,\infty) \setminus \{2\}$.*

Proof. To show (i) we apply again Lemma 4.1.25. For any $F \in \mathcal{F}$, $s > 0$ define the function $\psi\colon \mathbb{R}^k \to \mathbb{R}$, $\psi(v) = \bar{S}(T(F) + sv, F)$. Then

$$\psi_{F,s}(v) = \frac{1}{2}\bar{q}(F)s^2 \sum_{m=1}^{k} v_m^2 - \frac{1}{2\bar{q}(F)} \sum_{m=1}^{k} \bar{p}_m(F)^2. \qquad (4.1.17)$$

Define $\mathbb{S}_p^{k-1} := \{x \in \mathbb{R}^k \,|\, \|x\|_p = 1\}$, $p \in [1,\infty)$ where $\|\cdot\|_p$ is the ℓ^p-norm on $\mathbb{R}^k$. Then $\psi_{F,s}$ is constant on $\mathbb{S}_2^{k-1}$.

We prove (ii) and (iii) together. Assume there is a scoring function S^* satisfying the conditions above, so in particular, it is strictly metrically $\mathcal{F}$-order-sensitive with respect to the ℓ^p-norm for $p \in [1,\infty)$. Invoking Lemma 4.1.8(i), S^* is strictly componentwise $\mathcal{F}$-order-sensitive for T. Thanks to Proposition 4.1.22, S^* is additively separable. Then, using Corollary 4.1.28, S^* is equivalent to S defined at (4.1.16) almost everywhere. In case $p = 2$, assertion (i) already shows that S is strictly metrically $\mathcal{F}$-order-sensitive. For $p \neq 2$ recall that strict metrical order-sensitivity is preserved by equivalence. However, we will show that S is nowhere metrically $\mathcal{F}$-order-sensitive. To this end, fix any $F \in \mathcal{F}$ and $s > 0$. Assume that $\psi_{F,s}$ defined at (4.1.17) is constant on $\mathbb{S}_p^{k-1}$. Then necessarily for all $v \in \mathbb{S}_p^{k-1}$

$$\psi_{F,s}(v) = \psi_{F,s}(e_m) = \frac{1}{2}\bar{q}(F)s^2 - \frac{1}{2\bar{q}(F)}\sum_{m=1}^k \bar{p}_m(F)^2,$$

where e_m denotes the mth standard basis vector of $\mathbb{R}^k$. But for any vector $v \in \mathbb{S}_p^{k-1} \setminus \{e_1, \ldots, e_m\}$ we have that $\sum_{m=1}^k v_m^2 = \|v\|_2^2 \neq 1$ if $p \neq 2$. □

Remark 4.1.31. Savage (1971, Section 5) has already shown that in case of the mean, the squared loss is essentially the only symmetric loss in the sense that it is the only metrically order-sensitive loss for the mean. See also Patton (2015, Section 2.1) for a discussion that symmetry – or metrical order-sensitivity – is not necessary for strict consistency of scoring functions with respect to the mean.

Quantiles

The following proposition provides a similar characterization for the one-dimensional case $k = 1$ of an α-quantile.

Proposition 4.1.32. *Let $\alpha \in (0,1)$ and $\mathcal{F}$ be a family of distribution functions on $\mathbb{R}$ with unique α-quantiles T_α. Assume that for any $F \in \mathcal{F}$, its translation $F_\lambda(\cdot) = F(\cdot - \lambda)$ is also an element of $\mathcal{F}$ for all $\lambda \in \mathbb{R}$. Consequently, $T_\alpha\colon \mathcal{F} \to \mathsf{A} = \mathbb{R}$ is surjective. Under assumptions (V1) in Fissler and Ziegel (2016) with respect to the strict identification function $V_\alpha\colon \mathbb{R}\times\mathbb{R} \to \mathbb{R}$, $V_\alpha(x,y) = \mathbb{1}\{y \le x\} - \alpha$, there is no strictly metrically $\mathcal{F}$-order-sensitive scoring function for T_α satisfying Assumption (S1) in Fissler and Ziegel (2016).*

Proof. Assume that there exists a strictly metrically $\mathcal{F}$-order-sensitive scoring function $S_\alpha\colon \mathbb{R}\times\mathbb{R} \to \mathbb{R}$ satisfying Assumption (S1) in Fissler and Ziegel (2016). Due to Lemma 4.1.25, for any $F \in \mathcal{F}$ and any $x \in \mathbb{R}$

$$\bar{S}_\alpha(T_\alpha(F) + x, F) = \bar{S}_\alpha(T_\alpha(F) - x, F).$$

Using Osband's principle (Fissler and Ziegel, 2016, Theorem 3.2) and taking the derivative with respect to x on both sides, this yields

$$h(T_\alpha(F)+x)\bar{V}_\alpha(T_\alpha(F)+x,F) = -h(T_\alpha(F)-x)\bar{V}_\alpha(T_\alpha(F)-x,F) \tag{4.1.18}$$

for some positive function $h\colon \mathbb{R}\to\mathbb{R}$ (the fact that $h\ge 0$ follows from the strict consistency of S_α and the surjectivity of T_α, and $h>0$ follows like in the proof of Proposition 4.1.26). Equation (4.1.18) implies

$$\frac{h(q_\alpha(F)+x)}{h(q_\alpha(F)-x)} = -\frac{F(q_\alpha(F)-x)-\alpha}{F(q_\alpha(F)+x)-\alpha} \overset{x\to\infty}{\longrightarrow} \frac{\alpha}{1-\alpha}. \tag{4.1.19}$$

As a consequence, for $\alpha\neq 1/2$, the function h cannot be constant.

W.l.o.g. let $F_0\in\mathcal{F}$ with $T_\alpha(F_0)=0$. For any $\lambda\in\mathbb{R}$ define $F_\lambda(\cdot)=F_0(\cdot-\lambda)$. Hence, $T_\alpha(F_\lambda)=\lambda$ for any $\lambda\in\mathbb{R}$. Then (4.1.19) implies that for any $\lambda\in\mathbb{R}$ and for any $x\in\mathbb{R}$

$$\frac{h(\lambda+x)}{h(\lambda-x)} = -\frac{F_\lambda(\lambda-x)-\alpha}{F_\lambda(\lambda+x)-\alpha} = -\frac{F_0(-x)-\alpha}{F_0(x)-\alpha}. \tag{4.1.20}$$

Identity (4.1.20) implies four asymptotic identities:

(i) For any fixed $\lambda\in\mathbb{R}$
$$\lim_{x\to\infty}\frac{h(\lambda+x)}{h(\lambda-x)} = \frac{\alpha}{1-\alpha}.$$

(ii) Setting $\lambda=x$ and letting $x\to\infty$, one can see that $h(+\infty):=\lim_{x\to\infty}h(x)$ exists and that
$$h(+\infty)=h(0)\frac{\alpha}{1-\alpha}.$$

(iii) Setting $\lambda=-x$ and letting $x\to\infty$, one can see that $h(-\infty):=\lim_{x\to\infty}h(-x)$ exists and that
$$h(-\infty)=h(0)\frac{1-\alpha}{\alpha}.$$

(iv) For any fixed $x\in\mathbb{R}$, one has that
$$1=\frac{h(+\infty)}{h(+\infty)} = \lim_{\lambda\to\infty}\frac{h(\lambda+x)}{h(\lambda-x)} = -\frac{F_0(-x)-\alpha}{F_0(x)-\alpha}. \tag{4.1.21}$$

Recalling that $h>0$, (ii) and (iii) imply that

$$\frac{h(+\infty)}{h(-\infty)}=1,$$

whereas (i) implies that

$$\frac{h(+\infty)}{h(-\infty)}=\frac{\alpha}{1-\alpha}.$$

As a consequence, necessarily $\alpha = 1/2$. So under the conditions of the proposition, there is no strictly metrically $\mathcal{F}$-order-sensitive scoring function for the α-quantile if $\alpha \neq 1/2$. (4.1.21) together with (4.1.20) yield that h must be constant. On the other hand, for $\alpha = 1/2$, equation (4.1.21) yields that any $F \in \mathcal{F}$ must be symmetric around its median, that means

$$F(T_{1/2}(F) + x) = 1 - F((T_{1/2}(F) - x)-) \qquad \forall x \in \mathbb{R},$$

where $F(y-)$ denotes the left-sided limit of F at y. However, if F_0 is symmetric around its median, then any translation F_λ of F_0 is symmetric around its median. But then, there is a convex combination of F_0 and F_λ with mixture-parameter $\beta \in (0,1)$, $\beta \neq 1/2$ such that $\beta F_0 + (1-\beta)F_\lambda$ is not symmetric around its median if $\lambda \neq 0$. Consequently, the conditions of the proposition are violated such that a strictly metrically $\mathcal{F}$-order-sensitive function for the median does not exist in this setting. □

It is crucial to recapitulate that the reasons for the non-existence of a strictly metrically order-sensitive scoring function for the α-quantile are of different nature in the two cases that $\alpha \neq 1/2$ and that $\alpha = 1/2$ in the proof of Proposition 4.1.32. In both cases, we used the assumptions of the proposition to use Osband's principle to derive a representations of the derivative of the expected score. Assuming that the derivative has the form as stated in Osband's principle, one can directly derive a contradiction for $\alpha \neq 1/2$. However, for $\alpha = 1/2$, this form merely implies that the distributions in $\mathcal{F}$ must be symmetric around their medians. This is not contradictory to the form of the gradient derived via Osband's principle, but only to the assumptions stated in Osband's principle, in particular, the convexity of $\mathcal{F}$. But refraining from assuming the convexity needed in Osband's principle, this leads the way to a sufficiency condition for strict metrical order-sensitivity as shown in the following lemma.

Lemma 4.1.33. *Let $\mathcal{F}$ be a family of distribution functions on $\mathbb{R}$ with unique medians $T_{1/2}\colon \mathcal{F} \to \mathbb{R}$ and finite first moments. If all distributions in $\mathcal{F}$ are symmetric around their medians in the sense that*

$$F(T_{1/2}(F) + x) = 1 - F((T_{1/2}(F) - x)-)$$

for all $F \in \mathcal{F}$, $x \in \mathbb{R}$, then any scoring function that is equivalent to the absolute loss $S\colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, $S(x,y) = |x-y|$, is strictly metrically $\mathcal{F}$-order-sensitive with respect to the median.

Proof. It is known that S is strictly $\mathcal{F}$-consistent for $T_{1/2}$. Then, we can use Lemma 4.1.25. For $F \in \mathcal{F}$ with $t = T_{1/2}(F)$ and for $x \in \mathbb{R}$ one has

$$\begin{aligned}\bar{S}(t+x,F) &= \mathbb{E}_F|t+x-Y| = \mathbb{E}_F|x-(Y-t)| \\ &= \mathbb{E}_F|x+(Y-t)| = \mathbb{E}_F|t-x-Y| = \bar{S}(t-x,F).\end{aligned} \tag{4.1.22}$$

Note again that equivalence preserves strict metrical order-sensitivity. □

As mentioned above, under the conditions of Lemma 4.1.33, the necessary characterization of strictly consistent scoring functions via Osband's principle is not available. In particular, this means that we cannot use Proposition 4.1.26. Indeed, if the distributions in $\mathcal{F}$ are symmetric around their medians in the sense of (4.1.21) and under the integrability condition that all elements in $\mathcal{F}$ have a finite first moment, the median and the mean coincide. Hence, any convex combination of a strictly consistent scoring function for the mean and the median provides a strictly consistent scoring function; see also Patton (2015, Section 2.6). *A fortiori*, any scoring function which is equivalent to $S(x, y) = (1 - \lambda)|x - y| + \lambda|x - y|^2$, $\lambda \in [0, 1]$ is strictly metrically $\mathcal{F}$-order-sensitive. However, the class of strictly metrically $\mathcal{F}$-order-sensitive scoring functions is even bigger. In the sequel, we will denote the median / mean merely by the 'center of symmetry' or merely 'center' of the distribution if (4.1.21) is satisfied for all $F \in \mathcal{F}$. The following lemma shows that – in contrast to the median – the center of symmetry of a distribution is unique if it exists.

Lemma 4.1.34. *Let F be a symmetric distribution on $\mathbb{R}$ in the sense that there exists a 'center' $C = C(F) \in \mathbb{R}$ such that*

$$F(C + x) = 1 - F((C - x)-) \qquad \forall x \in \mathbb{R}. \tag{4.1.23}$$

Then, the center C is unique.

Proof. Let $C_1 < C_2 \in \mathbb{R}$ be two centers for a distribution F, such that C_1 and C_2 satisfy (4.1.23). Then, for any $x \in \mathbb{R}$

$$\begin{aligned} F(C_1 + x) &= 1 - F((C_1 - x)-) \\ &= 1 - F((C_2 - (C_2 - C_1 + x))-) \\ &= F(2C_2 - C_1 + x). \end{aligned}$$

Since F is increasing, it is constant on the interval $[C_1 + x, 2C_2 - C_1 + x]$. As this holds for all $x \in \mathbb{R}$, F must be globally constant and thus fails to be a distribution function. $\square$

Proposition 4.1.35. *Let $\mathcal{F}$ be a family of symmetric distributions on $\mathbb{R}$ in the sense that for all $F \in \mathcal{F}$ there exists a center $C(F) \in \mathbb{R}$ satisfying (4.1.23). Let $\Phi\colon \mathbb{R} \to \mathbb{R}$ be even and convex. Define the scoring function*

$$S\colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}, \qquad (x, y) \mapsto S(x, y) = \Phi(x - y). \tag{4.1.24}$$

Then, the following assertions hold:

(i) S is $\mathcal{F}$-consistent for the center functional $C\colon \mathcal{F} \to \mathbb{R}$ and S is metrically $\mathcal{F}$-order-sensitive for C.

(ii) If Φ is strictly convex, then S is strictly $\mathcal{F}$-consistent for C (and consequently strictly metrically $\mathcal{F}$-order-sensitive for C).

(iii) If $\Phi(0) < \Phi(x)$ for all $x \in \mathbb{R} \setminus \{0\}$ and $\mathbb{P}(Y \in (C(F) - \varepsilon, C(F) + \varepsilon)) > 0$ for all $\varepsilon > 0$ and for all $F \in \mathcal{F}$ where $Y \sim F$, then S is strictly $\mathcal{F}$-consistent for C (and consequently strictly metrically $\mathcal{F}$-order-sensitive for C).

Proof. (i) One can directly verify that $x \mapsto S(x, y)$ is convex for any $y \in \mathbb{R}$. Hence, the expected score $x \mapsto \bar{S}(x, F)$ is convex for any $F \in \mathcal{F}$. In particular, for any $F \in \mathcal{F}$ with center $c = C(F) \in \mathbb{R}$, the function $x \mapsto \bar{S}(c+x, F)$ is convex. Moreover, a computation similar to (4.1.22) yields that the function $x \mapsto \bar{S}(c + x, F)$ is even. Consequently, S is $\mathcal{F}$-consistent for C and, with respect to Lemma 4.1.25, it is metrically $\mathcal{F}$-order-sensitive.

(ii) Let Φ be strictly convex. Fix some $x \in \mathbb{R} \setminus \{0\}$ and define the function,

$$\Psi_x \colon \mathbb{R} \to \mathbb{R}, \quad y \mapsto \Psi_x(y) = \frac{1}{2}\big(\Phi(x - y) + \Phi(-x - y)\big) - \Phi(-y). \tag{4.1.25}$$

The function Ψ_x is even and due to the strict consistency of Φ, Ψ_x is strictly positive. Let $F \in \mathcal{F}$ with center $c = C(F) \in \mathbb{R}$. Then

$$\begin{aligned}\bar{S}(c + x, F) - \bar{S}(c, F) &= \frac{1}{2}\big(\bar{S}(c + x, F) + \bar{S}(c - x, F)\big) - \bar{S}(c, F) \\ &= \mathbb{E}_F\left[\Psi_x(Y - c)\right] > 0.\end{aligned}$$

So S is strictly $\mathcal{F}$-consistent for C. With Lemma 4.1.25 one concludes that S is also strictly metrically $\mathcal{F}$-order-sensitive for C.

(iii) Let $x \in \mathbb{R} \setminus \{0\}$ and consider the function Ψ_x defined at (4.1.25). Since Φ is convex, it is continuous. Hence, also Ψ_x is continuous. Due to the assumptions, $\Psi_x \geq 0$ and $\Psi_x(0) > 0$. Therefore, there is some $\varepsilon > 0$ such that $\Psi_x(y) > 0$ for all $y \in (-\varepsilon, \varepsilon)$. Now, let $F \in \mathcal{F}$ with center $c = C(F) \in \mathbb{R}$. Then,

$$\begin{aligned}\bar{S}(c + x, F) - \bar{S}(c, F) &= \mathbb{E}_F\left[\Psi_x(Y - c)\right] \\ &\geq \mathbb{E}_F[\mathbb{1}\{Y \in (c - \varepsilon, c + \varepsilon)\}\Psi_x(Y - c)] > 0.\end{aligned}$$

That means again that S is strictly $\mathcal{F}$-consistent for C. So Lemma 4.1.25 yields that S is strictly metrically $\mathcal{F}$-order-sensitive.

□

Example 4.1.36. One prominent example of a scoring function besides the linear or the squared loss is the so-called *Huber loss* which was presented in Huber (1964) and arises upon taking

$$\Phi(t) = \begin{cases} \frac{1}{2}t^2, & \text{for } |t| < k \\ k|t| - \frac{1}{2}k^2, & \text{for } |t| \geq k \end{cases}$$

in (4.1.24), where $k \in \mathbb{R}$, $k \geq 0$ is a tuning parameter.

Remark 4.1.37. If the a distribution is symmetric, then not only the mean and the median coincide with the center of symmetry, but also the α-trimmed mean and the α-Winsorized mean, $\alpha \in (0, 1/2)$; see Huber and Ronchetti (2009, pp. 57–59) for a definition of the two functionals.

Due to the negative result of Proposition 4.1.32 we dispense with an investigation of scoring functions that are metrically order-sensitive for vectors of different quantiles.

Expectiles

As a last example of a popular elicitable functional, let us draw attention to expectiles. The special situation of the 1/2-expectile, such that it coincides with the mean functional, was already considered in the subsection on ratios of expectations with the same denominator. So the main purpose is to explore the case $\tau \neq 1/2$. It is obvious that the canonical scoring function for the τ-expectile, that is, the asymmetric squared loss

$$S_\tau(x, y) = |\mathbb{1}\{y \le x\} - \tau|(x - y)^2$$

is not a metrically order-sensitive scoring function since $x \mapsto S_\tau(x + y, y)$ is not an even function. *A fortiori*, it turns out that under some smoothness conditions on the distribution functions $F \in \mathcal{F}$ and some richness assumption on $\mathcal{F}$ there is no metrically $\mathcal{F}$-order-sensitive scoring function for the τ-expectile for $\tau \neq 1/2$. Strictly speaking, we have the following proposition where we recall that the function $V_\tau \colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$

$$V_\tau(x, y) = 2|\mathbb{1}\{y \le x\} - \tau|\,(x - y) \tag{4.1.26}$$

is an oriented strict identification function for the τ-expectile.

Proposition 4.1.38. *Let $\tau \in (0, 1)$, $\tau \neq 1/2$, and $T = \mu_\tau \colon \mathcal{F} \to \mathsf{A} \subseteq \mathbb{R}$ be the τ-expectile, such that Assumption (V1) in Fissler and Ziegel (2016) holds with respect to the strict $\mathcal{F}$-identification function at (4.1.26), and assume that T is surjective. Assume that $\bar{V}(\cdot, F)$ is twice differentiable for all $F \in \mathcal{F}$ and that there is a strictly $\mathcal{F}$-consistent scoring function $S \colon \mathsf{A} \times \mathbb{R} \to \mathbb{R}$ such that $\bar{S}(\cdot, F)$ is three times differentiable for all $F \in \mathcal{F}$. In particular, let each $F \in \mathcal{F}$ be differentiable with derivative $f = F'$.*

If there is a $t \in \mathsf{A}$ and $F_1, F_2 \in \mathcal{F}$ such that $T(F_1) = T(F_2) = t$, $F_1(t) = F_2(t)$, but $F_1'(t) = f_1(t) \neq f_2(t) = F_2'(t)$, then S is not metrically $\mathcal{F}$-order-sensitive.

Proof. Under the assumptions, Osband's principle implies that there is a function $h \colon \mathrm{int}(\mathsf{A}) \to \mathbb{R}$, $h > 0$ (by an argument like in the proof of Proposition 4.1.26) such that for all $F \in \mathcal{F}$, $x \in \mathrm{int}(\mathsf{A})$

$$\frac{\mathrm{d}}{\mathrm{d}x}\bar{S}(x, F) = h(x)\bar{V}_\tau(x, F).$$

Using the same argument as in the proof of Osband's principle (Fissler and Ziegel, 2016, Theorem 3.2), h is twice differentiable. Assume that S is metrically $\mathcal{F}$-order sensitive. Then, due to Lemma 4.1.25 for any $F \in \mathcal{F}$ the function $g_F \colon \mathsf{A} \ni x \mapsto g_F(x) = \bar{S}(T(F) + x, F)$ is an even function. Hence, invoking the smoothness assumptions, the third derivative of g_F must be odd. So necessarily $g_F'''(0) = 0$. Denoting $t_F = T(F)$, some tedious calculations lead to

$$g_F'''(0) = 2h'(t_F)\big(F(t_F)(1-2\tau) + \tau\big) + 2h(t_F)f(t_F)(1-2\tau). \qquad (4.1.27)$$

Recalling that $h > 0$ and $\tau \neq 1/2$ implies $g_{F_1}'''(0) \neq g_{F_2}'''(0)$. So S cannot be metrically $\mathcal{F}$-order-sensitive. $\square$

Remark 4.1.39. Starting with some differentiable $F_1 \in \mathcal{F}$ it is possible to construct an F_2 satisfying the conditions of Proposition 4.1.38 by changing F_1 in a neighborhood of $t = T(F_1)$.

Remark 4.1.40. Inspecting the proof of Proposition 4.1.38, equation (4.1.27) yields for $\tau = 1/2$

$$g_F'''(0) = h'(t_F)F(t_F)$$

for any $F \in \mathcal{F}$, $t_F = T(F)$. With the surjectivity of T this proves that $h' = 0$, such that h is necessarily constant. Hence, we get an alternative proof that the squared loss is the only metrically order-sensitive scoring function for the mean, up to equivalence.

Connections to (quasi-)convexity

It is interesting to examine the connection between metrical order-sensitivity and the convexity of the expected score. However, generally, neither implies the other – there are examples of metrically order-sensitive scoring functions, which are not convex and vice versa. However, one can show the following weaker result.

Lemma 4.1.41. *Let S be a metrically $\mathcal{F}$-order-sensitive scoring function for some functional $T \colon \mathcal{F} \to \mathsf{A}$ where the metric is induced by a norm $\|\cdot\|$. Then for any $F \in \mathcal{F}$, the expected score $\bar{S}(\cdot, F)$ is quasi-convex.*

Proof. Let $F \in \mathcal{F}$, $t = T(F)$, $x, z \in \mathsf{A}$, $\lambda \in [0,1]$ with $x' = \lambda x + (1-\lambda)z \in \mathsf{A}$.

$$\begin{aligned}
\|x' - t\| &= \|\lambda(x-t) + (1-\lambda)(z-t)\| \\
&\le \|\lambda(x-t)\| + \|(1-\lambda)(z-t)\| \\
&= \lambda\|x-t\| + (1-\lambda)\|z-t\| \le \max\{\|x-t\|, \|z-t\|\}.
\end{aligned}$$

Using the metrical $\mathcal{F}$-order-sensitivity of S with respect to $\|\cdot\|$, one concludes with

$$\bar{S}(x', F) \le \max\{\bar{S}(x,F), \bar{S}(z,F)\}.$$

$\square$

Remark 4.1.42. In Corollary 4.1.48 we will show that quasi-convexity of a scoring function in turn implies order-sensitivity on line segments.

Example 4.1.43. Let $\mathsf{A} = \mathsf{O} = \mathbb{R}$, $S(x, y) = |x - y|^{1/2}$ and $T\colon \mathcal{F} \to \mathbb{R}$ be a functional implicitly defined such that S becomes a strictly $\mathcal{F}$-consistent scoring function[5]. Then S is quasi-convex, but not convex in x. However, if $\mathcal{F}$ contains only distributions which are symmetric around $T(F)$ and such that $T(F)$ is uniquely defined, an argument similar as in the proof of Proposition 4.1.35 yields that S is metrically $\mathcal{F}$-order-sensitive.

Example 4.1.44. There are convex scoring functions which are not metrically order-sensitive. For example, the α-pinball loss $S(x, y) = (\mathbb{1}\{y \le x\} - \alpha)(x - y)$ is convex in x. However, due to Proposition 4.1.32, for $\alpha \neq 1/2$, S cannot be metrically order-sensitive for the α-quantile.

4.1.7. Order-sensitivity on line segments

Recalling Lemma 4.1.8, every componentwise order-sensitive scoring function is also order-sensitive on line segments. With the help of the findings of Subsection 4.1.3 it is possible to derive the following Corollary.

Corollary 4.1.45. *If $\mathcal{F}$ is convex and $T\colon \mathcal{F} \to \mathsf{A} \subseteq \mathbb{R}^k$ is linear and surjective, then any strictly $\mathcal{F}$-consistent scoring function for T is strictly $\mathcal{F}$-order-sensitive on line segments.*

Proof. The linearity of T implies that T is mixture-continuous. Then the assertion follows directly by Proposition 4.1.14 and the special form of the image of the path γ in the proof therein, which is a line segment. □

Corollary 4.1.45 immediately leads the way to the result that the class of strictly order-sensitive scoring functions on line segments is strictly bigger than the class of strict componentwise order-sensitive scoring functions (for some functionals with dimension $k \ge 2$.) E.g. consider a vector of expectations satisfying the conditions of Proposition 4.1.22 which are the same as the one in Proposition 4.4 in Fissler and Ziegel (2016). Due to the latter result, there are strictly consistent scoring functions – and hence, with Corollary 4.1.45, strictly order-sensitive on line segments – which are not additively separable. By Proposition 4.1.22 they cannot be strictly componentwise order-sensitive.

The next Lemma asserts that order-sensitivity on line segments is stable under applying an isomorphism via the revelation principle (Proposition 2.3.2).

[5] Besides integrability assumptions on $\mathcal{F}$ one has to pay attention since $\arg\min_{x\in\mathbb{R}} \bar{S}(x, F)$ is generally not a singleton. In particular, even if one excludes those distributions F where the arg min is not unique, the resulting functional is not mixture-continuous.

Lemma 4.1.46. *Let $S\colon \mathsf{A}\times\mathsf{O}\to\mathbb{R}$ be a (strictly) $\mathcal{F}$-order-sensitive scoring function on line segments for a functional $T\colon\mathcal{F}\to\mathsf{A}\subseteq\mathbb{R}^k$. Let $g\colon\mathsf{A}\to\mathsf{A}'\subseteq\mathbb{R}^k$ be an isomorphism where A' is the image of A under g. Then $S_g\colon\mathsf{A}'\times\mathsf{O}\to\mathbb{R}$ defined as $S_g(x',y)=S(g^{-1}(x'),y)$ is a (strictly) $\mathcal{F}$-order-sensitive scoring function on line segments for the functional $T_g=g\circ T\colon\mathcal{F}\to\mathsf{A}'$.*

Proof. Let $F\in\mathcal{F}$, $t=T(F)$ and $t_g=T_g(F)=g(t)$. Let $v\in\mathbb{S}^{k-1}$ and $s\in[0,\infty)$. Using the linearity of g^{-1} we get

$$\bar{S}_g(t_g+sv,F)=\bar{S}\big(g^{-1}(g(t)+sv),F\big)=\bar{S}(t+sg^{-1}(v),F).$$

Since also g is an isomorphism, we have that $g^{-1}(v)/\|g^{-1}(v)\|\in\mathbb{S}^{k-1}$. Hence, the map $s\mapsto\bar{S}_g(t_g+sv,F)$ is (strictly) increasing for all $v\in\mathbb{S}^{k-1}$ if S is (strictly) order-sensitive on line segments. $\square$

Connections to quasi-convexity

It is obvious that quasi-convexity of an expected strictly consistent score implies order-sensitivity on line segments; see Section 4.2 for an introduction and a formal definition of quasi-convexity. More precisely, we have the following equivalent characterization of order-sensitivity on line segments.

Lemma 4.1.47. *Let $S\colon\mathsf{A}\times\mathsf{O}\to\mathbb{R}$ be a strictly $\mathcal{F}$-consistent scoring function for a functional $T\colon\mathcal{F}\to\mathsf{A}\subseteq\mathbb{R}^k$ where A is convex. Then S is (strictly) $\mathcal{F}$-order-sensitive on line segments if and only if for all $F\in\mathcal{F}$, $t=T(F)$, and for all $v\in\mathbb{S}^{k-1}$ the map*

$$\psi\colon D'=\{s\in\mathbb{R}\colon t+sv\in\mathsf{A}\}\to\mathbb{R},\qquad s\mapsto\bar{S}(t+sv,F)\tag{4.1.28}$$

is (strictly) quasi-convex.

The proof is straight forward and therefore omitted. In particular, we obtain the following corollary.

Corollary 4.1.48. *Let $S\colon\mathsf{A}\times\mathsf{O}\to\mathbb{R}$ be a strictly $\mathcal{F}$-consistent scoring function for a functional $T\colon\mathcal{F}\to\mathsf{A}\subseteq\mathbb{R}^k$ where A is convex. If for all $F\in\mathcal{F}$ the expected score $\bar{S}(\cdot,F)\colon\mathsf{A}\to\mathbb{R}$ is (strictly) quasi-convex, then S is (strictly) $\mathcal{F}$-order-sensitive on line segments for T.*

From a theoretical point of view, quasi-convexity is strictly stronger than order-sensitivity on line segments. In terms of sublevel sets of the expected score, quasi-convexity is equivalent to the convexity of the sublevel sets, whereas order-sensitivity on line segments implies merely that the sublevel sets are star domains with star point $T(F)$. And indeed, we shall give examples of scoring functions that are order-sensitive on line segments, but not quasi-convex; compare Example 4.1.50 with Proposition 4.2.28.

Scoring functions for ratios of expectations with the same denominator

We can extend the result of Corollary 4.1.45 to the case of ratios of expectations with the same denominator.

Lemma 4.1.49. *Let $T\colon \mathcal{F} \to \mathsf{A} \subseteq \mathbb{R}^k$ be a ratio of expectations with the same denominator, that is, $T(F) = \bar{p}(F)/\bar{q}(F)$ for some $\mathcal{F}$-integrable functions $p\colon \mathsf{O} \to \mathbb{R}^k$, and $q\colon \mathsf{O} \to \mathbb{R}$ where we assume that $\bar{q}(F) > 0$ for all $F \in \mathcal{F}$. Assume that A is open. Then any strictly $\mathcal{F}$-consistent scoring function of the form given at equation (4.5) in Fissler and Ziegel (2016) is strictly $\mathcal{F}$-order-sensitive on line segments.*

Proof. We use the notation of Proposition 4.4 in Fissler and Ziegel (2016). Let $F \in \mathcal{F}$, $t = T(F)$, $v \in \mathbb{S}^{k-1}$ and $s \in \mathbb{R}$ such that $t + sv \in \mathsf{A}$. Then, consider the function ψ given at (4.1.28). One obtains

$$\psi'(s) = sv^\top h(t+sv)\bar{V}(t+sv,F) \ = s\,\bar{q}(F)v^\top h(t+sv)v \begin{cases} > 0, & \text{if } s > 0 \\ = 0, & \text{if } s = 0 \\ < 0, & \text{if } s < 0, \end{cases}$$

where we used the fact that $h(x)$ is positive definite for all $x \in \mathsf{A}$. $\square$

Scoring functions for the pair (VaR, ES)

In this subsection, we shall give examples of strictly consistent scoring functions for the pair (VaR, ES) which are strictly order-sensitive on line segments. To this end, we assume the conditions of Corollary 5.5 in Fissler and Ziegel (2016). So we fix a (small) $\alpha \in (0,1)$ and assume that $\mathcal{F}$ is a class of distribution functions on $\mathbb{R}$ with finite first moments and unique α-quantiles. Moreover, to simplify the analysis, suppose that any distribution $F \in \mathcal{F}$ has a continuous derivative f (which then is also its density). Recall that $\mathrm{ES}_\alpha \le \mathrm{VaR}_\alpha$, such that it is no restriction to consider the action domain $\mathsf{A}_0 = \{x \in \mathbb{R}^2 \colon x_1 \ge x_2\}$. Then, a scoring function $S\colon \mathsf{A}_0 \times \mathbb{R} \to \mathbb{R}$ of the form

$$\begin{aligned} S(x_1,x_2,y) = {} & \big(\mathbb{1}\{y \le x_1\} - \alpha\big)G_1(x_1) - \mathbb{1}\{y \le x_1\}G_1(y) \\ & + G_2(x_2)\left(x_2 - x_1 + \frac{1}{\alpha}\mathbb{1}\{y \le x_1\}(x_1 - y)\right) - \mathcal{G}_2(x_2), \end{aligned} \tag{4.1.29}$$

where $G_1, G_2, \mathcal{G}_2\colon \mathbb{R} \to \mathbb{R}$, $\mathcal{G}_2' = G_2$, $\mathbb{1}_{(-\infty,x_1]}G_1$ is $\mathcal{F}$-integrable for all $x_1 \in \mathbb{R}$, is strictly $\mathcal{F}$-consistent for $T = (\mathrm{VaR}_\alpha, \mathrm{ES}_\alpha)$, if G_1 is increasing and $\mathcal{G}_2$ is strictly increasing and strictly convex. Throughout this paragraph, assume moreover that G_1 and G_2 are twice differentiable.

For some $F \in \mathcal{F}$, the first and second order partial derivatives of the expected score $\bar{S}(x_1, x_2, F)$ have the form

$$\left.\begin{aligned}
\partial_1 \bar{S}(x_1,x_2,F) &= (F(x_1)-\alpha)\left(G_1'(x_1)+\tfrac{1}{\alpha}G_2(x_2)\right),\\
\partial_2 \bar{S}(x_1,x_2,F) &= G_2'(x_2)\left(x_2 - x_1 + \tfrac{1}{\alpha}F(x_1)x_1 - \tfrac{1}{\alpha}\int_{-\infty}^{x_1} y\,\mathrm{d}F(y)\right),\\
\partial_1\partial_1 \bar{S}(x_1,x_2,F) &= f(x_1)\left(G_1'(x_1)+\tfrac{1}{\alpha}G_2(x_2)\right) + (F(x_1)-\alpha)G_1''(x_1),\\
\partial_1\partial_2 \bar{S}(x_1,x_2,F) &= \partial_2\partial_1 \bar{S}(x_1,x_2,F) = \tfrac{1}{\alpha}G_2'(x_2)(F(x_1)-\alpha),\\
\partial_2\partial_2 \bar{S}(x_1,x_2,F) &= G_2''(x_2)\left(x_2 - x_1 + \tfrac{1}{\alpha}F(x_1)x_1 - \tfrac{1}{\alpha}\int_{-\infty}^{x_1} y\,\mathrm{d}F(y)\right) + G_2'(x_2).
\end{aligned}\right\} \tag{4.1.30}$$

Now, for $v \in \mathbb{S}^1$, consider the function $\psi\colon D' \to \mathbb{R}$ defined at (4.1.28). Then, for $s \in D'$, $t = T(F)$ and $\bar{s} = t + sv$, one obtains

$$\begin{aligned}
\psi'(s) = v_1(F(\bar{s}_1)-\alpha)&\left(G_1'(\bar{s}_1)+\frac{1}{\alpha}G_2(\bar{s}_2)\right) \\
&+ v_2 G_2'(\bar{s}_2)\left(sv_2 + \frac{1}{\alpha}\left(\bar{s}_1(F(\bar{s}_1)-\alpha) - \int_{t_1}^{\bar{s}_1} y\,\mathrm{d}F(y)\right)\right).
\end{aligned} \tag{4.1.31}$$

We immediately see that $\psi'(0) = 0$. If $v_2 = 0$, then $v_1 \in \{-1, 1\}$ such that

$$\psi'(s) = v_1(F(\bar{s}_1)-\alpha)\left(G_1'(\bar{s}_1)+\frac{1}{\alpha}G_2(\bar{s}_2)\right)\begin{cases} >0, & s>0\\ =0, & s=0\\ <0, & s<0.\end{cases}$$

Similarly, we get for $v_2 \in \{-1, 1\}$, $v_1 = 0$, that $\psi'(s) = sG_2'(\bar{s}_2)$. Now, suppose without loss of generality that $v_2 > 0$ and $v_1 \neq 0$. Then, in a similar way as performed in a preprint version of Fissler and Ziegel (2015)[6], equation (4.1.31) can be rewritten as $\psi'(s) = v_2 G_2'(\bar{s}_2)(R(s) - L(s))$, where

$$\begin{aligned}
R(s) &= \frac{1}{\alpha}\left(\bar{s}_1(F(\bar{s}_1)-\alpha) - \int_{t_1}^{\bar{s}_1} y\,\mathrm{d}F(y)\right),\\
L(s) &= -sv_2 - v_1(F(\bar{s}_1)-\alpha)\,\frac{G_1'(\bar{s}_1)+\frac{1}{\alpha}G_2(\bar{s}_2)}{v_2 G_2'(\bar{s}_2)}.
\end{aligned}$$

Due to our assumptions it holds that $G_2' > 0$, so $\psi'(s) = 0$ if and only if $R(s) = L(s)$. Since

$$R'(s) = \frac{v_1}{\alpha}(F(\bar{s}_1)-\alpha)$$

is increasing with $R'(0) = 0$, R is a convex function, decreasing for $s < 0$ and increasing for $s > 0$. The first summand, $-sv_2$, of L is a linear function with slope $-v_2 < 0$. The second summand vanishes for $s = 0$, it is ≤ 0 for $s > 0$ and ≥ 0 for $s < 0$. Consequently, $R(s) > L(s)$ for $s > 0$, and hence $\psi'(s) > 0$ for $s > 0$.

[6]This preprint version we are referring to corresponds to version 1 on arXiv.

Due to the form of L, one has $L'(0) \leq v_2 < 0$. With a continuity argument, one obtains that $\psi'(s) < 0$ in a neighborhood of 0 for $s < 0$. This implies that ψ has a local minimum at $s = 0$ (alternatively, one can also see that $\psi''(0) > 0$). However, it is not clear whether there exists an $s^* < 0$ such that $R(s^*) = L(s^*)$. If such an $s^* < 0$ exists with $\psi'(s^*) = 0$ then ψ is quasi-convex if and only if s^* is not a strict local maximum, but an inflection point. A necessary condition is that $\psi''(s^*) \geq 0$ and a sufficient condition is that $\psi''(s^*) > 0$; see Proposition 4.2.3. For any $s \in D'$, ψ'' takes the form

$$\begin{aligned}\psi''(s) &= v_1^2 f(\bar{s}_1)\left(G_1'(\bar{s}_1) + \frac{1}{\alpha}G_2(\bar{s}_2)\right) + v_1(F(\bar{s}_1) - \alpha)\left(v_1 G_1''(\bar{s}_1) + \frac{v_2}{\alpha}G_2'(\bar{s}_2)\right)\\ &\quad + v_2^2 G_2'(\bar{s}_2) + v_2^3 G_2''(\bar{s}_2)s + \frac{v_2^2}{\alpha}G_2''(\bar{s}_2)\left(\bar{s}_1(F(\bar{s}_1) - \alpha) - \int_{t_1}^{\bar{s}_1} y\,\mathrm{d}F(y)\right)\\ &\quad + \frac{v_2}{\alpha}G_2'(\bar{s}_2)v_1(F(\bar{s}_1) - \alpha).\end{aligned}$$

Indeed, one can see that $\psi''(0) = v_1^2 f(\bar{s}_1)\left(G_1'(\bar{s}_1) + \frac{1}{\alpha}G_2(\bar{s}_2)\right) + v_2^2 G_2'(\bar{s}_2) > 0$. Now, computing $\psi''(s^*)$ using the condition that $R(s^*) = L(s^*)$, one obtains

$$\begin{aligned}\psi''(s^*) &= v_1^2 f(\bar{s}_1)\left(G_1'(\bar{s}_1^*) + \frac{1}{\alpha}G_2(\bar{s}_2^*)\right) + v_2^2 G_2'(\bar{s}_2^*)\\ &\quad + v_1(F(\bar{s}_1^*) - \alpha)\left(v_1 G_1''(\bar{s}_1^*) + v_2\frac{2}{\alpha}G_2'(\bar{s}_2^*) - v_2 G_2''(\bar{s}_2^*)\frac{G_1'(\bar{s}_1^*) + \frac{1}{\alpha}G_2(\bar{s}_2^*)}{G_2'(\bar{s}_2^*)}\right).\end{aligned} \tag{4.1.32}$$

The first summand in (4.1.32) is ≥ 0, and the second > 0. Since $v_2 > 0$ and $s^* < 0$, we have $v_1(F(\bar{s}_1^*) - \alpha) < 0$. Consequently, one needs to control the sign of the last bracket. We give a sufficient criterium such that also the third summand is ≥ 0:

$$G_1'' \equiv 0 \qquad \text{and} \qquad G_2(x)G_2''(x) - 2(G_2'(x))^2 \geq 0 \quad \forall x. \tag{4.1.33}$$

We give two examples for functions G_1, G_2 satisfying (4.1.33) and the conditions of Fissler and Ziegel (2016, Corollary 5.5), also stated after equation (4.1.29).

Example 4.1.50. Consider the restricted action domain $\mathsf{A}_0^- = \{x \in \mathbb{R}^2 \colon x_1 \geq x_2,\ x_2 < 0\} \subset \mathsf{A}_0$. So the domain of G_2 is $(-\infty, 0)$. Then choose G_1 linear with a non-negative slope and G_2 in the family

$$G_2(x) = |x|^{-b}, \quad b \in (0, 1], \quad x < 0.$$

Indeed, then $G_2'(x) = b|x|^{-b-1} > 0$ for $x < 0$ and $b > 0$, and

$$G_2(x)G_2''(x) - 2(G_2'(x))^2 = |x|^{-2(b+1)}(b - b^2) \geq 0,$$

for $b \leq 1$.

Connections to oriented identification functions

At the beginning of Subsection 4.1.2 we have discussed the notion of *orientation* of a strict identification function for a real-valued functional, which is due to Steinwart et al. (2014), but has already been mentioned by Lambert et al. (2008). Giving an alternative, but equivalent formulation of (4.1.5), which also corresponds to the formulation in Steinwart et al. (2014), one can say that a strict $\mathcal{F}$-identification function $V\colon \mathsf{A}\times\mathsf{O}\to\mathbb{R}$ for a functional $T\colon \mathcal{F}\to\mathsf{A}\subseteq\mathbb{R}$ is oriented for T if

$$\bar V(x,F) > 0 \quad\Longleftrightarrow\quad x > T(F) \tag{4.1.34}$$

for all $F\in\mathcal{F}$, $x\in\mathsf{A}$. In our preprint Fissler and Ziegel (2015), we gave a generalization of orientation for higher-dimensional functionals.

Definition 4.1.51 (Orientation)**.** Let $T\colon\mathcal{F}\to\mathsf{A}\subseteq\mathbb{R}^k$ be a functional with a strict $\mathcal{F}$-identification function $V\colon\mathsf{A}\times\mathsf{O}\to\mathbb{R}^k$. Then V is called an *oriented* strict $\mathcal{F}$-identification function for T if

$$v^\top \bar V(T(F)+sv,F) > 0 \quad\Longleftrightarrow\quad s>0$$

for all $v\in\mathbb{S}^{k-1} := \{x\in\mathbb{R}^k\colon \|x\|=1\}$, for all $F\in\mathcal{F}$ and for all $s\in\mathbb{R}$ such that $T(F)+sv\in\mathsf{A}$.

Remark 4.1.52. (i) Indeed, the one-dimensional definition of orientation at (4.1.34) is nested in Definition 4.1.51 upon recalling that $\mathbb{S}^0=\{-1,1\}$.

(ii) Our notion of orientation differs from the one proposed by Frongillo and Kash (2015a). In contrast to their definition, our definition is *per se* independent of a (possibly non-existing) strictly consistent scoring function for T. Moreover, whereas their definition has connections to the convexity of the expected score, our definition shows strong ties to order-sensitivity on line segments.

If the gradient of an expected score induces an oriented identification function, then the scoring function is strictly order-sensitive on line segments and vice versa. However, the existence of an oriented identification function is not sufficient for the existence of a strictly order-sensitive scoring function on line segments. The reason is that – due to integrability conditions – the identification function is not necessarily the gradient of some (scoring) function.

Likewise, under Assumption (S1) in Fissler and Ziegel (2016), the gradient of an expected score induces a *locally oriented* identification function in the sense of the following definition.

Definition 4.1.53 (Local orientation)**.** Let $T\colon\mathcal{F}\to\mathsf{A}\subseteq\mathbb{R}^k$ be a functional with an $\mathcal{F}$-identification function $V\colon\mathsf{A}\times\mathsf{O}\to\mathbb{R}^k$. Then V is called a *locally* oriented $\mathcal{F}$-identification function for T if for all $F\in\mathcal{F}$ and for all $v\in\mathbb{S}^{k-1}$ there is an $\varepsilon>0$ such that for all $s\in(-\varepsilon,\varepsilon)$ with $T(F)+sv\in\mathsf{A}$

$$v^\top \bar V(T(F)+sv,F) > 0 \quad\Longleftrightarrow\quad s>0.$$

We end this paragraph by mentioning that, apparently, there is a flaw in the proof of Steinwart et al. (2014, Lemma 6). Assuming convexity of $\mathcal{F}$, continuity of $T\colon \mathcal{F} \to \mathbb{R}$ (actually, they only use mixture-continuity) and the fact that T has convex level sets, they claim that if V is a strict $\mathcal{F}$-identification function for T, then either V or $-V$ is oriented. However, considering the mean functional, any identification function $V\colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ of the form

$$V(x,y) = h(x)(x-y)$$

is a strict identification function if $h \neq 0$. But for

$$h\colon \mathbb{R} \to \mathbb{R}, \qquad h(x) = \begin{cases} 1, & x \geq 0 \\ -1, & x < 0, \end{cases}$$

V is not oriented. Of course, multiplying with a discontinuous function h, will also cause V to be discontinuous in x; cf. Remark 3.2.2.

4.1.8. Nested information sets

Corollary 2 in Holzmann and Eulert (2014)[7], which corresponds to Lemma 2.2.6 in this thesis, can also be considered as a notion of order-sensitivity or monotonicity. However, it argues on the level of the prediction space setting and not on the 'conditional' level, assuming that the forecast is deterministic. So it should not be mixed with the notions presented above in this section.

4.2. Convexity of scoring functions

In this section, we investigate the notions of convexity and quasi-convexity of scoring functions. To this end, consider a generic scoring function $S\colon \mathsf{A} \times \mathsf{O} \to \mathbb{R}$ where $\mathsf{A} \subseteq \mathbb{R}^k$ is convex, $\mathsf{O} \subseteq \mathbb{R}^d$. Let $\mathcal{F}$ be a class of distributions on O such that S is $\mathcal{F}$-integrable. We start with some definitions.

Definition 4.2.1 (Convex scoring function)**.** A scoring function $S\colon \mathsf{A} \times \mathsf{O} \to \mathbb{R}$ is called convex if $S(\cdot, y)\colon \mathsf{A} \to \mathbb{R}$ is convex for all $y \in \mathsf{O}$. That is, if for all $y \in \mathsf{O}$

$$S((1-\lambda)x_0 + \lambda x_1, y) \leq (1-\lambda)S(x_0,y) + \lambda S(x_1,y) \qquad \forall x_0, x_1 \in \mathsf{A},\ \forall \lambda \in [0,1].$$

If the inequality is strict for $x_0 \neq x_1$ and $\lambda \in (0,1)$, S is strictly convex. S is called $\mathcal{F}$-convex if the expected score $\bar{S}(\cdot, F)\colon \mathsf{A} \to \mathbb{R}$ is convex for all $F \in \mathcal{F}$. That is, if for all $F \in \mathcal{F}$

$$\bar{S}((1-\lambda)x_0 + \lambda x_1, F) \leq (1-\lambda)\bar{S}(x_0,F) + \lambda \bar{S}(x_1,F) \qquad \forall x_0, x_1 \in \mathsf{A},\ \forall \lambda \in [0,1].$$

If the inequality is strict for $x_0 \neq x_1$ and $\lambda \in (0,1)$, S is strictly $\mathcal{F}$-convex.

[7]See Patton (2015) for a partial converse of this result.

Definition 4.2.2 (Quasi-convex scoring function)**.** A scoring function $S\colon \mathsf{A}\times\mathsf{O} \to \mathbb{R}$ is called quasi-convex if $S(\cdot, y)\colon \mathsf{A} \to \mathbb{R}$ is quasi-convex for all $y \in \mathsf{O}$. That is, if for all $y \in \mathsf{O}$

$$S((1-\lambda)x_0 + \lambda x_1, y) \le \max\{S(x_0, y), S(x_1, y)\} \qquad \forall x_0, x_1 \in \mathsf{A},\ \forall \lambda \in [0,1].$$

If the inequality is strict for $x_0 \neq x_1$ and $\lambda \in (0,1)$, S is strictly quasi-convex. S is called $\mathcal{F}$-quasi-convex if the expected score $\bar{S}(\cdot, F)\colon \mathsf{A} \to \mathbb{R}$ is quasi-convex for all $F \in \mathcal{F}$. That is, if for all $F \in \mathcal{F}$

$$\bar{S}((1-\lambda)x_0 + \lambda x_1, F) \le \max\{\bar{S}(x_0, F), \bar{S}(x_1, F)\} \qquad \forall x_0, x_1 \in \mathsf{A},\ \forall \lambda \in [0,1].$$

If the inequality is strict for $x_0 \neq x_1$ and $\lambda \in (0,1)$, S is strictly $\mathcal{F}$-quasi-convex.

Let us recall that a function $f\colon \mathbb{R}^k \supseteq C \to \mathbb{R}$, where C is convex, is quasi-convex if and only if its lower-level sets are convex, that is, if for all $z \in \mathbb{R}$ the set $\{x \in C\colon f(x) \le z\}$ is convex. For a good introduction into the theory of quasi-convex (or quasi-concave) functions, we refer the reader to the books of Schaible and Ziemba (1981) and Avriel et al. (2010) as well as to the survey article of Greenberg and Pierskalla (1971) and the two research articles of Diewert et al. (1981), and Crouzeix and Ferland (1982). We cite two equivalent characterization of quasi-convexity. The first one is given in terms of second order conditions and corresponds to Proposition 9 and Proposition 11 in Schaible and Ziemba (1981, pp. 36–39).

Proposition 4.2.3 (Schaible and Ziemba (1981))**.** *Let $C \subseteq \mathbb{R}^k$ be an open convex set and let $f\colon C \to \mathbb{R}$ be a twice continuously differentiable function with gradient ∇f and Hessian $\nabla^2 f$. Then f is quasi-convex if and only if for all $x \in C$ and for all $v \in \mathbb{S}^{k-1}$,*

$$v^\top \nabla f(x) = 0$$

implies

(i) $v^\top \nabla^2 f(x) v > 0$; or

(ii) $v^\top \nabla^2 f(x) v = 0$ and $\psi\colon \{s \in \mathbb{R}\colon x + sv \in C\} \to \mathbb{R}$, $\psi(s) = f(x + sv)$ is quasi-convex.

Moreover, f is strictly quasi-convex if and only if this implication holds where in (ii) ψ does not attain a local maximum at $s = 0$.

The next result corresponds to Theorem 3.15 in Avriel et al. (2010, p. 69) and the second part concerning strict quasi-convexity is due to Proposition 10 in Schaible and Ziemba (1981, p. 38). But before, we give a definition of a semistrict local maximum, which is due to Definition 3.3 in Avriel et al. (2010, p. 60).

Definition 4.2.4 (Avriel et al. (2010))**.** A function $f\colon C \to \mathbb{R}$ where $C \subseteq \mathbb{R}$ is an open interval, is said to attain a semistrict local maximum at a point $x_0 \in C$ if there exist $x_1, x_2 \in C$, with $x_1 < x_0 < x_2$ such that

$$f(x_0) \ge f((1-\lambda)x_1 + \lambda x_2) \qquad \forall \lambda \in [0,1]$$

and

$$f(x_0) > \max\{f(x_1), f(x_2)\}.$$

Proposition 4.2.5 (Avriel et al. (2010))**.** *Let $C \subseteq \mathbb{R}^k$ be an open convex set and let $f\colon C \to \mathbb{R}$ be a differentiable function with gradient ∇f. Then f is quasi-convex if and only if for all $x \in C$ and for all $v \in \mathbb{S}^{k-1}$*

$$v^\top \nabla f(x) = 0$$

implies that the function $\psi\colon \{s \in \mathbb{R}\colon x + sv \in C\} \to \mathbb{R}$, $\psi(s) = f(x + sv)$ does not attain a semistrict local maximum at $s = 0$.

Moreover, f is strictly quasi-convex if and only if this implication holds where the function ψ does not attain a local maximum at $s = 0$.

Since the space of convex functions over the same domain forms a convex cone, the convexity of a scoring function implies its $\mathcal{F}$-convexity with respect to any class of probability distributions $\mathcal{F}$ on O. However, the analogue result is not true for quasi-convex scoring functions, since the sum of quasi-convex functions does not need to be quasi-convex.[8] Nevertheless, in the one-dimensional case, order-sensitivity ensures quasi-convexity. The reason is that the class of (strictly) quasi-convex functions defined on an interval in $\mathbb{R}$ can be easily described. It consists exactly of those functions that are either (strictly) monotone or there is a point x in their domain such that they are (strictly) decreasing up to x and afterwards (strictly) increasing.

Lemma 4.2.6. *Let $\mathcal{F}$ be convex and $T\colon \mathcal{F} \to \mathsf{A} \subseteq \mathbb{R}$ be surjective and mixture-continuous. Then, any $\mathcal{F}$-consistent scoring function for T is quasi-convex.*

Proof. Let S be $\mathcal{F}$-consistent for T. Then, by Proposition 4.1.10, S is $\mathcal{F}$-order-sensitive for T. Let $F \in \mathcal{F}$, and $x \in \operatorname{int}(\mathsf{A})$ and without loss of generality $x \geq T(F)$. Then there are $x_0 \leq x \leq x_1$ such that x is a convex combination of x_0 and x_1. But due to the order-sensitivity $\bar{S}(x, F) \leq \bar{S}(x_1, F)$. □

Remark 4.2.7. It is worth mentioning that the definitions of ($\mathcal{F}$-)(quasi-)convexity of a scoring function is not relative to a functional, but an absolute notion. Moreover, if a scoring function is strictly $\mathcal{F}$-convex, it induces a functional $T\colon \mathcal{F} \to \mathbb{R}$, via $T(F) = \arg\min_{x\in\mathbb{R}} \bar{S}(x, F)$.

Remark 4.2.8. (Quasi-)convexity is also preserved by equivalence of scoring functions.

[8] To illustrate this fact, consider the family of functions $f_y\colon \mathbb{R} \to \mathbb{R}$, $f_y(x) = |x - y|^{1/2}$ for $y \in \mathbb{R}$. For any $y \in \mathbb{R}$, f_y is quasi-convex. However, $f_{-1} + f_1$ has two global minima at $x \in \{-1, 1\}$ and a strict local maximum at $x = 0$. Hence, $f_{-1} + f_1$ is not quasi-convex.

4.2.1. Motivation

The use of quasi-convex, and in particular convex scoring functions might be beneficial in many different areas. Subsequently, we shall give examples related to learning and regression, show how convex scoring functions can incentivize 'sharper' forecasts and cooperations between competing forecasters, and finally show a connection to the question of 'backtestability' of risk measures in the context of quantitative finance.

Learning

Let us repeat the framework of learning and M-estimation already described in Subsection 2.2.1. If we have n observations of an i.i.d. sequence $(Y_t)_{t\in\mathbb{N}}$ where Y_t has some unknown distribution $F \in \mathcal{F}$ (actually, ergodicity is sufficient) and one wants to estimate a functional $T\colon \mathcal{F} \to \mathsf{A} \subseteq \mathbb{R}^k$, one can use any strictly $\mathcal{F}$-consistent scoring function S to do M-estimation in the sense that

$$\arg\min_{x\in\mathsf{A}} \frac{1}{n}\sum_{t=1}^{n} S(x, Y_t) \tag{4.2.1}$$

is a consistent estimator for

$$\arg\min_{x\in\mathsf{A}} \mathbb{E}[S(x,Y)] = \arg\min_{x\in\mathsf{A}} \bar{S}(x,F) = T(F) \tag{4.2.2}$$

under some regularity conditions detailed in Huber and Ronchetti (2009, Chapter 6). Then, if S is ($\mathcal{F}$-)convex, both (4.2.1) and (4.2.2) are convex optimization problems. Furthermore, if S is quasi-convex, then (4.2.1) is a quasi-convex optimization problem, whereas (4.2.2) is a quasi-convex optimization problem if S is $\mathcal{F}$-quasi-convex.

Note that not only convexity is beneficial in the context of optimization, but also quasi-convexity; see for example Chapter 3 in Diewert et al. (1981).

Regression

We use the notation introduced in Subsection 2.2.2. That is, we consider n observations of an i.i.d. sequence $(Z_t, Y_t)_{t\in\mathbb{N}}$ where $(Z_t, Y_t) \sim (Z, Y)$. Recall that Z takes values in $\mathbb{R}^\ell$ and consists of the explanatory factors for the output variable Y, taking values in $\mathsf{O} \subseteq \mathbb{R}^d$. Assume that $\mathcal{G}$ is a class of models, where a model $g \in \mathcal{G}$ is a (measurable) function $g\colon \mathbb{R}^\ell \to \mathbb{R}^d$. Let us further assume that the class $\mathcal{G}$ is *convex*. Let $T\colon \mathcal{F} \to \mathsf{A} \subseteq \mathbb{R}^k$ be a functional such that (i) the (regular version of) the conditional distribution $\mathcal{L}(Y|Z)$ is in $\mathcal{F}$ (almost surely); and (ii) for each $g \in \mathcal{G}$, $g(Z) \in \mathsf{A}$ almost surely. Let $S\colon \mathsf{A}\times\mathsf{O} \to \mathbb{R}$ be a strictly $\mathcal{F}$-consistent scoring function for T. Then, the goal of regression is to determine the 'oracle function'[9]

$$\arg\min_{g\in\mathcal{G}} \frac{1}{n}\sum_{t=1}^{n} S(g(Z_t), Y_t). \tag{4.2.3}$$

If S is a convex scoring function, (4.2.3) is clearly a convex optimization problem. Moreover, the law of large numbers yields that the objective function in (4.2.3) converges (almost surely or in probability) to $\mathbb{E}[S(g(Z), Y)]$. Under some regularity conditions detailed in Huber and Ronchetti (2009, Chapter 6), then also the arg min defined at (4.2.3) converges (almost surely or in probability) to

$$\arg\min_{g\in\mathcal{G}} \mathbb{E}[S(g(Z), Y)]. \tag{4.2.4}$$

So still the convexity of S is sufficient for (4.2.4) to be a convex optimization problem. *A fortiori*, the identity

$$\mathbb{E}[S(g(Z), Y)] = \mathbb{E}\big[\mathbb{E}[S(g(Z), Y) \mid Z]\big] = \mathbb{E}\left[\bar{S}(g(Z), \mathcal{L}(Y|Z))\right]$$

shows that the $\mathcal{F}$-convexity of S is sufficient for (4.2.4) to be a convex optimization problem.

On the other hand, neither the $\mathcal{F}$-quasi-convexity nor the quasi-convexity of S imply that (4.2.3) or (4.2.4) is a quasi-convex optimization problem (unless $\mathcal{F}$ contains all necessary discrete distributions). The difficulty is again that the sum of quasi-convex functions (over the same domain or over different domains) is not necessarily quasi-convex. So it appears that, whereas convexity of scoring functions is highly desirable in the context of regression, quasi-convexity is of limited use.

Incentives to maximize 'sharpness' subject to calibration

In the context of the evaluation of probabilistic forecasts, Gneiting et al. (2007, pp. 245–246) proposed the

> "paradigm of *maximizing the sharpness of the predictive distributions subject to calibration*. Calibration refers to the statistical consistency between the distributional forecasts and the observations and is a joint property of the predictions and the observed values. Sharpness refers to the concentration of the predictive distributions and is a property of the forecasts only. The more concentrated the predictive distributions are, the sharper the forecasts, and the sharper the better, subject to calibration."

If one strives to establish a similar notion of this sharpness principle in the context of point forecasts, first, it is important to think of a new definition of sharpness in this context. *Prima facie*, point forecasts are maximally sharp in the sense that they are concentrated at one point. However, moving one level upwards and

[9] We use this terminology in line with the machine learning community; see the contribution of H. Liu in the discussion article Ehm et al. (2016, p. 549).

taking the perspective of the prediction space setting (actually, for our considerations, the one-period-prediction-space-setting due to Gneiting and Ranjan (2013) is sufficient), one could argue that a forecast is the sharper the more unnecessary variance it avoids. Clearly, this definition is a bit vague, and we shall come up with a precise definition. But before, we want to give an illustrative example in the easiest possible setting, that is, when observation and forecast are independent.

Note that within this motivation, we will not focus too much on technicalities such as measurability or integrability assumptions. Moreover, we confine ourselves to an exposition of the ideas for the one-dimensional setting, such that $\mathsf{A} \subseteq \mathbb{R}$. Consider a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, and two random variables X, Y where Y is the observation and X the forecast. Let X be measurable with respect to some sub-σ-algebra $\mathcal{A}_0 \subset \mathcal{A}$. Then, if the goal is to predict some functional T of the (conditional) distribution of Y, the ideal forecast, knowing $\mathcal{A}_0$, would be $T(\mathcal{L}(Y|\mathcal{A}_0))$. Let us assume for a moment that $\mathcal{A}_0$ does not contain any information about Y in the sense that $\mathcal{A}_0$ and $\sigma(Y)$ are independent. Then the ideal forecast is clearly $T(\mathcal{L}(Y))$, wich is a constant. Consequently, X does not have a chance to be the ideal forecast (with respect to his (useless) information $\mathcal{A}_0$) unless X is a constant. Even if one does not know which constant is the ideal forecast, at least, the 'closest' constant to X should be deemed a better forecast than X itself. The term 'close' can be understood in the L^2-sense, implying that the L^2-projection of X onto the space of constants, that is merely its expectation $\mathbb{E}[X]$, can be regarded a better forecast and should be preferred over X. Therefore, it can be a reasonable requirement for a strictly consistent scoring function for T to reflect this preference. If a scoring function is $\mathcal{F}$-convex (where, as usual, $\mathcal{F}$ is the domain of T), this property is automatically satisfied due to the Jensen inequality. Indeed,

$$\mathbb{E}[S(X,Y)] = \mathbb{E}\big[\bar{S}(X, \mathcal{L}(Y))\big] \geq \bar{S}\big(\mathbb{E}[X], \mathcal{L}(Y)\big) = \mathbb{E}\big[S(\mathbb{E}[X], Y)\big], \tag{4.2.5}$$

where we obtain the first identity by conditioning on $\mathcal{A}_0$ and taking the expectation with respect to Y. It is remarkable that (4.2.5) holds irrespective whether S is consistent for T or not. The following example shows that (4.2.5) is generally not satisfied if S is not convex.

Example 4.2.9. Let $T\colon \mathcal{F} \to \mathbb{R}$ be the mean functional and $\mathcal{F}$ be the class of probability distributions with finite first moments. Due to Gneiting (2011, Theorem 7) the scoring function $S\colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, $S_\phi(x, y) = -\phi(x) + \phi'(x)(x - y)$ is strictly $\mathcal{F}$-consistent for T if ϕ is strictly convex. Now consider the convex function $\phi(x) = x^2/(1 + |x|)$ with derivative

$$\phi'(x) = \frac{2x(1 + |x|) - \operatorname{sgn}(x)x^2}{(1 + |x|)^2}.$$

Clearly (and consistently with Corollary 4.2.17), S_ϕ is not $\mathcal{F}$-convex.

Now let Y have distribution F with mean $T(F) = 0$, and let X be independent of Y with distribution $\mathbb{P}(X = 10) = \mathbb{P}(X = 0) = \frac{1}{2}$. Then

$$\bar{S}(0, F) = S(0, 0) = 0, \quad \bar{S}(5, F) = S(5, 0) = \frac{25}{36}, \quad \bar{S}(10, F) = S(10, 0) = \frac{100}{121},$$

such that

$$\mathbb{E}[S(X, Y)] = \frac{50}{121} < \frac{25}{36} = \mathbb{E}\big[S(\mathbb{E}[X], Y)\big],$$

and (4.2.5) is violated.

Clearly, the assumption that $\mathcal{A}_0$ and Y are independent is quite restrictive, but is fulfilled in some situations. One example is the case of i.i.d. observations $(Y_t)_{t\in\mathbb{N}}$, and the information set $\mathcal{A}_0$ for a fixed time t is generated by the past observations $Y_1, \ldots, Y_{t-1}$; see Section 2.2 and the arguments around equation (2.2.2).

For the general situation where $\mathcal{A}_0$ and Y are not independent, one can apply a similar reasoning. Then, the ideal forecast $T(\mathcal{L}(Y|\mathcal{A}_0))$ is not necessarily constant, but a $\sigma(\mathcal{L}(Y|\mathcal{A}_0))$-measurable random variable. Again, one can argue that the 'closest' $\sigma(\mathcal{L}(Y|\mathcal{A}_0))$-measurable random variable to X, that is the L^2-projection of X to the space of $\sigma(\mathcal{L}(Y|\mathcal{A}_0))$-measurable (and square-integrable) random variables, which is the conditional expectation

$$\mathbb{E}\big[X|\sigma(\mathcal{L}(Y|\mathcal{A}_0))\big],$$

should be deemed a better forecast. Similarly to (4.2.5), this is automatically reflected by $\mathcal{F}$-convex scoring functions and the conditional Jensen inequality.[10] Indeed,

$$\begin{aligned}\mathbb{E}[S(X,Y)] &= \mathbb{E}\big[\mathbb{E}[S(X,Y)\,|\,\mathcal{A}_0]\big] = \mathbb{E}\big[\bar{S}(X, \mathcal{L}(Y|\mathcal{A}_0))\big] \\ &\geq \mathbb{E}\left[\bar{S}\big(\mathbb{E}[X|\sigma(\mathcal{L}(Y|\mathcal{A}_0))], \mathcal{L}(Y|\mathcal{A}_0)\big)\right] = \mathbb{E}\big[S(\mathbb{E}[X|\sigma(\mathcal{L}(Y|\mathcal{A}_0))], Y)\big]. \qquad (4.2.6)\end{aligned}$$

In the following example, we illustrate this more general situation, and show that the objects X, $\mathbb{E}[X|\mathcal{A}_0]$, and $T(\mathcal{L}(Y|\mathcal{A}_0))$ do generally not coincide.

Example 4.2.10. Let T be again the mean functional, μ be some random variable taking values in $\mathbb{Z}$ with a non-symmetric distribution and which is such that μ takes positive and negative values with positive probability, τ be independent of μ with distribution $\mathbb{P}(\tau = 1) = \mathbb{P}(\tau = -1) = \frac{1}{2}$ and Y be an observation with conditional distribution

$$\mathcal{L}(Y|\mu, \tau) = \mathcal{N}(\tau\mu, 1).$$

Let X be a forecast, having only access to μ, that is, X is measurable with respect to $\mathcal{A}_0 = \sigma(\mu)$. Then

$$\mathcal{L}(Y|\mathcal{A}_0) = \mathcal{L}(Y|\mu) = \frac{1}{2}\mathcal{N}(\mu, 1) + \frac{1}{2}\mathcal{N}(-\mu, 1).$$

[10] We assume that the regular version of $\mathcal{L}(Y|\mathcal{A}_0)$ is in $\mathcal{F}$ with probability 1.

Clearly, the ideal forecast with respect to $\mathcal{A}_0$ is

$$T(\mathcal{L}(Y|\mathcal{A}_0)) = \frac{1}{2}(\mu - \mu) = 0.$$

Assume that, for some reason, the forecaster ignores the possibility that $\tau = -1$. Therefore, he issues the misspecified forecast $X = \mu$. Now, let us determine the conditional expectation of X given $\sigma(\mathcal{L}(Y|\mathcal{A}_0))$. It holds that

$$\mathcal{L}(Y|\mathcal{A}_0) = \mathcal{L}(Y|X) = \mathcal{L}(Y|-X),$$

so $\sigma(\mathcal{L}(Y|\mathcal{A}_0))$ does not contain information about the sign of X. On the other hand

$$\mathbb{E}\big[Y^2|\mathcal{A}_0\big] = \mathbb{E}\big[Y^2|X\big] = \frac{1}{2}\big(1 + X^2 + 1 + (-X)^2\big) = 1 + X^2.$$

Hence, $\sigma(\mathcal{L}(Y|\mathcal{A}_0)) = \sigma(\mathcal{L}(Y|X)) = \sigma(X^2) = \sigma(|X|)$. Moreover,

$$\mathbb{E}\big[X|\sigma(|X|)\big] = |X|\,\big(\mathbb{P}(X = |X|) - \mathbb{P}(X = -|X|)\big),$$

which is, due to our assumptions, different from X and different from $T(\mathcal{L}(Y|\mathcal{A}_0)) = 0$.

Recall that, given some information set $\mathcal{A}_0$, the ultimate goal of point forecasting is to issue the ideal forecast $T(\mathcal{L}(Y|\mathcal{A}_0))$. One can equivalently define this goal as finding the minimizer of

$$\mathbb{E}\big[(X - T(\mathcal{L}(Y|\mathcal{A}_0)))^2\big]$$

over all ($\mathcal{A}_0$-measurable) random variables. Now, having defined this L^2-distance as an objective function, one can consider the usual bias-variance-decomposition. We suggest the following definition of calibration in the context of point forecasts.

Definition 4.2.11 (Calibration)**.** Given some observation Y, a functional T and an information set $\mathcal{A}_0$, a ($\mathcal{A}_0$-measurable) forecast X is

(a) *calibrated on average* if

$$\mathbb{E}[X - T(\mathcal{L}(Y|\mathcal{A}_0))] = 0; \tag{4.2.7}$$

(b) *conditionally calibrated* if

$$\mathbb{E}\big[X - T(\mathcal{L}(Y|\mathcal{A}_0))\,|\,\sigma(T(\mathcal{L}(Y|\mathcal{A}_0)))\big] = 0; \tag{4.2.8}$$

(c) *strongly conditionally calibrated* if

$$\mathbb{E}\big[X - T(\mathcal{L}(Y|\mathcal{A}_0))\,|\,\sigma(\mathcal{L}(Y|\mathcal{A}_0))\big] = 0. \tag{4.2.9}$$

Consistently, the left hand side of (4.2.7) is called *calibration bias*, and the left hand side of (4.2.8) ((4.2.9)) is called *(strong) conditional calibration bias.*

Clearly, strong conditional calibration implies conditional calibration which in turn implies calibration on average.

Definition 4.2.12 (Sharpness)**.** Given a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, a forecast $X^{(1)}$ is sharper than a forecast $X^{(2)}$ if

$$\operatorname{Var}(X^{(1)}) \le \operatorname{Var}(X^{(2)}).$$

Clearly, the sharpness relation defines a total preorder on the set of L^2-integrable random variables on $(\Omega, \mathcal{A}, \mathbb{P})$ where the constants are the sharpest forecasts.[11] Having established this terminology, it is easy to check that the ideal forecast is the unique forecast maximizing sharpness, subject to conditional calibration.[12] Similarly to the concept in probabilistic forecasting, calibration is a joint property of the prediction and the observation whereas sharpness is a property of the forecast, only.

As demonstrated at (4.2.6), under all forecasts with the same strong calibration bias, a convex scoring function obtains the smallest expected score for the sharpest of all these forecasts.[13]

Incentives for cooperation between forecasters

Another motivation to consider convex scores could be that they incentivize cooperations between different forecasters. To illustrate this, suppose there is a competition between m different forecasters with predictions $X^{(1)}, \dots, X^{(m)}$ for some outcome Y. Let $S\colon \mathsf{A} \times \mathsf{O} \to \mathbb{R}$ be a convex scoring function. If the forecasters were to be penalized according to their realized scores by the organizer of the competition and they were allowed to negotiate and combine their respective forecasts, they could enter the following game: They agree to a payment to the group according to their individual realized scores with respect to the observation Y, that is, $S(X^{(i)}, Y)$ for $i \in \{1, \dots, m\}$, and to redistribute equal shares to everybody, that is

$$\frac{1}{m}\sum_{i=1}^m S(X^{(i)}, Y)\,.$$

However, they agree to quote the same forecast to the organizer which is just the average of the individual forecasts $\bar{X} = \frac{1}{m}\sum_{i=1}^m X^{(i)}$. Then, they are able to realize the saving

$$\frac{1}{m}\sum_{i=1}^m S(X^{(i)}, Y) - S(\bar{X}, Y) \ge 0$$

[11] One could also think of different definitions of sharpness, e.g. in terms of generated σ-algebras. Then $X^{(1)}$ would be sharper than $X^{(2)}$ if $\sigma(X^{(1)}) \subseteq \sigma(X^{(2)})$. This in turn would induce a partial order on the space of random variables on $(\Omega, \mathcal{A}, \mathbb{P})$ where, again, the constants are the sharpest forecasts.

[12] Indeed, this assertion is true for both notions of sharpness.

[13] Again, this assertion is still true with the alternative concept of sharpness.

in contrast to their individual strategies.

If S is merely quasi-convex, there are still possibilities for a joint insurance between the forecasters. Suppose they are rather risk averse and fear to be penalized by the worst fine, that is $\max_{i\in\{1,\dots,m\}} S(X^{(i)}, Y)$, they can again agree to issue a joint linearly combined forecast $\bar{X}_c \in \operatorname{conv}\{X^{(1)}, \dots, X^{(m)}\}$. Then, they are better of by the amount

$$\max_{i\in\{1,\dots,m\}} S(X^{(i)}, Y) - S(\bar{X}_c, Y) \geq 0$$

in contrast to the worst case scenario (but not with respect to their individual strategies). Such insurance type properties related to quasi-convex scores are also discussed e.g. in Ehm et al. (2016, p. 558) or in Kascha and Ravazzolo (2010, p. 237).

Depending on the particular goal at hand, these observations show that the use of convex scoring functions can be double edged. Taking a forecaster's perspective, this property can be desirable because it allows for a kind of arbitrage by cooperating with his competitors. On the other hand, taking the organizer's angle, such cooperations between and combinations of competing forecasts dilute the information given by the individual quotes. Consequently, the organizer is not able to distinguish between the different forecasters. And more severe, the best forecast is not visible any more. So an organizer using convex scoring functions should be aware of this problematic and should prevent competing forecasters from cooperation.

A definition of backtestability proposed by Acerbi and Székely (2017)

Acerbi and Székely (2017) proposed a new definition of backtestability. Using our notation, they say that a real-valued functional $T\colon \mathcal{F} \to \mathbb{R}$ is $\mathcal{F}$-backtestable if there is an oriented and strict $\mathcal{F}$-identification function $V : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ for T such that the expected identification function $\bar{V}(\cdot, F)\colon \mathbb{R} \to \mathbb{R}$ is strictly increasing for all $F \in \mathcal{F}$. They call such a particular identification function a *backtest function*. Consequently, if an $\mathcal{F}$-backtestable functional T is elicitable (and in the setting of Steinwart et al. (2014), this is the case), it has a strictly $\mathcal{F}$-convex strictly $\mathcal{F}$-consistent scoring function, which can be obtained by integration of the backtest function. *Vice versa*, if an elicitable functional has a strictly $\mathcal{F}$-convex and strictly $\mathcal{F}$-consistent scoring function being sufficiently smooth (the expected score must be differentiable for all $F \in \mathcal{F}$), then the functional is $\mathcal{F}$-backtestable. That means that backtestability in the sense of Acerbi and Székely (2017) and convex elicitability almost coincide (in the one-dimensional setting).

Acerbi and Székely (2017) motivate this definition of backtestability arguing that an identification function fulfilling the additional properties of a backtest function does not only allow for model validation in the sense of traditional backtests (see Fissler et al. (2016) or Nolde and Ziegel (2016, Section 2.2)), but also allows for model selection in the sense of comparative backtests (see *ibidem*).

What remains unclear so far is how to generalize this notion to the case of higher-dimensional functionals. There appear to be (at least) three natural choices: (i) One could require the functional $T\colon \mathcal{F} \to \mathbb{R}^k$ to have an identification function such that each of its components is a backtest function of one argument only (which induces a convex and componentwise order-sensitive scoring function); (ii) one could say that a functional $T\colon \mathcal{F} \to \mathbb{R}^k$ is backtestable if it possesses a strictly $\mathcal{F}$-consistent and $\mathcal{F}$-convex scoring function. A bit weaker could be the requirement that (iii) T possesses a strictly $\mathcal{F}$-consistent scoring function $S\colon \mathbb{R}^k \times \mathsf{O} \to \mathbb{R}$ such that

$$\psi\colon \mathbb{R} \to \mathbb{R}, \qquad \psi(s) = \bar{S}(T(F) + sv, F)$$

is convex for all $F \in \mathcal{F}$ and for all $v \in \mathbb{S}^{k-1}$. Clearly, this would be very close to (and a bit stronger than) our definition of order-sensitivity on line segments.

4.2.2. Determining convex scoring functions for popular functionals

It is known that the α-pinball loss $S_\alpha(x,y) = (\mathbb{1}\{y \le x\} - \alpha)(x-y)$ (or asymmetric linear loss) and the asymmetric squared loss $S_\tau(x,y) = |\mathbb{1}\{y \le x\} - \tau|(x-y)^2$ are convex scoring functions that are strictly consistent for the α-quantile (relative to the class of strictly increasing distributions with a finite mean), and for the τ-expectile (relative to the class of distributions having a finite second moment). It is natural to ask if those scoring functions are the only strictly consistent and convex scoring functions for these functionals. In this section, we determine the classes of $\mathcal{F}$-convex strictly $\mathcal{F}$-consistent scoring functions for quantiles and expectiles as well as for ratios of expectations, where we also give an extension to the higher-dimensional setting, subject to smoothness and richness assumptions on the respective class $\mathcal{F}$.

The main message is that the flexibility one has in building convex scoring functions crucially depends on the richness of the class $\mathcal{F}$ and in particular on the image of the respective functional $T(\mathcal{F})$. Roughly speaking, if $T(\mathcal{F}) = \mathbb{R}$ $(= \mathbb{R}^k)$, then there is only one $\mathcal{F}$-convex and strictly $\mathcal{F}$-consistent scoring function (up to equivalence). On the other hand, if the image $T(\mathcal{F})$ is bounded, there are other $\mathcal{F}$-convex scoring functions.

Towards the end of this section, we also consider functionals having non-elicitable components. To illuminate this situation, we consider the prominent example of the pair (VaR, ES), determine an $\mathcal{F}$-quasi-convex scoring function and show that under some richness assumptions on $\mathcal{F}$, there is no $\mathcal{F}$-convex scoring function.

In many parts this section, we assume the regularity assumptions of Fissler and Ziegel (2016, Proposition 3.4) such that the strictly consistent scoring functions are necessarily of the form given in the latter proposition.

Quantiles

Steinwart et al. (2014, Corollary 10) state that the α-pinball loss (modulo inessential transformations) is the only convex strictly consistent scoring function for the α-quantile. While they consider the situation of a class of distributions $\mathcal{F}$ where each density f of a distribution $F \in \mathcal{F}$ is bounded away from 0, it is possible to construct other convex scoring functions if one assumes that the densities are *uniformly* bounded away from 0. To make the analysis a bit easier, we will assume a bit more regularity conditions.

Proposition 4.2.13. *Let $\alpha \in (0,1)$ and let $\mathcal{F}$ be a class of continuously differentiable probability distribution functions F on $\mathbb{R}$ with densities $f = F'$ having a unique α-quantile $q_\alpha(F)$. Let $T = q_\alpha \colon \mathcal{F} \to \mathsf{A} = (a,b)$ where $-\infty \le a < b \le \infty$. Assume that T is surjective on (a,b) and that the densities are uniformly bounded from below in (a,b) in the sense that there is an $\varepsilon \ge 0$ such that for all $F \in \mathcal{F}$ with density $f = F'$ it holds that $\inf_{x\in(a,b)} f(x) \ge \varepsilon$. Then, any scoring function $S\colon (a,b) \times \mathsf{O} \to \mathbb{R}$ of the form*

$$S(x,y) = \int_z^x h(w)(\mathbb{1}\{y \le w\} - \alpha)\,\mathrm{d}w, \tag{4.2.10}$$

where $z \in (a,b)$, $h\colon (a,b) \to \mathbb{R}$ is positive, differentiable and satisfies

$$|h'(x)| \le \frac{\varepsilon\, h(x)}{\max(\alpha, 1-\alpha)} \qquad \forall x \in (a,b) \tag{4.2.11}$$

is an $\mathcal{F}$-convex strictly $\mathcal{F}$-consistent scoring function for T.

Proof. Let $F \in \mathcal{F}$ with density $f = F'$. If S is of the form given at (4.2.10) such that (4.2.11) is satisfied, the second derivative of the expected score $\bar{S}(x,F)$, $x \in (a,b)$, takes the form

$$h'(x)(F(x) - \alpha) + h(x)f(x). \tag{4.2.12}$$

Recall that the second summand $h(x)f(x)$ is non-negative while $h'(x)$ can have both signs and $F(x)-\alpha$ definitely has both signs depending on whether $x < q_\alpha(F)$ or $x > q_\alpha(F)$. However, under (4.2.11) one obtains that

$$0 \le |h'(x)(F(x) - \alpha)| \le |h'(x)| \max(\alpha, 1-\alpha) \le h(x)\varepsilon \le h(x)f(x).$$

Consequently, (4.2.12) is non-negative and the the expected score $\bar{S}(\cdot, F)$ is $\mathcal{F}$-convex. □

Example 4.2.14. There is a number of possible candidates for h satisfying (4.2.11). For example, one can take

$$h(x) = c\exp(\delta x)$$

with $c > 0$ and $\delta \in [-\varepsilon, \varepsilon]$.

Remark 4.2.15. Obviously, the ε appearing in Proposition 4.2.13 can only be strictly positive if A is bounded, that is, if $a, b \in \mathbb{R}$. Otherwise, $\varepsilon = 0$ and the only scoring functions of the form (4.2.10) satisfying (4.2.11) are equivalent to the α-pinball loss.

Ratios of expectations

In Corollary 4.2.17, we retrieve the result of Caponnetto (2005) that the quadratic loss (up to equivalence) is the only convex strictly consistent scoring function for the mean when the class of distributions is sufficiently rich in the sense that it is convex and the mean functional is surjective on $\mathbb{R}$. However, up to our knowledge, it has not been studied yet if (a) a similar statement is true for ratios of expectations, and (b) if other strictly consistent scoring functions are convex if the mean (and more generally, a ratio of expectations) is not surjective on $\mathbb{R}$.

We first consider the one-dimensional case $k = 1$. Let $p, q\colon \mathsf{O} \to \mathbb{R}$ and define the functional $T\colon \mathcal{F} \to \mathsf{A} \subseteq \mathbb{R}$, $T(F) = \mathbb{E}_F[p(Y)]/\mathbb{E}_F[q(Y)] = \bar{p}(F)/\bar{q}(F)$. Throughout this section, without loss of generality, we shall assume that $\bar{q}(F) > 0$ for all $F \in \mathcal{F}$; see Remark 4.1.23. In the following Proposition, we will work under the assumptions of Fissler and Ziegel (2016, Proposition 4.4(i)). Note that the mixture-continuity of T implies that $\mathrm{int}(T(\mathcal{F})) = (a, b)$ for some $-\infty \le a \le b \le \infty$.[14]

Proposition 4.2.16. *Let the assumptions of Fissler and Ziegel (2016, Proposition 4.4 (i)) hold with $k = 1$ such that $\mathrm{int}(T(\mathcal{F})) = \mathrm{int}(\mathsf{A}) = (a, b)$ for some $-\infty \le a \le b \le \infty$. Suppose additionally that $\bar{S}(\cdot, F)\colon (a, b) \to \mathbb{R}$ is C^2 for all $F \in \mathcal{F}$. Then $h\colon (a, b) \to \mathbb{R}$ is C^1 and $\bar{S}(\cdot, F) : (a, b) \to \mathbb{R}$ is convex for all $F \in \mathcal{F}$ if and only if*

$$h(x) + h'(x)\big(x - a\mathbb{1}\{h'(x) < 0\} - b\mathbb{1}\{h'(x) > 0\}\big) = w(x) \qquad \forall x \in (a, b) \quad (4.2.13)$$

for some non-negative C^0-function $w\colon (a, b) \to \mathbb{R}$. Moreover, if w is strictly positive, then $\bar{S}(\cdot, F) : (a, b) \to \mathbb{R}$ is strictly convex for all $F \in \mathcal{F}$.

Proof. Fix some $F \in \mathcal{F}$. Due to Theorem 3.2 and Proposition 4.4 in Fissler and Ziegel (2016) there is a positive function $h\colon (a, b) \to \mathbb{R}$ such that $\frac{\partial}{\partial x}\bar{S}(x, F) = h(x)(x\bar{q}(F) - \bar{p}(F))$. The expected score $\bar{S}(x, F)$ is convex in x if and only if $\frac{\partial^2}{(\partial x)^2}\bar{S}(x, F) \ge 0$ for all $x \in (a, b)$ and this in turn is equivalent to

$$h(x) + h'(x)\,(x - T(F)) \ge 0 \qquad \forall x \in (a, b). \qquad (4.2.14)$$

The inequality at (4.2.14) holds for all $F \in \mathcal{F}$ if and only if

$$h(x) + h'(x)\big(x - a\mathbb{1}\{h'(x) < 0\} - b\mathbb{1}\{h'(x) > 0\}\big) \ge 0 \qquad \forall x \in (a, b)$$

with the convention $0 \cdot \infty = 0 \cdot (-\infty) = 0$. □

[14] Note that, actually, one does not have to impose Assumption (V3) in Proposition 4.4 in Fissler and Ziegel (2016). The standard identification function $V(x, y) = q(y)x - p(y)$ is a polynomial in x, so in particular, it is smooth.

It is easy to see that (4.2.13) implies that $h' \le 0$ if $b = \infty$ and $h' \ge 0$ if $a = -\infty$.

Corollary 4.2.17. *Let $T\colon \mathcal{F} \to \mathbb{R}$, $T(F) = \mathbb{E}_F[p(Y)]/\mathbb{E}_F[q(Y)]$ be a ratio of expectations and $\mathcal{F}$ convex and containing all point measures. If T is surjective, any convex strictly $\mathcal{F}$-consistent C^2-scoring function $S\colon \mathbb{R}\times\mathsf{O} \to \mathbb{R}$ for T is of equivalent form as*

$$S_0(x,y) = \frac{1}{2}x^2 q(y) - xp(y). \tag{4.2.15}$$

Proof. Let $S\colon \mathbb{R}\times\mathsf{O} \to \mathbb{R}$ be a convex and strictly $\mathcal{F}$-consistent C^2-scoring function for T. Due to Osband's principle (Fissler and Ziegel, 2016, Theorem 3.2), there is a C^1-function $h\colon \mathbb{R} \to \mathbb{R}$ such that

$$\frac{\mathrm{d}}{\mathrm{d}x}S(x,y) = h(x)(xq(y) - p(y)) \qquad \forall (x,y) \in \mathbb{R}\times\mathbb{R}.$$

With respect to Proposition 4.2.16, h must be constant such that S is of equivalent form as S_0 at (4.2.15). $\square$

This means in particular that the quadratic score is – up to equivalence – the only convex strictly consistent scoring function for the mean functional provided that the mean functional is surjective on $\mathbb{R}$ and the class of distributions is sufficiently rich.

Given some non-negative C^0-function $w\colon (a,b) \to \mathbb{R}$ we are interested in C^1-solutions h for (4.2.13). We remark that due to the presence of the indicators, the operator mapping h to the left hand side of (4.2.13) is generally non-additive (but it is additive if $h' \ge 0$ or $h' \le 0$). Let $h \in C^1((a,b))$ and $h > 0$. Then the continuity of h' yields that

$$\begin{aligned}
\mathsf{A}_- &:= \{x \in (a,b)\colon h'(x) < 0\} = \bigcup_{i\in\mathcal{I}} (a_i, b_i),\\
\mathsf{A}_+ &:= \{x \in (a,b)\colon h'(x) > 0\} = \bigcup_{j\in\mathcal{J}} (a_j, b_j),\\
\mathsf{A}_0 &:= \{x \in (a,b)\colon h'(x) = 0\}
\end{aligned}$$

for some index sets $\mathcal{I}, \mathcal{J}$ and for $a \le a_i < b_i \le b$, $a \le a_j < b_j \le b$. For notational convenience, we will assume that $\mathcal{I}$ and $\mathcal{J}$ are disjoint. To ensure a unique representation, we assume that $\mathcal{I}$ and $\mathcal{J}$ are minimal implying that for all $i \in \mathcal{I}, j \in \mathcal{J}$ and $a_i, a_j \ne a$, $b_i, b_j \ne b$ one has that $a_i, a_j, b_i, b_j \in \mathsf{A}_0$.[15] Recall that with respect to (4.2.13), if $a = -\infty$ then $\mathcal{I} = \emptyset$, and if $b = \infty$ then $\mathcal{J} = \emptyset$. For $i \in \mathcal{I}$, $j \in \mathcal{J}$ respectively, (4.2.13) is satisfied on (a_i, b_i), (a_j, b_j) respectively, if and only if

$$h(x) = \frac{W_i(x)}{x-a} \quad x \in (a_i, b_i), \qquad \text{resp. } h(x) = \frac{W_j(x)}{x-b} \quad x \in (a_j, b_j) \tag{4.2.16}$$

[15] Actually, due to a compactness argument, $\mathcal{I}$ and $\mathcal{J}$ are then finite if $a, b \in \mathbb{R}$.

for some antiderivatives W_i, W_j of w. Since $w \ge 0$, W_i, W_j are increasing. The fact that $h > 0$ implies the following: If $\mathcal{I} \neq \emptyset$, define $a' := \inf_{i\in\mathcal{I}} a_i > -\infty$. Then, for any sequence $(i(n))_{n\ge1} \subset \mathcal{I}$ such that $a_{i(n)} \to a'$ and for any sequence $(x_n)_{n\ge1}$ such that $x_n \in (a_{i(n)}, b_{i(n)})$ and $x_n \downarrow a'$, the limit $\lim_{n\to\infty} W_{i(n)}(x_n) =: W(a')$ must exist and must be greater or equal to 0. Analogously, if $\mathcal{J} \neq \emptyset$, set $b' := \sup_{j\in\mathcal{J}} b_j < \infty$. Then, for any sequence $(j(n))_{n\ge1} \subset \mathcal{J}$ such that $b_{j(n)} \to b'$ and for any sequence $(x_n)_{n\ge1}$ such that $x_n \in (a_{j(n)}, b_{j(n)})$ and $x_n \uparrow b'$, the limit $\lim_{n\to\infty} W_{j(n)}(x_n) =: W(b')$ must exist and must be less or equal to 0. For the sake of uniqueness, let W_0 be a certain antiderivative of w chosen due to some (not specified) normalization condition. Then we shall write $W_i(x) = W_0(x) + c_i$ and $W_j(x) = W_0(x)+c_j$ with constants $c_i, c_j \in \mathbb{R}$. In summary, the sign condition $h > 0$ yields that

$$W_0(a_i) + c_i \ge 0, \quad W_0(b_j) + c_j \le 0 \quad \forall i \in \mathcal{I}, j \in \mathcal{J}. \tag{4.2.17}$$

Moreover, for all $i \in \mathcal{I}$, the fact that $h'(x) < 0$ for $x \in (a_i, b_i)$ implies that

$$0 < (W_0(x) + c_i) + (a - x)w(x) \qquad x \in (a_i, b_i). \tag{4.2.18}$$

Analogously for all $j \in \mathcal{J}$, the fact that $h'(x) > 0$ for $x \in (a_j, b_j)$ implies that

$$0 < (-W_0(x) - c_j) + (x - b)w(x) \qquad x \in (a_j, b_j). \tag{4.2.19}$$

Condition (4.2.17) plus the fact that W_0 is increasing yields that the first summand in brackets is non-negative in (4.2.18) and (4.2.19), and the fact that $w \ge 0$ implies that the second summand is non-positive, respectively. Moreover, it is important to notice that on any open subset of A_0, h can be of both forms at (4.2.16) (of course, in these open sets w is necessarily constant). In the border cases, meaning that $x_0 \in \{a_i : i \in \mathcal{I}\} \cup \{a_j : j \in \mathcal{J}\} \cup \{b_i : i \in \mathcal{I}\} \cup \{b_j : j \in \mathcal{J}\} \subset \mathsf{A}_0$ one has for $x_0 \in \{a_i : i \in \mathcal{I}\} \cup \{b_i : i \in \mathcal{I}\} \subset \mathsf{A}_0$

$$c_i = -W_0(x_0) + (x_0 - a)w(x_0),$$

such that

$$h(x_0) = \frac{W_0(x_0) + c_i}{x_0 - a} = w(x_0).$$

And on the other hand, for $x_0 \in \{a_j : j \in \mathcal{J}\} \cup \{b_j : j \in \mathcal{J}\} \subset \mathsf{A}_0$

$$c_j = -W_0(x_0) + (x_0 - b)w(x_0),$$

such that

$$h(x_0) = \frac{W_0(x_0) + c_j}{x_0 - b} = w(x_0).$$

This shows that it is in principle possible to construct a global C^1-solution h on (a, b) satisfying (4.2.16), (4.2.17), (4.2.18), and (4.2.19).

Example 4.2.18. Let $a = 0$, $b = 1$ and $w(x) = \min(x, 1-x)$. Then a solution of (4.2.13) is

$$h(x) = \begin{cases} \frac{1}{x}\left(\frac{1}{8} + \frac{x^2}{2}\right) = \frac{1}{8x} + \frac{x}{2}, & \text{if } x \in [0, 1/2] \\ \frac{1}{1-x}\left(\frac{1}{8} + \frac{(1-x)^2}{2}\right) = \frac{1}{8-8x} + \frac{1-x}{2}, & \text{if } x \in (1/2, 1]. \end{cases}$$

Remark 4.2.19. If $w \equiv 0$ and $a, b \in \mathbb{R}$, then a solution of (4.2.13) with $h > 0$ is either globally of the form $h(x) = c_i/(x-a)$, $c_i > 0$ or $h(x) = c_j/(x-b)$, $c_j < 0$.

Example 4.2.20. Let $a = 0$, $b = \infty$ and consider the case of monomials for w, that is $w(x) = x^\alpha$ for $x > 0$ where $\alpha \in \mathbb{R}$. So we are in the situation where $h' \le 0$ and $h(x) = (W_0(x) + c)/x$ for some antiderivative W_0 of w. Since $\lim_{x \downarrow 0} W_0(x)$ must be finite, necessarily $\alpha > -1$. And due to condition (4.2.18), necessarily $\alpha \le 0$. Finally, integration yields that S is of equivalent form as $S'_\alpha(x, y) = S_\alpha(x, y) + c\, S_{\log}(x, y)$ where

$$S_\alpha(x, y) = \frac{1}{\alpha + 2} x^{\alpha+2} q(y) - \frac{1}{\alpha + 1} x^{\alpha+1} p(y) \qquad \alpha \in (-1, 0], \tag{4.2.20}$$

$$S_{\log}(x, y) = x q(y) - p(y) \log(x). \tag{4.2.21}$$

Remark 4.2.21. It is worth mentioning that for the mean-functional, meaning $q \equiv 1$ and $p(y) = y$ with the action domain $\mathsf{A} = (0, \infty)$, the scoring function S_α, $\alpha \in (-1, 0]$, given at (4.2.20) is positively homogeneous of degree $\alpha + 2$. Moreover, for $\mathsf{O} = \mathsf{A} = (0, \infty)$ (which is a reasonable assumption in the case of the mean functional), $S_{\log}$ given at (4.2.21) is of equivalent form as

$$S'_{-1}(x, y) = x - y \log\left(\frac{x}{y}\right),$$

which is positively homogeneous of degree 1.[16] All those scoring functions mentioned above can be found (in an equivalent form) in Patton (2011, Proposition 4). However, note that not all positively homogeneous functions given in that Proposition are convex (only those with $b \in [-1, 0]$).

Remark 4.2.22. We see that for the mean functional and the case of $\mathsf{A} = \mathbb{R}$, the only convex scoring function and the only metrically order-sensitive scoring function coincide (where 'only' should be understood in the 'up-to-equivalence'-sense). However, for the second moment, a natural assumption is $\mathsf{A} = [0, \infty)$ or $\mathsf{A} = (0, \infty)$ and there are indeed more convex scoring functions than the only metrically order-sensitive scoring function given at (4.1.19).

[16] If $\mathsf{O} \cap (-\infty, 0] \neq \emptyset$, then for $S(x, y) = x - y \log(x)$ still the score *difference* $(0, \infty) \times (0, \infty) \times \mathbb{R} \to \mathbb{R}$, $(x, x', y) \mapsto S(x, y) - S(x', y)$ is homogeneous of degree 1.

Expectiles

As it is quite often the case, expectiles share common features both with quantiles and with ratios of expectations (and more precisely, with expectations). It is obvious that the canonical scoring function for the τ-expectile, that is, the asymmetric quadratic loss

$$S_\tau(x,y) = |\mathbb{1}\{y \le x\} - \tau|(x-y)^2 \tag{4.2.22}$$

is a convex scoring function. We shall show that under some richness assumptions on the class of distributions $\mathcal{F}$ which imply that the image of the τ-expectile corresponds to whole $\mathbb{R}$, the asymmetric piecewise loss S_τ is the only $\mathcal{F}$-convex scoring function that is strictly $\mathcal{F}$-consistent for the τ-expectile, up to equivalence. On the other hand, a smaller class of $\mathcal{F}$ can result in more flexibility in the same spirit as Proposition 4.2.16. However, the analysis is more involved compared to the case of ratios of expectations.

We begin with the following lemma (which also holds for the case $\tau = 1/2$). Recall the standard identification function $V_\tau \colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ for the τ-expectile, defined as

$$V_\tau(x,y) = 2|\mathbb{1}\{y \le x\} - \tau|(x-y). \tag{4.2.23}$$

Lemma 4.2.23. *Let $\tau \in (0,1)$ and let $\mathcal{F}$ be a class of distribution functions on $\mathbb{R}$ with finite first moments and let $T = \mu_\tau \colon \mathcal{F} \to \mathbb{R}$ be the τ-expectile. If there is an $F_0 \in \mathcal{F}$ such that $F_0(\cdot - \lambda) \in \mathcal{F}$ for all $\lambda \in \mathbb{R}$, then $T(\mathcal{F}) = \{T(F) \colon F \in \mathcal{F}\} = \mathbb{R}$ and for each $x \in \mathbb{R}$*

$$\bar{V}(x, \mathcal{F}) := \{\bar{V}(x,F) \colon F \in \mathcal{F}\} = \mathbb{R}.$$

Proof. Note that the τ-expectile is translation equivariant. Consequently, $T(F(\cdot - \lambda)) = T(F) + \lambda$. Hence, $T(\mathcal{F}) = \mathbb{R}$. Moreover, for fixed $x \in \mathbb{R}$ the function $y \mapsto V_\tau(x,y)$ is piecewise linear with strictly negative slope. Hence, the second claim follows. □

For the next proposition, we assume that the strictly consistent scoring function is of the form given via Propositions 3.4 and 4.2 in Fissler and Ziegel (2016).

Proposition 4.2.24. *Let $\tau \in (0,1)$ and $\mathcal{F}$ be a class of differentiable distribution functions on $\mathbb{R}$ with finite mean. Let $T = \mu_\tau \colon \mathcal{F} \to \mathsf{A} \subseteq \mathbb{R}$ be the τ-expectile, and assume that T is surjective and A is an open interval. Define the functions $a, b \colon \mathsf{A} \to \mathbb{R} \cup \{-\infty, \infty\}$*

$$a(x) := \inf\{\bar{V}(x,F) \colon F \in \mathcal{F}\} \in \mathbb{R} \cup \{-\infty\}, \tag{4.2.24}$$
$$b(x) := \sup\{\bar{V}(x,F) \colon F \in \mathcal{F}\} \in \mathbb{R} \cup \{\infty\}. \tag{4.2.25}$$

Define the strictly $\mathcal{F}$-consistent scoring function $S \colon \mathsf{A} \times \mathbb{R} \to \mathbb{R}$ via

$$S(x,y) = \int_z^x h(w) V(w,y) \, \mathrm{d}w, \tag{4.2.26}$$

where $z \in \mathsf{A}$, $h \colon \mathsf{A} \to \mathbb{R}$ is positive and differentiable. Then, the following two assertions hold.

(i) If S is $\mathcal{F}$-convex, then

$$h(x)\max\{2\tau, 2-2\tau\} + h'(x)\big(\mathbb{1}\{h'(x)<0\}b(x) + \mathbb{1}\{h'(x)>0\}a(x)\big) \ge 0$$

for all $x \in \mathsf{A}$.

(ii) If

$$h(x)\min\{2\tau, 2-2\tau\} + h'(x)\big(\mathbb{1}\{h'(x)<0\}b(x) + \mathbb{1}\{h'(x)>0\}a(x)\big) \ge 0 \tag{4.2.27}$$

for all $x \in \mathsf{A}$, then S is $\mathcal{F}$-convex.

(iii) If $a(x) = -\infty$ and $b(x) = \infty$ for all $x \in \mathsf{A}$, then S is $\mathcal{F}$-convex if and only if h is constant.

Proof. Let S be of the form (4.2.26). For some $F \in \mathcal{F}$, the second derivative of the expected score $\bar{S}(x, F)$ is

$$\frac{\partial^2}{(\partial x)^2}\bar{S}(x,F) = h(x)\frac{\partial}{\partial x}\bar{V}(x,F) + h'(x)\bar{V}(x,F).$$

Now,

$$\frac{\partial}{\partial x}\bar{V}(x,F) = 2F(x)(1-2\tau) + 2\tau \in \operatorname{conv}(\{2\tau, 2-2\tau\})$$

for all $x \in \mathsf{A}$ and for all $F \in \mathcal{F}$. S is $\mathcal{F}$-convex if and only if for all $x \in \mathsf{A}$ and $F \in \mathcal{F}$,

$$\frac{\partial^2}{(\partial x)^2}\bar{S}(x,F) \ge 0.$$

Observe that

$$\inf_{G\in\mathcal{F}} h'(x)\bar{V}(x,G) = h'(x)\big(\mathbb{1}\{h'(x)<0\}b(x) + \mathbb{1}\{h'(x)>0\}a(x)\big).$$

Moreover, for any $x \in \mathsf{A}$ and any $F \in \mathcal{F}$

$$\begin{aligned}
&h(x)\max\{2\tau, 2-2\tau\} + h'(x)\bar{V}(x,F)\\
&\ge \frac{\partial^2}{(\partial x)^2}\bar{S}(x,F)\\
&\ge \inf_{G\in\mathcal{F}}\frac{\partial^2}{(\partial x)^2}\bar{S}(x,G)\\
&\ge \inf_{G\in\mathcal{F}} h(x)\frac{\partial}{\partial x}\bar{V}(x,G) + \inf_{G\in\mathcal{F}} h'(x)\bar{V}(x,G)\\
&\ge h(x)\min\{2\tau, 2-2\tau\} + h'(x)\big(\mathbb{1}\{h'(x)<0\}b(x) + \mathbb{1}\{h'(x)>0\}a(x)\big).
\end{aligned}$$

Hence, assertions (i) and (ii) follow.

For (iii), one can combine (i) and (ii) such that S is $\mathcal{F}$-convex if and only if $h \equiv 0$. Then the claim follows by the fact that A is an open interval. $\square$

One can see that under the assumptions of Lemma 4.2.23 and Proposition 4.2.24, part (iii) yields that S_τ defined at (4.2.22) is the only $\mathcal{F}$-convex scoring function that is strictly $\mathcal{F}$-consistent for the τ-expectile, of course, up to equivalence.

Remark 4.2.25. In practice, it might be somewhat involved to determine the functions $a(x), b(x)$ appearing at (4.2.24), (4.2.25). However, assuming that they are known and that they satisfy sufficient regularity conditions, one can, in principle, construct solutions to (4.2.27) in the same spirit as after Corollary 4.2.17 in the case of the ratio of expectations.

Higher-dimensional functionals: Ratios of expectations with the same denominator

From an abstract point of view, the characterization of the convexity of the (expected) scores is clear: Under sufficient regularity conditions, the (expected) score is convex if and only if its Hessian is positive semi-definite. In the case where a strictly consistent scoring function is the sum of strictly consistent scoring functions for each component – for example if the functional is a vector of quantiles and expectiles; see Fissler and Ziegel (2016, Proposition 4.2) – the convexity of the expected score is equivalent to the fact that all summands are convex. Confining our attention for a moment to functionals with elicitable components, the remaining relevant examples are ratios of expectations with the same denominator. Even though the computations are a bit more involved, the tenor of the one-dimensional setting remains the same when passing to the higher-dimensional case: if the image of the functional T is the whole space $\mathbb{R}^k$, a strictly consistent scoring function for T is convex if and only if the matrix valued function h from Osband's principle is constant (when working with the canonical identification function for the ratio of expectations with the same denominator). On the other hand, if the image of T is contained in the intersection of half-spaces (including the particular case that the image is bounded), then there is a bit more flexibility and there are also non-constant matrix-valued functions h which yield a convex score.

Proposition 4.2.26. *Let $T\colon \mathcal{F} \to \mathsf{A} \subseteq \mathbb{R}^k$, $T(F) = \bar{p}(F)/\bar{q}(F)$, be surjective and a ratio of expectations with the same (positive) denominator. Let the assumptions and the notation of Fissler and Ziegel (2016, Proposition 4.4 (i)) prevail. If $\mathsf{A} = \mathbb{R}^k$, then a strictly $\mathcal{F}$-consistent scoring function is $\mathcal{F}$-convex if and only if h is constant.*

Proof. Let S be of the form given at Fissler and Ziegel (2016, Proposition 4.4 (i)) and assume that it is $\mathcal{F}$-convex. Then, the Hessian of the expected score $\bar{S}(x, F)$, $x \in \operatorname{int}(\mathsf{A}), F \in \mathcal{F}$, is

$$\begin{aligned}\nabla^2 \bar{S}(x, F) &= \bar{q}(F)\big(h(x) + \operatorname{mat}(\partial_1 h(x)(x - T(F)), \dots, \partial_k h(x)(x - T(F)))\big) \\ &=: \bar{q}(F) A(x, F).\end{aligned}$$

Here, $\partial_m h$ denotes the $(k \times k)$-matrix consisting of all partial derivatives of the entries of h with respect to x_m. Moreover, for column-vectors $a_1, \dots, a_k \in \mathbb{R}^k$, $\mathrm{mat}(a_1, \dots, a_k)$ is the $(k \times k)$-matrix consisting of column-vectors $a_1, \dots, a_k$. A necessary criterion for the positive semi-definiteness of the Hessian is that the diagonal elements of $A(x, F)$ are non-negative for all $x \in \mathrm{int}(\mathsf{A})$ and all $F \in \mathcal{F}$. Let e_m, $m \in \{1, \dots, k\}$, be the mth canonical basis vector of $\mathbb{R}^k$. Then

$$\langle e_m, A(x, F)e_m \rangle = h_{mm}(x) + \langle \partial_m h_{m\cdot}(x), x - T(F) \rangle,$$

where $\partial_m h_{m\cdot}$ denotes the mth row-vector of $\partial_m h$. Since for all $x \in \mathbb{R}^k$

$$\{x - T(F) \colon F \in \mathcal{F}\} = \{T(F) \colon F \in \mathcal{F}\} = \mathbb{R}^k,$$

whence

$$\{\langle \partial_m h_{m\cdot}(x), x - T(F) \rangle \colon F \in \mathcal{F}\} = \begin{cases} \{0\}, & \text{if } \partial_m h_{m\cdot}(x) = 0 \\ \mathbb{R}, & \text{else.} \end{cases}$$

So the $\mathcal{F}$-convexity of S implies that $\partial_m h_{m\cdot} \equiv 0$. Now, let $m, n \in \{1, \dots, k\}$, $m \neq n$. Then

$$\begin{aligned} &\langle e_m + e_n, A(x, F)(e_m + e_n) \rangle \\ &\qquad = h_{mm}(x) + h_{nn}(x) + 2h_{mn}(x) + 2\langle \partial_m h_{n\cdot}(x), x - T(F) \rangle. \end{aligned}$$

With the same argument as before, one can conclude that $\partial_m h_{n\cdot} \equiv 0$. So we have shown that h is necessarily locally constant. Since $\mathsf{A} = \mathbb{R}^k$ is connected, h is constant. For sufficiency, recall that h is positive definite, so S is even strictly convex if h is constant. □

Let $\mathcal{F}' \subseteq \mathcal{F}$ such that $T(\mathcal{F}') \subseteq (a, b)$ for some $a, b \in \bar{\mathbb{R}}^k$ where $(a, b) = \prod_{m=1}^k (a_m, b_m)$. If $(a, b) \neq \mathbb{R}^k$, then there are $\mathcal{F}'$-convex strictly $\mathcal{F}$-consistent scoring functions for T such that the link function h from Osband's principle is non-constant. The proof of this fact is straight forward: one can consider the sum of strictly $\mathcal{F}$-consistent scoring functions S_m for each component T_m where S_m is constructed according to Proposition 4.2.16. For $\mathcal{F}'$-convex scoring functions with a non-diagonal and non-constant function h, the construction is considerably more involved, but in principle possible.

Remark 4.2.27. In case of the prominent examples for higher-dimensional functionals functionals consisting of elicitable components – ratios of expectations with the same denominator, vectors of quantiles and / or expectiles – it would be an interesting question if there are (strictly consistent) quasi-convex scoring functions besides the convex ones. Regretfully, this question seems to be more involved than one would think, and hence, it is deferred to future work. However, it is possible to answer that question in the case of strictly consistent scoring functions for the pair (Value at Risk, Expected Shortfall) which will be done in the subsequent paragraph.

(Quasi-)Convexity of strictly consistent scoring functions for (VaR, ES)

We consider again the class of strictly consistent scoring functions for the pair $(\mathrm{VaR}_\alpha, \mathrm{ES}_\alpha)$, $\alpha \in (0,1)$, given in Corollary 5.5 in Fissler and Ziegel (2016). As a first result, we give an example of a class of $\mathcal{F}$-quasi-convex scoring functions.

Proposition 4.2.28. *Let $\alpha \in (0,1)$. Let $\mathcal{F}$ be a class of distribution functions on $\mathbb{R}$ with finite first moments, unique α-quantiles, continuous densities, and $\mathrm{ES}_\alpha(F) < 0$ for all $F \in \mathcal{F}$. Let $\mathsf{A}_0^- = \{x \in \mathbb{R}^2 \colon x_1 \ge x_2, x_2 < 0\}$. Then, any scoring function $S\colon \mathsf{A}_0^- \times \mathbb{R} \to \mathbb{R}$ which is of equivalent form as*

$$S_0(x_1, x_2, y) = \frac{x_1}{x_2} - \frac{1}{\alpha}\mathbb{1}\{y \le x_1\}\frac{x_1 - y}{x_2} + \log(-x_2) \tag{4.2.28}$$

is strictly $\mathcal{F}$-consistent for $T = (\mathrm{VaR}_\alpha, \mathrm{ES}_\alpha)$ and $\mathcal{F}$-quasi-convex.

Moreover, if the elements of $\mathcal{F}$ have densities which are strictly positive at their α-quantiles, any scoring function equivalent to S_0 is strictly $\mathcal{F}$-quasi-convex.

Remark 4.2.29. Nolde and Ziegel (2016) showed that S_0 defined at (4.2.28) is such that the score difference $\mathsf{A}_0^- \times \mathsf{A}_0^- \times \mathbb{R} \to \mathbb{R}$, $(x, x', y) \mapsto S_0(x, y) - S_0(x', y)$ is positively homogeneous of degree 0.

Proof. First of all, note and recall that equivalence of scoring functions preserves strict consistency and quasi-convexity. So it is sufficient to consider the score S_0. This score is of the form given in Fissler and Ziegel (2016, Corollary 5.5) where $a \equiv -1$, $G_1 \equiv 0$ is increasing and $\mathcal{F}$-integrable, and $\mathcal{G}(x_2) = -\log(-x_2)$ is strictly increasing and strictly convex (on $(-\infty, 0)$). Hence, S_0 is strictly $\mathcal{F}$-consistent for T.

To check if S_0 is $\mathcal{F}$-quasi-convex, we check the criterion given in Proposition 4.2.3. However, to be more transparent, we perform the calculation with the general form of the scoring function given in Fissler and Ziegel (2016, Corollary 5.5). Let $F \in \mathcal{F}$ with density / derivative f, $v = (v_1, v_2)^\top \in \mathbb{S}^1$, $x = (x_1, x_2)^\top \in \mathsf{A}_0^-$ and $(v_1, v_2)\nabla \bar S_0(x, F) = 0$. Using (4.1.30), this is equivalent to

$$\begin{aligned} v_1(F(x_1) - \alpha)\left(G_1'(x_1) + \frac{1}{\alpha}G_2(x_2)\right) \\ = -v_2 G_2'(x_2)\left(x_2 - \frac{1}{\alpha}\int_{-\infty}^{x_1} y \,\mathrm{d}F(y) + \frac{1}{\alpha}x_1(F(x_1) - \alpha)\right). \end{aligned} \tag{4.2.29}$$

Using this identity, one obtains

$$\begin{aligned} (v_1, v_2)\nabla^2 \bar S_0(x_1, x_2, F)\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} &= v_1^2 f(x_1)\left(G_1'(x_1) + \frac{1}{\alpha}G_2(x_2)\right) + v_2^2 G_2'(x_2) \\ &+ v_1(F(x_1) - \alpha)\left(v_1 G_1''(x_1) + v_2\frac{2}{\alpha}G_2'(x_2) - v_2 G_2''(x_2)\frac{G_1'(x_1) + \frac{1}{\alpha}G_2(x_2)}{G_2'(x_2)}\right). \end{aligned} \tag{4.2.30}$$

Due to the special choice of G_1 and G_2, the third summand vanishes. In particular, one obtains

$$(v_1, v_2)\nabla^2 \bar{S}_0(x_1, x_2, F) \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = -\frac{v_1^2 f(x_1)}{\alpha x_2} + \frac{v_2^2}{x_2^2},$$

which is strictly positive for $v_2 \neq 0$. For $v_2 = 0$, the function $\psi(s) = \bar{S}(x_1+s, x_2, F)$ has the second derivative $\psi''(s) = -\frac{f(x_1+s)}{\alpha x_2} \geq 0$. So ψ is convex and in particular quasi-convex and the claim follows with Proposition 4.2.3.

Moreover, observe that (4.2.29) can hold for $v_2 = 0$ if and only if x_1 equals the α-quantile of F. So then $\psi''(0) = -\frac{f(x_1)}{\alpha x_2}$ is strictly positive if the density is strictly positive at the α-quantile. Hence, the claim follows again by Proposition 4.2.3. □

Remark 4.2.30. The similarity between the proof of Proposition 4.2.28 and the determination of scoring functions which are order-sensitive on line segments around Example 4.1.50 is obvious. In particular, the terms at (4.1.32) and (4.2.30) are almost the same, and in both situations, one has to control its sign and show that it is not negative. However, the slight difference is that for the first situation at equation (4.1.32), one knows that $v_1(F(\bar{s}^*) - \alpha) < 0$, whereas one cannot control the sign of the corresponding expression at (4.2.30). That is also the very reason why there is more flexibility for the choice of G_1 when one wants to merely arrive at order-sensitive scoring functions on line segments. One only needs to guarantee that the bracket in the third summand is not positive, whereas the corresponding term needs to vanish in the situation of Proposition 4.2.28.

As a second result, we show that there is no $\mathcal{F}$-convex strictly $\mathcal{F}$-consistent scoring function for $(\mathrm{VaR}_\alpha, \mathrm{ES}_\alpha)$ if the class $\mathcal{F}$ is sufficiently rich. In particular, we make the following assumption.

Assumption (A1). Let $\alpha \in (0, 1)$. Let $\mathcal{F}$ be a class of distribution functions on $\mathbb{R}$ which are continuously differentiable, have unique α-quantiles and finite first moments. Assume moreover that for every $x \in \mathbb{R}$ there exist two distributions $F_1, F_2 \in \mathcal{F}$ with densities / derivatives f_1, f_2 such that

$$F_1(x) < \alpha, \qquad F_2(x) > \alpha, \qquad f_1(x) = f_2(x) = 0.$$

Proposition 4.2.31. *Let $\alpha \in (0, 1)$ and $\mathcal{F}$ be a class of distributions satisfying Assumption (A1). Then, any strictly $\mathcal{F}$-consistent scoring function for $T = (\mathrm{VaR}_\alpha, \mathrm{ES}_\alpha)$ of the form given in Corollary 5.5 in Fissler and Ziegel (2016) fails to be $\mathcal{F}$-convex.*

Proof. Assume that S is of the form given in Corollary 5.5 in Fissler and Ziegel (2016) (which can also be found at (4.1.29)). Assume that S is $\mathcal{F}$-convex. Then necessarily $\partial_1\partial_1\bar{S}(x_1, x_2, F) \geq 0$ and $\partial_2\partial_2\bar{S}(x_1, x_2, F) \geq 0$ for all $(x_1, x_2) \in \mathsf{A}$ and for all $F \in \mathcal{F}$. From the form given at (4.1.30) and using Assumption (A1), one concludes that $G_1'' \equiv 0$. Consider the determinant of the Hessian of $\bar{S}(x_1, x_2, F)$, that is,

$$f(x_1)\left(G_1'(x_1) + \frac{1}{\alpha}G_2(x_2)\right)\partial_2\partial_2\bar{S}(x_1, x_2, F) - \left(\frac{1}{\alpha}G_2'(x_2)(F(x_1) - \alpha)\right)^2.$$

Due to Assumption (A1) and the fact that $G_2' > 0$ there is an element in $\mathcal{F}$ such that the determinant of the Hessian becomes negative which is a contradiction to the $\mathcal{F}$-convexity of S. □

Remark 4.2.32. One can pose the question how restrictive Assumption (A1) actually is. The idea how to ensure that this assumption is valid is that if for a distribution F and some $x \in \mathbb{R}$, $F(x) < \alpha$ and $f(x) > 0$, it is possible to change the density only locally around x such that the resulting density vanishes at x, but for the resulting distribution $\tilde{F}$, still $\tilde{F}(x) < \alpha$.

For other higher-dimensional functionals consisting of at least one component which is not elicitable, it is interesting how (quasi-)convexity behaves under the revelation principle. The following result shows that under a linear bijection, (quasi-)convexity is preserved.

Lemma 4.2.33 ((Quasi-)Convexity under the revelation principle)**.** *If $S\colon \mathsf{A}\times\mathsf{O} \to \mathbb{R}$ is an $\mathcal{F}$(-quasi)-convex and strictly $\mathcal{F}$-consistent scoring function for some functional $T\colon \mathcal{F} \to \mathsf{A} \subseteq \mathbb{R}^k$ and g is a* linear *bijection from A to A', then $S_g(x', y) = S(g^{-1}(x'), y)$ is an $\mathcal{F}$(-quasi)-convex strictly $\mathcal{F}$-consistent scoring function for $T_g = g \circ T$.*

Proof. Let $F \in \mathcal{F}$, $x_1', x_2' \in \mathsf{A}'$ and $\lambda \in [0, 1]$. Due to the linearity of g^{-1} and the $\mathcal{F}$-convexity of S, one obtains

$$\begin{aligned}\bar{S}_g((1-\lambda)x_1' + \lambda x_2', F) &= \bar{S}((1-\lambda)g^{-1}(x_1') + \lambda g^{-1}(x_2'), F)\\ &\le (1-\lambda)\bar{S}(g^{-1}(x_1'), F) + \lambda\bar{S}(g^{-1}(x_2'), F)\\ &= (1-\lambda)\bar{S}_g(x_1', F) + \lambda\bar{S}_g(x_2', F).\end{aligned}$$

If S is $\mathcal{F}$-quasi-convex, one obtains

$$\begin{aligned}\bar{S}_g((1-\lambda)x_1' + \lambda x_2', F) &= \bar{S}((1-\lambda)g^{-1}(x_1') + \lambda g^{-1}(x_2'), F)\\ &\le \max\{\bar{S}(g^{-1}(x_1'), F), \bar{S}(g^{-1}(x_2'), F)\}\\ &= \max\{\bar{S}_g(x_1', F), \bar{S}_g(x_2', F)\}.\end{aligned}$$

This proves the claim. □

Final remark

So far, given a functional $T\colon \mathcal{F} \to \mathsf{A}$, we have strived to determine strictly $\mathcal{F}$-consistent and $\mathcal{F}$-(quasi-)convex scoring functions for T. We have seen many examples where the richness of $\mathcal{F}$ imposes harsh constraints on the class of $\mathcal{F}$-(quasi-)convex scoring functions. However, in most situations described above, there are other scoring functions that are $\mathcal{F}'$-convex for a subclass $\mathcal{F}' \subset \mathcal{F}$. However, they remain at least strictly $\mathcal{F}$-consistent for whole $\mathcal{F}$. So one can also consider to use them if one cannot exclude distributions in $\mathcal{F} \setminus \mathcal{F}'$ *a priori*, but nevertheless, the main focus lies on distributions in $\mathcal{F}'$.

4.3. Equivariant functionals and order-preserving scoring functions

Many statistical functionals have an invariance or equivariance property. For example, the mean is a linear functional, and hence, it is equivariant under linear transformations. So $\mathbb{E}[\varphi(X)] = \varphi(\mathbb{E}[X])$ for any random variable X and any linear map $\varphi\colon \mathbb{R} \to \mathbb{R}$ (of course, the same is true for the higher-dimensional setting). On the other hand, the variance is invariant under translations, that is $\mathrm{Var}(X - c) = \mathrm{Var}(X)$ for any $c \in \mathbb{R}$, but scales quadratically, so $\mathrm{Var}(\lambda X) = \lambda^2 \mathrm{Var}(X)$ for any $\lambda \in \mathbb{R}$. The next definition strives to formalize such notions.

Definition 4.3.1 (π-equivariance)**.** Let $\mathcal{F}$ be a class of probability distributions on O and A be an action domain. Let Φ be a group of bijective transformations $\varphi\colon \mathsf{O} \to \mathsf{O}$, Φ^* another group of bijective transformations $\varphi^*\colon \mathsf{A} \to \mathsf{A}$, and $\pi\colon \Phi \to \Phi^*$ be a map. A functional $T\colon \mathcal{F} \to \mathsf{A}$ is *π-equivariant* if for all $\varphi \in \Phi$

$$T(\mathcal{L}(\varphi(Y))) = (\pi\varphi)(T(\mathcal{L}(Y)))$$

for all Y such that $\mathcal{L}(Y) \in \mathcal{F}$.

From now on, we tacitly assume that Φ be a group of bijective transformations on O and Φ^* be a group of bijective transformations on A. We illustrate the notion of π-equivariance with some examples.

Example 4.3.2. (i) For $\mathsf{A} = \mathsf{O} = \mathbb{R}$, the mean functional is π-equivariant for $\Phi = \Phi^* = \{x \mapsto x + c, c \in \mathbb{R}\}$ the translation group and π the identity map, or for $\Phi = \Phi^* = \{x \mapsto \lambda x, \lambda \in \mathbb{R} \setminus \{0\}\}$ the multiplicative group and again π the identity map.

(ii) For $\mathsf{A} = \mathsf{O} = \mathbb{R}$, Value at Risk at level α, Expected Shortfall at level α and the τ-expectile are π-equivariant for $\Phi = \Phi^* = \{x \mapsto x + c, c \in \mathbb{R}\}$ the translation group and π the identity map, or for $\Phi = \Phi^* = \{x \mapsto \lambda x, \lambda > 0\}$ the multiplicative group and again π the identity map.

(iii) For $\mathsf{A} = [0, \infty)$ and $\mathsf{O} = \mathbb{R}$, the variance is π-equivariant for $\Phi = \{x \mapsto x+c, c \in \mathbb{R}\}$ the translation group and $\Phi^* = \{\mathrm{id}_{\mathsf{A}}\}$ the trivial group consisting of the identity on A, such that π is the constant map.

(iv) For $\mathsf{A} = [0, \infty)$ and $\mathsf{O} = \mathbb{R}$, the variance is π-equivariant for $\Phi = \Phi^* = \{x \mapsto \lambda x, \lambda \in \mathbb{R} \setminus \{0\}\}$ the multiplicative group, and $\pi((x \mapsto \lambda x)) = (x \mapsto \lambda^2 x)$.

(v) Let $\mathsf{A} = \mathbb{R}^k$, $\mathsf{O} = \mathbb{R}$ and T be the functional whose mth component is the mth moment. Then T is π-equivariant with $\Phi = \{y \mapsto \lambda y, \lambda \in \mathbb{R} \setminus \{0\}\}$, $\Phi^* = \{x \mapsto (\lambda^m x_m)_{m=1}^k, \lambda \in \mathbb{R}\setminus\{0\}\}$, and $\pi((y \mapsto \lambda y)) = (x \mapsto (\lambda^m x_m)_{m=1}^k)$.

If a functional T is elicitable, π-equivariance can also be expressed in terms of strictly consistent scoring functions; see also Gneiting (2011, p. 750).

Lemma 4.3.3. *Let $S\colon \mathsf{A} \times \mathsf{O} \to \mathbb{R}$ be a strictly $\mathcal{F}$-consistent scoring function for a functional $T\colon \mathcal{F} \to \mathsf{A}$ and let $\pi\colon \Phi \to \Phi^*$. Then, T is π-equivariant if and only if for all $\varphi \in \Phi$*

$$\operatorname*{arg\,min}_{x\in\mathsf{A}} \bar{S}(x, \mathcal{L}(\varphi(Y))) = (\pi\varphi)\big(\operatorname*{arg\,min}_{x\in\mathsf{A}} \bar{S}(x, \mathcal{L}(Y))\big) \tag{4.3.1}$$

for all Y such that $\mathcal{L}(Y) \in \mathcal{F}$.

The proof of Lemma 4.3.3 is direct. Note that (4.3.1) is equivalent to the assertion that for all $\varphi \in \Phi$

$$\operatorname*{arg\,min}_{x\in\mathsf{A}} \bar{S}((\pi\varphi)(x), \mathcal{L}(\varphi(Y))) = \operatorname*{arg\,min}_{x\in\mathsf{A}} \bar{S}(x, \mathcal{L}(Y)) \tag{4.3.2}$$

for all Y such that $\mathcal{L}(Y) \in \mathcal{F}$. This means that the scoring function

$$S_{\pi,\varphi}\colon \mathsf{A} \times \mathsf{O} \to \mathbb{R}, \quad (x, y) \mapsto S_{\pi,\varphi}(x, y) = S((\pi\varphi)(x), \varphi(y)) \tag{4.3.3}$$

is also strictly $\mathcal{F}$-consistent for T. The fact that the map $(S, \varphi) \mapsto S_{\pi,\varphi}$ preserves strict $\mathcal{F}$-consistency can be expressed as

$$\operatorname{sgn}\big(\bar{S}(x, F) - \bar{S}(x', F)\big) = \operatorname{sgn}\big(\bar{S}_{\pi,\varphi}(x, F) - \bar{S}_{\pi,\varphi}(x', F)\big)$$

for all $F \in \mathcal{F}$ and for all $x, x' \in \mathsf{A}$ whenever $x = T(F)$ or $x' = T(F)$. So the ranking of the correct forecast and any deliberately misspecified model is the same in terms of S and $S_{\pi,\varphi}$. In the same spirit as we have motivated order-sensitivity of scoring functions, for fixed $\pi\colon \Phi \to \Phi^*$, it is a natural requirement on a scoring function S that for all $\varphi \in \Phi$ the ranking of *any two* forecasts is the same in terms of S and in terms of $S_{\pi,\varphi}$. This is encoded in the following definition.

Definition 4.3.4 (π-order-preservingness)**.** Let $\pi\colon \Phi \to \Phi^*$. A scoring function $S\colon \mathsf{A} \times \mathsf{O} \to \mathbb{R}$ is *π-order-preserving* with respect to $\mathcal{F}$ if for all $\varphi \in \Phi$ one has

$$\operatorname{sgn}\big(\bar{S}(x, F) - \bar{S}(x', F)\big) = \operatorname{sgn}\big(\bar{S}_{\pi,\varphi}(x, F) - \bar{S}_{\pi,\varphi}(x', F)\big)$$

for all $F \in \mathcal{F}$ and for all $x, x' \in \mathsf{A}$, where $S_{\pi,\varphi}$ is defined at (4.3.3).

We give a pointwise criterion which implies that S is π-order-preserving.

Definition 4.3.5. Let $\pi\colon \Phi \to \Phi^*$. A scoring function $S\colon \mathsf{A} \times \mathsf{O} \to \mathbb{R}$ is *linearly π-order-preserving* if for all $\varphi \in \Phi$ there is a $\lambda > 0$ such that

$$\lambda\big(S(x,y) - S(x',y)\big) = S_{\pi,\varphi}(x,y) - S_{\pi,\varphi}(x',y) \tag{4.3.4}$$

for all $y \in \mathsf{O}$ and for all $x, x' \in \mathsf{A}$.

Lemma 4.3.6. *Let $\pi\colon \Phi \to \Phi^*$. If a scoring function $S\colon \mathsf{A} \times \mathsf{O} \to \mathbb{R}$ is linearly π-order-preserving, it is π-order-preserving with respect to any class $\mathcal{F}$ of probability distributions on O.*

Proof. The proof of Lemma 4.3.6 is standard. $\square$

Now, we have settled the abstract theoretical framework. The two practically most relevant examples are translation invariance and positive homogeneity of scoring functions. They are described in the two subsequent subsections.

4.3.1. Translation invariance

Let $\mathsf{A} = \mathsf{O} = \mathbb{R}^k$. Then a scoring function $S\colon \mathbb{R}^k \times \mathbb{R}^k \to \mathbb{R}$ is *translation invariant* if for all $x, y, z \in \mathbb{R}^k$

$$S(x - z, y - z) = S(x, y).$$

Analogously, an identification function $V\colon \mathbb{R}^k \times \mathbb{R}^k \to \mathbb{R}^k$ is translation invariant if for all $x, y, z \in \mathbb{R}^k$

$$V(x - z, y - z) = V(x, y).$$

Clearly, a translation invariant scoring function is linearly π-order-preserving with $\lambda = 1$ where $\Phi = \Phi^* = \{x \mapsto x - z, z \in \mathbb{R}^k\}$ is the translation group and π is the identity. Let π be fixed throughout this subsection. Then, given a certain functional, one can wonder about the class of strictly consistent scoring functions that are translation invariant. Clearly, with respect to Lemma 4.3.3 and Lemma 4.3.6, this class is empty if the functional T is not π-equivariant. On the other hand, the following proposition yields that, under the conditions of Osband's principle, a π-equivariant functional possesses essentially at most one translation invariant strictly consistent scoring function (the scoring function is unique up to scaling and adding a constant, but the existence is not guaranteed). Instead of writing π-equivariant, we shall customarily write translation equivariant in the sequel.

Proposition 4.3.7. *Let $T\colon \mathcal{F} \to \mathbb{R}^k$ be an identifiable functional with a translation invariant strict $\mathcal{F}$-identification function $V\colon \mathbb{R}^k \times \mathbb{R}^k \to \mathbb{R}^k$. Then T is translation equivariant.*

If additionally T is elicitable and has two translation invariant strictly $\mathcal{F}$-consistent scoring functions $S, S'\colon \mathbb{R}^k \times \mathbb{R}^k \to \mathbb{R}$ such that Assumptions (V1), (V2), (S1), *(F1), and (VS1) in Fissler and Ziegel (2016) hold, then there is some* $\lambda > 0$ *and* $c \in \mathbb{R}$ *such that*

$$S'(x,y) = \lambda S(x,y) + c$$

for almost all $(x,y) \in \mathbb{R}^k \times \mathbb{R}^k$*.*

In case of the mean functional on $\mathbb{R}$, this Proposition was already shown by Savage (1971) who showed that the squared loss is essentially the only strictly consistent scoring function for the mean that is of prediction error form[17].

Proof of Proposition 4.3.7. We use the following customary notation. If a random variable Y has distribution F with $F \in \mathcal{F}$, we write $F - z$ for the distribution of $Y - z$ where $z \in \mathbb{R}^k$.

For the first claim, consider any $F \in \mathcal{F}$ and $z \in \mathbb{R}^k$. Then

$$0 = \mathbb{E}_F[V(T(F), Y) = \mathbb{E}_F[V(T(F) - z, Y - z)].$$

Since V is a strict $\mathcal{F}$-identification function for T, $T(F - z) = T(F) - z$.

For the second claim, we can directly apply Fissler and Ziegel (2016, Proposition 3.4). Assume that S and S' are almost everywhere of the form given there with corresponding matrix-valued functions $h, h' \colon \mathbb{R}^k \to \mathbb{R}^{k\times k}$ and functions $a, a' \colon \mathbb{R}^k \to \mathbb{R}$. Since

$$\bar{S}(x, F) = \bar{S}(x - z, F - z)$$

for all $x, z \in \mathbb{R}^k$ and $F \in \mathcal{F}$, this yields for the gradient with respect to x the identity

$$h(x)\bar{V}(x,F) = h(x-z)\bar{V}(x-z, F-z) = h(x-z)\bar{V}(x,F), \tag{4.3.5}$$

where the second identity is due to the translation invariance of V. So (4.3.5) is equivalent to

$$\bar{V}(x,F) \in \ker\big(h(x-z) - h(x)\big).$$

Now, one can use Assumption (V1) which implies that

$$\ker\big(h(x-z) - h(x)\big) = \mathbb{R}^k.$$

Since $x, z \in \mathbb{R}^k$ were arbitrary, the function h is constant. Similarly, one can show that h' is constant. Consequently, there is a $\lambda > 0$ such that almost everywhere

$$S(x,y) - a(y) = \lambda(S'(x,y) - a'(y)).$$

This is equivalent to

$$S(x,y) - \lambda S'(x,y) = a(y) - \lambda a'(y).$$

Since the left-hand side is translation invariant, the right-hand side must be constant, which yields the claim. □

[17] That means that the scoring function is a function in $x - y$ only.

Of course, there are also other more general notions of translation invariance for scoring functions, especially when A and O do not coincide. One prominent example is the functional $T = (\mathrm{VaR}_\alpha, \mathrm{ES}_\alpha)$ for some $\alpha \in (0,1)$. Here, a natural choice is $\mathsf{A} = \mathsf{A}_0 = \{x \in \mathbb{R}^2 \colon x_1 \ge x_2\}$ and $\mathsf{O} = \mathbb{R}$. The functional T is π-equivariant for $\Phi = \{y \mapsto y + c,\ c \in \mathbb{R}\}$ the translation group on $\mathbb{R}$,

$$\Phi^* = \{(x_1, x_2) \mapsto (x_1 + c, x_2 + c),\ c \in \mathbb{R}\}$$

and

$$\pi((y \mapsto y + c)) = ((x_1, x_2) \mapsto (x_1 + c, x_2 + c)). \tag{4.3.6}$$

Then, one could define translation invariance of a scoring function as being linearly π-order-preserving for π defined at (4.3.6). In the following proposition, we give a family of strictly consistent scoring functions for the pair $(\mathrm{VaR}_\alpha, \mathrm{ES}_\alpha)$ which are linearly π-order-preserving with $\lambda = 1$ in (4.3.4). That means, their score differences are translation invariant.

Proposition 4.3.8. *Let $\alpha \in (0,1)$. Let $\mathcal{F}$ be a class of distribution functions with finite first moments and unique α-quantiles. Define $T = (\mathrm{VaR}_\alpha, \mathrm{ES}_\alpha) \colon \mathcal{F} \to \mathbb{R}^2$. The following assertions hold:*

(i) Suppose there is some $c > 0$ such that

$$\mathrm{ES}_\alpha(F) + c > \mathrm{VaR}_\alpha(F) \qquad \forall F \in \mathcal{F}. \tag{4.3.7}$$

That is, $T(\mathcal{F}) \subseteq \mathsf{A}_c := \{(x_1, x_2) \in \mathbb{R}^2 \colon x_2 + c > x_1\}$. Then, any scoring function $S \colon \mathsf{A}_c \times \mathbb{R} \to \mathbb{R}$, which is equivalent to

$$\begin{aligned} S_c(x_1, x_2, y) = {} & (\mathbb{1}\{y \le x_1\} - \alpha) c x_1 - \mathbb{1}\{y \le x_1\} c y \\ & + \alpha(x_2^2/2 + x_1^2/2 - x_1 x_2) \\ & + \mathbb{1}\{y \le x_1\}(-x_2(y - x_1) + y^2/2 - x_1^2/2), \end{aligned} \tag{4.3.8}$$

is strictly $\mathcal{F}$-consistent for T and has translation invariant score differences in the sense that

$$\begin{aligned} S_c(x_1 + z, x_2 + z, y + z) - S_c(x_1' + z, x_2' + z, y + z) \\ = S_c(x_1, x_2, y) - S_c(x_1', x_2', y) \end{aligned} \tag{4.3.9}$$

for all $(x_1, x_2), (x_1', x_2') \in \mathsf{A}_c$ and for all $z, y \in \mathbb{R}$.

(ii) Under the conditions of Theorem 5.2(iii) in Fissler and Ziegel (2016), there are strictly $\mathcal{F}$-consistent scoring functions for T with translation invariant score differences if and only if there is some $c > 0$ such that (4.3.7) holds. Then, any such scoring function is necessarily equivalent to S_d defined at (4.3.8) almost everywhere with $d \ge c$.

Proof. (i) First, one can see that S_c is of the form given at (5.2) in Fissler and Ziegel (2016) with $G_1(x_1) = -x_1^2/2 + cx_1$ and $\mathcal{G}_2(x) = (\alpha/2)x_2^2$. This means that $\mathcal{G}_2$ is strictly convex and the function $x_1 \mapsto x_1 G_2(x_2)/\alpha + G_1(x_1)$ is strictly increasing if and only if $x_2 + c > x_1$, that is, if and only if $(x_1, x_2) \in \mathsf{A}_c$. Invoking Theorem 5.2(ii) in Fissler and Ziegel (2016), this shows the strict $\mathcal{F}$-consistency of S_c.

Now, consider the score

$$\begin{aligned} S_c'(x_1, x_2, y) &= S_c(x_1, x_2, y) + \alpha c y \\ &= (\mathbb{1}\{y \le x_1\} - \alpha) c (x_1 - y) \\ &\quad + \alpha(x_2^2/2 + x_1^2/2 - x_1 x_2) \\ &\quad + \mathbb{1}\{y \le x_1\}(-x_2(y - x_1) + y^2/2 - x_1^2/2). \end{aligned} \tag{4.3.10}$$

A direct computation yields that $S_c'(x_1 + z, x_2 + z, y + z) = S_c'(x_1, x_2, y)$ for all $(x_1, x_2) \in \mathsf{A}_c$, $y \in \mathbb{R}$ and for all $z \in \mathbb{R}$. This proves (4.3.9).

(ii) Under the conditions of Theorem 5.2(iii) in Fissler and Ziegel (2016), any strictly $\mathcal{F}$-consistent scoring function $S\colon \mathsf{A} \times \mathbb{R} \to \mathbb{R}$, where $\mathsf{A} = T(\mathcal{F})$, is almost everywhere of the form

$$\begin{aligned} S(x_1, x_2, y) = {} & \big(\mathbb{1}\{y \le x_1\} - \alpha\big) G_1(x_1) - \mathbb{1}\{y \le x_1\} G_1(y) \\ & + G_2(x_2)\left(x_2 - x_1 + \frac{1}{\alpha}\mathbb{1}\{y \le x_1\}(x_1 - y)\right) - \mathcal{G}_2(x_2) + a(y). \end{aligned} \tag{4.3.11}$$

Here, $\mathcal{G}_2$ is strictly convex, $\mathcal{G}_2' = G_2$, for all $x_2 \in \mathsf{A}_2'$, and the function

$$\mathsf{A}_{1,x_2}' \to \mathbb{R}, \qquad x_1 \mapsto x_1 G_2(x_2)/\alpha + G_1(x_1)$$

is strictly increasing. Moreover, the proof of Theorem 5.2(iii) in Fissler and Ziegel (2016) yields that G_1 and G_2 are differentiable (under the condition of that theorem). The translation invariance of the score differences is equivalent to the fact that the function $\Psi\colon \mathbb{R} \times \mathsf{A} \times \mathsf{A} \times \mathbb{R} \to \mathbb{R}$ defined as

$$\begin{aligned} \Psi(z, x_1, x_2, x_1', x_2', y) = {} & S(x_1 + z, x_2 + z, y + z) - S(x_1' + z, x_2' + z, y + z) \\ & - S(x_1, x_2, y) + S(x_1', x_2', y) \end{aligned}$$

is constant. Let $z, y \in \mathbb{R}$ and $(x_1, x_2), (x_1', x_2') \in \mathsf{A}$. Then

$$\begin{aligned} & \frac{\partial}{\partial x_2}\Psi(z, x_1, x_2, x_1', x_2', y) \\ & \qquad = \big(x_2 - x_1 + \frac{1}{\alpha}\mathbb{1}\{y \le x_1\}(x_1 - y)\big)\big(G_2'(x_2 + z) - G_2'(x_2)\big). \end{aligned}$$

But since $\frac{\partial}{\partial x_2}\Psi(z, x_1, x_2, x_1', x_2', y)$ must vanish for all $z, y \in \mathbb{R}$ and (x_1, x_2), $(x_1', x_2') \in \mathsf{A}$, one obtains that necessarily $G_2' \equiv 0$. That is, $\mathcal{G}_2(x_2) = d_1 x_2^2 +$

$d_2x_2 + d_3$ with $d_1 > 0$ (ensuring the strict convexity of $\mathcal{G}_2$) and $d_2, d_3 \in \mathbb{R}$. Similarly, the derivative of Ψ with respect to z must vanish for all $z, y \in \mathbb{R}$ and (x_1, x_2), $(x_1', x_2') \in \mathsf{A}$. One can calculate that

$$\begin{aligned}
&\frac{\partial}{\partial z}\Psi(z, x_1, x_2, x_1', x_2', y) \\
&\qquad = \big(\mathbb{1}\{y \le x_1\} - \alpha\big)G_1'(x_1 + z) - \mathbb{1}\{y \le x_1\}G_1'(y + z) \\
&\qquad - \big(\mathbb{1}\{y \le x_1'\} - \alpha\big)G_1'(x_1' + z) + \mathbb{1}\{y \le x_1'\}G_1'(y + z) \\
&\qquad + 2d_1\big(\frac{1}{\alpha}\mathbb{1}\{y \le x_1\}(x_1 - y) - x_1\big) \\
&\qquad - 2d_1\big(\frac{1}{\alpha}\mathbb{1}\{y \le x_1'\}(x_1' - y) - x_1'\big).
\end{aligned}$$

This implies that necessarily $G_1'(x_1) = (-2d_1/\alpha)x_1 + d_4$ for some $d_4 \in \mathbb{R}$. Hence, $G_1(x_1) = (-d_1/\alpha)x_1^2 + d_4x_1 + d_5$ for some $d_5 \in \mathbb{R}$. Now, consider the function

$$\begin{aligned}
\psi_{x_2}(x_1) &= x_1G_2(x_2)/\alpha + G_1(x_1) \\
&= x_1(2d_1x_2 + d_2)/\alpha - d_1x_1^2/\alpha + d_4x_1 + d_5.
\end{aligned}$$

The function ψ_{x_2} is strictly increasing if and only if

$$x_2 + \frac{d_2 + d_4\alpha}{2d_1} > x_1.$$

This condition is satisfied for all $(x_1, x_2) \in \mathsf{A} = T(\mathcal{F})$ if and only there is a $c > 0$ such that $T(\mathcal{F}) \subseteq \mathsf{A}_c$ and

$$d := \frac{d_2 + d_4\alpha}{2d_1} \ge c. \tag{4.3.12}$$

Then, the resulting scoring function at (4.3.11) with

$$\begin{aligned}
\mathcal{G}_2(x_2) &= d_1x_2^2 + d_2x_2 + d_3, && d_1 > 0,\ d_2, d_3 \in \mathbb{R} && (4.3.13)\\
G_1(x_1) &= (-d_1/\alpha)x_1^2 + d_4x_1 + d_5, && d_4, d_5 \in \mathbb{R} && (4.3.14)
\end{aligned}$$

is equivalent to S_d defined at (4.3.8), which concludes the proof.

□

Remark 4.3.9. The scoring function S_c at (4.3.7) upon choosing $d_1 = \alpha/2$, $d_2 = d_3 = 0$, $d_4 = c$ and $d_5 = 0$ in (4.3.13) and (4.3.14). Obviously, the condition at (4.3.12) is satisfied.

Remark 4.3.10. The scoring function S_c' defined at (4.3.10) has a close relationship to the class of scoring functions proposed by Acerbi and Székely (2014); see equation (5.6) in Fissler and Ziegel (2016). Indeed, $S_c'(x_1, x_2, y) = c\big(\mathbb{1}\{y \le x_1\} - \alpha\big)(x - y) + S^W(x_1, x_2, y)$ with $W = 1$ where S^W is defined at equation (5.6) in Fissler and Ziegel (2016). That means it is the sum of the standard α-pinball loss for VaR_α – which is translation invariant – and S^1. Similarly, the condition at (4.3.7) is similar to the one at equation (5.7) in Fissler and Ziegel (2016). Since $\mathrm{ES}_\alpha \le \mathrm{VaR}_\alpha$, the maximal action domain where S_c or S'_c are strictly consistent is the stripe $\{(x_1,x_2)\in\mathbb{R}^2\colon x_1\ge x_2 > x_1 - c\}$. Of course, by letting $c\to\infty$, one obtains the maximal sensible action domain $\mathsf{A}_0 = \{(x_1,x_2)\in\mathbb{R}^2\colon x_1\ge x_2\}$ for the pair $(\mathrm{VaR}_\alpha, \mathrm{ES}_\alpha)$. However, considering the properly normalized version S_c/c or S'_c/c, this converges to a strictly consistent scoring function for VaR_α as $c\to\infty$, but which is independent of the forecast for ES_α. That means, there is a *caveat* concerning the tradeoff between the size of the action domain and the sensitivity in the ES-forecast. This might cast doubt on the usage of scoring functions with translation invariant score differences for $(\mathrm{VaR}_\alpha, \mathrm{ES}_\alpha)$ in general.

4.3.2. Homogeneity

For $\mathsf{A}\subseteq\mathbb{R}^k$, $\mathsf{O}\subseteq\mathbb{R}^d$ such that for any $\lambda>0$, and any $x\in\mathsf{A}$, $y\in\mathsf{O}$, one has $\lambda x\in\mathsf{A}$, $\lambda y\in\mathsf{O}$ (for example, A and O can be convex cones), a scoring function $S\colon \mathsf{A}\times\mathsf{O}\to\mathbb{R}$ is called positively homogeneous of degree $b\in\mathbb{R}$ if

$$S(\lambda x, \lambda y) = \lambda^b S(x,y)$$

for all $(x,y)\in\mathsf{A}\times\mathsf{O}$ and $\lambda>0$. Nolde and Ziegel (2016, p. 14) summarize arguments in favor of using positively homogeneous scoring functions given in the literature:

> "Efron (1991) argues that it is a crucial property of a scoring function to be positive homogeneous in estimation problems such as regression. Patton (2011) underlines the importance of positive homogeneity of the scoring function for forecast ranking. Acerbi and Székely (2014) argue in favor of using positive homogeneous scoring functions because they are so-called "unit consistent". That is, if r and x are given in U.S. dollars, for example, then, for a positive homogeneous scoring function S, the score $S(r,x)$ will have unit (U.S. dollars)^b."

Furthermore – and in line with our argumentation – they argue that for forecast ranking and comparison, only the positive homogeneity of the score difference is important. Clearly, this amounts to π-order-preservingness for the canonical isomorphism between the respective multiplicative groups

$$\pi((y\mapsto\lambda y)) = (x\mapsto\lambda x).$$

For the most important functionals in practice, the class of strictly consistent positively homogeneous scoring functions (strictly consistent scoring functions inducing a positively homogeneous score difference) has already been determined. We confine ourselves to give references for the corresponding results in the literature:

- **mean**: Patton (2011, Proposition 4), Gneiting (2011, Equations (19) and (20));
- **quantiles**: Gneiting (2011, Equation (26)), Nolde and Ziegel (2016, Theorem 2.4);
- **expectiles**: Nolde and Ziegel (2016, Theorem 2.5);
- **(VaR, ES)**: Nolde and Ziegel (2016, Theorem 2.6).

4.4. Possible applications

Recall that the usage of strictly consistent scoring functions in applications is two-fold: On the one hand, they can be used to rank and compare competing forecasts; on the other hand, they can serve as a tool in the context of learning via M-estimation, and regression. Clearly, the main focus of the preceding results concerning applications typically lies on either of the sides.

The notion of π-order-preserving scoring functions introduced in Section 4.3 is clearly motivated from the context of forecast comparison and is mainly beneficial there. On the contrary, (linear) π-order-preserving scoring functions can be also advantageous for M-estimation. Clearly, if the functional to be estimated is π-equivariant, then by equation (4.3.2) any M-estimator induced by a strictly consistent scoring function is π-equivariant as well. However, suppose the minimization procedure of M-estimation has some inaccuracies (maybe due to numerical imprecision), then it is not clear if the resulting inaccurate M-estimator is π-equivariant. But if the scoring function is strongly linearly π-order-preserving in the sense that equation (4.3.4) holds not only for score differences, but for the score itself, then the inaccurately calculated M-estimator still is π-equivariant.

Convexity of scoring functions discussed in Section 4.2 has a multitude of different motivations which we have already detailed in Subsection 4.2.1. These motivations and possible applications stem from both realms – learning and forecast ranking. However, the advantages in the context of learning / regression might seem to be more obvious.

Order-sensitivity, to which we have devoted Section 4.1 of this thesis, is certainly helpful in the context of forecast comparison. Nevertheless, there appear to be two exceptions from that rule: First, self-calibration (see Subsection 4.1.4) can ensure the asymptotic consistency of the M-estimator in the context of learning. Second, Subsection 4.1.3 with its main result, Proposition 4.1.16, asserting that the expected score of any strictly consistent scoring function for a mixture-continuous functional has only one local minimum, can turn out to be numerically attractive for M-estimation. It opens the way to use different minimization algorithms to numerically determine the minimizer of the expected score. We briefly describe one method we have in mind, namely *steepest descent* with a sort of perturbation: One can use the steepest descent method. Then, one eventually reaches the global minimum, or a local maximum, or a saddle-point. Then, if one applies a small perturbation and and calculates some iterations of the steepest descent method, one will again approach the global minimum in the first case, and in the two latter cases, one gets away from the critical points and will eventually end up also at the global minimum.

5. Implications for Backtesting

The content of this chapter is the joint article Fissler et al. (2016), which appeared in *Risk Magazine*. It shows how one can use the result of Fissler and Ziegel (2016, Corollary 5.5) that Expected Shortfall is jointly elicitable with Value at Risk in the context of backtesting. More generally, we commented in this note on the role of elicitability for backtesting problems in general, thereby suggesting the usage of *comparative backtests* of Diebold-Mariano style. Two other insightful references on the relationship between elicitability and backtesting are the articles Acerbi and Székely (2014), Acerbi and Székely (2017) and Nolde and Ziegel (2016).

Due to copyright reasons, the article is not in the journal style. However, its content corresponds literally to the same version which appeared in the journal. Again, two preprint versions can be found at `https://arxiv.org/abs/1507.00244`.

Expected Shortfall is jointly elicitable with Value at Risk – Implications for backtesting

Tobias Fissler* Johanna F. Ziegel* Tilmann Gneiting†

September 1, 2015

Abstract

In this note, we comment on the relevance of elicitability for backtesting risk measure estimates. In particular, we propose the use of Diebold-Mariano tests, and show how they can be implemented for Expected Shortfall (ES), based on the recent result of Fissler and Ziegel (2015) that ES is jointly elicitable with Value at Risk (VaR).

Joint elicitability of ES and VaR

There continues to be lively debate about the appropriate choice of a quantitative risk measure for regulatory purposes or internal risk management. In this context, it has been shown by Weber (2006) and Gneiting (2011) that Expected Shortfall (ES) is not elicitable. Specifically, there is no strictly consistent scoring (or loss) function $S : \mathbb{R}^2 \to \mathbb{R}$ such that, for any random variable X with finite mean, we have

$$\mathrm{ES}_\alpha(X) = \arg\min_{e \in \mathbb{R}} \mathbb{E}[S(e, X)].$$

Recall that ES of X at level $\alpha \in (0, 1)$ is defined as

$$\mathrm{ES}_\alpha(X) = \frac{1}{\alpha} \int_0^\alpha \mathrm{VaR}_\beta(X)\, \mathrm{d}\beta,$$

where Value at Risk (VaR) is given by $\mathrm{VaR}_\alpha(X) = \inf\{x \in \mathbb{R} : \mathbb{P}(X \le x) \ge \alpha\}$. In contrast, VaR at level $\alpha \in (0, 1)$ is elicitable for random variables with a unique α-quantile. The possible strictly consistent scoring functions for VaR are of the form

$$S_V(v, x) = (\mathbb{1}\{x \le v\} - \alpha)(G(v) - G(x)), \tag{1}$$

where G is a strictly increasing function such that the expectation $\mathbb{E}[G(X)]$ exists.

*University of Bern, Department of Mathematics and Statistics, Institute of Mathematical Statistics and Actuarial Science, Sidlerstrasse 5, 3012 Bern, Switzerland, e-mail: tobias.fissler@stat.unibe.ch; johanna.ziegel@stat.unibe.ch

†Heidelberg Institute for Theoretical Studies and Karlsruhe Institute of Technology, HITS gGmbH, Schloß-Wolfsbrunnenweg 35, 69118 Heidelberg, Germany, e-mail: tilmann.gneiting@h-its.org

However, it turns out that ES is elicitable of higher order in the sense that the pair (VaR_α, ES_α) is jointly elicitable. Indeed, we have that

$$(\mathrm{VaR}_\alpha(X), \mathrm{ES}_\alpha(X)) = \arg\min_{(v,e)\in\mathbb{R}^2} \mathbb{E}[S_{V,E}(v, e, X)],$$

where possible choices of $S_{V,E}$ are given by

$$\begin{aligned} S_{V,E}(v, e, x) = {} & (\mathbb{1}\{x \le v\} - \alpha)(G_1(v) - G_1(x)) \\ & + \frac{1}{\alpha} G_2(e)\mathbb{1}\{x \le v\}(v - x) + G_2(e)(e - v) - \mathcal{G}_2(e), \end{aligned} \tag{2}$$

with G_1 and G_2 being strictly increasing continuously differentiable functions such that the expectation $\mathbb{E}[G_1(X)]$ exists, $\lim_{x\to-\infty} G_2(x) = 0$ and $\mathcal{G}_2' = G_2$; see Fissler and Ziegel (2015, Corollary 5.5). One can nicely see the structure of $S_{V,E}$: The first summand in (2) is exactly a strictly consistent scoring function for VaR_α given at (1) and hence only depends on v, whereas the second summand cannot be split into a part depending only on v and one depending only on e, respectively, hence illustrating the fact that ES_α itself is not elicitable. Acerbi and Székely (2014) were first to suggest that the pair (VaR_α, ES_α) is jointly elicitable. They proposed a strictly consistent scoring function under the additional assumption that there exists a real number w such that $\mathrm{ES}_\alpha(X) > w\,\mathrm{VaR}_\alpha(X)$ for all assets X under consideration. Despite encouraging simulation results, there is currently no formal proof available of the strict consistency of their proposal. In contrast, the scoring functions given at (2) do not require additional assumptions, and it has been formally proven that they provide a class of strictly consistent scoring functions.

The role of elicitability in backtesting

The lack of elicitability of ES (of first order) has led to a lively discussion about whether or not and how it is possible to backtest ES forecasts; see, for example, Acerbi and Székely (2014), Carver (2014), and Emmer et al. (2015). It is generally accepted that elicitability is useful for model selection, estimation, generalized regression, forecast comparison, and forecast ranking. Having provided strictly consistent scoring functions for (VaR_α, ES_α), we take the opportunity to comment on the role of elicitability in backtesting.

The traditional approach to backtesting aims at model verification. To this end, one tests the null hypothesis:

$$H_0^C : \text{“The risk measure estimates at hand are correct.”}$$

Specifically, suppose we have sequences $(x_t)_{t=1,\ldots,N}$ and $(v_t, e_t)_{t=1,\ldots,N}$, where x_t is the realized value of the asset at time point t, and v_t and e_t denote the estimated VaR_α and ES_α given at time $t-1$ for time point t, respectively. A backtest uses some test statistic T_1, which is a function of $(v_t, e_t, x_t)_{t=1,\ldots,N}$, such that we know the distribution of T_1 (at least approximately) if the null hypothesis of correct risk measure estimates holds. If we reject H_0^C at some small level, the model or the estimation procedure for the risk measure is deemed inadequate. Traditional backtests have a wealth of legitimate and important uses, particularly when a single internal risk model is checked and monitored over time. For this approach of model verification, elicitability of the risk measure is not relevant, as pointed out by Acerbi and Székely (2014) and Davis (2014). However, tests of this type can be problematic as a basis of decision making for adapting or changing an internal model, or, in regulatory practice, notably in view of the anticipated revised standardised approach (Bank for International Settlements, 2013, pp. 5–6), which "should provide a credible fall-back in the event that a bank's internal market risk model is deemed inadequate". If the internal model fails the backtest, the standardised approach may fail the test, too, and in fact it might be inferior to the internal model. Conversely, the internal model might pass the backtest, despite a more informative standardised model being available, as illustrated in the numerical example below. Generally, tests of the hypothesis H_0^C are not aimed at, and do not allow for, model comparison and model ranking.

Alternatively, one could use the following null hypothesis in backtesting:

H_0^- : "The risk measure estimates at hand are *at least as good* as the ones from the standard procedure."

Here, the standard procedure could be a method specified by the regulator, or it could be a technique that has been used in the past. An important caveat is that the standard procedure needs to provide daily VaR_α and ES_α predictions on a bank's portfolio. For internal use at banks, this requirement should not pose problems. In the regulatory setting, the revised standardised approach proposed by the Bank for International Settlements (2013, 2014) only gives the capital charge based on exposures, and thus the approach would need to be adapted for VaR_α and ES_α predictions to become available. Let us now write $(v_t^*, e_t^*)_{t=1,\ldots,N}$ for the sequence of VaR_α and ES_α estimates by the standard procedure. Making use of the elicitability of $(\mathrm{VaR}_\alpha, \mathrm{ES}_\alpha)$, we take one of the scoring functions $S_{V,E}$ given at (2) to define the test statistic

$$T_2 = \frac{\bar{S}_{V,E} - \bar{S}^*_{V,E}}{\sigma_N}, \tag{3}$$

where

$$\bar{S}_{V,E} = \frac{1}{N}\sum_{t=1}^{N} S_{V,E}(v_t, e_t, x_t), \quad \bar{S}^*_{V,E} = \frac{1}{N}\sum_{t=1}^{N} S_{V,E}(v_t^*, e_t^*, x_t),$$

and σ_N is a suitable estimate of the respective standard deviation. Under H_0^-, the test statistic T_2 has expected value less than or equal to zero. Following the lead of Diebold and Mariano (1995), comparative tests that are based on the asymptotic normality of the test statistics T_2 have been employed in a wealth of applications.

Under both H_0^C and H_0^-, the backtest is passed if the null hypothesis fails to be rejected. However, as Fisher (1949, p. 16) noted, "the null hypothesis is never proved or established, but it is possibly disproved, in the course of experimentation." In other words, a passed backtest does not imply the validity of the respective null hypothesis. Passing the backtest simply means that the hypothesis of correctness (H_0^C) or superiority (H_0^-), respectively, could not be falsified.

In the case of comparative backtests, a more conservative approach could be based on the following null hypothesis:

H_0^+ : "The risk measure estimates at hand are *at most as good* as the ones from the standard procedure."

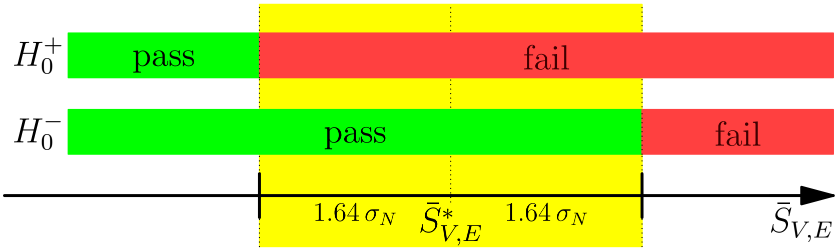

Figure 1: Decisions taken in comparative backtests under the null hypotheses H_0^- and H_0^+ at level 0.05. In the yellow region the two backtests entail distinct decisions. See the text for details.

This can also be tested using the statistic T_2 in (3), which has expected value greater than or equal to zero under H_0^+. The backtest now is passed when H_0^+ is rejected. The decisions taken in comparative backtesting under H_0^- and H_0^+ are illustrated in Figure 1, where the colours relate to the three-zone approach of the Bank for International Settlements (2013, pp. 103–108). In regulatory practice, the distinction between Diebold-Mariano tests under the two hypotheses amounts to a reversed onus of proof. In the traditional setting, it is the regulator's burden to show that the internal model is incorrect. In contrast, if a backtest is passed when H_0^+ is rejected, banks are obliged to demonstrate the superiority of the internal model. Such an approach to backtesting may entice banks to improve their internal models, and is akin to regulatory practice in the health sector, where market authorisation for medicinal products hinges on comparative clinical trials. In the health context, decision-making under H_0^- corresponds to equivalence or non-inferiority trials, which are "not conservative in nature, so that many flaws in the design or conduct of the trial will tend to bias the results", whereas "efficacy is most convincingly established by demonstrating superiority" under H_0^+ (European Medicines Agency, 1998, p. 17). Technical detail is available in a specialized strand of the biomedical literature; for a concise review, see Lesaffre (2008).

Numerical example

We now give an illustration in the simulation setting of Gneiting et al. (2007). Specifically, let $(\mu_t)_{t=1,\ldots,N}$ be a sequence of independent standard normal random variables. Conditional on μ_t, the return X_t is normally distributed with mean μ_t and variance 1, denoted $\mathcal{N}(\mu_t, 1)$. Under our Scenario A, the standard method for estimating risk measures uses the unconditional distribution $\mathcal{N}(0, 2)$ of X_t, whereas the internal procedure takes advantage of the information contained in μ_t and uses the conditional distribution $\mathcal{N}(\mu_t, 1)$. Therefore,

$$(v_t, e_t) = (\mathrm{VaR}_\alpha(\mathcal{N}(\mu_t, 1)), \mathrm{ES}_\alpha(\mathcal{N}(\mu_t, 1))) = \left(\mu_t + \Phi^{-1}(\alpha),\, \mu_t - \frac{1}{\alpha}\,\varphi(\Phi^{-1}(\alpha))\right)$$

Table 1: Percentage of decisions in the green, yellow, and red zone in traditional and comparative backtests. Under Scenario A, a decision in the green zone is desirable and in the joint interest of banks and regulators. Under Scenario B, the red zone corresponds to a decision in the joint interest of all stakeholders.

Scenario A		**Green**	Yellow	Red
Traditional	$\mathrm{VaR}_{0.01}$	89.35	10.65	0.00
Traditional	$\mathrm{ES}_{0.025}$	93.62	6.36	0.02
Comparative	$\mathrm{VaR}_{0.01}$	88.23	11.77	0.00
Comparative	$(\mathrm{VaR}_{0.025}, \mathrm{ES}_{0.025})$	87.22	12.78	0.00
Scenario B		Green	Yellow	**Red**
Traditional	$\mathrm{VaR}_{0.01}$	89.33	10.67	0.00
Traditional	$\mathrm{ES}_{0.025}$	93.80	6.18	0.02
Comparative	$\mathrm{VaR}_{0.01}$	0.00	11.77	88.23
Comparative	$(\mathrm{VaR}_{0.025}, \mathrm{ES}_{0.025})$	0.00	12.78	87.22

and

$$(v_t^*, e_t^*) = (\mathrm{VaR}_\alpha(\mathcal{N}(0,2)), \mathrm{ES}_\alpha(\mathcal{N}(0,2))) = \left(\sqrt{2}\,\Phi^{-1}(\alpha),\ -\frac{\sqrt{2}}{\alpha}\,\varphi(\Phi^{-1}(\alpha))\right),$$

where φ and Φ denote the density and the cumulative distribution function of the standard normal distribution, respectively. Under Scenario B, the roles of the standard method and the internal procedure are interchanged: The standard model now uses the conditional distribution $\mathcal{N}(\mu_t, 1)$, whereas the internal model uses the unconditional distribution $\mathcal{N}(0, 2)$. Consequently, under Scenario A the desired decision is to pass the backtest. Under Scenario B the backtest should be failed.

As tests of traditional type, we consider the coverage test for $\mathrm{VaR}_{0.01}$ described by the Bank for International Settlements (2013, pp. 103–108) and the generalized coverage test for $\mathrm{ES}_{0.025}$ proposed by Costanzino and Curran (2015). As shown by Clift et al. (2015), the latter performs similarly to the approaches of Wong (2008) and Acerbi and Székely (2014), but is easier to implement. The outcome of the test is structured into green, yellow, and red zones, as described in the aforementioned references. For the comparative backtest for $(\mathrm{VaR}_{0.025}, \mathrm{ES}_{0.025})$, we use the functions $G_1(v) = v$ and $G_2(e) = \exp(e)/(1 + \exp(e))$ in (2) and define the zones as implied by Figure 1. Finally, our comparative backtest for $\mathrm{VaR}_{0.01}$ uses the function $G(v) = v$ in (1), which is equivalent to putting $G_1(v) = v$ and $G_2(e) = 0$ in (2). For σ_N in the test statistic T_2 in (3) we use the standard estimator.

Table 1 summarizes the simulation results under Scenario A and B, respectively. We use sample size $N = 250$ and repeat the experiment 10,000 times. The traditional backtests are performed for the internal model in the scenario at hand. Under Scenario A, the four tests give broadly equivalent results. The benefits of the comparative approach become apparent under Scenario B, where the traditional approach yields highly undesirable decisions, in that the simplistic internal model passes the backtest, while a more informative standard model would be available. This can neither be in banks' nor in regulators' interests. We emphasize that this problem will arise with any traditional backtest, as a traditional backtest assesses optimality only with respect to the information used for providing the risk measure estimates.

Comparative tests based on test statistics of the form T_2 in (3) can be used to compare forecasts in the form of full predictive distributions, provided a proper scoring rule is used (Gneiting and Raftery, 2007), or to compare risk assessments, provided the risk measure admits a strictly consistent scoring function, so elicitability is crucial. In particular, proper scoring rules and consistent scoring functions are sensitive to increasing information utilized for prediction; see Holzmann and Eulert (2014). However, as consistent scoring functions are not unique, a question of prime practical interest is which functions ought to be used in regulatory settings or internally. In the present context, initial numerical experiments show that the choice of the functions G_1 and G_2 in (2) affects the discrimination ability of the backtest but does not lead to contradictory decisions. That is, with a different choice of G_1, G_2, the number of decisions in the yellow zone may vary but it was observed extremely rarely that a decision was changed from green to red or vice versa. A comprehensive analysis is beyond the scope of our note and left to future work.

Discussion

We have raised the idea of comparative backtesting. In contrast to the traditional approach, comparative backtests provide a credible fall-back in case a risk measure estimate does not pass the backtest. Elicitability of the risk measure is not relevant for traditional backtesting, whereas it is crucial for comparative backtesting. Hence, the recent result of Fissler and Ziegel (2015) that the pair $(\mathrm{VaR}_\alpha, \mathrm{ES}_\alpha)$ is jointly elicitable open the possibility for comparative backtests of risk measure estimates for this pair.

Arguably, now may be the time to revisit and investigate fundamental statistical issues in banking supervision. Chances are that comparative backtests, where a bank's internal risk model is held accountable relative to an agreed-upon standardised approach, turn out to be beneficial to all stakeholders, including banks, regulators, and society at large.

Acknowledgements

We thank two anonymous referees, Carlo Acerbi, Paul Embrechts, Fernando Fasciati, Fabian Krüger, Alexander McNeil, Alexander Schied, Patrick Schmidt, and the organiser, Imre Kondor, and participants of the "International Workshop on Systemic Risk and Regulatory Market Risk Measures" in Pullach for inspiring discussions and helpful comments. Tobias Fissler acknowledges funding by the Swiss National Science Foundation (SNF) via grant 152609, and Tilmann Gneiting by the European Union Seventh Framework Programme under grant agreement no. 290976.

References

C. Acerbi and B. Székely. Back-testing expected shortfall. *Risk*, December, 2014.

Bank for International Settlements. *Consultative Document: Fundamental review of the trading book: A revised market risk framework*. 2013.

Bank for International Settlements. *Consultative Document: Fundamental review of the trading book: Outstanding issues*. 2014.

L. Carver. Back-testing expected shortfall: mission possible? *Risk*, November, 2014.

S. S. Clift, N. Costanzino, and M. Curran. Empirical performance of backtesting methods for expected shortfall. 2015. URL `ssrn.com/abstract=2618345`.

N. Costanzino and M. Curran. Backtesting general spectral risk measures with application to expected shortfall. *The Journal of Risk Model Validation*, 9:21–33, 2015.

M. Davis. Consistency of internal risk measure estimates. 2014. URL `arXiv:1410.4382`.

F. X. Diebold and R. S. Mariano. Comparing predictive accuracy. *Journal of Business and Economic Statistics*, 13:253–263, 1995.

S. Emmer, M. Kratz, and D. Tasche. What is the best risk measure in practice? A comparison of standard measures. 2015. URL `arXiv:1312.1645v4`.

European Medicines Agency. ICH Topic E 9: Statistical Principles for Clinical Trials. Note for Guidance on Statistical Principles for Clinical Trials, 1998. URL `www.ema.europa.eu`.

R. A. Fisher. *The Design of Experiments*. Oliver and Boyd, London, 5th edition, 1949.

T. Fissler and J. F. Ziegel. Higher order elicitability and Osband's principle. 2015. URL `arXiv:1503.08123`.

T. Gneiting. Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106:746–762, 2011.

T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102:359–378, 2007.

T. Gneiting, F. Balabdaoui, and A. E. Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society Series B*, 69:243–268, 2007.

H. Holzmann and M. Eulert. The role of the information set for forecasting – with applications to risk management. *Annals of Applied Statistics*, 8:79–83, 2014.

E. Lesaffre. Superiority, equivalence, and non-inferiority trials. *Bulletin of the NYU Hospital for Joint Diseases*, 66:150–154, 2008.

S. Weber. Distribution-invariant risk measures, information, and dynamic consistency. *Mathematical Finance*, 16:419–441, 2006.

W. K. Wong. Backtesting trading risk of commercial banks using expected shortfall. *Journal of Banking & Finance*, 32:1404–1415, 2008.

6. Discussion

6.1. Reception in the literature

Our two joint articles Fissler and Ziegel (2016) and Fissler et al. (2016), discussed in the first part of this thesis, have gained considerable attention, on the one hand for the theoretical issues they are concerned with (Frongillo and Kash, 2015a,c; Frongillo et al., 2016; Ehm et al., 2016; Grushka-Cockayne et al., 2016; Maume-Deschamps et al., 2016; Embrechts et al., 2016; Liu and Wang, 2016; Zwingmann and Holzmann, 2016; Bayer and Dimitriadis, 2017), on the other hand due to the practical implications they have on the discussion about the best choice of a risk measure (Kou and Peng, 2015; Jakobsons and Vanduffel, 2015; Lerch et al., 2015; Davis, 2016; Nolde and Ziegel, 2016; Allen et al., 2016; Asimit et al., 2016; Burzoni et al., 2016; Corbetta and Peri, 2016; Clift et al., 2016; Pitera and Schmidt, 2016).

6.2. Outlook for possible future research projects

Besides the questions that could be answered in this part of the thesis, there is certainly a plenitude of remaining or newly arisen open questions. We confine ourselves to a list of possible future projects in this area, tackling some of these questions.

6.2.1. Vector-valued risk measures in the context of systemic risk

Most of the vector-valued functionals investigated in the literature on elicitability are on the one hand vectors consisting of elicitable components (such as the vector of ratios of expectations with the same denominator) or bijections thereof; and on the other hand, functionals have been studied that consist of a non-elicitable component one is actually interested in and some other components being elicitable, but only having the character of auxiliary components making the whole vector elicitable. A prominent example for the second case is the pair (VaR, ES); see Fissler and Ziegel (2016), and for a general explanation of this phenomenon Frongillo and Kash (2015c).

An interesting class of functionals is given by vector-valued risk measures which naturally take values in $\mathbb{R}^k$ or which are set-vector-valued taking values in $2^{\mathbb{R}^k}$. These risk measures often appear in the assessment of systemic risk (Jouini et al., 2004; Armenti et al., 2016; Feinstein et al., 2016). After getting an overview of the most important choices of such risk measures one can tackle the canonical question if they are elicitable and/or identifiable which would open the way to classical and comparative backtesting (Fissler et al., 2016; Nolde and Ziegel, 2016). While the findings of project 6.2.3 can help to give a characterization of the class of strictly consistent scoring functions for systemic risk measures, the results of Chapter 4 could lead the way to characterize the subclass of order-sensitive and (quasi-)-convex strictly consistent scoring functions.

6.2.2. Testing the tail of the P&L-distribution

It is often argued in the literature that probabilistic forecasts are superior to point forecasts since they account for the uncertain nature of future events. For the evaluation and ranking of competing probabilistic forecasts, one commonly uses strictly proper scoring rules.

However, in the framework of quantitative risk management, one is usually not interested in assessing the whole Profit&Loss-distribution. Instead, the risk of an asset is only related to one tail of the P&L-distribution; see Liu and Wang (2016). Here, the definition of tail can have different forms, e.g. an absolute threshold (for example, only losses should be interesting for assessing the risk), or a relative threshold like an α-quantile or -expectile. This induces a tail operator $T\colon \mathcal{F} \to \mathcal{F}_0$ where $\mathcal{F}, \mathcal{F}_0$ are some adequate classes of probability distributions. One can also regard this tail operator as an infinite dimensional functional. Then it is natural to investigate the classical questions of point forecasts, namely, *mutatis mutandis*, is T elicitable, what is the class of strictly consistent scoring functions, and are there some particularly 'good' choices of scoring functions. It is clear that a strictly consistent scoring function for the tail operator induces a proper scoring rule which is not strictly proper on purpose (if T is not the identity operator).

Gneiting and Ranjan (2011) proposed the 'threshold-weighted continuous ranked probability score'

$$S(F,y) = \int_{-\infty}^{\infty} \big(F(z) - \mathbb{1}\{y \le z\}\big)^2 w(z)\,\mathrm{d}z, \tag{6.2.1}$$

whereas Diks et al. (2011) suggested the 'censored likelihood score function'

$$S^{\mathrm{csl}}(F,y) = -w(y)\log(f(y)) - (1-w(y))\log\Big(1 - \int w(z)f(z)\,\mathrm{d}z\Big), \tag{6.2.2}$$

where $w(y) = \mathbb{1}\{y \le a\}$ for some threshold $a \in \mathbb{R}$ and f is the density of the predictive tail F; see also Lerch et al. (2015). Both suggestions are strictly consistent scoring functions for the threshold-tail. As already indicated, a natural question is whether these suggestions are the only strictly consistent scoring functions. It appears that one can generalize the form of S^{csl} to a whole class of strictly consistent scoring functions for the threshold-tail which are also more flexible in terms of regularity assumptions. The key is that one needs a joint scoring function for the infinite dimensional tail and the one dimensional functional $F \mapsto \mathbb{E}_F[\mathbb{1}\{Y \le a\}]$.

Gneiting and Ranjan (2011) also proposed the 'quantile-weighted version of the continuous ranked probability score'

$$S(F, y) = \int_0^1 2\big(\mathbb{1}\{y \le F^{-1}(\beta)\} - \beta\big)(F^{-1}(\beta) - y)\nu(\beta)\,\mathrm{d}\beta, \tag{6.2.3}$$

where ν is a nonnegative weight function on $[0,1]$ and F^{-1} is the generalized inverse of the predictive tail F. Upon choosing $\nu(\beta) = \mathbb{1}\{\beta \le \alpha\}$, one obtains a strictly consistent scoring function for the α-quantile tail. Also in this case, it is clear that there are further strictly consistent scoring functions of a similar form. Moreover, it is an open question whether there is also an approach similar to the censored likelihood score function in that one elicits jointly the α-tail and the α-quantile. Moreover, it is not clear if there exists an 'expectile-weighted' analogon of (6.2.3). Another interesting approach could consist in considering scoring functions of the form (6.2.2) and choosing the threshold a as some quantile or expectile. Then, one would need a joint scoring function for the quantile- or expectile-tail and the corresponding quantile or expectile, respectively.

Note that the recent approach of Emmer et al. (2015) aims at evaluating quantiles at different levels, thus approximating the α-tail. Moreover, in practical situations, institutions usually first calculate a predictive distribution and then a risk measure based on that distributional forecast (Cont et al., 2010; Pitera and Schmidt, 2016). Consequently, quite often the predictive tail is already available such that it can be directly evaluated.

In some sense, this project could fill the gap between point forecasts, where one considers finite dimensional functionals, and probabilistic forecasts, where one considers one particular infinite dimensional functional, namely the identity operator; see the recent paper Grushka-Cockayne et al. (2016).

Some advances in this direction have been achieved by the very recent paper Holzmann and Klar (2016).

6.2.3. A sufficient condition for higher order elicitability

Since Osband's (1985) doctoral thesis it is well known that the convexity of the level sets of a functional is a necessary condition for its elicitability, both in the one and in the higher-dimensional setting. That means if a functional $T\colon \mathcal{F} \to \mathbb{R}^k$ is elicitable, then for every $F_0, F_1 \in \mathcal{F}$ such that $T(F_0) = T(F_1) = t$ and for all $\lambda \in (0,1)$, $T((1-\lambda)F_0 + \lambda F_1) = t$ whenever $(1-\lambda)F_0 + \lambda F_1 \in \mathcal{F}$. For $k = 1$, Steinwart et al. (2014) showed that under continuity assumptions on T, the convexity of the level sets of T is also sufficient for the elicitability. Their proof uses a separation theorem and hence relies heavily on the total order of $\mathbb{R}$. Moreover, the continuity assumptions also appear to be crucial since the mode functional is generally not elicitable despite having convex level sets; see Heinrich (2014).

It is an open and highly non trivial question if the convexity of the level sets of T continues to be a sufficient criterion for elicitability under appropriate regularity assumptions when $k > 1$. And if not, is there another sufficient condition on T which can be verified efficiently?

Part II.

Limit Theorems on the Poisson and Wiener Space

7. Introduction

Whereas Part I of this thesis belongs to the field of mathematical statistics and in many parts has its inspirations from direct applications, it is doubtless that Part II is more probabilistic in nature. Its main content are the papers Fissler and Thäle (2016a) in Chapter 8 and Fissler and Thäle (2016b) in Chapter 9. In both articles, the objects of research are (non-linear) Gaussian and Poisson functionals. Even though the two articles are self-contained, we shall give a very brief overview and introduction into (i) the chaos representation of Gaussian and Poisson functionals in terms of multiple Wiener-Itô integrals, and (ii) Stein's method (for the sake of brevity and clarity just for the normal approximation, the Gamma approximation being similar in nature) and how the Malliavin Calculus and Stein's method can be fruitfully combined. This introduction is not intended to give a detailed exposition of this subject – which could, indeed, be the content of a whole thesis on its own. Instead, it gives many references to the literature and also to the main articles Fissler and Thäle (2016a) and Fissler and Thäle (2016b) where we detailed on it.

Please note that the notations of Part I and Part II of this thesis are completely independent (of course, with the exception of standard notation in stochastics). Since the focus of the two articles Fissler and Thäle (2016a) and Fissler and Thäle (2016b) is slightly different, also the notation differs in some aspects between Chapter 8 and Chapter 9. However, since the articles are self-contained, the correspondences should be clear. For this introduction, we stick to the notation of Chapter 8.

Part II concludes with a short discussion in Chapter 10 providing an outlook to future research.

7.1. Chaos representation of Gaussian and Poisson functionals

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be an underlying common probability space which is rich enough to support all subsequent random objects. A Gaussian (Poisson) functional is a real-valued random variable which is measurable with respect to the σ-algebra generated by a Gaussian process G (a Poisson process η) on $(\Omega, \mathcal{F}, \mathbb{P})$. As mentioned in Fissler and Thäle (2016b), some examples for Gaussian functionals are the power and bi-power variation of Gaussian processes, which can be important in the context of risk quantification, the Euler characteristic of a Gaussian excursion set, or the statistics appearing around the Breuer-Major theorem. On the other hand, most examples for Poisson functionals are from the fields of geometric probability or stochastic geometry, but also U-statistics; see e.g. Hug and Reitzner (2016), Lachièze-Rey and Reitzner (2016), Schulte and Thäle (2016), and the references in Fissler and Thäle (2016a).

A both unifying and generalizing perspective on Gaussian and Poisson functionals can be provided in terms of so-called *completely random measures*, commonly denoted by φ; see Peccati and Taqqu (2011, Section 5.1) and Fissler and Thäle (2016a, Section 2.1). In a nutshell, one assumes that the underlying (Gaussian or Poisson) process is indexed by the family $\mathscr{Z}_\mu = \{Z \in \mathscr{Z} \colon \mu(Z) < \infty\}$ of Borel-measurable sets with finite measure of an underlying Polish space $\mathcal{Z}$ with a Borel σ-algebra $\mathscr{Z}$, equipped with a non-atomic and σ-finite reference measure μ. For the most interesting cases that φ coincides with a centered Gaussian measure G or a compensated Poisson measure $\hat{\eta}$, one can show that the space of square-integrable functionals of φ, denoted by $L^2(\sigma(\varphi), \mathbb{P})$, admits a *chaos decomposition* or *chaos representation* (Peccati and Taqqu, 2011, Section 5.9). That is, the space $L^2(\sigma(\varphi), \mathbb{P})$ has a representation in terms of a direct orthogonal sum

$$L^2(\sigma(\varphi), \mathbb{P}) = \bigoplus_{q=0}^{\infty} W_q^{\varphi}\,, \tag{7.1.1}$$

where W_q^{φ} is the *Wiener chaos* of order q. More precisely, W_q^{φ} is the image of the *multiple Wiener-Itô integral* of order q with respect to φ, that is a map

$$I_q^{\varphi} \colon L_s^2(\mathcal{Z}^q, \mathscr{Z}^{\otimes q}, \mu^{\otimes q}) \to L^2(\sigma(\varphi), \mathbb{P}),$$

which is linear and continuous. The elements of $L_s^2(\mathcal{Z}^q, \mathscr{Z}^{\otimes q}, \mu^{\otimes q})$ are symmetric and square-integrable functions on $\mathcal{Z}^q$ in the sense that the arguments are $\mu^{\otimes q}$-a.e. invariant under permutations. Moreover, I_0^{φ} is defined as the identity map on $\mathbb{R}$. For a construction of the multiple Wiener-Itô integrals with respect to a completely random measure we refer to Peccati and Taqqu (2011, Section 5), as well as to Nualart (2006) for the Gaussian case and to Last (2016) for the Poisson case. The decomposition (7.1.1) corresponds to the fact that for any square-integrable functional $F \in L^2(\sigma(\varphi), \mathbb{P})$ there are ($\mu^{\otimes q}$-a.e.) uniquely determined kernels $f_q \in L_s^2(\mathcal{Z}^q, \mathscr{Z}^{\otimes q}, \mu^{\otimes q})$ such that

$$F = \mathbb{E}[F] + \sum_{q=1}^{\infty} I_q^{\varphi}(f_q), \tag{7.1.2}$$

where the series converges in $L^2(\mathbb{P})$. The orthogonality relation in (7.1.1) is reflected by the so-called *Itô isometry*, asserting that

$$\mathbb{E}[I_p^{\varphi}(f) I_q^{\varphi}(g)] = \begin{cases} 0 & \text{if } p \neq q \\ q!\langle f, g\rangle_{L_s^2(\mathcal{Z}^q, \mathscr{Z}^{\otimes q}, \mu^{\otimes q})} & \text{if } p = q \end{cases} \tag{7.1.3}$$

for any integers $p, q \geq 1$ and any $f \in L_s^2(\mathcal{Z}^p, \mathscr{Z}^{\otimes p}, \mu^{\otimes p})$ and $g \in L_s^2(\mathcal{Z}^q, \mathscr{Z}^{\otimes q}, \mu^{\otimes q})$.

Not only the expectation of the product of two multiple integrals can be neatly expressed, but also the product $I_p^{\varphi}(f)I_q^{\varphi}(g)$ itself in terms of a chaos representation in the flavor of (7.1.2). The corresponding formulae are called *multiplication formulae* and can be found in Fissler and Thäle (2016a, Lemma 2.1) for the Poisson case and Fissler and Thäle (2016a, Remark 2.4) for the Gaussian case. It is remarkable that (i) the two formulae are similar in nature, but the one for the Poisson case has a more involved structure; (ii) in both cases, the chaotic representation is finite in the sense that the representation of the product contains integrals only up to order $\min\{p, q\}$; and (iii) the integrands of the respective integrals are so-called contractions of f and g; see Fissler and Thäle (2016a, Section 2.5).

Malliavin Calculus

An important toolbox when dealing with Gaussian and Poisson functionals is the so-called *Malliavin Calculus*, which can be considered is an infinite-dimensional differential calculus on the Wiener space $L^2(\sigma(\varphi), \mathbb{P})$. Its main ingredients are the following operators: the Malliavin derivative D, its adjoint operator, the divergence δ (also called Skorohod integral), as well as the Ornstein-Uhlenbeck generator L and its pseudo-inverse L^{-1}. The action of these operators can be expressed in terms of the chaos representation (7.1.2) and the operators are closely related. For the case of Gaussian integrals, we gave an exposition of the Malliavin operators in Fissler and Thäle (2016b, Section 2). For a detailed introduction, we refer to Nualart (2006, Chapter 1) for the Gaussian case, and to Bourguin and Peccati (2016) for the Poisson case.

7.2. Stein's method

The paper Fissler and Thäle (2016a) is concerned with qualitative non-central limit theorems for a sequence of Poisson functionals living inside a fixed chaos towards a centered Gamma limit. In particular, it investigates the question when the convergence of the first four moments of such a sequence to the 'correct' moments of the limiting distribution is equivalent to the sequence converging in distribution. On the other hand, in Fissler and Thäle (2016b), we considered quantitative central limit theorems for a sequence of Gaussian functionals with a possibly infinite chaos representation at (7.1.2). That means, we established rates of convergence for the distance of the respective distribution of a sequence of Gaussian functionals $(F_n)_{n\in\mathbb{N}}$ from a normal law in terms of different probability metrics such as the total variation distance, the Kolmogorov distance or the Wasserstein distance. Besides the Malliavin Calculus, the results of both papers deeply rely on *Stein's method* for the normal and the Gamma approximation (and, more precisely, on the combination of the two methods). Valuable references for Stein's method are Nourdin and Peccati (2012, Chapter 3) as well as Chen et al. (2011) and Bourguin and Peccati (2016) for the normal approximation; for the Gamma approximation, we refer to the PhD-theses Luk (1994), Pickett (2004) as well as to Chen et al. (2011, Example 13.2), and to the articles Gaunt et al. (2016) and Ley et al. (2017). Since we did not elaborate on Stein's method in the two subsequent papers Fissler and Thäle (2016a,b), we give here a brief summary of the heuristic behind Stein's method restricting on the case of the normal approximation and following the exposition in Nourdin and Peccati (2012, Chapter 3), remarking that the rationale behind the Gamma approximation via Stein's method is similar in nature.

Stein's method for normal approximation

The starting point of Stein's method for normal approximation is the following characterization of the law of a standard normal random variable, known as *Stein's lemma*.

Lemma 7.2.1 (Stein's lemma)**.** *A real-valued random variable X has the distribution of a standard normal random variable $N \sim \mathcal{N}(0,1)$ if and only if, for every differentiable function $f\colon \mathbb{R} \to \mathbb{R}$ such that $\mathbb{E}[|f'(N)|] < \infty$, the expectations $\mathbb{E}[|Xf(X)|]$ and $\mathbb{E}[|f'(X)|]$ are finite, and*

$$\mathbb{E}[f'(X) - Xf(X)] = 0. \tag{7.2.1}$$

A natural question is whether there is a quantitative version of Stein's lemma in the sense that if the left hand side of (7.2.1) is 'close' to zero for a sufficiently rich class of test functions f, then the law of X is also 'close' to the law of $N \sim \mathcal{N}(0,1)$. To give this vague statement a clearer meaning, recall the notion of a probability metric $d_{\mathcal{H}}$ between the laws of two random variables X, Y

$$d_{\mathcal{H}}(X,Y) := \sup_{h\in\mathcal{H}} |\mathbb{E}[h(X) - \mathbb{E}[h(Y)]| \,. \tag{7.2.2}$$

Here, $\mathcal{H}$ is a certain *separating* class of real-valued test functions h such that $\mathbb{E}[|h(X)|] < \infty$ and $\mathbb{E}[|h(Y)|] < \infty$ for all $h \in \mathcal{H}$. Recall that $\mathcal{H}$ is called separating if $d_{\mathcal{H}}(X,Y) = 0$ implies that the laws of X and Y coincide. Many well known metrics can be written in such a form, amongst them

- the *total variation distance* $d_{TV} := d_{\mathcal{H}_{TV}}$, where $\mathcal{H}_{TV} = \{\mathbb{1}_B \colon B \subseteq \mathbb{R} \text{ a Borel set}\}$;
- the *Kolmogorov distance* $d_K := d_{\mathcal{H}_K}$, where $\mathcal{H}_K = \{\mathbb{1}_{(-\infty,x]} \colon x \in \mathbb{R}\}$;
- the *Wasserstein distance* $d_W := d_{\mathcal{H}_W}$, where $\mathcal{H}_W = \{h\colon \|h\|_{Lip} \le 1\}$ with

$$\|h\|_{Lip} := \sup_{x\neq y} \frac{|h(x) - h(y)|}{|x-y|}.$$

To give a link between these probability distances and the left hand side of (7.2.1), *Stein's equation* comes into play. This is the ordinary differential equation

$$f'(x) - xf(x) = h(x) - \mathbb{E}[h(N)], \qquad x \in \mathbb{R}. \tag{7.2.3}$$

where $N \sim \mathcal{N}(0,1)$, $h\colon \mathbb{R} \to \mathbb{R}$ is a Borel-measurable function such that $\mathbb{E}[|h(N)|] < \infty$. For h fixed, the solution to Stein's equation (7.2.3) is denoted by f_h. A solution f_h which is such that $\mathbb{E}[|f_h(N)|] < \infty$ for $N \sim \mathcal{N}(0,1)$, is of the form

$$f_h(x) = e^{x^2/2} \int_{-\infty}^{x} \big(h(y) - \mathbb{E}[h(N)]\big)\, e^{-y^2/2}\, \mathrm{d}y.$$

For a certain class $\mathcal{H}$, one can determine a class of possible solutions $\mathscr{F}_{\mathcal{H}} \supseteq \{f_h \colon h \in \mathcal{H}\}$, such that one ends up with the estimate

$$\begin{aligned} d_{\mathcal{H}}(X,N) &= \sup_{h\in\mathcal{H}} |\mathbb{E}[h(X)] - \mathbb{E}[h(N)]| \\ &= \sup_{h\in\mathcal{H}} |\mathbb{E}[f_h'(X) - Xf_h(X)]| \\ &\le \sup_{f\in\mathscr{F}_{\mathcal{H}}} |\mathbb{E}[f'(X) - Xf(X)]|, \end{aligned} \tag{7.2.4}$$

where X is a generic random variable and $N \sim \mathcal{N}(0,1)$. A key observation is that for $\mathcal{H} \in \{\mathcal{H}_{TV}, \mathcal{H}_K, \mathcal{H}_W\}$, there is a constant $K > 0$ such that

$$\mathscr{F}_{\mathcal{H}} = \{f \colon \|f'\|_\infty \le K\}. \tag{7.2.5}$$

Combining Stein's method with the Malliavin Calculus

The inequality at (7.2.4) gains its power by combining it with the Malliavin Calculus. E.g. if $X = F \in L^2(\sigma(G), \mathbb{P})$ is a Gaussian functional satisfying certain regularity conditions, one can show the identity

$$\mathbb{E}[Ff(F)] = \mathbb{E}\big[f'(F)\langle DF, -DL^{-1}F\rangle_{L^2(\mu)}\big]. \tag{7.2.6}$$

Hence, combining (7.2.4) and (7.2.5), one obtains for $d_{\mathcal{H}} \in \{d_{TV}, d_K, d_W\}$ (Nourdin and Peccati, 2012, Proposition 5.1.1)

$$d_{\mathcal{H}}(F,N) \le K\,\mathbb{E}\big[|1 - \langle DF, -DL^{-1}F\rangle_{L^2(\mu)}|\big]. \tag{7.2.7}$$

The power of the estimate (7.2.7) is that on its right hand side, no test function h or f_h shows up, which makes it easier to discuss this bound.

For the case that $F \in L^2(\sigma(\eta), \mathbb{P})$ is a Poisson functional, one can derive a similar bound like (7.2.7) in terms of the Wasserstein distance. In particular, if $\mathbb{E}[F] = 0$ and $F \in \operatorname{dom} D$, Peccati et al. (2010, Theorem 3.1) yields that

$$d_W(F,N) \le \mathbb{E}\big[|1 - \langle DF, -DL^{-1}F\rangle_{L^2(\mu)}|\big] + \int_{\mathcal{Z}} \mathbb{E}\big[|D_zF|^2\,|D_zL^{-1}F|\big]\mu(\mathrm{d}z). \tag{7.2.8}$$

We remark that similar estimates like (7.2.7) and (7.2.8) can be obtained also for the Gamma approximation; see Nourdin and Peccati (2009, Theorem 3.11) for Gaussian functionals and Peccati and Thäle (2013, Theorem 2.1) for Poisson functionals.

The two subsequent research articles are a modest part of overall more than 250 research papers since 2004, dedicated to and relying on the combination of the Malliavin Calculus and Stein's method. For a bibliographical overview, we refer to the constantly updated webpage https://sites.google.com/site/malliavinstein/home.

8. A four moments theorem for Gamma limits on a Poisson chaos

The content of this chapter is the joint research article Fissler and Thäle (2016a). The main concern of this article is to establish a non-central 'four moments theorem' for Poisson functionals belonging to a fixed chaos. That is, we investigated under which conditions the convergence in distribution of a sequence of Poisson functionals on a fixed chaos to a Gamma limit is equivalent to the convergence of the first four moments of that sequence to the corresponding moments of the Gamma limit.

We included the journal version which appeared in *ALEA*, the *Latin American Journal of Probability and Mathematical Statistics*, which is available at `http://alea.impa.br/english/index_v13.htm`.

After the publication (and just before finishing this thesis, in January 2017), Giovanni Peccati and Christian Döbler pointed out to us that condition (b) of the main Theorem 3.5 implies that assertion (i) of the stated equivalence can actually never be true. That is, the sequence of multiple Wiener-Ito integrals with respect to a Poisson measure cannot converge to a centered Gamma limit, if the kernels are non-positive. Strictly speaking, the assertion of Theorem 3.5(b) is still true in the sense that the equivalence still holds. However, the result lost its original motivation to give a sufficient condition for convergence in distribution to a Gamma limit in terms of the convergence of the first four moments to the 'correct' moments. To avoid misunderstandings, we submitted an erratum for that paper which can be found directly after the paper, giving a new condition under which the convergence can actually hold and which is in line with the recent findings of Döbler and Peccati (2017).

ALEA, Lat. Am. J. Probab. Math. Stat. 13 (1), 163–192 (2016)

A four moments theorem for Gamma limits on a Poisson chaos

Tobias Fissler[1] **and Christoph Thäle**[2]

University of Bern
Institute of Mathematical Statistics
and Actuarial Science
Sidlerstrasse 5
CH–3012 Bern, Switzerland.
E-mail address: tobias.fissler@stat.unibe.ch
URL: http://www.imsv.unibe.ch/about_us/staff/fissler_tobias/index_eng.html

Ruhr-University Bochum
Faculty of Mathematics, NA 3/68
D–44781 Bochum, Germany.
E-mail address: christoph.thaele@rub.de
URL: http://www.ruhr-uni-bochum.de/ffm/Lehrstuehle/Thaele/index.html

Abstract. This paper deals with sequences of random variables belonging to a fixed chaos of order q generated by a Poisson random measure on a Polish space. The problem is investigated whether convergence of the third and fourth moment of such a suitably normalized sequence to the third and fourth moment of a centred Gamma law implies convergence in distribution of the involved random variables. A positive answer is obtained for $q = 2$ and $q = 4$. The proof of this four moments theorem is based on a number of new estimates for contraction norms. Applications concern homogeneous sums and U-statistics on the Poisson space.

1. Introduction

Probabilistic limit theorems for sequences of multiple stochastic integrals have found considerable attention during the last decade. One of the most remarkable results in this direction is the fourth moment theorem of Nualart and Peccati (2005). It asserts that a sequence of suitably normalized multiple stochastic integrals of order $q \geq 2$ with respect to a Gaussian random measure on a Polish space

Received by the editors February 5, 2015; accepted February 2, 2016.

2010 *Mathematics Subject Classification.* Primary: 60F05, 60H05, 60G57; Secondary: 60G55, 62E20.

Key words and phrases. Contraction, four moments theorem, Gamma distribution, homogeneous sums, multiple stochastic integrals, non-central limit theorem, Poisson chaos, Poisson random measure, universality, U-statistics.

[1]Supported by the Swiss National Science Foundation (SNF) via grant 152609.

[2]Supported by the German Research Foundation (DFG) via SFB-TR 12.

satisfies a central limit theorem if and only if the sequence of their fourth moments converges to 3, the fourth moment of a standard Gaussian distribution. This drastic simplification of the method of moments has stimulated a large number of applications, for example to Gaussian random processes or fields, mathematical statistics, random matrices or random polynomials (we refer the reader to the monograph Nourdin and Peccati (2012) and also to the constantly updated webpage `https://sites.google.com/site/malliavinstein/home` for further details and references).

Besides the fourth moment theorem mentioned above, there is also a 'non-central' version dealing with the approximation of a sequence of multiple stochastic integrals by a centred Gamma-distributed random variable, cf. Nourdin and Peccati (2009). Again, the result is a drastic simplification of the method of moments as it delivers convergence in distribution if and only if a certain linear combination of the third and the fourth moment of the involved random variables converges to the corresponding expression for centred Gamma random variables. In view of normalization conditions we see that in fact the first four moments of the random variables are involved, which gives rise to the name 'four moments theorem' for such a result. To simplify the terminology, we will also speak about a four moments theorem in the case of normal approximation.

The present paper asks whether a similar non-central limit theorem is available for sequences of multiple stochastic integrals with respect to a Poisson random measure on a Polish space. In this set-up, a central four moments theorem has been derived in Lachièze-Rey and Peccati (2013) under an additional sign condition (see also Eichelsbacher and Thäle, 2014), which, on the Poisson space, seems to be unavoidable. While Gamma approximation on the Poisson space in the spirit of the Malliavin-Stein method has been dealt with in Peccati and Thäle (2013), the problem of a four moments theorem similar to that for Gaussian multiple stochastic integrals mentioned above remained open in general. The main result of our paper, Theorem 3.5, delivers a four moments theorem for sequences of Poisson stochastic integrals of order $q = 2$ and $q = 4$. For this reason, the present work can be seen as a natural continuation of Peccati and Thäle (2013), where the case $q = 2$ has already been settled under additional assumptions, which we are able to overcome. The proof of our four moments theorem relies on a couple of new estimates for norms of so-called contraction kernels and the combinatorially involved multiplication formula for stochastic interals on the Poisson space. It is precisely this combinatorial complexity which allows us to obtain positive result only for sequences of Poisson stochastic integrals of order $q = 2$ and $q = 4$. However, all intermediate steps in our proof will be formulated for general $q \geq 2$ to make as transparent as possible and to highlight, in which argument the restrictive condition on the order of the integrals arises. The main difference between the central and the non-central version of the four moments theorem is that in the non-central case one has to deal with a linear combination of the third and the fourth moment of the stochastic integrals, while the central case only requires an analysis of the fourth moment. Even under additional conditions on the integrands, this leads to difficulties, which we can overcome only for $q = 2$ and $q = 4$. We have to leave it as an open problem for future research to extend our result to arbitrary even q by other methods. In the case of Gaussian stochastic integrals, one can *a priori* exclude that an integral of *odd* order converges in distribution to a Gamma-type limit; see Nourdin and Peccati

(2009, Remark 1.3). However, it remains unclear whether a Poisson integral of odd order can or cannot converge to such a Gamma-type limit. We also have to leave this issue as another open problem.

The main result of our paper is applied to a universality question for homogeneous sums on a Poisson chaos as well as to a non-central analogue of de Jong's theorem for completely degenerate U-statistics of order two and four. This partially complements the results for Gamma and normal approximation obtained in Eichelsbacher and Thäle (2014), Peccati and Thäle (2013) and Peccati and Zheng (2014). We emphasize in this context that limit theorems for non-linear functionals of Poisson random measures have recently found numerous applications especially in geometric probability or stochastic geometry; see Eichelsbacher and Thäle (2014), Lachièze-Rey and Peccati (2013), Last et al. (2014), Last et al. (2015+), Peccati and Thäle (2013), Schulte and Thäle (2012), Schulte and Thäle (2014) and in the theory of Lévy processes Eichelsbacher and Thäle (2014), Last et al. (2015+), Peccati et al. (2010), Peccati and Zheng (2010).

Our paper is structured as follows. In Section 2, we introduce and collect necessary background material. To contrast our results with those available for Gaussian multiple stochastic integrals, we shall present them in the context of completely random measures, which captures both settings. Our main results are the content of Section 3, while Section 4 contains applications to homogeneous sums and U-statistics. The proof of Theorem 3.5 is presented in the final Section 5.

2. Preliminaries

In this section, we introduce the basic definitions, mainly related to Poisson stochastic integrals. For further details and background material we refer the reader to the monograph Peccati and Taqqu (2011) as well as to the papers Nualart and Vives (1990) and Peccati et al. (2010).

2.1. *Completely random measures.* Without loss of generality, we assume that all objects are defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let $\mathcal{Z}$ denote a Polish space with Borel σ-field $\mathscr{Z}$, which is equipped with a non-atomic σ-finite measure μ. We define the class $\mathscr{Z}_\mu = \{B \in \mathscr{Z} : \mu(B) < \infty\}$ and let $\varphi = \{\varphi(B) : B \in \mathscr{Z}_\mu\}$ indicate a *completely random measure* on $(\mathcal{Z}, \mathscr{Z})$ with control measure μ. That is, φ is a set of random variables such that

(i) for every collection of pairwise disjoint elements $B_1, \dots, B_n \in \mathscr{Z}_\mu$, the random variables $\varphi(B_1), \dots, \varphi(B_n)$ are independent;

(ii) for every $B, C \in \mathscr{Z}_\mu$, one has the identity $\mathbb{E}[\varphi(B)\varphi(C)] = \mu(B \cap C)$.

If $\mathbb{E}[\varphi(B)] = 0$ and $\varphi(B) \in L^2(\mathbb{P})$ (i.e., $\varphi(B)$ is square-integrable with respect to $\mathbb{P}$) for every $B \in \mathscr{Z}_\mu$, then the mapping $\mathscr{Z}_\mu \to L^2(\mathbb{P}), B \mapsto \varphi(B)$, is σ-additive in the sense that for every sequence $(B_n)_{n\geq 1}$ of pairwise disjoint elements of $\mathscr{Z}_\mu$, one has that

$$\varphi\left(\bigcup_{n=1}^{\infty} B_n\right) = \sum_{n=1}^{\infty} \varphi(B_n) \qquad \mathbb{P}\text{-a.s.,} \tag{2.1}$$

where the right-hand side converges in $L^2(\mathbb{P})$. By $\sigma(\varphi)$ we denote the σ-field generated by φ.

In this paper, we shall deal with two special and prominent instances of completely random measures, namely a centred Gaussian and a compensated Poisson measure.

(a) A *centred Gaussian measure* with control measure μ is denoted by G and is a completely random measure such that the elements of G are jointly Gaussian and centred.

(b) A *compensated Poisson measure* with control measure μ is indicated by $\hat{\eta}$ and is a completely random measure such that for every $B \in \mathscr{Z}_\mu$, $\hat{\eta}(B) \overset{d}{=} \eta(B) - \mu(B)$, where $\eta(B)$ is a Poisson random variable with mean $\mu(B)$.

By definition, both G and $\hat{\eta}$ are centred families in $L^2(\mathbb{P})$, implying that (2.1) is satisfied. Moreover, for $\mathbb{P}$-almost every $\omega \in \Omega$, $\hat{\eta}(\cdot, \omega)$ is a signed measure on $(\mathcal{Z}, \mathscr{Z})$, while G does not satisfy this property, cf. Peccati and Taqqu (2011, Example 5.1.7 (iii)).

2.2. *L^2-spaces.* Let $q \geq 1$ be an integer. We shall use the shorthand notation $L^2(\mu^q)$ for the space $L^2(\mathcal{Z}^q, \mathscr{Z}^q, \mu^q)$ of (deterministic) functions that are square-integrable with respect to μ^q. The symbol $L^2_s(\mu^q)$ stands for the subspace of $L^2(\mu^q)$ consisting of symmetric functions, i.e. functions that are μ^q-a.e. invariant under permutations of their arguments. For $f, g \in L^2(\mu^q)$ we define the scalar product $\langle f, g\rangle_{L^2(\mu^q)} = \int_{\mathcal{Z}^q} fg \, \mathrm{d}\mu^q$ and the norm $\|f\|_{L^2(\mu^q)} = \langle f, f\rangle^{1/2}_{L^2(\mu^q)}$. If there is no risk of confusion, we suppress in what follows the dependency on q and μ, and merely write $\langle \cdot, \cdot \rangle$ and $\|\cdot\|$, respectively. Moreover, let $L^2(\sigma(\varphi), \mathbb{P})$ denote the space of all square-integrable functionals of φ, where φ is either a Poisson measure $\hat{\eta}$ or a Gaussian measure G. If $F \in L^2(\sigma(\varphi), \mathbb{P})$, we shall sometimes write $F = F(\varphi)$ in order to underpin the dependency of F on φ. As a convention, we shall use lower case variables for elements of $L^2(\mu^q)$ and capitals for elements of $L^2(\sigma(\varphi), \mathbb{P})$. Finally, we introduce the space $L^2(\mathbb{P}, L^2(\mu)) = L^2(\Omega \times \mathcal{Z}, \mathcal{F} \otimes \mathscr{Z}, \mathbb{P} \otimes \mu)$ as the space of all jointly square-integrable measurable mappings $u \colon \Omega \times \mathcal{Z} \to \mathbb{R}$. If $u, v \in L^2(\mathbb{P}, L^2(\mu))$, their scalar product is defined as $\langle u, v\rangle_{L^2(\mathbb{P}, L^2(\mu))} = \int_\Omega \int_{\mathcal{Z}} u(\omega, z) v(\omega, z) \mu(\mathrm{d}z) \mathbb{P}(\mathrm{d}\omega)$ and we denote by $\| \cdot \|_{L^2(\mathbb{P}, L^2(\mu))}$ the norm induced by it.

2.3. *Multiple stochastic integrals.* Let $\varphi = \hat{\eta}$ or $\varphi = G$. For every integer $q \geq 1$ we denote the *multiple stochastic integral* of order q with respect to φ by I^φ_q. It is a mapping $I^\varphi_q \colon L^2_s(\mu^q) \to L^2(\sigma(\varphi), \mathbb{P})$, which is linear and continuous. Additionally, for $f \in L^2_s(\mu^q)$, the random variable $I^\varphi_q(f)$ is centred. Moreover, the multiple stochastic integral satisfies the *Itô isometry*

$$\mathbb{E}[I^\varphi_p(f) I^\varphi_q(g)] = \begin{cases} 0 & \text{if } q \neq p \\ q! \langle f, g\rangle_{L^2_s(\mu^q)} & \text{if } q = p \end{cases} \tag{2.2}$$

for any integers $p, q \geq 1$ and $f \in L^2_s(\mu^p)$, $g \in L^2_s(\mu^q)$. For general $f \in L^2(\mu^q)$, we put $I^\varphi_q(f) = I^\varphi_q(\widetilde{f})$, where

$$\widetilde{f}(z_1, \ldots, z_q) = \frac{1}{q!} \sum_{\pi \in \Pi_q} f(z_{\pi(1)}, \ldots, z_{\pi(q)})$$

is the canonical symmetrization of f, and Π_q is the group of all $q!$ permutations π of the set $\{1, \ldots, q\}$. We emphasize that due to Jensen's inequality and the convexity of norms, we have the inequality $\|\widetilde{f}\| \leq \|f\|$. As a convention, we set $I_0^\varphi \colon \mathbb{R} \to \mathbb{R}$ equal to the identity map on $\mathbb{R}$.

Since this article is mostly concerned with Poisson integrals, we shall write I_q instead of $I_q^{\hat{\eta}}$.

2.4. *Chaos decomposition.* The Itô isometry in (2.2) formalizes an orthogonality relation between multiple stochastic integrals of different order. Even more, one has the following so-called *chaos decomposition* (see Nualart and Vives, 1990):

$$L^2(\sigma(\varphi), \mathbb{P}) = \bigoplus_{q=0}^{\infty} W_q^\varphi \,, \tag{2.3}$$

where $W_0^\varphi = \mathbb{R}$ and $W_q^\varphi = \{I_q^\varphi(f) \colon f \in L_s^2(\mu^q)\}$ for $\varphi = \hat{\eta}$ or $\varphi = G$, $q \geq 1$. Depending on the choice of φ, we shall often use the terms *Poisson chaos* and *Gaussian chaos of order* q for W_q^φ, respectively.

A consequence of (2.3) is that any $F \in L^2(\sigma(\varphi), \mathbb{P})$, with $\varphi = \hat{\eta}$ or $\varphi = G$, admits a chaotic decomposition

$$F = \mathbb{E}[F] + \sum_{q=1}^{\infty} I_q^\varphi(f_q) \,,$$

where the kernels $f_q \in L_s^2(\mu^q)$ are unique μ^q-a.e. and the series converges in $L^2(\mathbb{P})$.

2.5. *Contractions.* Fix integers $p, q \geq 1$ and functions $f \in L_s^2(\mu^p)$, $g \in L_s^2(\mu^q)$. For any $r \in \{0, \dots, p \wedge q\}$, $\ell \in \{1, \dots, r\}$ we define the *contraction* $f \star_r^\ell g \colon \mathcal{Z}^{p+q-r-\ell} \to \mathbb{R}$ which acts on the tensor product $f \otimes g$ and reduces the number of variables from $p+q$ to $p+q-r-\ell$ in the following way: r variables are identified and among these, ℓ are integrated out with respect to μ. More formally,

$$\begin{aligned} f \star_r^\ell g(\gamma_1, \dots, \gamma_{r-\ell}, t_1, \dots, t_{p-r}, s_1, \dots, s_{q-r}) \\ = \int_{\mathcal{Z}^\ell} f(z_1, \dots, z_\ell, \gamma_1, \dots, \gamma_{r-\ell}, t_1, \dots, t_{p-r}) \\ \times g(z_1, \dots, z_\ell, \gamma_1, \dots, \gamma_{r-\ell}, s_1, \dots, s_{q-r}) \, \mu^\ell(\mathrm{d}(z_1, \dots, z_\ell)) \,, \end{aligned}$$

and for $\ell = 0$ we put

$$\begin{aligned} f \star_r^0 g(\gamma_1, \dots, \gamma_r, t_1, \dots, t_{p-r}, s_1, \dots, s_{q-r}) \\ = f(\gamma_1, \dots, \gamma_r, t_1, \dots, t_{p-r}) g(\gamma_1, \dots, \gamma_r, s_1, \dots, s_{q-r}) \,. \end{aligned}$$

Note that even if f and g are symmetric, the contraction $f \star_r^\ell g$ is not necessarily symmetric. We denote the canonical symmetrization of $f \star_r^\ell g$ by

$$f \,\widetilde{\star}_r^\ell g(z_1, \dots, z_{p+q-r-\ell}) = \frac{1}{(p+q-r-\ell)!} \sum_{\pi \in \Pi_{p+q-r-\ell}} f \star_r^\ell g(z_{\pi(1)}, \dots, z_{\pi(p+q-r-\ell)}).$$

We also emphasize that for $f \in L_s^2(\mu^p)$ and $g \in L_s^2(\mu^q)$, the contraction $f \star_r^\ell g$ is neither necessarily well-defined nor necessarily an element of $L^2(\mu^{p+q-r-\ell})$. At least, by using the Cauchy-Schwarz inequality, we can deduce that $f \star_r^r g \in L^2(\mu^{p+q-2r})$ for any $r \in \{0, \dots, p \wedge q\}$. For this reason and to circumvent any complications in the calculations, we make the following technical assumptions.

2.6. *Technical assumptions (A).* We use the same set of technical assumptions as in Lachièze-Rey and Peccati (2013), Peccati et al. (2010) and Peccati and Thäle (2013). For a detailed explanation of the conditions and their consequences, we refer to these works.

For a sequence $F_n = I_q(f_n)$, $n \geq 1$, of multiple integrals of fixed order $q \geq 1$ with $f_n \in L^2(\mu_n^q)$ for every $n \geq 1$ (we allow the non-atomic and σ-finite measure to vary with n), we assume that the following three technical conditions are satisfied:

(a) for any $r \in \{1, \ldots, q\}$, the contraction $f_n \star_q^{q-r} f_n$ is an element of $L^2(\mu_n^r)$;

(b) for any $r \in \{1, \ldots, q\}$, $\ell \in \{1, \ldots, r\}$ and $(z_1, \ldots, z_{2q-r-\ell}) \in \mathcal{Z}^{2q-r-\ell}$, the quantity $(|f_n| \star_r^\ell |f_n|)(z_1, \ldots, z_{2q-r-\ell})$ is well-defined and finite;

(c) for any $k \in \{0, \ldots, 2(q-1)\}$ and any r and ℓ satisfying $k = 2(q-1) - r - \ell$, we have that

$$\int_{\mathcal{Z}} \left(\int_{\mathcal{Z}^k} \big(f_n(z, \cdot) \star_r^\ell f_n(z, \cdot) \big)^2 \mathrm{d}\mu_n^k \right)^{1/2} \mu_n(\mathrm{d}z) < \infty \,.$$

2.7. *Multiplication formula.* A very convenient property of multiple stochastic integrals is that one can express the product of two such integrals as a linear combination of multiple integrals of contraction kernels. More precisely, we have the following multiplication formula for Poisson integrals, which is taken from Last (2014, Proposition 6.1), but see also Peccati and Taqqu (2011, Proposition 6.5.1) for a version that holds under stronger integrability assumptions and for diffuse controls μ only.

Lemma 2.1 (Multiplication formula for Poisson integrals)**.** *Let $f \in L_s^2(\mu^p)$ and $g \in L_s^2(\mu^q)$, $p, q \geq 1$. Suppose that $f \star_r^\ell g \in L^2(\mu^{p+q-r-\ell})$ for every $r \in \{0, \ldots, p \wedge q\}$ and every $\ell \in \{0, \ldots, r\}$. Then*

$$I_p(f) I_q(g) = \sum_{r=0}^{p \wedge q} r! \binom{p}{r} \binom{q}{r} \sum_{\ell=0}^{r} I_{p+q-r-\ell}(f \,\widetilde{\star}_r^\ell\, g). \tag{2.4}$$

We remark that if a kernel $f \in L_s^2(\mu^q)$ satisfies the technical assumptions (A), the assumptions of Lemma 2.1 are automatically satisfied if $g = f$, implying that $I_q(f)^2 \in L^2(\sigma(\hat{\eta}), \mathbb{P})$. To simplify our notation, for $f \in L_s^2(\mu^q)$ we put $G_0^q f = q! \|f\|^2$ and

$$G_p^q f = \sum_{r=0}^{q} \sum_{\ell=0}^{r} \mathbf{1}(2q - r - \ell = p) r! \binom{q}{r}^2 \binom{r}{\ell} f \,\widetilde{\star}_r^\ell\, f \tag{2.5}$$

for $p \in \{1, \ldots, 2q\}$. In other words, the operator G_p^q turns a function of q variables into a function of p variables. We can now re-write (2.4) in a simplified form as

$$I_q(f)^2 = \sum_{p=0}^{2q} I_p(G_p^q f)$$

with $I_0(G_0^q f) = G_0^q f = q! \|f\|^2$.

The multiplication formula paves the way for the computation of moments of multiple stochastic integrals. In particular, we have the following expressions for the third and the fourth moment of a multiple Poisson integral.

Lemma 2.2 (Third and fourth moment of Poisson integrals)**.** *Fix an integer $q \geq 1$. Let $f \in L^2_s(\mu^q)$ such that the technical assumptions (A) are satisfied. Then $I_q(f) \in L^4(\mathbb{P})$. Moreover, we have that*

$$\mathbb{E}[I_q(f)^3] = q! \sum_{r=0}^{q} \sum_{\ell=0}^{r} \mathbf{1}(q = r+\ell) r! \binom{q}{r}^2 \binom{r}{\ell} \langle f \,\widetilde{\star}_r^\ell f, f \rangle, \tag{2.6}$$

$$\mathbb{E}[I_q(f)^4] = \sum_{p=0}^{2q} p! \|G_p^q f\|^2. \tag{2.7}$$

Proof: The technical assumptions (A) ensure that all symmetrized contraction kernels $f \,\widetilde{\star}_r^\ell f$ appearing in (2.6) and (2.7) are elements of $L^2(\mu^{2q-r-\ell})$, which implies that the third and the fourth moment of $I_q(f)$ are finite. The explicit formulae in (2.6) and (2.7) follow directly from the isometry property (2.2) and the multiplication formula (2.4). □

Remark 2.3. Note that for even $q \geq 2$, (2.6) reduces to

$$\mathbb{E}[I_q(f)^3] = q! \sum_{r=q/2}^{q} r! \binom{q}{r}^2 \binom{r}{q-r} \langle f \,\widetilde{\star}_r^{q-r} f, f \rangle \,. \tag{2.8}$$

Remark 2.4. There is also a multiplication formula for the Gaussian case. It reads

$$I_p^G(f) I_q^G(g) = \sum_{r=0}^{p \wedge q} r! \binom{p}{r} \binom{q}{r} I^G_{p+q-2r}(f \,\widetilde{\star}_r^r g),$$

where $p, q \geq 1$ and $f \in L^2_s(\mu^p)$, $g \in L^2_s(\mu^q)$, see Peccati and Taqqu (2011, Proposition 6.4.1). As a consequence, we see that the third and fourth moment of a Gaussian multiple integral have a more compact form compared to the Poisson case. Indeed, for an integer $q \geq 1$ and $f \in L^2_s(\mu^q)$, one has that

$$\mathbb{E}[I_q^G(f)^3] = \frac{(q!)^3}{(q/2!)^2} \langle f \,\widetilde{\star}_{q/2}^{q/2} f, f \rangle \, \mathbf{1}(q \text{ is even}) \,, \tag{2.9}$$

$$\mathbb{E}[I_q^G(f)^4] = \sum_{r=0}^{q} (r!)^2 \binom{q}{r}^4 (2q-2r)! \|f \,\widetilde{\star}_r^r f\|^2 \,.$$

In particular, the third moment of a Gaussian integral of odd order vanishes, while this is in general not the case for a Poisson integral.

3. Four moments theorems

This section contains the main results of our paper, namely a four moments theorem for Gamma approximation on a Poisson chaos of fixed order. To allow for an easier comparison with the existing literature, we first recall known results on a Gaussian chaos and also a version of the four moments theorem for normal approximation on a Poisson chaos.

3.1. *Four moments theorems on a Gaussian chaos.* The classical method of moments yields a central limit theorem for a normalized sequence of random variables under the condition that all moments converge to those of the standard Gaussian distribution. The four moments theorem on a Gaussian chaos is a drastical simplification of the method of moments as it provides a central limit theorem for a sequence of normalized Gaussian multiple stochastic integrals under the much weaker condition that only the fourth moment converges to 3 (which is the fourth moment of the standard Gaussian distribution). Alternatively, this statement can be re-formulated in terms of the convergence of norms of contractions. In what follows we write $X \sim \mathcal{L}$ if a random variable X has distribution $\mathcal{L}$.

Theorem 3.1 (see Theorem 1 in Nualart and Peccati, 2005)**.** *Fix an integer $q \geq 2$ and let $\{f_n : n \geq 1\} \subset L_s^2(\mu^q)$ be such that*

$$\lim_{n\to\infty} q!\|f_n\|^2 = \lim_{n\to\infty} \mathbb{E}[I_q^G(f_n)^2] = 1\,.$$

Further, let $N \sim \mathcal{N}(0,1)$ be a standard Gaussian random variable. Then the following three assertions are equivalent:

(i) As $n \to \infty$, the sequence $\{I_q^G(f_n) : n \geq 1\}$ converges in distribution to N;

(ii) $\lim_{n\to\infty} \mathbb{E}[I_q^G(f_n)^4] = 3$;

(iii) $\lim_{n\to\infty} \|f_n \star_r^r f_n\| = 0$ for every $r \in \{1,\ldots,q-1\}$.

In the subsequent work Nourdin and Peccati (2009), the authors have shown a 'non-central' version of Theorem 3.1 where the limiting distribution is a centred Gamma distribution. To state the result properly, let us recall the formal definition of the latter limit law.

Definition 3.2 (Centred Gamma distribution)**.** A random variable Y has a centred Gamma distribution $\overline{\Gamma}_\nu$ with parameter $\nu > 0$, if

$$Y \stackrel{d}{=} 2X - \nu,$$

where X has the usual Gamma law with mean and variance both equal to $\nu/2$ and where $\stackrel{d}{=}$ stands for equality in distribution. The probability density of $\overline{\Gamma}_\nu$ is given by

$$g_\nu(x) = \frac{2^{-\nu/2}}{\Gamma(\nu/2)}(x+\nu)^{\nu/2-1}e^{-(x+\nu)/2}\mathbf{1}(x > -\nu),$$

and the the first four moments of Y are

$$\mathbb{E}[Y] = 0\,, \quad \mathbb{E}[Y^2] = 2\nu\,, \quad \mathbb{E}[Y^3] = 8\nu\,, \quad \mathbb{E}[Y^4] = 12\nu^2 + 48\nu\,.$$

We are now in the position to re-phrase the following non-central analogue of Theorem 3.1.

Theorem 3.3 (see Theorem 1.2 in Nourdin and Peccati, 2009)**.** *Let $\nu > 0$ and fix an even integer $q \geq 2$. Let $\{f_n : n \geq 1\} \subset L_s^2(\mu^q)$ be such that*

$$\lim_{n\to\infty} q!\|f_n\|^2 = \lim_{n\to\infty} \mathbb{E}[I_q^G(f_n)^2] = 2\nu\,.$$

Further, let $Y \sim \overline{\Gamma}_\nu$ be a centred Gamma-distributed random variable with parameter ν. Then the following three assertions are equivalent:

(i) As $n \to \infty$, the sequence $\{I_q^G(f_n) : n \geq 1\}$ converges in distribution to Y;

(ii) $\lim_{n\to\infty} \mathbb{E}[I_q^G(f_n)^4] - 12\mathbb{E}[I_q^G(f_n)^3] = 12\nu^2 - 48\nu$;

(iii) $\lim_{n\to\infty} \|f_n \star_r^r f_n\| = 0$ for every $r \in \{1,\ldots,q-1\} \setminus \{q/2\}$, and
$\lim_{n\to\infty} \|f_n \widetilde{\star}_{q/2}^{q/2} f_n - c_q\, f_n\| = 0$ with $c_q = \frac{4}{(q/2)!\binom{q}{q/2}^2}$.

It is a characterizing feature of the centred Gamma-distribution that the so-called 'middle-contraction' $f_n \star_{q/2}^{q/2} f_n$ plays a special role in condition (iii). The fact that the middle-contraction does not vanish goes hand in hand with the appearance of the third moment in condition (ii), recall (2.9).

3.2. *Four moments theorems on a Poisson chaos.* We now turn to four moments theorems on a Poisson chaos of fixed order $q \geq 2$. To this end, let, for each $n \geq 1$, μ_n be a σ-finite non-atomic measure on $(\mathcal{Z}, \mathscr{Z})$ and denote by $\hat{\eta}_n$ a compensated Poisson random measure with control μ_n. Further let $\{f_n : n \geq 1\}$ be a sequence of symmetric function such that f_n is square-integrable with respect to μ_n^q for each $n \geq 1$, in short $f_n \in L_s^2(\mu_n^q)$. In this set-up, $\|f_n\|$ denotes the norm of f_n with respect to μ_n^q, and $f_n \star_r^\ell f_n$ stands for the contraction taken with respect to μ_n. Finally, define $F_n = I_q(f_n)$, where for each n the stochastic integral is defined with respect to $\hat{\eta}_n$.

As in the Gaussian case discussed in the previous section, we start with the case of a standard normal limiting distribution.

Theorem 3.4 (see Theorem 3.12 in Lachièze-Rey and Peccati, 2013)**.** *Let $\{\mu_n : n \geq 1\}$ be a sequence of σ-finite and non-atomic measures such that $\lim_{n\to\infty} \mu_n(\mathcal{Z}) = \infty$ and fix $q \geq 2$. Let $f_n \in L_s^2(\mu_n^q)$, $n \geq 1$, be a sequence such that for each $n \geq 1$ either $f_n \geq 0$ or $f_n \leq 0$. Suppose that the technical assumptions (A) and the normalization condition*

$$\lim_{n\to\infty} q!\|f_n\|^2 = \lim_{n\to\infty} \mathbb{E}[I_q(f_n)^2] = 1 \tag{3.1}$$

are satisfied. Further, suppose that $\{I_q(f_n)^4 : n \geq 1\}$ is uniformly integrable and let $N \sim \mathcal{N}(0,1)$ be a standard Gaussian random variable. Then the following three assertions are equivalent:

(i) As $n \to \infty$, the sequence $\{I_q(f_n) : n \geq 1\}$ converges in distribution to N;
(ii) $\lim_{n\to\infty} \mathbb{E}[I_q(f_n)^4] = 3$;
(iii) $\lim_{n\to\infty} \|f_n \star_r^\ell f_n\| = 0$ for all $r \in \{1, \ldots, q\}$ and $\ell \in \{1, \ldots, r \wedge (q-1)\}$, and $\lim_{n\to\infty} \|f_n\|_{L^4(\mu_n^q)} = 0$.

Let us comment on the differences between Theorem 3.1 and Theorem 3.4.

(1) In the Poisson case, one has to ensure that the involved control measures are infinite measures, at least in the limit, as $n \to \infty$. The reason for this is that otherwise, the normalization (3.1) and the condition that $\lim_{n\to\infty} \|f_n\|_{L^4(\mu_n^q)} = 0$ are mutually exclusive, see also the remark after Assumption N in Peccati and Taqqu (2008) for a brief discussion of this problem.
(2) One has to assume that the functions f_n have a constant sign, that is for each $n \geq 1$ either $f_n \geq 0$ or $f_n \leq 0$. The reason for this is that in the Poisson case, besides of the contraction norms $\|f_n \star_r^\ell f_n\|$, also scalar products of the form $\langle f_n \star_{r_1}^{\ell_1} f_n, f_n \star_{r_2}^{\ell_2} f_n \rangle$ enter the expression of the fourth moments $\mathbb{E}[I_q(f_n)^4]$. The sign condition then allows to control the signs of these scalar products, which rules out cancellation effects.
(3) In the Poisson case, one also has to assume that the sequence $\{I_q(f_n)^4 : n \geq 1\}$ is uniformly integrable, while in the Gaussian case, this condition is automatically fulfilled thanks to the hypercontractivity property of Gaussian integrals (see e.g. Nourdin and Peccati, 2012, Theorem 2.7.2). This is

needed to ensure that the convergence in distribution of $I_q(f_n)$ to N implies the convergence of the first four moments.

For general $q \geq 2$ and general sequences $f_n \in L^2_s(\mu_n^q)$, $n \geq 1$, there is no known version of a four moments theorem on a Poisson chaos relaxing one of the conditions discussed above. However, for $q = 2$ the sign condition is not necessary as shown by Theorem 2 in Peccati and Taqqu (2008). Moreover, for general $q \geq 2$ and if the sequence $\{f_n : n \geq 1\}$ is *tamed* (see Definition 4.2 below), Theorem 3.2 in Peccati and Zheng (2014) provides a four moments theorem without a sign condition. In this case, also condition (iii) can be relaxed by assuming – besides the condition on the L^4-norm of f_n – only that $\lim_{n\to\infty} \|f_n \star_r^r f_n\| = 0$ for all $r \in \{1, \ldots, q-1\}$.

After having discussed the four moments theorem for normal approximation on the Poisson space, we now turn to the main result of the present work, namely a version of Theorem 3.3 for Poisson integrals of order $q = 2$ and $q = 4$. The reason for this rather restrictive condition on the order of the involved integrals will be discussed below.

Theorem 3.5 (Four moments theorem for Poisson integrals)**.** *Fix $\nu > 0$. Let $q \geq 2$ be even and $f_n \in L^2_s(\mu_n^q)$, $n \geq 1$, be a sequence satisfying the technical assumptions (A) and the normalization condition*

$$\lim_{n\to\infty} q!\|f_n\|^2 = \lim_{n\to\infty} \mathbb{E}[I_q(f_n)^2] = 2\nu\,.$$

Furthermore, let the sequence $\{I_q(f_n)^4 : n \geq 1\}$ be uniformly integrable and let $Y \sim \overline{\Gamma}_\nu$ be a random variable following a centred Gamma distribution with parameter ν. If one of the conditions

(a) *$q = 2$ and $\lim_{n\to\infty} \|f_n^2\| = 0$,*

(b) *$q = 4$ and $f_n \leq 0$ for all $n \geq 1$*

is satisfied, then the following three assertions are equivalent:

(i) As $n \to \infty$, the sequence $\{I_q(f_n) : n \geq 1\}$ converges in distribution to Y;

(ii) $\lim_{n\to\infty} \mathbb{E}[I_q(f_n)^4] - 12\mathbb{E}[I_q(f_n)^3] = 12\nu^2 - 48\nu$;

(iii) $\lim_{n\to\infty} \|f_n \star_r^\ell f_n\| = 0$ for all $r \in \{1, \ldots, q\}$ and $\ell \in \{1, \ldots, r \wedge (q-1)\}$ such that $(r, \ell) \neq (q/2, q/2)$, $\lim_{n\to\infty} \|f_n\|_{L^4(\mu_n^q)} = 0$, and $\lim_{n\to\infty} \|f_n \,\widetilde{\star}_{q/2}^{q/2} f_n - c_q\, f_n\| = 0$ with $c_q = \frac{4}{(q/2)!\binom{q}{q/2}^2}$.

Remark 3.6. Under condition (a), Theorem 3.5 is a version of Proposition 2.9 in Peccati and Thäle (2013). However, in that paper one has to assume that for each $n \geq 1$ the reference measure μ_n is finite. As discussed earlier in this section, this is a quite restrictive assumption. We provide a proof which circumvents this technicality.

The implication (i) $\Longrightarrow$ (ii) of Theorem 3.5 is a direct consequence of the uniform integrability assumption. That (iii) implies (i) follows from a generalization of Theorem 2.6 in Peccati and Thäle (2013) stated as Proposition 5.1 below. Showing the implication (ii) $\Longrightarrow$ (iii) is the main part of the proof. While the proof of the corresponding implication in Theorem 3.4 is rather straight forward and works for arbitrary $q \geq 2$, the proof here is based on a couple of new estimates and arguments. They are of independent interest and might also be helpful beyond the context of the present paper. In sharp contrast to Theorem 3.4, our arguments show that the 'usual' technique (relying on the multiplication formula for Poisson integrals similar as in the proofs of Theorems 3.1, 3.3 or 3.4) for proving the implication (ii) $\Longrightarrow$ (iii) only works in case that $q = 2$ and $q = 4$ and cannot be improved. The main reason for this is the involved combinatorial structure on a Poisson chaos implied by the multiplication formula (2.4). The proof of Theorem 3.5 is the content of Section 5 below.

Theorem 3.5 has a counterpart in a free probability setting, see Bourguin (2015). Here, one studies the approximation of the law of a sequence of elements belonging to a fixed chaos of order $q \geq 1$ of the so-called *free* Poisson algebra by the Marchenko-Pastur law (also called free Poisson law). It is interesting to see that in this case, the proof works for arbitrary $q \geq 1$ and does not need a sign condition on the kernels. This is explained by the relatively simple combinatorial structure on a free Poisson chaos, which is inherited from the free multiplication formula in which all combinatorial coefficients are equal to one. This causes that the expressions for the third and fourth moment are much simpler compared to the classical set-up of the present paper and implies that a proof of the corresponding free four moments theorem works in full generality.

Comparing Theorem 3.4 and Theorem 3.5, it is natural to ask whether there exists a version of Theorem 3.5 dealing with a sequence of non-negative kernels. Indeed, Corollary 3.8 below provides such a version, but it deals with a different limiting law, namely what we call a *centred reflected Gamma distribution*. In case of a limiting Gaussian law, this phenomonon is not visible, since a Gaussian law is symmetric, see also the discussion in Remark 5.10.

Definition 3.7 (Centred reflected Gamma distribution)**.** A random variable Y has a centred reflected Gamma distribution $\widehat{\Gamma}_\nu$ with parameter $\nu > 0$, if $-Y \sim \overline{\Gamma}_\nu$.

Note that if $Y \sim \widehat{\Gamma}_\nu$ follows a centred reflected Gamma distribution with parameter ν, the first four moments of Y are given by

$$\mathbb{E}[Y] = 0\,, \quad \mathbb{E}[Y^2] = 2\nu\,, \quad \mathbb{E}[Y^3] = -8\nu\,, \quad \mathbb{E}[Y^4] = 12\nu^2 + 48\nu\,.$$

Moreover, while the centred Gamma distribution has support $[-\nu, \infty)$, the centred reflected Gamma distribution is supported on $(-\infty, \nu]$. The next result is an immediate consequence of Theorem 3.5 and the definition of $\widehat{\Gamma}_\nu$.

Corollary 3.8 (Four moments theorem for Poisson integrals with non-negative kernels)**.** *Fix $\nu > 0$. Let $q \geq 2$ be an even integer and $f_n \in L^2_s(\mu^q_n)$, $n \geq 1$, be a sequence of kernels satisfying the technical assumptions (A) and the normalization condition*

$$\lim_{n\to\infty} q!\|f_n\|^2 = \lim_{n\to\infty} \mathbb{E}[I_q(f_n)^2] = 2\nu\,.$$

Let the sequence $\{I^4_q(f_n) \colon n \geq 1\}$ be uniformly integrable and suppose that $Y \sim \widehat{\Gamma}_\nu$ is a random variable having a centred reflected Gamma distribution with parameter ν. If one of the conditions

(a) $q = 2$ and $\lim\limits_{n\to\infty} \|f^2_n\| = 0$,

(b) $q = 4$ and $f_n \geq 0$ for all $n \geq 1$

is satisfied, then the following three assertions are equivalent:

(i) As $n \to \infty$, the sequence $\{I_q(f_n) \colon n \geq 1\}$ converges in distribution to Y;

(ii) $\lim\limits_{n\to\infty} \mathbb{E}[I_q(f_n)^4] + 12\mathbb{E}[I_q(f_n)^3] = 12\nu^2 - 48\nu$;

(iii) $\lim_{n\to\infty} \|f_n \star_r^\ell f_n\| = 0$ *for all* $r \in \{1,\dots,q\}$, $\ell \in \{1,\dots,r\wedge(q-1)\}$ *such that* $(r,\ell) \neq (q/2, q/2)$, $\lim_{n\to\infty}\|f_n^2\| = 0$, *and* $\lim_{n\to\infty}\|f_n \widetilde{\star}_{q/2}^{q/2} f_n + c_q f_n\| = 0$ *with* $c_q = \frac{4}{(q/2)!\binom{q}{q/2}^2}$.

Remark 3.9. We emphasize that one could derive our main result, Theorem 3.5, also for the *two-parametric* centred Gamma distribution $\overline{\Gamma}_{a,\lambda}$, $a, \lambda > 0$, with probability density

$$h_{a,\lambda}(x) = \frac{\lambda^a}{\Gamma(a)}(x + a/\lambda)^{a-1}e^{-(\lambda x + a)}\,\mathbf{1}(x > -a/\lambda).$$

The one-parametric centred Gamma distribution $\overline{\Gamma}_\nu$ then arises by putting $a = \nu/2$ and $\lambda = 1/2$. In order to allow for a better comparison with the existing literature Nourdin and Peccati (2009) and Peccati and Thäle (2013), and to keep the presentation transparent, we have decided to restrict to the one-parametric case.

4. Application to homogeneous sums and U-statistics

4.1. *Homogeneous sums.* According to Peccati and Zheng (2014) a universality result is a 'mathematical statement implying that the asymptotic behaviour of a large random system does not depend on the distribution of its components'. Such results are at the heart of modern probability and the class of examples comprises the classical central limit theorem or the semicircular law in free probability. In this section, we shall derive a universality result for so-called homogeneous sums based on a sequence of independent centred Poisson random variables. For further background material concerning universality results for homogeneous sums we refer to the monograph Nourdin and Peccati (2012) as well as to the original papers Nourdin et al. (2010) and Peccati and Zheng (2014).

We start by introducing the notion of a particularly well-behaved class of kernels.

Definition 4.1 (Index functions)**.** Fix an integer $q \geq 1$. A function $h\colon \mathbb{N}^q \to \mathbb{R}$ is an *index function* of order q, if

(a) h is symmetric in the sense that $h(i_1,\dots,i_q) = h(i_{\pi(1)},\dots,i_{\pi(q)})$ for all $(i_1,\dots,i_q) \in \mathbb{N}^q$ and all permutations $\pi \in \Pi_q$;
(b) h vanishes on diagonals meaning that for $(i_1,\dots,i_q) \in \mathbb{N}^q$, $h(i_1,\dots,i_q) = 0$ whenever $i_k = i_\ell$ for some $k \neq \ell$.

Fix an integer $N \geq 1$. If g and h are two index functions of order q, we define their scalar product by

$$\langle g, h\rangle_{(N,q)} = \sum_{1\leq i_1,\dots,i_q\leq N} g(i_1,\dots,i_q)h(i_1,\dots,i_q)$$

and write $\|h\|_{(N,q)} = \langle h, h\rangle_{(N,q)}^{1/2}$ for the corresponding norm. We frequently suppress the subscript (N,q) if it is clear from the context.

As in Section 3, we denote by $\{\mu_n : n \geq 1\}$ a sequence of σ-finite non-atomic measures on some Polish space $(\mathcal{Z}, \mathscr{Z})$. The following definition should not be confused with the definition of a tamed sequence given in Bourguin (2015) or Bourguin and Peccati (2014).

Definition 4.2 (Tamed sequences)**.** Fix an integer $q \geq 1$. A sequence $f_n \in L^2_s(\mu_n^q)$, $n \geq 1$, is *tamed* if there exists a sequence of integers $\{N_n : n \geq 1\}$ with $N_n \to \infty$, as $n \to \infty$, and an infinite measurable partition $\{B_i : i \geq 1\}$ of $\mathcal{Z}$ verifying the following conditions:

(a) there exists an $\alpha \in (0, \infty)$ such that $\alpha < \mu_n(B_i) < \infty$ for every $i, n \geq 1$,

(b) there is a sequence of index functions $\{h_n : n \geq 1\}$ of order q, such that f_n has the representation

$$f_n(z_1, \dots, z_q) = \sum_{1 \leq i_1, \dots, i_q \leq N_n} h_n(i_1, \dots, i_q) \prod_{k=1}^{q} \frac{\mathbf{1}_{B_{i_k}}(z_k)}{\sqrt{\mu_n(B_{i_k})}} . \tag{4.1}$$

Remark 4.3. (a) It follows from the definition that if a sequence $f_n \in L^2_s(\mu_n^q)$, $n \geq 1$, is tamed, we necessarily must have that $\mu_n(\mathcal{Z}) = \infty$ for every $n \geq 1$.

(b) If $f_n \in L^2_s(\mu_n^q)$, $n \geq 1$, is a tamed sequence with a representation as at (4.1), we have that $\|h_n\|_{(N_n, q)} = \|f_n\|_{L^2(\mu_n^q)} < \infty$.

(c) One easily verifies that tamed sequences automatically satisfy the technical assumptions (A).

Definition 4.4 (Homogeneous sums)**.** Fix integers $N, q \geq 1$ and let $\mathbf{X} = \{X_i : i \geq 1\}$ be a sequence of random variables. Let h be an index function of order q. Then

$$Q_q(N, h, \mathbf{X}) = \sum_{1 \leq i_1, \dots, i_q \leq N} h(i_1, \dots, i_q) X_{i_1} \cdots X_{i_q}$$

is the *homogeneous sum* of h of order q based on the first N elements of $\mathbf{X}$.

If $\mathbf{X} = \{X_i : i \geq 1\}$ is a sequence of independent and centred random variables with unit variance, then

$$\mathbb{E}[Q_q(N, h, \mathbf{X})] = 0, \qquad \mathbb{E}[Q_q(N, h, \mathbf{X})^2] = q! \|h\|^2_{(N,q)}.$$

In what follows, two particular classes of random variables play a special role. By $\mathbf{G} = \{G_i : i \geq 1\}$ we indicate a sequence of independent and identically distributed random variables, such that $G_i \sim \mathcal{N}(0, 1)$ for every $i \geq 1$. Moreover, we shall write $\mathbf{P} = \{P_i : i \geq 1\}$ for a sequence of independent random variables verifying

$$P_i \stackrel{d}{=} \frac{\mathrm{Po}(\lambda_i) - \lambda_i}{\sqrt{\lambda_i}} , \qquad i \geq 1 ,$$

where $\mathrm{Po}(\lambda_i)$ indicates a Poisson random variable with mean λ_i, such that $\alpha = \inf\{\lambda_i : i \geq 1\} > 0$.

There is a close connection between homogeneous sums based on $\mathbf{P}$ (or $\mathbf{G}$) and multiple stochastic integrals with respect to a centred Poisson measure $\hat{\eta}_n$ (or a Gaussian measure G_n) of tamed sequences. Namely, if $q \geq 1$ is a fixed integer and $f_n \in L^2_s(\mu_n^q)$, $n \geq 1$, is a tamed sequence with representation (4.1), then there is a sequence of centred Poisson measures $\{\hat{\eta}_n : n \geq 1\}$ (or a sequence of Gaussian measures $\{G_n : n \geq 1\}$) such that

$$I_q^{\hat{\eta}_n}(f_n) = Q_q(N_n, h_n, \mathbf{P}), \qquad I_q^{G_n}(f_n) = Q_q(N_n, h_n, \mathbf{G}). \tag{4.2}$$

Vice versa, given a sequence of index functions $\{h_n : n \geq 1\}$ of order $q \geq 1$ and a sequence of integers $\{N_n : n \geq 1\}$ diverging to infinity, as $n \to \infty$, such that $\|h_n\|_{(N_n,q)} < \infty$ for every $n \geq 1$, then there is a tamed sequence $\{f_n : n \geq 1\}$ with representation (4.1) and sequences of centred Poisson measures $\{\hat{\eta}_n : n \geq 1\}$ and Gaussian measures $\{G_n : n \geq 1\}$ such that (4.2) holds.

The following result is a version of Nourdin et al. (2010, Theorem 1.8) and Nourdin et al. (2010, Theorem 1.12). Notice that there, the results are stated for integer-valued parameters $\nu \geq 1$, but they continue to hold for any $\nu > 0$.

Theorem 4.5 (Gamma universality of homogeneous sums on a fixed Gaussian chaos)**.** *Fix $\nu > 0$, let $q \geq 2$ be even and $f_n \in L^2_s(\mu_n^q)$, $n \geq 1$, be a tamed sequence with representation (4.1) that satisfies the normalization condition*

$$\lim_{n\to\infty} q!\|f_n\|^2 = \lim_{n\to\infty} \mathbb{E}[I_q^G(f_n)^2] = \lim_{n\to\infty} \mathbb{E}[Q_q(N_n, h_n, \mathbf{G})^2] = 2\nu.$$

Let $Y \sim \overline{\Gamma}_\nu$ be a centred Gamma random variable with parameter ν. Then the following five assertions are equivalent:

(i) As $n \to \infty$, the sequence $\{Q_q(N_n, h_n, \mathbf{G})\colon n \geq 1\}$ converges in distribution to Y;

(ii) $\lim_{n\to\infty} \mathbb{E}[Q_q(N_n, h_n, \mathbf{G})^4] - 12\mathbb{E}[Q_q(N_n, h_n, \mathbf{G})^3] = 12\nu^2 - 48\nu$;

(iii) $\lim_{n\to\infty} \|f_n \star_r^r f_n\| = 0$ for every $r \in \{1, \ldots, q-1\} \setminus \{q/2\}$, and $\lim_{n\to\infty} \|f_n \widetilde{\star}_{q/2}^{q/2} f_n - c_q\, f_n\| = 0$ with $c_q = \frac{4}{(q/2)!\binom{q}{q/2}^2}$;

(iv) for every sequence $\mathbf{X} = \{X_i\colon i \geq 1\}$ of independent centred random variables with unit variance which is such that $\sup_i \mathbb{E}|X_i|^{2+\varepsilon} < \infty$ for some $\varepsilon > 0$, the sequence $\{Q_q(N_n, h_n, \mathbf{X})\colon n \geq 1\}$ converges in distribution to Y, as $n \to \infty$;

(v) for every sequence $\mathbf{X} = \{X_i\colon i \geq 1\}$ of i.i.d. centred random variables with unit variance, the sequence $\{Q_q(N_n, h_n, \mathbf{X})\colon n \geq 1\}$ converges in distribution to Y, as $n \to \infty$.

The following result answers the question whether Theorem 4.5 continues to hold if in (i) and (ii) the class $\mathbf{G}$ is replaced by $\mathbf{P}$. Due to the discussion in Section 3.2, we cannot avoid additional assumptions in the Poisson case. In particular, we have to assume that either $q = 2$ or $q = 4$.

Theorem 4.6 (Gamma universality of homogeneous sums on a fixed Poisson chaos)**.** *Fix $\nu > 0$ and let $q \geq 2$ be even and $f_n \in L^2_s(\mu_n^q)$, $n \geq 1$, be a tamed sequence with representation (4.1) that satisfies the normalization condition*

$$\lim_{n\to\infty} q!\|f_n\|^2 = \lim_{n\to\infty} \mathbb{E}[I_q(f_n)^2] = \lim_{n\to\infty} \mathbb{E}[Q_q(N_n, h_n, \mathbf{P})^2] = 2\nu. \tag{4.3}$$

Let $Y \sim \overline{\Gamma}_\nu$ be a random variable following a centred Gamma distribution with parameter ν. If one of the conditions

(a) *$q = 2$ and $\lim_{n\to\infty} \|f_n^2\| = 0$,*

(b) *$q = 4$ and $f_n \leq 0$ for all $n \geq 1$*

is satisfied, then the following five assertions are equivalent:

(i) As $n \to \infty$, the sequence $\{Q_q(N_n, h_n, \mathbf{P})\colon n \geq 1\}$ converges in distribution to Y;

(ii) $\lim_{n\to\infty} \mathbb{E}[Q_q(N_n, h_n, \mathbf{P})^4] - 12\mathbb{E}[Q_q(N_n, h_n, \mathbf{P})^3] = 12\nu^2 - 48\nu$;

(iii) $\lim_{n\to\infty} \|f_n \star_r^r f_n\| = 0$ for all $r \in \{1, \ldots, q-1\} \setminus \{q/2\}$, and $\lim_{n\to\infty} \|f_n \widetilde{\star}_{q/2}^{q/2} f_n - c_q\, f_n\| = 0$ with $c_q = \frac{4}{(q/2)!\binom{q}{q/2}^2}$;

(iv) for every sequence $\mathbf{X} = \{X_i\colon i \geq 1\}$ of independent centred random variables with unit variance which is such that $\sup_i \mathbb{E}|X_i|^{2+\varepsilon} < \infty$ for some $\varepsilon > 0$, the sequence $\{Q_q(N_n, h_n, \mathbf{X})\colon n \geq 1\}$ converges in distribution to Y, as $n \to \infty$;

(v) *for every sequence* $\mathbf{X} = \{X_i : i \geq 1\}$ *of i.i.d. centred random variables with unit variance, the sequence* $\{Q_q(N_n, h_n, \mathbf{X}) : n \geq 1\}$ *converges in distribution to* Y*, as* $n \to \infty$.

Proof: At first, we observe that due to Theorem 4.5, the assertions (iii), (iv) and (v) are equivalent. In Peccati and Zheng (2014, Subsection 4.2), it has been argued that

$$\sup_{i \geq 1} \mathbb{E}|P_i|^p < \infty \tag{4.4}$$

for all $p \geq 1$. This means that $\mathbf{P}$ is a special instance of a sequence with the properties in assertion (iv) such that we obtain the implication (iv) $\Longrightarrow$ (i). Moreover, (4.4) implies together with the normalization condition (4.3) and Nourdin et al. (2010, Lemma 4.2) that the sequence $\{Q_q(N_n, h_n, \mathbf{P})^4 : n \geq 1\}$ is uniformly integrable such that we get the implication (i) $\Longrightarrow$ (ii).

To prove (ii) $\Longrightarrow$ (iii), we apply Theorem 3.5. For this, one has to observe that assertion (iii) in Theorem 3.5 implies assertion (iii) in Theorem 4.6. $\square$

Remark 4.7. Theorem 4.6 shows that one can dispense with the assumption on the uniform integrability of the sequence $\{I_q(f_n)^4 : n \geq 1\}$ in Theorem 3.5 whenever the sequence $f_n \in L^2_s(\mu^q_n)$, $n \geq 1$, is tamed.

Remark 4.8. Replacing in (b) the condition that $f_n \leq 0$ by $f_n \geq 0$, in (ii) the moment condition by $\lim_{n\to\infty} \mathbb{E}[Q_q(N_n, h_n, \mathbf{P})^4] + 12\mathbb{E}[Q_q(N_n, h_n, \mathbf{P})^3] = 12\nu^2 - 48\nu$ and in (iii) the condition on the middle-contraction by $\|f_n \,\widetilde{\star}^{q/2}_{q/2} f_n + c_q\, f_n\| \to 0$, one arrives at a version of Theorem 4.6 with a centred reflected Gamma limiting random variable $Y \sim \widehat{\Gamma}_\nu$ in assertion (i), (iv) and (v).

4.2. *U-statistics.* Our second application is concerned with U-statistics. To introduce them, fix an integer $d \geq 1$ and let $\mathbf{Y} = \{Y_i : i \geq 1\}$ be a sequence of i.i.d. random vectors in $\mathbb{R}^d$, whose distribution has a density $p(\cdot)$ with respect to the Lebesgue measure on $\mathbb{R}^d$. Next, for any $n \geq 1$, let N_n be a Poisson random variable with mean n and define

$$\eta_n = \sum_{i=1}^{N_n} \delta_{Y_i} \,. \tag{4.5}$$

Clearly, η_n is a Poisson random measure on $\mathbb{R}^d$ with control measure $\mu_n(\mathrm{d}x) = np(x)\,\mathrm{d}x$, implying that $\mu_n(\mathbb{R}^d) = n \to \infty$, as $n \to \infty$. Now, we put $\hat{\eta}_n = \eta_n - \mu_n$ and set $\mu = \mu_1$ for the sake of convenience. By a *Poisson U-statistic* of order $q \geq 2$ based on η_n we mean in this paper a random variable of the form

$$U_n = \sum_{1 \leq i_1 < \cdots < i_q \leq N_n} h_n(Y_{i_1}, \ldots, Y_{i_q}) \,, \qquad n \geq 1 \,,$$

where the kernel $h_n : (\mathbb{R}^d)^q \to \mathbb{R}$ is an element of $L^1_s(\mu^q)$. On the other hand, a *classical U-statistic* is a random variable $\hat{U}_n$ such that

$$\hat{U}_n = \sum_{1 \leq i_1 < \cdots < i_q \leq n} h_n(Y_{i_1}, \ldots, Y_{i_q}) \,, \qquad n \geq 1 \,.$$

The difference between U_n and $\hat{U}_n$ is that U_n involves a random number $\binom{N_n}{q}$ of summands, while the number of summands in the definition of $\hat{U}_n$ is fixed (namely

$\binom{n}{q}$). We say that a (Poisson or classical) U-statistic is *completely degenerate* if

$$\int_{\mathbb{R}^d} h_n(x, z_1, \ldots, z_{q-1})\, p(x)\, \mathrm{d}x = 0$$

for μ^{q-1}-almost every $(z_1, \ldots, z_{q-1}) \in (\mathbb{R}^d)^{q-1}$. In particular, this implies that $\mathbb{E}[U_n] = \mathbb{E}[\hat{U}_n] = 0$. Moreover, we suppose that U_n and $\hat{U}_n$ are square-integrable.

We recall the following particular case of a celebrated theorem of de Jong, which provides a simple moment condition under which a central limit theorem for a sequence of completely degenerate U-statistics is guaranteed.

Theorem 4.9 (de Jong, 1987, 1990)**.** *Let $q \geq 2$ and $\{h_n : n \geq 1\}$ be a sequence of non-zero elements of $L_s^4(\mu^q)$. Suppose that the U-statistics U_n and $\hat{U}_n$ are completely degenerate and define $\sigma^2(n) = \mathrm{Var}(U_n)$. Then the moment condition $\lim_{n\to\infty} \mathbb{E}[U_n^4]/\sigma(n)^4 = 0$ implies that, as $n \to \infty$, the sequences $U_n/\sigma(n)$ and $\hat{U}_n/\sigma(n)$ converge in distribution to a standard Gaussian random variable.*

In our paper, we are interested in the Gamma approximation of Poisson and classical U-statistics. The next result generalizes Theorem 2.13 (B) in Peccati and Thäle (2013), where the authors had to restrict to the case $q = 2$. Here, we add a corresponding limit theorem in case that $q = 4$ under an additional sign condition. It can be seen as a non-central version of de Jong's theorem, Theorem 4.9. We shall see that in the non-central case a similar result is true under a suitable condition involving only the third and the fourth moment.

Theorem 4.10. *Suppose that $q \in \{2, 4\}$. For each $n \geq 1$ let $h_n \in L_s^4(\mu^q)$ be a function such that*

$$\sup_{n\geq 1} \frac{\int h_n^4\, \mathrm{d}\mu_n^q}{(\int h_n^2\, \mathrm{d}\mu_n^q)^2} < \infty$$

and suppose that the U-statistics U_n and $\hat{U}_n$ are completely degenerate. Further assume that there exists $\nu > 0$ such that $\lim_{n\to\infty} \mathbb{E}[U_n^2] = 2\nu$ and that

(a) $\lim_{n\to\infty} \|h_n^2\| = 0$ *if* $q = 2$,

(b) $f_n \leq 0$ *for all* $n \geq 1$ *if* $q = 4$.

Then the moment condition $\lim_{n\to\infty} \mathbb{E}[U_n^4] - 12\mathbb{E}[U_n^3] = 12\nu^2 - 48\nu$ implies that both random variables U_n and $\hat{U}_n$ converge in distribution to $Y \sim \overline{\Gamma}_\nu$, as $n \to \infty$.

Proof: Using the fact that the Poisson U-statistics U_n is an element of the sum of the first q Poisson chaoses with respect to $\hat{\eta}_n$ as introduced after (4.5) (see Reitzner and Schulte, 2013, Theorem 3.6), as well as the fact that U_n is completely degenerate, one obtains that $U_n = I_q(h_n)$ for every $n \geq 1$. The result for the Poisson U-statistics U_n then follows immediately from Theorem 3.5. Moreover, it is known from Dynkin and Mandelbaum (1983) that $\mathbb{E}[(U_n - \hat{U}_n)^2] = O(n^{-1/2})$, as $n \to \infty$. This yields the result also for $\hat{U}_n$. □

Remark 4.11. Using Theorem 2.6 in Peccati and Thäle (2013) or its generalization Proposition 5.1 below, one can add a rate of convergence (for a certain smooth probability distance) between U_n or $\hat{U}_n$ and the limiting random variable Y. However, we do not pursue such quantitative results in this paper.

Remark 4.12. In assumption (b) of Theorem 4.10 one can replace the sign condition $f_n \le 0$ by $f_n \ge 0$ and at the same time the moment condition $\mathbb{E}[U_n^4] - 12\mathbb{E}[U_n^3] \to 12\nu^2 - 48\nu$ by $\mathbb{E}[U_n^4] + 12\mathbb{E}[U_n^3] \to 12\nu^2 - 48\nu$. In this case, the limiting random variable Y has a centred reflected Gamma distribution $\widehat{\Gamma}_\nu$ with parameter $\nu > 0$.

5. Proof of Theorem 3.5

5.1. *Strategy of the proof.* Before entering the details of the proof of Theorem 3.5, let us briefly summarize the overall strategy.

First of all, the implication (i) $\Longrightarrow$ (ii) of Theorem 3.5 is a direct consequence of the uniform integrability of the sequence $\{I_q(f_n)^4 : n \ge 1\}$. Next, the implication (iii) $\Longrightarrow$ (i) will follow from a generalization of the main result of Peccati and Thäle (2013), which has been derived by the Malliavin-Stein method. It delivers a criterion in terms of contraction norms, which ensures centred Gamma convergence on a fixed Poisson chaos of even order and is presented as Proposition 5.1 below. The main part of proof of Theorem 3.5 consists in showing that (ii) implies (iii). It is based on the technical Lemmas 5.2 and 5.4, which establish new inequalities for norms of contraction kernels, that are also of independent interest. Next, in Lemma 5.6 we derive an asymptotic lower bound for the moment expression $\mathbb{E}[I_q(f_n)^4] - 12\mathbb{E}[I_q(f_n)^3]$ in terms of contraction norms. Finally, Lemma 5.7 shows under the conditions of Theorem 3.5 that if the lower bound for $\mathbb{E}[I_q(f_n)^4] - 12\mathbb{E}[I_q(f_n)^3]$ converges to the 'correct' quantity, the contraction conditions in (iii) are satisfied. Lemma 5.9 proves that this lower bound actually converges.

We emphasize that we state all intermediate steps of the proof of Theorem 3.5 as general as possible in order to highlight in which step the restrictive condition that $q = 2$ or $q = 4$ and the sign condition on the kernels arise.

5.2. *Preparatory steps.* We start our investigations with a generalization of Theorem 2.6 in Peccati and Thäle (2013). The main difference between that result and Proposition 5.1 is that for technical reasons it has been assumed in Peccati and Thäle (2013) that μ_n is a finite measure for each $n \ge 1$ such that $\mu_n(\mathcal{Z}) \to \infty$, as $n \to \infty$. Our next result shows that one can dispense with this assumption.

Proposition 5.1. *Fix $\nu > 0$ and an even integer $q \ge 2$. Let the sequence $f_n \in L_s^2(\mu_n^q)$, $n \ge 1$, satisfy the technical assumptions (A) and the normalization condition*

$$\lim_{n\to\infty} q!\|f_n\|^2 = \lim_{n\to\infty} \mathbb{E}[I_q(f_n)^2] = 2\nu\,.$$

Then, if

$$\begin{aligned}
&\lim_{n\to\infty} \|f_n \star_r^\ell f_n\| = 0 \text{ for all } r \in \{1,\dots,q\},\ \ell \in \{1,\dots,r\wedge(q-1)\} \\
&\text{with } (r,\ell) \neq (q/2,q/2)\,, \\
&\lim_{n\to\infty} \|f_n^2\| = 0\,, \\
&\lim_{n\to\infty} \|f_n \,\widetilde{\star}_{q/2}^{q/2} f_n - c_q\, f_n\| = 0 \text{ with } c_q = \frac{4}{(q/2)!\binom{q}{q/2}^2}\,,
\end{aligned}$$

the sequence $\{I_q(f_n) : n \ge 1\}$ converges in distribution to $Y \sim \overline{\Gamma}_\nu$, as $n \to \infty$.

Proof: In principle, one can follow the proof of Peccati and Thäle (2013, Theorem 2.6). The only part where the assumption about the finiteness of the measures μ_n enters is Peccati and Thäle (2013, Proposition 2.3). To circumvent this problem, one uses the modified integration-by-parts formula Schulte (2016, Lemma 2.3) and concludes as in the proof of Theorem 4.1 of Eichelsbacher and Thäle (2014). Since the computations are quite straight forward, we omit the details. □

We now present two estimates of the norm of a symmetrized contraction kernel in terms of non-symmetrized contraction norms. In particular, our first lemma generalizes Peccati and Taqqu (2011, Identity (11.6.30)). We recall for $f \in L^2_s(\mu^q)$, $q \geq 1$, that $\|f \,\widetilde{\star}_q^q f\|^2 = \|f \star_q^q f\|^2 = \|f\|^4$ and $\|f \star_0^0 f\|^2 = \|f\|^4$.

Lemma 5.2. *Let $q \geq 1$ be an integer and $f \in L^2_s(\mu^q)$ be a kernel satisfying the technical assumptions (A). Then*

$$\|f \,\widetilde{\star}_0^0 f\|^2 = \frac{(q!)^2}{(2q)!}\left(2\|f\|^4 + \sum_{p=1}^{q-1} \binom{q}{p}^2 \|f \star_p^p f\|^2\right). \tag{5.1}$$

Furthermore, for any $r \in \{1, \dots, q-1\}$ one has the inequality

$$\|f \,\widetilde{\star}_r^r f\|^2 \leq \frac{((q-r)!)^2}{(2(q-r))!}\left(2\|f \star_r^r f\|^2 + \sum_{p=1}^{q-r-1} \binom{q-r}{p}^2 \|f \star_p^p f\|^2\right). \tag{5.2}$$

If $q \geq 2$ is an even integer, Equation (5.2) yields that

$$\|f \,\widetilde{\star}_{q/2}^{q/2} f\|^2 \leq \frac{((q/2)!)^2}{q!}\left(2\|f \star_{q/2}^{q/2} f\|^2 + \sum_{p=1}^{q/2-1} \binom{q/2}{p}^2 \|f \star_p^p f\|^2\right). \tag{5.3}$$

This inequality will turn out to be crucial in what follows.

Before entering the proof of Lemma 5.2, we introduce some notation. Recall that for an integer $p \geq 1$, we denote the group of $p!$ permutations of the set $\{1, \dots, p\}$ by Π_p. For a kernel $g \in L^2(\mu^p)$ and a permutation $\pi \in \Pi_p$, we use the shorthand $g(\pi)$ for the mapping $\mathcal{Z}^p \ni (z_1, \dots, z_p) \mapsto g(\pi)(z_1, \dots, z_p) = g(z_{\pi(1)}, \dots, z_{\pi(p)})$. We can immediately see that $\|g\| = \|g(\pi)\|$ for all $\pi \in \Pi_p$ such that automatically $g(\pi) \in L^2(\mu^p)$. In the following, we use the convention that $\pi_0 \in \Pi_p$ is the identity map, meaning that $g(\pi_0) = g$.

For any integer $M \geq 1$, any two permutations $\pi, \sigma \in \Pi_{2M}$ and any $p \in \{0, \dots, M\}$ we shall use the notation

$$\pi \sim_p \sigma$$

if and only if

$$|\{\pi(1), \dots, \pi(M)\} \cap \{\sigma(1), \dots, \sigma(M)\}| = p\,,$$

where $|\cdot|$ stands for the cardinality of the argument set. If $\pi \sim_p \sigma$, then clearly $|\{\pi(M+1), \dots, \pi(2M)\} \cap \{\sigma(M+1), \dots, \sigma(2M)\}| = p$. In the proof of Peccati and Taqqu (2011, Proposition 11.2.2), there is an explanation that, given a permutation $\pi \in \Pi_{2M}$ and an integer $p \in \{0, \dots, M\}$, there are exactly $(M!)^2 \binom{M}{p}^2$ permutations $\sigma \in \Pi_{2M}$ such that $\pi \sim_p \sigma$.

Proof of Lemma 5.2: Let $q \geq 1$ be an integer and $f \in L_s^2(\mu^q)$ be a kernel satisfying the technical assumptions (A). Fix $r \in \{0, 1, \dots, q-1\}$. We have that

$$\begin{aligned} \|f \widetilde{\star}_r^r f\|^2 = \langle f \star_r^r f, f \widetilde{\star}_r^r f \rangle &= \frac{1}{(2q-2r)!} \sum_{\pi \in \Pi_{2q-2r}} \langle f \star_r^r f, f \star_r^r f(\pi) \rangle \\ &= \frac{1}{(2q-2r)!} \sum_{p=0}^{q-r} \sum_{\pi \sim_p \pi_0} \langle f \star_r^r f, f \star_r^r f(\pi) \rangle . \end{aligned} \tag{5.4}$$

To prove (5.1), let $r = 0$ and $\pi \sim_0 \pi_0$ or $\pi \sim_q \pi_0$. Then we get

$$\begin{aligned} &\langle f \star_0^0 f, f \star_0^0 f(\pi) \rangle \\ &= \int_{\mathcal{Z}^{2q}} f(z_1, \dots, z_q) f(z_{q+1}, \dots, z_{2q}) \\ &\qquad \times f(z_{\pi(1)}, \dots, z_{\pi(q)}) f(z_{\pi(q+1)}, \dots, z_{\pi(2q)}) \mu^{2q}(\mathrm{d}(z_1, \dots, z_{2q})) \\ &= \left(\int_{\mathcal{Z}^q} f(w_1, \dots, w_q)^2 \mu^q(\mathrm{d}(w_1, \dots, w_q)) \right)^2 \\ &= \|f\|^4 . \end{aligned}$$

Now, let $\pi \sim_p \pi_0$ with $p \in \{1, \dots, q-1\}$. Then

$$\begin{aligned} &\langle f \star_0^0 f, f \star_0^0 f(\pi) \rangle \\ &= \int_{\mathcal{Z}^{2q}} f(z_1, \dots, z_q) f(z_{q+1}, \dots, z_{2q}) \\ &\qquad \times f(z_{\pi(1)}, \dots, z_{\pi(q)}) f(z_{\pi(q+1)}, \dots, z_{\pi(2q)}) \mu^{2q}(\mathrm{d}(z_1, \dots, z_{2q})) \\ &= \int_{\mathcal{Z}^{2q-2p} \times \mathcal{Z}^p \times \mathcal{Z}^p} f(z_1, \dots, z_q) f(z_{\pi(1)}, \dots, z_{\pi(q)}) \\ &\qquad \times f(z_{q+1}, \dots, z_{2q}) f(z_{\pi(q+1)}, \dots, z_{\pi(2q)}) \mu^{2q}(\mathrm{d}(z_1, \dots, z_{2q})) \\ &\overset{(\star)}{=} \int_{\mathcal{Z}^{2q-2p}} f \star_p^p f(w_1, \dots, w_{2q-2p}) \\ &\qquad \times f \star_p^p f(w_1, \dots, w_{2q-2p}) \mu^{2q-2p}(\mathrm{d}(w_1, \dots, w_{2q-2p})) \\ &= \|f \star_p^p f\|^2 . \end{aligned}$$

We note that the assumption that f is symmetric is essential to get the identity highlighted by $(\star)$. In view of (5.4), we obtain

$$\begin{aligned} &\|f \widetilde{\star}_0^0 f\|^2 \\ &= \frac{1}{(2q)!} \Bigg(\sum_{\pi \sim_0 \pi_0} \langle f \star_0^0 f, f \star_0^0 f(\pi) \rangle + \sum_{\pi \sim_q \pi_0} \langle f \star_0^0 f, f \star_0^0 f(\pi) \rangle \\ &\qquad + \sum_{p=1}^{q-1} \sum_{\pi \sim_p \pi_0} \langle f \star_0^0 f, f \star_0^0 f(\pi) \rangle \Bigg) \\ &= \frac{1}{(2q)!} \left(2(q!)^2 \|f\|^4 + \sum_{p=1}^{q-1} (q!)^2 \binom{q}{p}^2 \|f \star_p^p f\|^2 \right) , \end{aligned}$$

such that (5.1) follows. Now, let $r \in \{1, \dots, q-1\}$ and observe that for $\pi \sim_{q-r} \pi_0$ one has that

$$
\begin{aligned}
&\langle f \star_r^r f, f \star_r^r f(\pi) \rangle \\
&= \int_{\mathcal{Z}^{2q-2r}} \Bigg(\int_{\mathcal{Z}^r} f(x_1, \dots, x_r, z_1, \dots, z_{q-r}) \\
&\qquad\qquad \times f(x_1, \dots, x_r, z_{q-r+1}, \dots, z_{2q-2r}) \mu^r(\mathrm{d}(x_1, \dots, x_r)) \Bigg) \\
&\quad \times \Bigg(\int_{\mathcal{Z}^r} f(y_1, \dots, y_r, z_{\pi(1)}, \dots, z_{\pi(q-r)}) \\
&\qquad\qquad \times f(y_1, \dots, y_r, z_{\pi(q-r+1)}, \dots, z_{\pi(2q-2r)}) \mu^r(\mathrm{d}(y_1, \dots, y_r)) \Bigg) \\
&\qquad\qquad \mu^{2q-2r}(\mathrm{d}(z_1, \dots, z_{2q-2r})) \\
&= \int_{\mathcal{Z}^{2q}} f(x_1, \dots, x_r, z_1, \dots, z_{q-r}) f(y_1, \dots, y_r, z_{\pi(1)}, \dots, z_{\pi(q-r)}) \\
&\qquad \times f(x_1, \dots, x_r, z_{q-r+1}, \dots, z_{2q-2r}) f(y_1, \dots, y_r, z_{\pi(q-r+1)}, \dots, z_{\pi(2q-2r)}) \\
&\qquad\qquad \mu^{2q}(\mathrm{d}(x_1, \dots, x_r, y_1, \dots, y_r, z_1, \dots, z_{2q-2r})) \\
&= \int_{\mathcal{Z}^{2r} \times \mathcal{Z}^{q-r} \times \mathcal{Z}^{q-r}} f(x_1, \dots, x_r, z_1, \dots, z_{q-r}) f(y_1, \dots, y_r, z_{\pi(1)}, \dots, z_{\pi(q-r)}) \\
&\qquad \times f(x_1, \dots, x_r, z_{q-r+1}, \dots, z_{2q-2r}) f(y_1, \dots, y_r, z_{\pi(q-r+1)}, \dots, z_{\pi(2q-2r)}) \\
&\qquad\qquad \mu^{2q}(\mathrm{d}(x_1, \dots x_r, y_1, \dots, y_r, z_1, \dots, z_{2q-2r})) \\
&= \int_{\mathcal{Z}^{2r}} \Big(f \star_{q-r}^{q-r} f(x_1, \dots, x_r, y_1, \dots, y_r) \Big)^2 \mu^{2r}(\mathrm{d}(x_1, \dots x_r, y_1, \dots, y_r)) \\
&= \| f \star_{q-r}^{q-r} f \|^2 \\
&= \| f \star_r^r f \|^2 .
\end{aligned}
$$

Similarly, we obtain for the case that $\pi \sim_0 \pi_0$,

$$\langle f \star_r^r f, f \star_r^r f(\pi) \rangle = \| f \star_r^r f \|^2 .$$

Now, let $\pi \sim_p \pi_0$ with $p \in \{1, \dots, q-r-1\}$. Then, there is a permutation $\sigma \in \Pi_{2q-2p}$ such that

$$
\begin{aligned}
&\langle f \star_r^r f, f \star_r^r f(\pi) \rangle \qquad (5.5) \\
&= \int_{\mathcal{Z}^{2q}} f(x_1, \dots, x_r, z_1, \dots, z_{q-r}) f(y_1, \dots, y_r, z_{\pi(1)}, \dots, z_{\pi(q-r)}) \\
&\qquad \times f(x_1, \dots, x_r, z_{q-r+1}, \dots, z_{2q-2r}) f(y_1, \dots, y_r, z_{\pi(q-r+1)}, \dots, z_{\pi(2q-2r)}) \\
&\qquad\qquad \mu^{2q}(\mathrm{d}(x_1, \dots, x_r, y_1, \dots, y_r, z_1, \dots, z_{2q-2r})) \\
&= \int_{\mathcal{Z}^{2q-2p} \times \mathcal{Z}^p \times \mathcal{Z}^p} f(x_1, \dots, x_r, z_1, \dots, z_{q-r}) f(y_1, \dots, y_r, z_{\pi(1)}, \dots, z_{\pi(q-r)}) \\
&\qquad \times f(x_1, \dots, x_r, z_{q-r+1}, \dots, z_{2q-2r}) f(y_1, \dots, y_r, z_{\pi(q-r+1)}, \dots, z_{\pi(2q-2r)}) \\
&\qquad\qquad \mu^{2q}(\mathrm{d}(x_1, \dots, x_r, y_1, \dots, y_r, z_1, \dots, z_{2q-2r}))
\end{aligned}
$$

$$
\begin{aligned}
&= \int_{\mathcal{Z}^{2q-2p}} f \star_p^p f(w_1, \dots, w_{2q-2p}) \\
&\qquad\qquad \times f \star_p^p f(w_{\sigma(1)}, \dots, w_{\sigma(2q-2p)}) \mu^{2q-2p}(\mathrm{d}(w_1, \dots, w_{2q-2p})) \\
&= \langle f \star_p^p f, f \star_p^p f(\sigma) \rangle \\
&\leq \| f \star_p^p f \| \, \| f \star_p^p f(\sigma) \| \\
&= \| f \star_p^p f \|^2 .
\end{aligned}
$$

Note that contrary to the case $r = 0$, σ shows up because of the appearance of the variables $x_1, \dots, x_r, y_1, \dots, y_r$. Therefore, we need to apply the Cauchy-Schwarz inequality once, which is the very reason for the inequality in (5.2). At this stage, (5.2) follows by (5.4) and

$$
\begin{aligned}
\| f \widetilde{\star}_r^r f \|^2 &= \frac{1}{(2q-2r)!} \Bigg(\sum_{\pi \sim_0 \pi_0} \langle f \star_r^r f, f \star_r^r f(\pi) \rangle + \sum_{\pi \sim_{q-r} \pi_0} \langle f \star_r^r f, f \star_r^r f(\pi) \rangle \\
&\qquad\qquad + \sum_{p=1}^{q-r-1} \sum_{\pi \sim_p \pi_0} \langle f \star_r^r f, f \star_r^r f(\pi) \rangle \Bigg) \\
&\leq \frac{1}{(2q-2r)!} \left(2((q-r)!)^2 \| f \star_r^r f \|^2 + \sum_{p=1}^{q-r-1} ((q-r)!)^2 \binom{q-r}{p}^2 \| f \star_p^p f \|^2 \right) .
\end{aligned}
$$

This completes the proof. □

Remark 5.3. A combinatorial argument shows that the permutation $\sigma \in \Pi_{2q-2p}$ appearing in (5.5) cannot be such that $f \star_p^p f(\sigma) = f \star_p^p f$ (in particular, σ cannot be the identity). Hence, we cannot omit applying Cauchy-Schwarz in this case.

In Lemma 5.2 no condition on the sign of f was necessary. However, if we assume that f has constant sign, we are able to deduce a 'reverse' counterpart of (5.2).

Lemma 5.4. *Let $q \geq 1$ be an integer and $f \in L_s^2(\mu^q)$ a kernel satisfying the technical assumptions (A). If $f \leq 0$ or $f \geq 0$, then, for any $r \in \{0, 1, \dots, q-1\}$, one has that*

$$
\| f \widetilde{\star}_r^r f \|^2 \geq \frac{2((q-r)!)^2}{(2q-2r)!} \| f \star_r^r f \|^2 . \tag{5.6}
$$

Proof: The left-hand side of (5.6) satisfies the identity at (5.4). Using the fact that f has constant sign, the right-hand side of (5.4) becomes smaller if we sum only over a subset of Π_{2q-2r}, namely over all $\pi \in \Pi_{2q-2r}$ such that $\pi \sim_0 \pi_0$ or $\pi \sim_{q-r} \pi_0$. Hence, we end up with

$$
\begin{aligned}
\| f \widetilde{\star}_r^r f \|^2 &\geq \frac{1}{(2q-2r)!} \left(\sum_{\pi \sim_0 \pi_0} \langle f \star_r^r f, f \star_r^r f(\pi) \rangle + \sum_{\pi \sim_{q-r} \pi_0} \langle f \star_r^r f, f \star_r^r f(\pi) \rangle \right) \\
&= \frac{2((q-r)!)^2}{(2(q-r))!} \| f \star_r^r f \|^2 ,
\end{aligned}
$$

which completes the proof. □

Remark 5.5. In view of Remark 5.3, inequality (5.6) is optimal under the conditions of Lemma 5.4.

5.3. *Proof of the implication (ii) $\Longrightarrow$ (iii).* Let us introduce some notation. We shall write $a_n \asymp b_n$ for two real-valued sequences $\{a_n : n \geq 1\}$, $\{b_n : n \geq 1\}$, whenever $\lim_{n\to\infty} a_n - b_n = 0$. Be aware that this does not necessarily imply that one of the individual sequences converges, but of course ensures the convergence of both sequences whenever one of them converges.

The next lemma establishes an asymptotic lower bound for the linear combination of the fourth and third moment $\mathbb{E}[I_q(f_n)^4] - 12\mathbb{E}[I_q(f_n)^3]$ of a sequence of Poisson integrals of even order $q \geq 2$ where $f_n \in L^2_s(\mu^q_n)$, $n \geq 1$. It is one of the main ingredients to show the implication (ii) $\Longrightarrow$ (iii) in Theorem 3.5. Note that this bound holds for general even $q \geq 2$. Moreover, at this point we do not need an assumption on the sign of the kernels.

Lemma 5.6. *Let $\nu > 0$ and $q \geq 2$ be an even integer. Let $f_n \in L^2_s(\mu^q_n)$, $n \geq 1$, be a sequence of kernels such that the technical assumptions (A) and the normalization condition*

$$\lim_{n\to\infty} q!\|f_n\|^2 = 2\nu$$

are satisfied. Then one has that

$$\mathbb{E}[I_q(f_n)^4] - 12\mathbb{E}[I_q(f_n)^3] \asymp 12\nu^2 - 48\nu + A(I_q(f_n)) + R(I_q(f_n))\,, \tag{5.7}$$

where the terms on the right-hand side of (5.7) satisfy $A(I_q(f_n)) \geq A'(I_q(f_n))$ with

$$\begin{aligned} A'(I_q(f_n)) = &\sum_{p=1}^{q/2-1} \frac{(q!)^4}{(p!)^2}\left(\frac{2}{(q-p)!^2} - \frac{1}{2\big((q/2)!(q/2-p)!\big)^2}\right)\|f_n \star_p^p f_n\|^2 \\ &+ \sum_{p=1,p\neq q}^{2q-1} p!\|G^q_p f_n\|^2 + q!\sum_{p=q/2+1}^{q}(p!)^2\binom{p}{q}^4\binom{p}{q-p}^2\|f_n \,\widetilde{\star}_p^{q-p} f_n\|^2 \\ &+ 24q!\|c_q^{-1} f_n \,\widetilde{\star}_{q/2}^{q/2} f_n - f_n\|^2 \end{aligned} \tag{5.8}$$

with $c_q = \frac{4}{(q/2)!\binom{q}{q/2}^2}$, and

$$\begin{aligned} R(I_q(f_n)) = q! &\sum_{\substack{r,p=q/2\\ r\neq p}}^{q} r!\,p!\binom{q}{r}^2\binom{q}{p}^2\binom{r}{q-r}\binom{p}{q-p}\langle f_n \,\widetilde{\star}_r^{q-r} f_n, f_n \,\widetilde{\star}_p^{q-p} f_n\rangle \\ &- 12q!\sum_{p=q/2+1}^{q} p!\binom{q}{p}^2\binom{p}{q-p}\langle f_n \,\widetilde{\star}_p^{q-p} f_n, f_n\rangle\,. \end{aligned} \tag{5.9}$$

Proof of Lemma 5.6: In view of Lemma 2.2 and since q is even, one has that

$$\begin{aligned}
&\mathbb{E}[I_q(f_n)^4] - 12\mathbb{E}[I_q(f_n)^3] \\
&= \sum_{p=0}^{2q} p!\|G_p^q f_n\|^2 - 12q! \sum_{p=q/2}^{q} p!\binom{q}{p}^2\binom{p}{q-p}\langle f_n \widetilde{\star}_p^{q-p} f_n, f_n\rangle \\
&= (q!)^2\|f_n\|^4 + (2q)!\|f_n \widetilde{\star}_0^0 f_n\|^2 + \sum_{p=1}^{2q-1} p!\|G_p^q f_n\|^2 \\
&\quad - 12q! \sum_{p=q/2}^{q} p!\binom{q}{p}^2\binom{p}{q-p}\langle f_n \widetilde{\star}_p^{q-p} f_n, f_n\rangle \\
&= 3(q!)^2\|f_n\|^4 + \sum_{p=1}^{q-1} \frac{(q!)^4}{\left(p!(q-p)!\right)^2}\|f_n \star_p^p f_n\|^2 + \sum_{p=1}^{2q-1} p!\|G_p^q f_n\|^2 \\
&\quad - 12q! \sum_{p=q/2}^{q} p!\binom{q}{p}^2\binom{p}{q-p}\langle f_n \widetilde{\star}_p^{q-p} f_n, f_n\rangle \\
&= 3(q!)^2\|f_n\|^4 + T_1(I_q(f_n)) + T_2(I_q(f_n)) + T_3(I_q(f_n)),
\end{aligned}$$

where the third equality stems from (5.1). The terms T_1, T_2, T_3 read as follows:

$$T_1(I_q(f_n)) = \sum_{\substack{p=1\\ p\neq q/2}}^{q-1} \frac{(q!)^4}{\left(p!(q-p)!\right)^2}\|f_n \star_p^p f_n\|^2 + \sum_{p=1, p\neq q}^{2q-1} p!\|G_p^q f_n\|^2,$$

$$T_2(I_q(f_n)) = \frac{(q!)^4}{(q/2)!^4}\|f_n \star_{q/2}^{q/2} f_n\|^2 + q!\|G_q^q f_n\|^2 - 12q!(q/2)!\binom{q}{q/2}^2\langle f_n \widetilde{\star}_{q/2}^{q/2} f_n, f_n\rangle,$$

$$T_3(I_q(f_n)) = -12q! \sum_{p=q/2+1}^{q} p!\binom{q}{p}^2\binom{p}{q-p}\langle f_n \widetilde{\star}_p^{q-p} f_n, f_n\rangle.$$

We use (5.3) to see that

$$\begin{aligned}
&\frac{(q!)^4}{(q/2)!^4}\|f_n \star_{q/2}^{q/2} f_n\|^2 \\
&\qquad \geq \frac{(q!)^5}{2(q/2)!^6}\|f_n \widetilde{\star}_{q/2}^{q/2} f_n\|^2 - \frac{1}{2}\sum_{p=1}^{q/2-1} \frac{(q!)^4}{\left((q/2)!p!(q/2-p)!\right)^2}\|f_n \star_p^p f_n\|^2.
\end{aligned}$$

Using the definition of $G_q^q f_n$ given at (2.5), we have the estimate

$$\begin{aligned}
T_2(I_q(f_n)) \geq q! \Bigg(&\Big\| \sum_{r=q/2}^{q} r!\binom{q}{r}^2\binom{r}{q-r} f_n \widetilde{\star}_r^{q-r} f_n\Big\|^2 + \frac{1}{2}\frac{(q!)^4}{(q/2)!^6}\|f_n \widetilde{\star}_{q/2}^{q/2} f_n\|^2 \\
&-12(q/2)!\binom{q}{q/2}^2\langle f_n \widetilde{\star}_{q/2}^{q/2} f_n, f_n\rangle \Bigg) \\
&- \frac{1}{2}\sum_{p=1}^{q/2-1} \frac{(q!)^4}{\left((q/2)!p!(q/2-p)!\right)^2}\|f_n \star_p^p f_n\|^2
\end{aligned}$$

$$\begin{aligned}
&= q!\left(\frac{3}{2}\frac{(q!)^4}{(q/2)!^6}\|f_n \,\widetilde{\star}_{q/2}^{q/2} f_n\|^2 - 12\frac{(q!)^2}{(q/2)!^3}\langle f_n \,\widetilde{\star}_{q/2}^{q/2} f_n, f_n\rangle\right)\\
&\quad + q!\sum_{r=q/2+1}^{q}(r!)^2\binom{q}{r}^4\binom{r}{q-r}^2\|f_n \,\widetilde{\star}_r^{q-r} f_n\|^2\\
&\quad + q!\sum_{\substack{r,p=q/2\\ r\neq p}}^{q} r!p!\binom{q}{r}^2\binom{q}{p}^2\binom{r}{q-r}\binom{p}{q-p}\langle f_n \,\widetilde{\star}_r^{q-r} f_n, f_n \,\widetilde{\star}_p^{q-p} f_n\rangle\\
&\quad - \frac{1}{2}\sum_{p=1}^{q/2-1}\frac{(q!)^4}{\big((q/2)!p!(q/2-p)!\big)^2}\|f_n \star_p^p f_n\|^2.
\end{aligned}$$

Using the relation $\|f_n \star_p^p f_n\| = \|f_n \star_{q-p}^{q-p} f_n\|$, valid for all $p \in \{1,\dots,q-1\}$, we obtain

$$\begin{aligned}
&\sum_{\substack{p=1\\ p\neq q/2}}^{q-1}\frac{(q!)^4}{\big(p!(q-p)!\big)^2}\|f_n \star_p^p f_n\|^2 - \frac{1}{2}\sum_{p=1}^{q/2-1}\frac{(q!)^4}{\big((q/2)!p!(q/2-p)!\big)^2}\|f_n \star_p^p f_n\|^2\\
&= \sum_{p=1}^{q/2-1}\frac{(q!)^4}{(p!)^2}\left(\frac{2}{(q-p)!^2} - \frac{1}{2\big((q/2)!(q/2-p)!\big)^2}\right)\|f_n \star_p^p f_n\|^2.
\end{aligned}$$

The proof is concluded by observing that

$$\begin{aligned}
&q!\left(\frac{3}{2}\frac{(q!)^4}{((q/2)!)^6}\|f_n \,\widetilde{\star}_{q/2}^{q/2} f_n\|^2 - 12\frac{(q!)^2}{((q/2)!)^3}\langle f_n \,\widetilde{\star}_{q/2}^{q/2} f_n, f_n\rangle\right)\\
&= \frac{3}{2}q!\left(\frac{(q!)^4}{((q/2)!)^6}\|f_n \,\widetilde{\star}_{q/2}^{q/2} f_n\|^2 - 2\times 4\frac{(q!)^2}{((q/2)!)^3}\langle f_n \,\widetilde{\star}_{q/2}^{q/2} f_n, f_n\rangle + 16\|f_n\|^2\right)\\
&\quad - 24q!\|f_n\|^2\\
&= 24q!\|c_q^{-1} f_n \,\widetilde{\star}_{q/2}^{q/2} f_n - f_n\|^2 - 24q!\|f_n\|^2
\end{aligned}$$

and by recalling condition (a), which implies that $3(q!)^2\|f_n\|^4 - 24q!\|f_n\|^2$ converges to $12\nu^2 - 48\nu$. □

While all previous results did not use the assumptions on the order of the integral and the sign of the kernels, in the next lemma we need that $q \in \{2,4\}$ and that the kernels have constant sign.

Lemma 5.7. *Let $\nu > 0$ and $q \in \{2,4\}$. Let $f_n \in L_s^2(\mu_n^q)$, $n \geq 1$, be a sequence of kernels such that the technical assumptions (A) and the normalization condition $\lim_{n\to\infty} q!\|f_n\|^2 = 2\nu$ are satisfied. Assume that for each $n \geq 1$ either $f_n \leq 0$ or $f_n \geq 0$. Then the following two assertions concerning the term $A'(I_q(f_n))$ defined at (5.8) are true:*

(1) $A'(I_q(f_n)) \geq 0$ for all $n \geq 1$;

(2) If $A'(I_q(f_n)) \to 0$, as $n \to \infty$, then

$$\lim_{n\to\infty}\|f_n \star_r^\ell f_n\| = 0 \tag{5.10}$$

for all $r \in \{1,\dots,q\}$ and $\ell \in \{1,\dots,r\wedge(q-1)\}$ such that $(r,\ell) \neq (q/2,q/2)$,

$$\lim_{n\to\infty}\|f_n^2\| = 0, \tag{5.11}$$

$$\lim_{n\to\infty} \|f_n \,\widetilde{\star}_{q/2}^{q/2}\, f_n - c_q f_n\| = 0 \qquad \textit{with } c_q = \frac{4}{(q/2)!\binom{q}{q/2}^2}\,. \tag{5.12}$$

Proof: We start by showing the first assertion of the lemma. The only term that might be negative on right-hand side of (5.8) is the first sum. For the case $q = 2$, this does not play any role, because then the sum vanishes. Hence, $A'(I_q(f_n))$ is a positive linear combination of non-negative terms.

Now, let $q \geq 4$ be even. Using the fact that f_n has constant sign, $\|f_n \star_p^p f_n\| = \|f_n \star_{q-p}^{q-p} f_n\|$ for all $p \in \{1, \ldots, q-1\}$ as well as Lemma 5.4, we obtain the estimate

$$\begin{aligned}
\sum_{p=1,\, p\neq q}^{2q-1} p!\|G_p^q f_n\|^2 &\geq \sum_{p=1,\, p\neq q}^{2q-1} p! \sum_{r=0}^{q}\sum_{\ell=0}^{r} \mathbf{1}(2q-r-\ell = p) r!^2 \binom{q}{r}^4 \binom{r}{\ell}^2 \|f_n \,\widetilde{\star}_r^\ell f_n\|^2 \\
&\geq \sum_{\substack{p=1,\, p\neq q,\\ p \text{ even}}}^{2q-1} p!((q-p/2)!)^2 \binom{q}{q-p/2}^4 \|f_n \,\widetilde{\star}_{q-p/2}^{q-p/2} f_n\|^2 \\
&= \sum_{p=1,\, p\neq q/2}^{q-1} (2p)!((q-p)!)^2 \binom{q}{q-p}^4 \|f_n \,\widetilde{\star}_{q-p}^{q-p} f_n\|^2 \\
&= \sum_{p=1,\, p\neq q/2}^{q-1} (2(q-p))!(p!)^2 \binom{q}{p}^4 \|f_n \,\widetilde{\star}_p^p f_n\|^2 \\
&\geq \sum_{p=1,\, p\neq q/2}^{q-1} 2((q-p)!)^2 (p!)^2 \binom{q}{p}^4 \|f_n \star_p^p f_n\|^2 \\
&= \sum_{p=1}^{q/2-1} \frac{4(q!)^4}{((q-p)!)^2(p!)^2} \|f_n \star_p^p f_n\|^2.
\end{aligned}$$

Hence, we end up with

$$\begin{aligned}
&\sum_{p=1}^{q/2-1} \frac{(q!)^4}{(p!)^2} \left(\frac{2}{(q-p)!^2} - \frac{1}{2\big((q/2)!(q/2-p)!\big)^2} \right) \|f_n \star_p^p f_n\|^2 + \sum_{p=1, p\neq q}^{2q-1} p!\|G_p^q f_n\|^2 \\
&\geq \sum_{p=1}^{q/2-1} \frac{(q!)^4}{(p!)^2} \left(\frac{6}{(q-p)!^2} - \frac{1}{2\big((q/2)!(q/2-p)!\big)^2} \right) \|f_n \star_p^p f_n\|^2.
\end{aligned} \tag{5.13}$$

For $q = 4, p = 1$ we have that

$$\frac{6}{(q-p)!^2} - \frac{1}{2\big((q/2)!(q/2-p)!\big)^2} = \frac{1}{24} > 0\,.$$

So, for $q = 4$ (and $q = 2$), the term $A'(I_q(f_n))$ is bounded from below by a linear combination with positive coefficients of the norms of the contraction kernels appearing in (5.10), (5.11) and (5.12) (while for all even $q \geq 6$ this cannot be guaranteed any more). This proves both statements of the lemma. □

Remark 5.8. As anticipated, for all even $q \geq 6$ there are combinatorial coefficients in (5.13) which are negative, implying that our proof cannot be generalized to Poisson integrals of arbitrary order. The reason is that one would need a sharper version of Lemma 5.4, which is in general not available as discussed in Remark 5.5. As a consequence, we have to leave it as an open problem to establish a four moments theorem for the Gamma approximation for Poisson integrals of order $q \geq 6$ by different methods.

It remains to check whether the conditions of Theorem 3.5 are sufficient to imply that $A'(I_q(f_n)) \to 0$. The following lemma shows that this is indeed the case.

Lemma 5.9. *Let $\nu > 0$ and $q \in \{2,4\}$. Let $f_n \in L^2_s(\mu_n^q)$, $n \geq 1$, be a sequence of kernels satisfying the technical assumptions (A) and the normalization condition*

$$\lim_{n\to\infty} q!\|f_n\|^2 = \lim_{n\to\infty} \mathbb{E}[I_q(f_n)^2] = 2\nu\,.$$

Let the sequence $\{I_q(f_n)^4 \colon n \geq 1\}$ be uniformly integrable. If one of the conditions

(a) *$q = 2$ and $\lim_{n\to\infty} \|f_n^2\| = 0$,*

(b) *$q = 4$ and $f_n \leq 0$ for all $n \geq 1$,*

is satisfied, then the following implication is true. If

$$\lim_{n\to\infty} \mathbb{E}[I_q(f_n)^4] - 12\mathbb{E}[I_q(f_n)^3] = 12\nu^2 - 48\nu$$

then

$$\lim_{n\to\infty} \|f_n \star_r^\ell f_n\| = 0 \tag{5.14}$$

for all $r \in \{1,\dots,q\}$ and $\ell \in \{1,\dots,r\wedge(q-1)\}$ such that $(r,\ell) \neq (q/2,q/2)$,

$$\lim_{n\to\infty} \|f_n^2\| = 0, \tag{5.15}$$

$$\lim_{n\to\infty} \|f_n \,\widetilde{\star}_{q/2}^{q/2} f_n - c_q f_n\| = 0 \qquad \text{with } c_q = \frac{4}{(q/2)!\binom{q}{q/2}^2}\,. \tag{5.16}$$

Proof: First apply Lemma 5.6 to deduce that $A(I_q(f_n)) + R(I_q(f_n)) \to 0$, as $n \to \infty$.

Assume that $q = 2$ and $\|f_n^2\| \to 0$. Then (5.15) is satisfied by assumption. Moreover,

$$R(I_2(f_n)) = 32\langle f_n \,\widetilde{\star}_1^1 f_n, f_n \,\widetilde{\star}_2^0 f_n\rangle - 48\langle f_n \,\widetilde{\star}_2^0 f_n, f_n\rangle.$$

By the Cauchy-Schwarz inequality, we see that

$$|\langle f_n \,\widetilde{\star}_1^1 f_n, f_n \,\widetilde{\star}_2^0 f_n\rangle| \leq \|f_n \,\widetilde{\star}_1^1 f_n\|\,\|f_n \,\widetilde{\star}_2^0 f_n\|, \qquad |\langle f_n \,\widetilde{\star}_2^0 f_n, f_n\rangle| \leq \|f_n\|\,\|f_n \,\widetilde{\star}_2^0 f_n\|.$$

With respect to the definition of the contractions, we see that $f_n \,\widetilde{\star}_2^0 f_n = f_n^2$. We shall argue now that the sequence $\|f_n \,\widetilde{\star}_1^1 f_n\|$ is bounded. For this, observe that for any fixed $(s,t) \in \mathcal{Z}^2$, we obtain by the Cauchy-Schwarz inequality that

$$\begin{aligned} |f_n \star_1^1 f_n(t,s)| &= \left| \int_{\mathcal{Z}} f_n(z,t) f_n(z,s) \mu(\mathrm{d}z) \right| \\ &\leq \left(\int_{\mathcal{Z}} f_n^2(z,t)\mu(\mathrm{d}z) \right)^{1/2} \left(\int_{\mathcal{Z}} f_n^2(z,s)\mu(\mathrm{d}z) \right)^{1/2}. \end{aligned}$$

Consequently,

$$\|f_n \,\widetilde{\star}_1^1 f_n\|^2 \leq \|f_n \star_1^1 f_n\|^2 = \int_{\mathcal{Z}^2} |f_n \star_1^1 f_n(t,s)|^2 \mu^2(\mathrm{d}(s,t)) \leq \|f_n\|^4.$$

By assumption, we have that $\|f_n\|^2 \to \nu$, so the sequence is bounded. Now, the fact that $\|f_n^2\| \to 0$ implies that $R(I_2(f_n)) \to 0$. Hence, $A(I_2(f_n)) \to 0$, which implies that $A'(I_2(f_n)) \to 0$ using Lemma 5.7(1). Now, we apply Lemma 5.7(2) to see that (5.14) and (5.16) follow.

Next, let $q = 4$ and suppose that $f_n \leq 0$. Recall that the tensor product is bi-linear and it is easily verified that the contraction operation preserves this bi-linearity. Now, the fact that the kernels are non-positive ensures that $R(I_q(f_n)) \geq 0$ and we can again apply Lemma 5.7(1) to see that $0 \leq A'(I_q(f_n)) \leq A(I_q(f_n))$. Hence, we deduce that $A(I_q(f_n)) \to 0$. This directly implies that $A'(I_q(f_n)) \to 0$, such that the claim follows again by Lemma 5.7(2). □

Remark 5.10. Let us explain in some more detail why in contrast to the case of normal approximation the kernels have to be non-positive for Gamma approximations. An inspection of the proof of Theorem 3.5 shows that a constant sign of the kernels is necessary to control the sign of scalar products. This is necessary in Lemma 5.4 and therefore also in Lemma 5.7 to control the signs of $A(I_q(f_n))$ and $A'(I_q(f_n))$, respectively. On the other hand, this is also necessary in part b) of Lemma 5.9, where one has to control the sign of $R(I_q(f_n))$. In this context, scalar products of the form $\langle f_n \widetilde{\star}_p^{q-p} f_n, f_n \rangle$, $p \in \{q/2+1, \dots, q\}$, appear. They are thrice-linear in f_n, such that $f_n \leq 0$ implies that $\langle f_n \widetilde{\star}_p^{q-p} f_n, f_n \rangle \leq 0$ and we can conclude that $R(I_q(f_n)) \geq 0$. In summary, knowing that $A'(I_q(f_n)) \geq 0$ and $R(I_q(f_n)) \geq 0$ enables us to use part (2) of Lemma 5.7 to get the implication (ii) $\Longrightarrow$ (iii) in Theorem 3.5. Note that the latter scalar products in $R(I_q(f_n))$ actually stem from the third moment in assertion (ii) of Theorem 3.5 (see also (2.8)).

It is worth mentioning that this asymmetry in the assertions for Theorem 3.5 (and also in Theorem 3.3) is actually an intrinsic property of the Gamma distribution which contrasts the normal case. For the central limit theorem in a Poisson chaos, it can be easily seen that if the law of the sequence $\{I_q(f_n) \colon n \geq 1\}$ converges to a standard normal law $\mathcal{N}(0,1)$, then also the law of $\{I_q(-f_n) \colon n \geq 1\} = \{-I_q(f_n) \colon n \geq 1\}$ converges to $\mathcal{N}(0,1)$, since the standard normal law is symmetric. Consistently, assertions (ii) and (iii) in the four moments theorem for normal approximation are invariant under a sign change of the kernels. In sharp contrast, the assertions for Gamma approximations are not invariant under such a sign change because of the lack of symmetry of the target distribution. This means that if the law of $\{I_q(f_n) \colon n \geq 1\}$ converges to $\overline{\Gamma}_\nu$ then that law of $\{I_q(-f_n) \colon n \geq 1\} = \{-I_q(f_n) \colon n \geq 1\}$ cannot converge to $\overline{\Gamma}_\nu$. Consistently, assertions (ii) and (iii) in Theorem 3.5 inherit this asymmetry, which is reflected by the appearance of the third moment in (ii) and the term $\| f_n \widetilde{\star}_{q/2}^{q/2} f_n - c_q f_n \|$ in (iii), both of them not being invariant under a change of the sign of f_n.

5.4. *An alternative approach to the four moments theorem.* In Remark 5.10 we explained that the sign condition on the kernels in part (b) of Theorem 3.5 ensures that $R(I_q(f_n)) \geq 0$. Together with $A'(I_q(f_n)) \geq 0$, this is sufficient in combination with part (2) of Lemma 5.7 to get the implication (ii) $\Longrightarrow$ (iii) in Theorem 3.5. On the other hand, for part (a) of Theorem 3.5, dealing with the case $q = 2$, the assumption that $\|f_n^2\| \to 0$ yields that $R(I_2(f_n)) \to 0$, an assertion also being sufficient in combination with $A'(I_q(f_n)) \geq 0$ to deduce the implication (ii) $\Longrightarrow$ (iii) in Theorem 3.5 from part (2) of Lemma 5.7. From this point of view, it is natural to ask whether the latter condition can be generalized to arbitrary $q \geq 2$. Our next result shows that this is indeed possible, but leads to a result which is weaker than Theorem 3.5. Moreover, the proof again only works for $q = 4$ and we still have to impose a sign condition on the sequence of kernels.

Proposition 5.11. *Fix $\nu > 0$. Let $f_n \in L_s^2(\mu_n^4)$, $n \geq 1$, be a sequence of kernels such that $f_n \geq 0$ for all $n \geq 1$ and such that the technical assumptions (A) and the normalization condition*

$$\lim_{n\to\infty} 4!\|f_n\|^2 = \lim_{n\to\infty} \mathbb{E}[I_4(f_n)^2] = 2\nu$$

are satisfied. Assume additionally that

$$\lim_{n\to\infty} \|f_n^2\| = 0, \ \ \textit{and} \ \lim_{n\to\infty} \|f_n \star_3^1 f_n\| = 0. \tag{5.17}$$

If the sequence $\{I_4(f_n)^4 \colon n \geq 1\}$ is uniformly integrable, then the equivalence stated in Theorem 3.5 remains valid.

Proof: The implication (i) $\Longrightarrow$ (ii) follows from the uniform integrability of the sequence $\{I_4(f_n)^4 \colon n \geq 1\}$ and (iii) $\Longrightarrow$ (i) is a consequence of Proposition 5.1. To establish the implication (ii) $\Longrightarrow$ (iii), we apply Lemma 5.6 and show that the term $R(I_4(f_n))$ defined at (5.9) converges to zero, as $n \to \infty$. With the Cauchy-Schwarz inequality we obtain for $p \in \{3, 4\}$ that

$$|\langle f_n \widetilde{\star}_p^{4-p} f_n, f_n\rangle| \leq \|f_n \star_p^{4-p} f_n\| \, \|f_n\| \to 0 \,,$$

since $\|f_n \star_p^{4-p} f_n\| \to 0$ and $\|f_n\|^2 \to \frac{\nu}{12}$. Moreover, for $p, r \in \{2, 3, 4\}$ with $p \neq r$ we also get

$$|\langle f_n \widetilde{\star}_p^{4-p} f_n, f_n \widetilde{\star}_r^{4-r} f_n\rangle| \leq \|f_n \widetilde{\star}_p^{4-p} f_n\| \, \|f_n \widetilde{\star}_r^{4-r} f_n\| \to 0 \,.$$

The convergence is ensured by condition (5.17) if $p, r > 2$. If otherwise $p \wedge r = 2$, we use condition (5.17) together with the observation that $\|f_n \star_2^0 f_n\| = \|f_n \star_4^2 f_n\|$ and $\|f_n \star_2^2 f_n\| \leq \|f_n \star_4^4 f_n\|$ as a consequence of Fubini's theorem and the Cauchy-Schwarz inequality. Summarizing, we see that $R(I_4(f_n)) \to 0$, which in turn implies that $A(I_q(f_n)) \to 0$ thanks to Lemma 5.6. We can then conclude as in the proof of part (b) of Lemma 5.9. $\square$

Acknowledgements

We would like to thank Johanna F. Ziegel for initiating this collaboration and Giovanni Peccati for helpful discussions and valuable comments.

References

S. Bourguin. Poisson convergence on the free Poisson algebra. *Bernoulli* **21** (4), 2139–2156 (2015). MR3378462.

S. Bourguin and G. Peccati. Semicircular limits on the free Poisson chaos: counterexamples to a transfer principle. *J. Funct. Anal.* **267** (4), 963–997 (2014). MR3217054.

E. B. Dynkin and A. Mandelbaum. Symmetric statistics, Poisson point processes, and multiple Wiener integrals. *Ann. Statist.* **11** (3), 739–745 (1983). MR707925.

P. Eichelsbacher and C. Thäle. New Berry-Esseen bounds for non-linear functionals of Poisson random measures. *Electron. J. Probab.* **19**, no. 102, 25 (2014). MR3275854.

P. de Jong. A central limit theorem for generalized quadratic forms. *Probab. Theory Related Fields* **75** (2), 261–277 (1987). MR885466.

P. de Jong. A central limit theorem for generalized multilinear forms. *J. Multivariate Anal.* **34** (2), 275–289 (1990). MR1073110.

R. Lachièze-Rey and G. Peccati. Fine Gaussian fluctuations on the Poisson space, I: contractions, cumulants and geometric random graphs. *Electron. J. Probab.* **18**, no. 32, 32 (2013). MR3035760.

G. Last. Stochastic analysis for Poisson processes. *ArXiv e-prints* (2014). 1405.4416.

G. Last, G. Peccati and M. Schulte. Normal approximation on Poisson spaces: Mehler's formula, second order Poincaré inequalities and stabilization (2015+). To appear in Probab. Theory Related Fields.

G. Last, M. D. Penrose, M. Schulte and C. Thäle. Moments and central limit theorems for some multivariate Poisson functionals. *Adv. in Appl. Probab.* **46** (2), 348–364 (2014). MR3215537.

I. Nourdin and G. Peccati. Noncentral convergence of multiple integrals. *Ann. Probab.* **37** (4), 1412–1426 (2009). MR2546749.

I. Nourdin and G. Peccati. *Normal approximations with Malliavin calculus*, volume 192 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge (2012). ISBN 978-1-107-01777-1. From Stein's method to universality. MR2962301.

I. Nourdin, G. Peccati and G. Reinert. Invariance principles for homogeneous sums: universality of Gaussian Wiener chaos. *Ann. Probab.* **38** (5), 1947–1985 (2010). MR2722791.

D. Nualart and G. Peccati. Central limit theorems for sequences of multiple stochastic integrals. *Ann. Probab.* **33** (1), 177–193 (2005). MR2118863.

D. Nualart and J. Vives. Anticipative calculus for the Poisson process based on the Fock space. In *Séminaire de Probabilités, XXIV, 1988/89*, volume 1426 of *Lecture Notes in Math.*, pages 154–165. Springer, Berlin (1990). MR1071538.

G. Peccati, J. L. Solé, M. S. Taqqu and F. Utzet. Stein's method and normal approximation of Poisson functionals. *Ann. Probab.* **38** (2), 443–478 (2010). MR2642882.

G. Peccati and M. S. Taqqu. Central limit theorems for double Poisson integrals. *Bernoulli* **14** (3), 791–821 (2008). MR2537812.

G. Peccati and M. S. Taqqu. *Wiener chaos: moments, cumulants and diagrams*, volume 1 of *Bocconi & Springer Series*. Springer, Milan; Bocconi University Press, Milan (2011). ISBN 978-88-470-1678-1. A survey with computer implementation, Supplementary material available online. MR2791919.

G. Peccati and C. Thäle. Gamma limits and U-statistics on the Poisson space. *ALEA Lat. Am. J. Probab. Math. Stat.* **10** (1), 525–560 (2013). MR3083936.

G. Peccati and C. Zheng. Multi-dimensional Gaussian fluctuations on the Poisson space. *Electron. J. Probab.* **15**, no. 48, 1487–1527 (2010). MR2727319.

G. Peccati and C. Zheng. Universal Gaussian fluctuations on the discrete Poisson chaos. *Bernoulli* **20** (2), 697–715 (2014). MR3178515.

M. Reitzner and M. Schulte. Central limit theorems for U-statistics of Poisson point processes. *Ann. Probab.* **41** (6), 3879–3909 (2013). MR3161465.

M. Schulte. Normal Approximation of Poisson Functionals in Kolmogorov Distance. *J. Theoret. Probab.* **29** (1), 96–117 (2016). MR3463079.

M. Schulte and C. Thäle. The scaling limit of Poisson-driven order statistics with applications in geometric probability. *Stochastic Process. Appl.* **122** (12), 4096–4120 (2012). MR2971726.

M. Schulte and C. Thäle. Distances between Poisson k-flats. *Methodol. Comput. Appl. Probab.* **16** (2), 311–329 (2014). MR3199049.

Erratum to: "A four moments theorem for Gamma limits on a Poisson chaos"

Tobias Fissler* Christoph Thäle†

24th January 2017

Abstract

This note corrects a condition in Theorem 3.5 in our paper Fissler and Thäle (2016).

It has been pointed out to us that the assertions of the equivalence stated in Theorem 3.5(b) of our paper Fissler and Thäle (2016) cannot be satisfied by a sequence of kernels $f_n \in L_s^2(\mu_n^q)$ in the case $q = 4$. Indeed, the sign condition $f_n \le 0$ implies that $\mathbb{E}[I_q(f_n)^3] \le 0$ in view of equation (2.8) in Fissler and Thäle (2016). On the other hand, if $I_q(f_n)$ converges in distribution to $Y \sim \overline{\Gamma}_\nu$, as $n \to \infty$, the uniform integrability of $\{I_q(f_n)^4 : n \ge 1\}$ implies that

$$\lim_{n\to\infty} \mathbb{E}[I_q(f_n)^3] = \mathbb{E}[Y^3] = 8\nu > 0,$$

which is a contradiction. In line with the equivalence, neither assertion (ii) nor (iii) can be satisfied if $f_n \le 0$. E.g. $c_q^2 \|f_n\|^2 \le \|f_n \,\widetilde{\star}_{q/2}^{q/2} f_n - c_q f_n\|^2 \to 0$ for $f_n \le 0$, but at the same time $q!\|f_n\|^2 \to 2\nu > 0$. This contradiction also affects the results based on Theorem 3.5(b), namely Corollary 3.8(b), Theorem 4.6(b), and Theorem 4.10(b).

In Section 5.4 and, in particular, in Proposition 5.11 of Fissler and Thäle (2016), we described an alternative way to a four moments theorem in the case $q = 4$ and for non-negative kernels under stronger conditions on the contraction norms of the kernels f_n. Against this background, Theorem 3.5 holds upon replacing condition (b) there by

(b') $q = 4$, $f_n \ge 0$ for all $n \ge 1$ and $\lim_{n\to\infty} \|f_n \star_p^{q-p} f_n\| = 0$ for all $p \in \{q/2+1, \dots, q\}$.

Mutatis mutandis, condition (b) in Corollary 3.8 should be replaced by

(b') $q = 4$, $f_n \le 0$ for all $n \ge 1$ and $\lim_{n\to\infty} \|f_n \star_p^{q-p} f_n\| = 0$ for all $p \in \{q/2+1, \dots, q\}$;

condition (b) in Theorem 4.6 should be replaced by

(b') $q = 4$, $f_n \ge 0$ for all $n \ge 1$ and $\lim_{n\to\infty} \|f_n \star_p^{q-p} f_n\| = 0$ for all $p \in \{q/2+1, \dots, q\}$;

and condition (b) in Theorem 4.10 should be replaced by

*University of Bern, Department of Mathematics and Statistics, Institute of Mathematical Statistics and Actuarial Science, Sidlerstrasse 5, CH-3012 Bern, Switzerland. E-mail: tobias.fissler@stat.unibe.ch

†Ruhr University Bochum, Faculty of Mathematics, D-44780 Bochum, Germany. E-mail: christoph.thaele@rub.de

(b') $h_n \geq 0$ for all $n \geq 1$ and $\lim_{n\to\infty} \|h_n \star_p^{q-p} h_n\| = 0$ for all $p \in \{q/2+1, \dots, q\}$ if $q = 4$.

We remark that also under the new condition (b') in Theorem 3.5, all the technical lemmas established in (Fissler and Thäle, 2016, Section 5) are needed to prove the result.

We finally remark that in (Fissler and Thäle, 2016, Lemma 5.7), we assumed that the kernels f_n have constant sign, i.e., that either $f_n \leq 0$ or $f_n \geq 0$. However, for $q = 2$, one can dispense with the sign condition. Indeed, using the notation from Fissler and Thäle (2016), we have that

$$\begin{aligned} A'(I_2(f_n)) &= \|G_1^2 f_n\|^2 + 6\|G_3^2 f_n\|^2 + 8\|f_n \widetilde{\star}_2^0 f_n\|^2 + 48\|f_n \widetilde{\star}_1^1 f_n - f_n\|^2 \\ &= 16\|f_n \star_2^1 f_n\|^2 + 96\|f_n \widetilde{\star}_1^0 f_n\|^2 + 8\|f_n^2\|^2 + 48\|f_n \widetilde{\star}_1^1 f_n - f_n\|^2. \end{aligned}$$

Hence, assertions (1) and (2) follow directly. This is of importance because in Lemma 5.9(a) we imposed no sign condition, but referred to Lemma 5.7.

Theorem 1.6 of the recent paper Döbler and Peccati (2017) is very close to establishing a four moments theorem for Poisson integrals with a Gamma limit. However, as discussed in Remark 1.7 ibidem, one sufficient condition which implies the four moments theorem is that certain contraction norms of the kernels converge to zero in the L^2-sense. This corresponds to our additional assumption from (b') above. However, the theory developed in Döbler and Peccati (2017) allows to remove our restrictive condition on the order of the integrals as well as the sign condition on the kernels f_n.

Acknowledgements

We would like to thank Giovanni Peccati and Christian Döbler for pointing out the contradiction to us.

References

C. Döbler and G. Peccati. The fourth moment theorem on the Poisson space. *ArXiv e-prints*, January 2017. URL `https://arxiv.org/abs/1701.03120`.

T. Fissler and C. Thäle. A four moments theorem for Gamma limits on a Poisson chaos. *ALEA Lat. Am. J. Probab. Math. Stat.*, 13(1):163–192, 2016. ISSN 1980-0436.

9. A new quantitative central limit theorem on the Wiener space with applications to Gaussian processes

This chapter contains the joint article Fissler and Thäle (2016b). It is devoted to establish quantitative central limit theorems for Gaussian functionals with a possibly *infinite* chaos decomposition. The applicability is demonstrated in terms of the Breuer-Major theorem.

The article has been submitted, but is currently under review. The version included in this thesis is identical in its content to the arXiv-version, which is available at `https://arxiv.org/abs/1610.01456`. The formatting is almost the same with some slight adaptations to allow for a better inclusion into the thesis.

A new quantitative central limit theorem on the Wiener space with applications to Gaussian processes

Tobias Fissler* Christoph Thäle†

Abstract

Quantitative limit theorems for non-linear functionals on the Wiener space are considered. Given the possibly infinite sequence of kernels of the chaos decomposition of such a functional, an estimate for different probability distances between the functional and a Gaussian random variable in terms of contraction norms of these kernels is derived. The applicability of this result is demonstrated by means of the Breuer-Major theorem, unfolding thereby a new connection between the Hermite rank of the considered function and a chaotic gap. Especially, power variations of the fractional Brownian motion and processes belonging to the Cauchy class are studied.

Keywords. Breuer-Major theorem, central limit theorem, chaos decomposition, chaotic gap, fractional Brownian motion, Gaussian process, Hermite rank, Malliavin-Stein method, power variation, Wiener space
MSC. Primary 60F05, 60G15; Secondary 60G22, 60H05, 60H07.

1 Introduction

Central limit theorems for non-linear functionals of Gaussian random processes (or measures) have triggered an enormous development in probability theory and mathematical statistics during the last decade. A cornerstone in this new area is the so-called fourth moment theorem of Nualart and Peccati. It says that a sequence $I_q(f^{(n)})$ of Gaussian multiple stochastic integrals of a fixed order $q \geq 2$ satisfying the normalization condition $\mathbb{E}[I_q(f^{(n)})^2] = 1$ for all $n \geq 1$ converges in distribution, as $n \to \infty$, to a standard Gaussian random variable Z if and only if the sequence $\mathbb{E}[I_q(f^{(n)})^4]$ of their fourth moments converges to 3, the fourth moment of Z. This qualitative limit theorem has been extended by Nourdin and Peccati in [7] to a quantitative statement in that the distance between the laws of $I_q(f^{(n)})$ and Z is measured in a suitable probability metric. For example, the total variation distance $d_{TV}(I_q(f^{(n)}), Z)$ between $I_q(f^{(n)})$ and the Gaussian variable Z can be bounded from above by

$$d_{TV}(I_q(f^{(n)}), Z) \leq C\sqrt{\mathbb{E}[I_q(f^{(n)})] - 3} \tag{1}$$

with a constant $C \in (0, \infty)$ only depending on q. More recently, Nourdin and Peccati [9] derived the optimal rate of convergence, removing thereby the square-root in (1). We

*University of Bern, Department of Mathematics and Statistics, Institute of Mathematical Statistics and Actuarial Science, Sidlerstrasse 5, CH-3012 Bern, Switzerland. E-mail: tobias.fissler@stat.unibe.ch
†Ruhr University Bochum, Faculty of Mathematics, D-44780 Bochum, Germany. E-mail: christoph.thaele@rub.de

emphasize that the proof of the estimate (1) is based on a combination of Stein's method for normal approximation with the Malliavin calculus of variations on the Wiener space. For further information and background material, we refer the reader to the monograph [8].

While the Malliavin-Stein approach provides useful estimates in case of a sequence of random elements living inside a fixed Wiener chaos or inside a finite sum of Wiener chaoses, the bounds become less tractable in cases in which the functionals belong to an infinite sum of Wiener chaoses, that is, if the functional F has the representation

$$F = \sum_{q=0}^{\infty} I_q(f_q) \tag{2}$$

with infinitely many of the functions f_q (called kernels in the sequel) being non-zero. On the other hand, functionals of this type often appear in concrete applications. Distinguished examples are the number of zeros of a random trigonometric polynomial [1], the power and the bi-power variation of a Gaussian random process [2], the Euler characteristic of a Gaussian excursion set [4] or the statistics appearing around the Breuer-Major theorem [3, 10], to name just a few. One way to obtain quantitative central limit theorems in these cases is to apply the so-called second-order Poincaré inequality developed by Nourdin, Peccati and Reinert [11]. This method has the advantage that it is not necessary to specify the chaos decomposition of F as at (2) explicitly, that is, to compute the functions f_q there explicitly. On the other hand, a major drawback of this approach is that it typically leads to a suboptimal rate of convergence. Moreover, in many situations the kernels f_q are in fact explicitly known and for this reason it is natural to ask for a purely analytical upper bound on the probability distance between F and Z in terms of the sequence of kernels f_q. The main goal of the present paper is to provide such an estimate (also for probability metrics different from the total variation distance) and to demonstrate its applicability by means of representative examples related to the classical Breuer-Major theorem. More precisely, we shall look at random variables of the form

$$F_n = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \{g(X_k) - \mathbb{E}[g(X_k)]\},$$

where $X = (X_k)_{k\in\mathbb{Z}}$ is a stationary Gaussian process and $g : \mathbb{R} \to \mathbb{R}$ is a suitable function. For example, X could be obtained from the increments of a fractional Brownian motion and $g(x) = |x|^p - \mathbb{E}|X_1|^p$, $p > 0$, in which case F_n becomes a so-called power variation of the fractional Brownian motion. In this context, our quantitative central limit theorem for F_n unfolds a new and unexpected feature, namely that the rates of convergence are influenced by the interplay of the Hermite rank of the function g and what we call the chaotic gap of F_n (in addition to the asymptotic behavior of the covariance function of X, of course). We would like to emphasize that in the context of power variations of a fractional Brownian motion we will show that the rate of convergence in the central limit theorem is universal, that is, independent of the parameter p, and coincides with the known rate for the quadratic variation, where $p = 2$. The same phenomenon also applies to processes that belong to the Cauchy class.

Our text is structured as follows. In Section 2, we summarize some basic elements of Gaussian analysis and, in particular, recall the definitions of the four basic operators from

Malliavin calculus that are crucial for our theory. Our main result, Theorem 3.1, is presented in Section 3. Our applications to Gaussian random processes are the content of Section 4. Finally, in Section 5 we present a multivariate exteions of our main result. The Appendix gathers some technical lemmas.

2 Elements of Gaussian analysis

2.1 Wiener chaos, chaos decomposition and multiplication formula

We let $\mathfrak{H}$ be a real separable Hilbert space with inner product $\langle \cdot, \cdot \rangle_{\mathfrak{H}}$ and norm $\| \cdot \|_{\mathfrak{H}}$. Moreover, for integers $q \geq 1$ we denote by $\mathfrak{H}^{\otimes q}$ the qth tensor power and by $\mathfrak{H}^{\odot q}$ the qth symmetric tensor power of $\mathfrak{H}$. The space $\mathfrak{H}^{\otimes q}$ is supplied with the canonical scalar product $\langle \cdot, \cdot \rangle_{\mathfrak{H}^{\otimes q}}$ and the canonical norm $\| \cdot \|_{\mathfrak{H}^{\otimes q}}$, while $\mathfrak{H}^{\odot q}$ is equipped with the norm $\sqrt{q!}\| \cdot \|_{\mathfrak{H}^{\otimes q}}$. An isonormal Gaussian process $W = \{W(h) : h \in \mathfrak{H}\}$ over $\mathfrak{H}$ is a family of Gaussian random variables defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and indexed by the elements of $\mathfrak{H}$ such that

$$\mathbb{E}[W(h)] = 0 \qquad \text{and} \qquad \mathbb{E}[W(h)W(h')] = \langle h, h' \rangle_{\mathfrak{H}}, \qquad h, h' \in \mathfrak{H}.$$

In what follows we will implicitly assume that the σ-field $\mathcal{F}$ is generated by W, that is, $\mathcal{F} = \sigma(W)$. Let us write $L^2(\Omega)$ for the space of square-integrable functions over Ω. For integers $q \geq 1$ we denote by $\mathfrak{C}_q$ the qth Wiener chaos over $\mathfrak{H}$. That is, $\mathfrak{C}_q$ is the closed linear subspace of $L^2(\Omega)$ generated by random variables of the form $H_q(W(h))$. Here, H_q is the qth Hermite polynomial and $h \in \mathfrak{H}$ satisfies $\|h\|_{\mathfrak{H}} = 1$. Recall that $H_0 \equiv 0$ and that

$$H_q(x) = (-1)^q \exp(x^2/2) \frac{\mathrm{d}^q}{\mathrm{d}x^q} \exp(-x^2/2), \qquad q \geq 1, \tag{3}$$

$$\mathbb{E}[H_q(X)H_p(Y)] = \begin{cases} p!\,(\mathbb{E}[XY])^p & : p = q \\ 0 & : \text{otherwise}, \end{cases} \tag{4}$$

for jointly Gaussian X, Y and integers $p, q \geq 1$. For convenience, we also define $\mathfrak{C}_0 := \mathbb{R}$. The mapping $h^{\otimes q} \mapsto H_q(W(h))$ can be extended to a linear isometry, denoted by I_q, from $\mathfrak{H}^{\odot q}$ to the qth Wiener chaos $\mathfrak{C}_q$, see Chapter 2 in [8]. We put $I_q(h) := I_q(\tilde{h})$ for general $h \in \mathfrak{H}^{\otimes q}$ where $\tilde{h} \in \mathfrak{H}^{\odot q}$ is the canonical symmetrization of h, and we use the convention that $I_0 \colon \mathbb{R} \to \mathbb{R}$ is the identity map. In particular, if $\mathfrak{H} = L^2(A)$ with a σ-finite non-atomic measure space $(A, \mathcal{A}, \mu)$, then I_q possesses an interpretation as a multiple stochastic integral of order q with respect to the Gaussian random measure on A with control measure μ. We refer to Chapter 2.7 in [8] for further details and explanations.

According to Theorem 2.2.4 in [8], every $F \in L^2(\Omega)$ admits a chaotic decomposition. In particular, this means that

$$F = \sum_{q=0}^{\infty} I_q(h_q)$$

with $h_0 = \mathbb{E}[F] := \int F \,\mathrm{d}\mathbb{P}$ and uniquely determined elements $h_q \in \mathfrak{H}^{\odot q}$, $q \geq 1$, that are called the kernels of the chaotic decomposition. We also mention that, for $q \geq 1$, $\mathbb{E}[I_q(h_q)] = 0$ and that

$$\mathbb{E}[I_p(h_p)I_q(h_q)] = \begin{cases} p!\langle h_p, h_q \rangle_{\mathfrak{H}^{\otimes q}} & : p = q \\ 0 & : \text{otherwise} \end{cases} \tag{5}$$

for $h_p \in \mathfrak{H}^{\odot p}$, $h_q \in \mathfrak{H}^{\odot q}$, $p, q \geq 1$, which implies that the variance of F satisfies

$$\mathrm{var}(F) := \mathbb{E}[F^2] - (\mathbb{E}[F])^2 = \sum_{q=1}^{\infty} q! \|h_q\|^2_{\mathfrak{H}^{\otimes q}} . \tag{6}$$

More generally, the covariance of $F = \sum_{q=0}^{\infty} I_q(h_q) \in L^2(\Omega)$ and $G = \sum_{q=0}^{\infty} I_q(h'_q) \in L^2(\Omega)$ is given by

$$\mathrm{cov}(F, G) := \mathbb{E}[FG] - \mathbb{E}[F]\mathbb{E}[G] = \sum_{q=1}^{\infty} q! \langle h_q, h'_q \rangle_{\mathfrak{H}^{\otimes q}} . \tag{7}$$

Another crucial fact is that the product of two multiple stochastic integrals can be expressed as a linear combination of multiple stochastic integrals. More generally, let $p, q \geq 1$ be integers and $h \in \mathfrak{H}^{\odot p}$, $h' \in \mathfrak{H}^{\odot q}$. Then one has the multiplication formula

$$I_q(h) I_p(h') = \sum_{r=0}^{\min(p,q)} r! \binom{q}{r} \binom{p}{r} I_{p+q-2r}(h \widetilde{\otimes}_r h'), \tag{8}$$

where $h \widetilde{\otimes}_r h' := \widetilde{h \otimes_r h'}$ stands for the canonical symmetrization of the contraction $h \otimes_r h' \in \mathfrak{H}^{\otimes p+q-2r}$. Note that for $h = h_1 \otimes \cdots \otimes h_p \in \mathfrak{H}^{\otimes p}$ and $h' = h'_1 \otimes \cdots \otimes h'_q \in \mathfrak{H}^{\otimes q}$ the contraction can be defined as

$$h \otimes_r h' := \langle h_1, h'_1 \rangle_{\mathfrak{H}} \cdots \langle h_r, h'_r \rangle_{\mathfrak{H}} \, [h_{r+1} \otimes \cdots \otimes h_p \otimes h'_{r+1} \otimes \cdots \otimes h'_q] . \tag{9}$$

By linearity, the contraction operation can be extended to any $h \in \mathfrak{H}^{\otimes p}$ and $h' \in \mathfrak{H}^{\otimes q}$. In the case that $\mathfrak{H} = L^2(A)$ with a σ-finite non-atomic measure space $(A, \mathcal{A}, \mu)$, we have that $\mathfrak{H}^{\otimes q} = L^2(A^q) := L^2(A^q, \mathcal{A}^{\otimes q}, \mu^{\otimes q})$ and that $\mathfrak{H}^{\odot q}$ coincides with the space $L^2_{\mathrm{sym}}(A^q)$ of $\mu^{\otimes q}$-almost everywhere symmetric functions on A^q. Moreover,

$$\begin{aligned}(f \otimes_r g)(y_1, \ldots, y_{p+q-2r}) := \int_{A^r} & f(x_1, \ldots, x_r, y_1, \ldots, y_{p-r}) \\ & \times g(x_1, \ldots, x_r, y_{p-r+1}, \ldots, y_{p+q-2r}) \, \mu^{\otimes r}(\mathrm{d}(x_1, \ldots, x_r))\end{aligned}$$

with $f \in L^2_{\mathrm{sym}}(A^p)$, $g \in L^2_{\mathrm{sym}}(A^q)$ and $y_1, \ldots, y_{p+q-2r} \in A$.

Convention. Through our paper, we will adopt the following convention that whenever the Hilbert space $\mathfrak{H}$ coincides with an $L^2(A)$-space we write f instead of h for an element of $L^2(A)$ to underline that we are dealing with functions. Furthermore, we use the shorthand notation $\| \cdot \|_q$ for $\| \cdot \|_{\mathfrak{H}^{\otimes q}}$ for all integers $q \geq 1$.

2.2 Malliavin operators

In this section, we recall the definition of the four basic operators from Malliavin calculus and summarize those properties which are needed later. For that purpose and to simplify our presentation we assume from now on that $\mathfrak{H} = L^2(A)$ with a σ-finite non-atomic measure space $(A, \mathcal{A}, \mu)$. Note that because of isomorphy of Hilbert spaces, this is no restriction of generality. For further details we direct the reader to the monographs [6, 8, 12].

Malliavin derivative. Suppose that $F \in L^2(\Omega)$ has a chaos decomposition

$$F = \sum_{q=0}^{\infty} I_q(f_q), \qquad f_q \in L^2_{\mathrm{sym}}(A^q), \tag{10}$$

and suppose that $\sum_{q=1}^{\infty} q\, q! \|f_q\|_q^2 < \infty$. In this case we say that F belongs to the domain of D, formally we indicate this by writing $F \in \mathrm{dom}(D)$. For $F \in \mathrm{dom}(D)$ and $x \in A$ we define the Malliavin derivative of F in direction x as

$$D_x F := \sum_{q=1}^{\infty} q I_{q-1}(f_q(x, \cdot)), \tag{11}$$

where $f_q(x, \cdot) \in L^2_{\mathrm{sym}}(A^{q-1})$ stands for the function f_q with one of its variables fixed to be x (which one is irrelevant, since the functions f_q are symmetric).
We further define for all integers $k \geq 1$ the iterated Malliavin derivative $D^k F$ as

$$D^k_{x_1,\ldots,x_k} F := \sum_{q=k}^{\infty} q(q-1)\cdots(q-k+1)\, I_{q-k}(f_q(x_1,\ldots,x_k,\cdot)), \qquad x_1,\ldots,x_k \in A,$$

whenever $F \in \mathrm{dom}(D^k)$, that is, if $F = \sum_{q=0}^{\infty} I_q(f_q)$ satisfies $\sum_{q=k}^{\infty} q(q-1)\cdots(q-k+1)\|f_q\|_q^2 < \infty$.
Finally, we introduce the subspace $\mathbb{D}^{1,4}$ of $\mathrm{dom}(D)$ containing all $F \in L^4(\Omega)$ such that

$$\mathbb{E}\|DF\|_1^4 = \mathbb{E}\left| \int_A |D_x F|^2 \,\mu(\mathrm{d}x) \right|^2 < \infty,$$

see Chapter 2.3 in [8] for a formal construction. Moreover, we recall that the Malliavin derivative can be used to compute the kernels f_q in the chaotic decomposition of a given functional F. Namely, assuming that $F \in \mathrm{dom}(D^q)$ for some $q \geq 1$, Stroock's formula [8, Corollary 2.7.8] says that

$$f_q = \frac{1}{q!}\, \mathbb{E}[D^q F]. \tag{12}$$

Divergence. We write $L^2(A \times \Omega) := L^2(A \times \Omega, \mathcal{A} \otimes \mathcal{F}, \mu \otimes \mathbb{P})$ for the space of square-integrable random processes $u = (u_x)_{x \in A}$ on A. Fix such a process $u \in L^2(A \times \Omega)$ and suppose that it satisfies

$$\left| \mathbb{E} \int_A (D_x F)\, u_x \,\mu(\mathrm{d}x) \right| \leq c\, \mathbb{E}[F^2]$$

for all $F \in \mathrm{dom}(D)$ and some constant $c > 0$ that is allowed to depend on u. We denote the class of such processes by $\mathrm{dom}(\delta)$ and define for $u \in \mathrm{dom}(\delta)$ the divergence $\delta(u)$ of u by the duality relation

$$\mathbb{E}[F\delta(u)] = \mathbb{E} \int_A (D_x F)\, u_x \,\mu(\mathrm{d}x), \qquad F \in \mathrm{dom}(D).$$

That is, δ is the operator which is adjoint to the Malliavin derivative D.

The divergence can also be defined in terms of chaotic decompositions. Suppose that $u \in \mathrm{dom}(\delta)$ such that $u_x \in L^2(A \times \Omega)$ for all $x \in A$. Then there are kernels $f_q \in L^2(A^{q+1})$, $q \geq 0$, such that

$$u_x = \sum_{q=0}^{\infty} I_q(f_q(x, \cdot\,)), \qquad x \in A,$$

and $f_q(x, \cdot\,) \in L^2_{\mathrm{sym}}(A^q)$. Moreover, $u \in \mathrm{dom}(\delta)$ if and only if $\sum_{q=0}^{\infty} (q+1)!\|\tilde{f}_q\|^2 < \infty$ and in this case $\delta(u)$ is given by

$$\delta(u) = \sum_{q=0}^{\infty} I_{q+1}(\tilde{f}_q),$$

where

$$\tilde{f}_q(x_1, \dots, x_{q+1}) := \frac{1}{(q+1)!} \sum_{\pi} f(x_{\pi(1)}, \dots, x_{\pi(q+1)})$$

denotes the canonical symmetrization of $f_q \in L^2(A^{q+1})$ with the sum running over all permutations π of $\{1, \dots, q+1\}$.

Ornstein-Uhlenbeck generator and its pseudo-inverse. Let $F \in L^2(\Omega)$ be a square integrable functional with chaos decomposition as at (10) and define

$$LF := -\sum_{q=0}^{\infty} q I_q(f_q),$$

whenever the series converges in $L^2(\Omega)$. The domain $\mathrm{dom}(L)$ of L is the set of those $F \in L^2(\Omega)$ for which $\sum_{q=1}^{\infty} q^2\, q! \|f_q\|_q^2 < \infty$. The operator L is called the generator of the Ornstein-Uhlenbeck semigroup associated with the Gaussian random measure on A having control measure μ. By L^{-1} we denote its pseudo-inverse acting on centred $F \in L^2(\Omega)$ as follows:

$$L^{-1}F := -\sum_{q=1}^{\infty} \frac{1}{q} I_q(f_q). \tag{13}$$

For non-centred $F \in L^2(\Omega)$ we put $L^{-1}F := L^{-1}(F - \mathbb{E}[F])$. Clearly, for centred $F \in \mathrm{dom}(L)$ one has that $LL^{-1}F = L^{-1}LF = F$. Moreover, the operators D, δ and L are related by

$$\delta(DF) = -LF, \qquad F \in \mathrm{dom}(L).$$

In fact, according to [12, Proposition 1.4.8], $F \in \mathrm{dom}(L)$ is equivalent to $F \in \mathrm{dom}(D)$ and $DF \in \mathrm{dom}(\delta)$.

3 A quantitative central limit theorem

Let W be an isonormal Gaussian process defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and over a Hilbert space $\mathfrak{H}$ as in the previous section. Further, let $F \in L^2(\Omega)$. Then, as we have seen above, F admits the chaos decomposition

$$F = \sum_{q=0}^{\infty} I_q(h_q) \tag{14}$$

with $h_0 = \mathbb{E}[F]$ and kernels $h_q \in \mathfrak{H}^{\odot q}$, $q \geq 1$.

Our aim is to measure the distance between F and a centred Gaussian random variable with the same variance as F. We do this in terms of different probability metrics. To define them, recall that a collection Φ of measurable functions $\varphi : \mathbb{R} \to \mathbb{R}$ is said to be separating if for any two random variables Y and Y', $\mathbb{E}[\varphi(Y)] = \mathbb{E}[\varphi(Y')]$ for all $\varphi \in \Phi$ with $\mathbb{E}[\varphi(Y)], \mathbb{E}[\varphi(Y')] < \infty$ implies that Y and Y' are identically distributed. For such a class of functions Φ we define the probability metric d_Φ by putting

$$d_\Phi(Y, Y') := \sup_{\varphi \in \Phi} \left| \mathbb{E}[\varphi(Y)] - \mathbb{E}[\varphi(Y')] \right|,$$

where Y and Y' are random variables satisfying $\mathbb{E}[\varphi(Y)], \mathbb{E}[\varphi(Y')] < \infty$ for all $\varphi \in \Phi$. Examples for such probability metrics are

- the total variation distance $d_{TV} := d_{\Phi_{TV}}$, where $\Phi_{TV} = \{\mathbf{1}_B : B \subset \mathbb{R} \text{ a Borel set}\}$,
- the Kolmogorov distance $d_K := d_{\Phi_K}$, where $\Phi_K = \{\mathbf{1}_{(-\infty, x]} : x \in \mathbb{R}\}$,
- the Wasserstein distance $d_W := d_{\Phi_W}$, where Φ_W is the class of Lipschitz functions $\varphi : \mathbb{R} \to \mathbb{R}$ with $\|\varphi\|_{Lip} \leq 1$, where $\|\varphi\|_{Lip} := \sup\{|\varphi(x) - \varphi(y)|/|x - y| : x, y \in \mathbb{R}, x \neq y\}$,
- the bounded Wasserstein distance $d_{bW} := d_{\Phi_{bW}}$, in which case Φ_{bW} is the class of functions $\varphi : \mathbb{R} \to \mathbb{R}$ with $\|\varphi\|_{Lip} + \|h\|_\infty \leq 1$, where $\|\varphi\|_\infty := \sup\{|\varphi(x)| : x \in \mathbb{R}\}$.

If F is as above and such that $\mathbb{E}[F] = 0$, and $Z \sim \mathcal{N}(0, \sigma^2)$ with $\sigma^2 := \mathbb{E}[F^2]$ denotes a Gaussian random variable, the main result of the seminal paper [7] (see also Chapter 5 in [8]) provides an upper bound for $d_\Phi(F, Z)$ by combining Stein's method for normal approximation with the Malliavin formalism as introduced in Section 2. Here, $d_\Phi \in \{d_{TV}, d_K, d_W, d_{bW}\}$ is one of the four probability distances introduced above. We use this bound to provide an estimate for $d_\Phi(F, Z)$ in terms of the kernels h_q appearing in the chaotic representation (14) of F.

Theorem 3.1. *Let $F \in L^2(\Omega)$ be centred and such that $\mathbb{E}[F^2] = \sigma^2 > 0$ and $F \in \mathbb{D}^{1,4}$. Let $Z \sim \mathcal{N}(0, \sigma^2)$ be a centred Gaussian random variable with variance σ^2 and denote by $h_q \in \mathfrak{H}^{\odot q}$, $q \geq 0$, the kernels in the chaotic decomposition* (14) *of F. Then,*

(15)

$$\begin{aligned} d_{bW}(F, Z) \leq d_W(F, Z) \leq \frac{c}{\sigma} \sum_{p=1}^{\infty} p \sum_{r=1}^{p-1} (r-1)! \binom{p-1}{r-1}^2 \sqrt{(2(p-r))!}\, \|h_p \otimes_r h_p\|_{\mathfrak{H}^{\otimes 2(p-r)}} \\ + \frac{c}{\sigma} \sum_{\substack{p,q=1 \\ p \neq q}}^{\infty} p \sum_{r=1}^{\min(p,q)} (r-1)! \binom{p-1}{r-1} \binom{q-1}{r-1} \sqrt{(p+q-2r)!}\, \|h_p \otimes_r h_q\|_{\mathfrak{H}^{\otimes p+q-2r}} \end{aligned}$$

with $c = \sqrt{2/\pi}$. In addition, if F has a density with respect to the Lebesgue measure on $\mathbb{R}$, then the same bound also holds with $c = 2/\sigma$ in case of the total variation distance and $c = 1/\sigma$ for the Kolmogorov distance.

Remark 3.2. The combination of Stein's method with techniques from Malliavin calculus has also been applied to functionals of Poisson random measures. In this context, a limit theorem that has the same spirit as our Theorem 3.1 was derived in [5] and, in fact, it was this paper that inspired us to consider a similar question for Gaussian functionals. Also our proof follows the principal idea developed in [5]. However, since such functionals are much easier from a combinatorial point of view, Theorem 3.1 has a much more neat form compared to its Poissonian analogue in [5].

Remark 3.3. By applying the Cauchy-Schwarz inequality, we have for $h_p \in \mathfrak{H}^{\odot p}$, $h_q \in \mathfrak{H}^{\odot q}$, $p, q \geq 1$, the estimate

$$\|h_p \otimes_r h_q\|_{\mathfrak{H}^{\otimes p+q-2r}} \leq \sqrt{\|h_p \otimes_{p-r} h_p\|_{\mathfrak{H}^{\otimes 2r}} \|h_q \otimes_{q-r} h_q\|_{\mathfrak{H}^{\otimes 2r}}} \tag{16}$$

$$\leq \frac{1}{2}\big(\|h_p \otimes_{p-r} h_p\|_{\mathfrak{H}^{\otimes 2r}} + \|h_q \otimes_{q-r} h_q\|_{\mathfrak{H}^{\otimes 2r}}\big) \tag{17}$$

$$= \frac{1}{2}\big(\|h_p \otimes_r h_p\|_{\mathfrak{H}^{\otimes 2(p-r)}} + \|h_q \otimes_r h_q\|_{\mathfrak{H}^{\otimes 2(q-r)}}\big),$$

see also [8, Equation (6.2.4)]. Hence, the right hand side of (15) can in principle be expressed solely in terms of the contraction norms $\|h_p \otimes_r h_p\|_{\mathfrak{H}^{\otimes 2(p-r)}}$, $p \geq 1$, $r = 1, \ldots, p$. However, in the course of such an approach, the term $\|h_p \otimes_p h_p\|_{\mathfrak{H}^{\otimes 0}} = \|h_p\|^2_{\mathfrak{H}^{\otimes p}}$, $p \geq 1$, shows up, which in turn is not present in (15). This term stems from the contraction norm $\|h_p \otimes_p h_q\|_{\mathfrak{H}^{\otimes q-p}}$, $p < q$, and it is precisely this term that forces us to deal with the chaotic gap arising in the context of Theorem 4.1 below.

Proof of Theorem 3.1. To simplify our presentation it is no loss of generality to assume that $\mathfrak{H} = L^2(A)$ for some σ-finite non-atomic measure space $(A, \mathcal{A}, \mu)$. In this case, we shall write $\|\cdot\|_q$ instead of $\|\cdot\|_{\mathfrak{H}^{\otimes q}}$ for integers $q \geq 1$. Moreover, in order to underline that the elements of the Hilbert space we are dealing with are functions, we use the symbols f and g instead of h and h' and denote the kernels of the chaotic decomposition of F by $f_q \in L^2_{\mathrm{sym}}(A^q)$, $q \geq 0$, building thereby on the notation aleady introduced in the previous section.

We prove the result only for the unit variance case $\sigma^2 = 1$, the general result then follows by a scaling argument exactly as in the proof of Theorem 5.1.3 in [8]. In this set-up, the same result provides an upper bound for $d_W(F, Z)$, $d_{bW}(F, Z)$, $d_K(F, Z)$ and $d_{TV}(F, Z)$ in terms of the Malliavin operators D and L^{-1}. Formally, due to the fact that $F \in \mathbb{D}^{1,4}$ implies that $\int_A (D_x F)(-D_x L^{-1} F)\, \mu(\mathrm{d}x) \in L^2(\Omega)$ as shown in Proposition 5.1.1 in [8], one has that

$$d_\Phi(F, Z) \leq c_\Phi \sqrt{\mathbb{E}\Big[\Big(1 - \int_A (D_x F)(-D_x L^{-1} F)\, \mu(\mathrm{d}x)\Big)^2\Big]} \tag{18}$$

with

$$c_\Phi = \begin{cases} \sqrt{\frac{2}{\pi}} & : \Phi = \Phi_W \text{ or } \Phi = \Phi_{bW} \\ 1 & : \Phi = \Phi_K \\ 2 & : \Phi = \Phi_{TV}, \end{cases} \tag{19}$$

where we implicitly used the assumption that F has a density in case of the Kolmogorov and the total variation distance, see [8]. Let us abbreviate the term under the above square-root by $T(F)$. Using the variance representation (6) and the definitions (11) and (13) of D and L^{-1}, respectively, $T(F)$ can be re-written as

$$T(F) = \mathbb{E}\Big[\Big(\int_A \sum_{p=1}^{\infty} pI_{p-1}(f_p(x,\,\cdot\,))\ \sum_{q=1}^{\infty} I_{q-1}(f_q(x,\,\cdot\,))\,\mu(\mathrm{d}x) - \sum_{n=1}^{\infty} n!\|f_n\|_n^2\Big)^2\Big]\,.$$

Thus, applying the inequality $\sqrt{\mathrm{var}(X+Y)} \le \sqrt{\mathrm{var}(X)} + \sqrt{\mathrm{var}(Y)}$ yields that $\sqrt{T(F)}$ can be estimated from above by

$$\begin{aligned}
&\sum_{p,q=1}^{\infty} p\left(\mathbb{E}\Big[\Big(\int_A I_{p-1}(f_p(x,\,\cdot\,))I_{q-1}(f_q(x,\,\cdot\,))\,\mu(\mathrm{d}x)\right.\\
&\qquad\qquad\left. - \mathbb{E}\int_A I_{p-1}(f_p(x,\,\cdot\,))I_{q-1}(f_q(x,\,\cdot\,))\,\mu(\mathrm{d}x)\Big)^2\Big]\right)^{1/2}\\
&= \sum_{p,q=1}^{\infty} p\left(\mathrm{var}\Big(\int_A I_{p-1}(f_p(x,\,\cdot\,))I_{q-1}(f_q(x,\,\cdot\,))\,\mu(\mathrm{d}x)\Big)\right)^{1/2},
\end{aligned}$$

where we used the Itô isometry (5) to get an alternative expression for the term $\sum\limits_{n=1}^{\infty} n!\|f_n\|_n^2$. Next, we compute the variance, using that

$$\mathrm{var}\Big(\int_A I_{p-1}(f_p(x,\,\cdot\,))I_{q-1}(f_q(x,\,\cdot\,))\,\mu(\mathrm{d}x)\Big) = T_1(F) - T_2(F)^2$$

with $T_1(F)$ and $T_2(F)$ given by

$$T_1(F) := \int_A\int_A \mathbb{E}\big[I_{p-1}(f_p(x,\,\cdot\,))I_{q-1}(f_q(x,\,\cdot\,))I_{p-1}(f_p(y,\,\cdot\,))I_{q-1}(f_q(y,\,\cdot\,))\big]\,\mu(\mathrm{d}y)\mu(\mathrm{d}x)$$

and

$$T_2(F) := \mathbb{E}\int_A I_{p-1}(f_p(x,\,\cdot\,))I_{q-1}(f_q(x,\,\cdot\,))\,\mu(\mathrm{d}x)\,.$$

To compute $T_1(F)$ we use twice the multiplication formula (8) together with the stochastic Fubini theorem [13, Theorem 5.13.1] and the isometry property (5). We obtain that

$$\begin{aligned}
T_1(F) &= \int_A\int_A \mathbb{E}\Bigg[\sum_{r=0}^{\min(p-1,q-1)}\sum_{s=0}^{\min(p-1,q-1)} r!s!\binom{p-1}{r}\binom{q-1}{r}\binom{p-1}{s}\binom{q-1}{s}\\
&\qquad\times I_{p+q-2(r+1)}(f_p(x,\,\cdot\,)\otimes_r f_q(x,\,\cdot\,))I_{p+q-2(s+1)}(f_p(y,\,\cdot\,)\otimes_s f_q(y,\,\cdot\,))\Bigg]\,\mu(\mathrm{d}y)\mu(\mathrm{d}x)\\
&= \sum_{r=0}^{\min(p-1,q-1)}\sum_{s=0}^{\min(p-1,q-1)} r!s!\binom{p-1}{r}\binom{q-1}{r}\binom{p-1}{s}\binom{q-1}{s}\\
&\qquad\qquad\times\mathbb{E}\big[I_{p+q-2(r+1)}(f_p\otimes_{r+1} f_q)I_{p+q-2(s+1)}(f_p\otimes_{s+1} f_q)\big]\\
&= \sum_{r=1}^{\min(p,q)} ((r-1)!)^2\binom{p-1}{r-1}^2\binom{q-1}{r-1}^2(p+q-2r)!\,\|f_p\widetilde{\otimes}_r f_q\|_{p+q-2r}^2\,.
\end{aligned}$$

On the other hand, we have

$$T_2(F) = \mathbf{1}(p=q) \int_A (p-1)! \|f_p(x,\cdot)\|_{p-1}^2 \, \mu(\mathrm{d}x) = \mathbf{1}(p=q)\,(p-1)!\|f_p\|_p^2 ,$$

which is just the square root of the last summand in the expression for $T_1(F)$ for $p = q$. Consequently, combining the expressions for $T_1(F)$ and $T_2(F)$ yields

$$\begin{aligned} &T_1(F) - T_2(F)^2 \\ &= \mathbf{1}(p=q) \sum_{r=1}^{p-1} ((r-1)!)^2 \binom{p-1}{r-1}^4 (2(p-r))! \|f_p \widetilde{\otimes}_r f_p\|_{2(p-r)}^2 \\ &\qquad + \mathbf{1}(p \neq q) \sum_{r=1}^{\min(p,q)} ((r-1)!)^2 \binom{p-1}{r-1}^2 \binom{q-1}{r-1}^2 (p+q-2r)! \, \|f_p \widetilde{\otimes}_r f_q\|_{p+q-2r}^2 . \end{aligned}$$

Together with the elementary inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, valid for all $a, b \geq 0$, and the fact that, by Jensen's inequality, $\|\widetilde{g}\|_p \leq \|g\|_p$ for all $g \in L^2(A^p)$, $p \geq 1$, this implies that

$$\begin{aligned} d_\Phi(F,Z) &\leq c_\Phi \sum_{p=1}^{\infty} p \sum_{r=1}^{p-1} (r-1)! \binom{p-1}{r-1}^2 \sqrt{(2(p-r))!} \, \|f_p \widetilde{\otimes}_r f_p\|_{2(p-r)} \\ &\qquad + c_\Phi \sum_{\substack{p,q=1 \\ p\neq q}}^{\infty} p \sum_{r=1}^{\min(p,q)} (r-1)! \binom{p-1}{r-1} \binom{q-1}{r-1} \sqrt{(p+q-2r)!} \, \|f_p \widetilde{\otimes}_r f_q\|_{p+q-2r} \\ &\leq c_\Phi \sum_{p=1}^{\infty} p \sum_{r=1}^{p-1} (r-1)! \binom{p-1}{r-1}^2 \sqrt{(2(p-r))!} \, \|f_p \otimes_r f_p\|_{2(p-r)} \\ &\qquad + c_\Phi \sum_{\substack{p,q=1 \\ p\neq q}}^{\infty} p \sum_{r=1}^{\min(p,q)} (r-1)! \binom{p-1}{r-1} \binom{q-1}{r-1} \sqrt{(p+q-2r)!} \, \|f_p \otimes_r f_q\|_{p+q-2r} . \end{aligned}$$

The proof is thus complete. □

Remark 3.4. It appears that the formula for the scalar product $\int_A (D_x F)(-D_x L^{-1} F)\, \mu(\mathrm{d}x)$ has already been computed in Equation (6.3.2) in [8]. However, it has not been used in [8] to derive a quantitative central limit theorem.

Note that the assumption $F \in \mathbb{D}^{1,4}$ in Theorem 3.1 justifies the inequality at (18), but does not necessarily imply that the sums on the right hand side of (15) converge. For this to hold, extra assumptions are needed. However, it turns out that they are not too restrictive in the applications we have in mind, see Theorem 4.1.

Let us briefly consider two special cases, namely that F belongs to a single Wiener chaos or to a finite sum of Wiener chaoses. Here, the result reduces to Proposition 3.2 or Proposition 3.7 in [7], respectively. Note that in these cases [8, Theorem 2.10.1] ensures that the functional F has a density with respect to the Lebesgue measure on $\mathbb{R}$. Moreover, one easily verifies that $F \in \mathbb{D}^{1,4}$. For simplicity we decided to restrict to the unit variance case only, which is, as explained above, no restriction of generality.

Corollary 3.5. *Let Z be a centred Gaussian random variable with unit variance and let d_Φ be one of the probability metrics d_{TV}, d_K, d_W or d_{bW}.*

(a) If $F = I_q(h)$ for some integer $q \geq 2$ and an element $h \in \mathfrak{H}^{\odot q}$ such that $\mathbb{E}[F^2] = 1$. Then there is a constant $c_1 \in (0, \infty)$ only depending on q and the choice of the probability metric such that

$$d_\Phi(F, Z) \leq c_1 \max_{r=1,\dots,q-1} \|h \otimes_r h\|_{\mathfrak{H}^{\otimes 2(q-r)}} .$$

(b) If $F = I_{q_1}(h_1) + \dots + I_{q_n}(h_n)$ for integers $n, q_1, \dots, q_n \geq 1$ and elements $h_i \in \mathfrak{H}^{\odot q_i}$, $i = 1, \dots, n$, such that $\mathbb{E}[F^2] = 1$. Then there are constants $c_1, c_2 \in (0, \infty)$ only depending on $q_1, \dots, q_n$ and on the choice of the probability metric such that

$$d_\Phi(F, Z) \leq c_1 \max_{\substack{r=1,\dots,q_i-1 \\ i=1,\dots,n}} \|h_i \otimes_r h_i\| + c_2 \max_{\substack{r=1,\dots,\min(q_i,q_j) \\ 1\leq i<j\leq n}} \|h_i \otimes_r h_j\| .$$

4 Application to Gaussian processes

4.1 The Breuer-Major Theorem

Let $(a_n)_{n\in\mathbb{N}}, (b_n)_{n\in\mathbb{N}}$ be two positive sequences. Then we write $a_n \lesssim b_n$ if a_n/b_n is bounded, and $a_n \sim b_n$ whenever $a_n \lesssim b_n$ and $b_n \lesssim a_n$.
Consider a one-dimensional centred and stationary Gaussian process $X = (X_k)_{k\in\mathbb{Z}}$ with unit variance and a covariance function

$$\rho(j) = \mathbb{E}[X_1 X_{1+j}], \qquad j \in \mathbb{Z} . \tag{20}$$

In what follows we will assume that

$$|\rho(j)| \sim |j|^{-\alpha} \tag{21}$$

for some $\alpha > 0$. Recall that the Cauchy-Schwarz implies that $|\rho(j)| \leq \rho(0) = 1$ for all $j \in \mathbb{Z}$. For technical reasons, we assume that for any $n \geq 1$ the vector $(X_1, \dots, X_n)$ is jointly Gaussian with a non-degenerate covariance matrix. As an example, one can think of X being obtained from the increments of a fractional Brownian motion $B^H = (B_t^H)_{t\in\mathbb{R}}$ with Hurst parameter $H \in (0, 1)$, that is, $X_k = B_{k+1}^H - B_k^H$ for all $k \in \mathbb{Z}$. In that case

$$\rho(j) = \frac{1}{2}\left(|j+1|^{2H} + |j-1|^{2H} - 2|j|^{2H}\right)$$

and thus (21) is satisfied with $\alpha = 2 - 2H$, see Chapter 7.4 in [8].
Let $g\colon \mathbb{R} \to \mathbb{R}$ be a non-constant measurable function with $\mathbb{E}|g(X_1)|^2 < \infty$ and consider the partial sum

$$F_n = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \{g(X_k) - \mathbb{E}[g(X_k)]\}, \qquad n \in \mathbb{N} . \tag{22}$$

We will assume that g has Hermite rank equal to $m \in \mathbb{N}$. That is, for each polynomial $p\colon \mathbb{R} \to \mathbb{R}$ with degree $\in \{1, \dots, m-1\}$ one has that $\mathbb{E}[(g(X_1) - \mathbb{E}[g(X_1)])p(X_1)] = 0$ and

$\mathbb{E}[g(X_1)H_m(X_1)] \neq 0$, where H_m is the mth Hermite polynomial. The Hermite rank can also be described in a different manner. Namely, due to the moment condition $\mathbb{E}|g(X_1)|^2 < \infty$, g has the following Hermite expansion:

$$g(x) - \mathbb{E}[g(X_1)] = \sum_{q=1}^{\infty} c_q H_q(x) \qquad \text{with} \qquad c_q = \frac{1}{q!}\mathbb{E}[g(X_1)H_q(X_1)]\,. \tag{23}$$

For g to have Hermite rank m means that $c_q = 0$ for all $q \in \{1, \ldots, m-1\}$ and that $c_m \neq 0$. (The existence of such an m is implied by the fact that g is non-constant.)
It turns out that the rates in Theorem 4.1 below do not only depend on α and the Hermite rank m, but also on a quantity $\gamma \in \mathbb{N} \cup \{\infty\}$ that we call the *chaotic gap* of g. Roughly speaking, it is the minimal distance between two active chaoses a functional lives in. More precisely, if $c_q = 0$ for all $q \neq m$, we set $\gamma = \infty$. If there is a $p \geq 1$ such that $c_p \neq 0$ and $c_{p+1} \neq 0$, we set $\gamma = 1$. Otherwise, $\gamma \geq 2$ and it is uniquely characterized by the following two conditions:

(i) for all $q \geq 1$: if $c_q \neq 0$ then $c_{q+1} = \cdots = c_{q+\gamma-1} = 0$,

(ii) there exists $p \geq 1$ such that $c_p \neq 0$ and $c_{p+\gamma} \neq 0$.

Of course, if $g = H_q$ for some $q \geq 1$, then $m = q$ and $\gamma = \infty$. Also if g is a linear combination of Hermite polynomials, one can directly determine the rank and the chaotic gap. More general examples involve even and odd functions, both having a chaotic gap of $\gamma = 2$, and the exponential function having a chaotic gap of $\gamma = 1$. Another interesting example is the indicator function $g = \mathbf{1}_{[0,\infty)}$, which satisfies $m = 1$ and $\gamma = 2$.
Using the parameters α, γ and m we can now formulate our quantitative central limit theorem for the random variables F_n defined at (22).

Theorem 4.1. *Let $g\colon \mathbb{R} \to \mathbb{R}$ be a measurable and non-constant function. Suppose that $\mathbb{E}|g(X_1)|^2 < \infty$, $\mathbb{E}|g((2+\varepsilon)X_1)|^2 < \infty$ for some $\varepsilon > 0$, that g has Hermite rank equal to $m \geq 2$, that $\rho(0) = 1$ and that $|\rho(j)| \sim |j|^{-\alpha}$ for some $\alpha > 1/m$. Further, denote by $\gamma \in \mathbb{N} \cup \{\infty\}$ the chaotic gap of g and let Z be a standard Gaussian variable. Then the following assertions are true for each of the probability metrics $d_\Phi \in \{d_W, d_{bW}\}$ and also for d_{TV} and d_K in the case that F_n has a density with respect to the Lebesgue measure on $\mathbb{R}$.*

(a) *If $m = 2$ and $\gamma = 1$ it holds that*

$$d_\Phi\left(\frac{F_n}{\sqrt{\mathrm{var}(F_n)}}, Z\right) \lesssim \frac{1}{\mathrm{var}(F_n)} \times \begin{cases} n^{-1/2} & : \alpha > 1 \\ n^{-\alpha/2} & : \alpha \in \left(\frac{2}{3}, 1\right) \\ n^{1-2\alpha} & : \alpha \in \left(\frac{1}{2}, \frac{2}{3}\right]. \end{cases} \tag{24}$$

(b) *If $m = 2$ and $\gamma \geq 2$ it holds that*

$$d_\Phi\left(\frac{F_n}{\sqrt{\mathrm{var}(F_n)}}, Z\right) \lesssim \frac{1}{\mathrm{var}(F_n)} \times \begin{cases} n^{-1/2} & : \alpha > \frac{3}{4} \\ n^{1-2\alpha} & : \alpha \in \left(\frac{1}{2}, \frac{3}{4}\right). \end{cases} \tag{25}$$

(c) *If $m \geq 3$ and $\gamma = 1$ it holds that*

$$d_\Phi\left(\frac{F_n}{\sqrt{\operatorname{var}(F_n)}}, Z\right) \lesssim \frac{1}{\operatorname{var}(F_n)} \times \begin{cases} n^{-1/2} & : \alpha > 1 \\ n^{-\alpha/2} & : \alpha \in \left(\frac{1}{m-\frac{1}{2}}, 1\right) \\ n^{1-m\alpha} & : \alpha \in \left(\frac{1}{m}, \frac{1}{m-\frac{1}{2}}\right]. \end{cases} \tag{26}$$

(d) *If $m \geq 3$ and $\gamma \geq 2$ it holds that*

$$d_\Phi\left(\frac{F_n}{\sqrt{\operatorname{var}(F_n)}}, Z\right) \lesssim \frac{1}{\operatorname{var}(F_n)} \times \begin{cases} n^{-1/2} & : \alpha > \frac{1}{2} \\ n^{-\alpha} & : \alpha \in \left(\frac{1}{m-1}, \frac{1}{2}\right) \\ n^{1-m\alpha} & : \alpha \in \left(\frac{1}{m}, \frac{1}{m-1}\right). \end{cases} \tag{27}$$

Moreover, as $n \to \infty$, one has that

$$\operatorname{var}(F_n) \to \sigma^2 := \operatorname{var}(g(X_1)) + 2\sum_{k=1}^{\infty} \operatorname{cov}[g(X_1), g(X_{1+k})] \in [0, \infty).$$

Remark 4.2. In Theorem 4.1 we have excluded some boundary cases, for example in part (b) the case that $\alpha = 3/4$. It is possible to fill these gaps and to derive rates of convergence, which involve logarithmic terms. For sake of simplicity we have excluded them from our discussion.

Let us briefly comment on the assumptions made in Theorem 4.1. At first, one might wonder whether the condition $\mathbb{E}|g((2+\varepsilon)X_1)|^2 < \infty$ for some $\varepsilon > 0$ is already implied by the condition that $\mathbb{E}|g(X_1)|^2 < \infty$. Whilst this is true for many choices of g such as $g(x) = |x|^p$, this is not generally the case as the following example shows. Due to our assumptions, $X_1 \sim \mathcal{N}(0,1)$ has a standard Gaussian distribution and we observe that

$$M(a) := \mathbb{E}\big[e^{aX_1^2}\big] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{(a-\frac{1}{2})x^2}\,\mathrm{d}x = \begin{cases} \frac{1}{\sqrt{1-2a}} & : a < \frac{1}{2} \\ \infty & : a \geq \frac{1}{2}. \end{cases}$$

Thus, taking $g(x) := e^{x^2/8}$ we conclude that $\mathbb{E}|g(X_1)|^2 = M(1/4) = \sqrt{2}$, while $\mathbb{E}|g((2+\varepsilon)X_1)|^2 = M(1+\varepsilon+\varepsilon^2/4) = \infty$ for all $\varepsilon > 0$. Moreover, the motivation to impose the moment condition $\mathbb{E}|g((2+\varepsilon)X_1)|^2 < \infty$ for some $\varepsilon > 0$ is to ensure the convergence of the corresponding sums in (15), thus also implying that $\langle DF_n, -DLF_n\rangle_{\mathcal{H}} \in L^2(\Omega)$ such that one can formally apply Theorem 3.1. As a consequence and in contrast to Section 6 in [11], we do not need moment assumptions involving derivatives of g such as $\mathbb{E}|g'(X_1)|^4 < \infty$, which would in turn imply that $F_n \in \mathbb{D}^{1,4}$. Hence, we can dispense with additional smoothness or regularity assumptions on g and are able to handle even non-continuous choices of g.
Theorem 4.1 also raises the question under which conditions the partial sums F_n have a density with respect to the Lebesgue measure on the real line. To give an answer to this question which goes beyond the (somehow restrictive) conditions of the transformation theorem for densities, we introduce the notion of a 0-measure-preserving map. A measurable function $f\colon \mathbb{R} \to \mathbb{R}$ is called 0-measure-preserving if for all Lebesgue null sets $B \subset \mathbb{R}$ also the preimage $f^{-1}(B) \subset \mathbb{R}$ is a Lebesgue null set. Using the technical assumption that

$(X_1, \dots, X_n)$ is jointly Gaussian with a non-degenerate covariance matrix, it turns out that F_n has a density with respect to the Lebesgue measure on $\mathbb{R}$ if and only if g is 0-measure-preserving. An argument is given in the Appendix. Examples of functions g where F_n has no density are given by locally constant functions, including, for example, the case of indicator functions. In turn, the power function $g(x) = |x|^p$ is 0-measure-preserving such that the functionals considered in Subsection 4.2 below possess a density for any choice of $p > 0$.

Some parts of the proof that involve convolutions of sequences with finite support and require an application of Young's inequality are inspired by [3]. We write $\ell^p(\mathbb{Z})$ for the space of sequences $u = (u_k)_{k\in\mathbb{Z}}$ such that $\|u\|_{\ell^p(\mathbb{Z})} := \left(\sum_{k\in\mathbb{Z}} |u_k|^p\right)^{1/p} < \infty$ if $p \in [1, \infty)$ and $\|u\|_{\ell^\infty(\mathbb{Z})} := \sup_{k\in\mathbb{Z}} |u_k| < \infty$ if $p = \infty$. Now, recall that the convolution of two sequences u, v on $\mathbb{Z}$ with finite support is defined as

$$(u * v)(k) := \sum_{j\in\mathbb{Z}} u(j)v(k-j), \qquad k \in \mathbb{Z},$$

and that $u * v$ has again a finite support. Due to Young's inequality, it holds that for $1 \le p, q, r \le \infty$ with $1/p + 1/q = 1 + 1/r$,

$$\|u * v\|_{\ell^r(\mathbb{Z})} \le \|u\|_{\ell^p(\mathbb{Z})} \|v\|_{\ell^q(\mathbb{Z})} \tag{28}$$

for sequences u and v with finite support.

Proof of Theorem 4.1. It is no loss of generality to assume that $\mathbb{E}[g(X_1)] = 0$. Then, since g is assumed to have Hermite rank equal to m, we have the unique Hermite expansion

$$g(x) = \sum_{q=m}^{\infty} c_q H_q(x) \qquad \text{with} \qquad c_q = \frac{1}{q!}\mathbb{E}[g(X_1)H_q(X_1)], \tag{29}$$

where the sum converges in the L^2-sense, meaning that $\mathbb{E}|g(X_1) - \sum_{q=m}^{n} c_q H_q(X_1)|^2 \to 0$, as $n \to \infty$. Moreover due to our assumption that $\mathbb{E}|g((2+\varepsilon)X_1)|^2 < \infty$ and thanks to (4), we have that

$$\sum_{q=m}^{\infty} q!\, c_q^2\, (2+\varepsilon)^{2q} < \infty, \tag{30}$$

see e.g. [8, Proposition 1.4.2].
The next is to observe that one can consider the Gaussian process $X = (X_k)_{k\in\mathbb{Z}}$ as a subset of an isonormal Gaussian process $\{W(h) : h \in \mathfrak{H}\}$, say, where $\mathfrak{H}$ is a real separable Hilbert space with scalar product $\langle \cdot, \cdot \rangle_{\mathfrak{H}}$. This means that for every $k \in \mathbb{Z}$ there exists an element $h_k \in \mathfrak{H}$ such that $X_k = W(h_k)$ and, consequently,

$$\langle h_k, h_\ell \rangle_{\mathfrak{H}} = \rho(k-\ell) \qquad \text{for all} \qquad k, \ell \in \mathbb{Z}, \tag{31}$$

see [8, Proposition 7.2.3] for details. Using that for $h \in \mathfrak{H}$, $I_q(h^{\otimes q}) = H_q(W(h))$ (H_q is again the qth Hermite polynomial), we see that $g(X_k) = \sum_{q=m}^{\infty} c_q I_q(h_k^{\otimes q})$ and hence

$$F_n = \sum_{q=m}^{\infty} I_q(f_{q,n}) \qquad \text{with} \qquad f_{q,n} = \frac{c_q}{\sqrt{n}} \sum_{k=1}^{n} h_k^{\otimes q} \in \mathfrak{H}^{\odot q}. \tag{32}$$

For $p, q \geq m$ and $r = 1, \ldots, \min(p, q)$ we compute

$$\begin{aligned}
\|f_{q,n}\|^2_{\mathfrak{H}^{\otimes q}} &= \frac{c_q^2}{n} \sum_{k,\ell=1}^{n} \rho^q(k-\ell), \\
f_{p,n} \otimes_r f_{q,n} &= \frac{c_p\, c_q}{n} \sum_{k,\ell=1}^{n} \rho^r(k-\ell)\, [h_k^{\otimes p-r} \otimes h_\ell^{\otimes q-r}], \\
\|f_{p,n} \otimes_r f_{q,n}\|^2_{\mathfrak{H}^{\otimes p+q-2r}} &= \frac{c_p^2\, c_q^2}{n^2} \sum_{i,j,k,\ell=1}^{n} \rho^r(i-j)\rho^r(k-\ell)\, \rho^{p-r}(i-k)\rho^{q-r}(j-\ell)\,.
\end{aligned}$$

Since $\sum_{k\in\mathbb{Z}} |\rho(k)|^q < \infty$ for all $q \geq m$, one has by dominated convergence that

$$\|f_{q,n}\|^2_{\mathfrak{H}^{\otimes q}} = c_q^2 \sum_{k=-(n-1)}^{n-1} \frac{n-|k|}{n} \rho^q(k) \to c_q^2 \sum_{k\in\mathbb{Z}} \rho^q(k) \leq c_q^2 \sum_{k\in\mathbb{Z}} |\rho(k)|^m < \infty\,,$$

as $n \to \infty$. With respect to the summability condition (30) and the variance representation (6), this implies that, as $n \to \infty$,

$$\sigma_n^2 := \operatorname{var}(F_n) = \sum_{q=m}^{\infty} q!\|f_{q,n}\|^2 \to \sigma^2 := \sum_{q=m}^{\infty} q!\, c_q^2 \sum_{k\in\mathbb{Z}} \rho^q(k) \in [0, \infty)\,. \tag{33}$$

In view of our main bound (15) we need to compute the asymptotic order of the quantity

$$A_n(p,q,r) := \frac{1}{n^2} \sum_{i,j,k,\ell=1}^{n} |\rho(i-j)|^r |\rho(k-\ell)|^r |\rho(i-k)|^{p-r} |\rho(j-\ell)|^{q-r}\,, \quad p, q \geq m \tag{34}$$

for $r = 1, \ldots, \min(p, q)$ if $p \neq q$ and for $r = 1, \ldots, q-1$ if $p = q$. We assume without loss of generality that $p \leq q$ and distinguish two cases. First, let $r = 1, \ldots, p-1$. Then

$$A_n(p,q,r) \leq A_n(p,p,r) = A_n(p,p,p-r) \leq A_n(m,m,\min(r, p-r, m-1))\,.$$

Second, assume that $m \leq r = p < q$ such that there exists some integer $t \geq 1$ with $p + t = q$. Note that we only have to consider those cases where t is greater than or equal to the chaotic gap γ. Then

$$A_n(p,q,p) \leq A_n(m, m+t, m) \leq A_n(m, m+\gamma, m)\,.$$

By index shifting, one obtains

$$\begin{aligned}
A_n(p,q,r) &= \frac{1}{n^2} \sum_{i,j,k,\ell=0}^{n-1} |\rho(|i-j|)|^r |\rho(|k-\ell|)|^r |\rho(|i-k|)|^{p-r} |\rho(|j-\ell|)|^{q-r} \\
&= \frac{1}{n^2} \sum_{\ell=0}^{n-1} \sum_{j=-l}^{n-1-\ell} \sum_{i=-j}^{n-1-j} \sum_{k=-\ell}^{n-1-\ell} |\rho(|i|)|^r |\rho(|k|)|^r |\rho(|i+j-k|)|^{p-r} |\rho(|j|)|^{q-r} \\
&\leq \frac{1}{n} \sum_{|j|\leq n-1} \sum_{i=-j}^{n-1-j} \sum_{|k|\leq n-1} |\rho(|i|)|^r |\rho(|k|)|^r |\rho(|k-(i+j)|)|^{p-r} |\rho(|j|)|^{q-r}\,.
\end{aligned}$$

This means that in the second case we get

$$\begin{aligned}A_n(p,p+t,p) \leq A_n(m,m+\gamma,m) &\leq \frac{1}{n}\left(\sum_{k\in\mathbb{Z}}|\rho(k)|^m\right)^2 \sum_{|j|\leq n-1}|\rho(j)|^\gamma \\ &\lesssim \begin{cases} n^{-1} & : \alpha > \frac{1}{\gamma} \\ n^{-\alpha\gamma} & : \alpha < \frac{1}{\gamma}\,. \end{cases}\end{aligned} \tag{35}$$

Now, let us come back to the first case and let $r = 1, \dots, m-1$. For any integer $s \geq 1$ introduce the truncated sequence

$$\rho_{s,n}(k) = |\rho(k)|^s\,\mathbf{1}(|k| \leq n-1)\,. \tag{36}$$

Then, using again a careful index shifting, we see that

$$\begin{aligned}A_n(m,m,r) &= \frac{1}{n^2}\sum_{i,j,k,\ell=0}^{n-1}\rho_{r,n}(|i-j|)\rho_{r,n}(|k-\ell|)\rho_{m-r,n}(|i-k|)\rho_{m-r,n}(|j-\ell|) \\ &\leq \frac{1}{n^2}\sum_{i,j=0}^{n-1}(\rho_{r,n}*\rho_{m-r,n})(|i-j|)^2 \\ &\leq \frac{1}{n}\sum_{|j|\leq n-1}(\rho_{r,n}*\rho_{m-r,n})(j)^2 \\ &\leq \frac{1}{n}\,\|\rho_{r,n}*\rho_{m-r,n}\|^2_{\ell^2(\mathbb{Z})}\,.\end{aligned}$$

Now, we apply Young's inequality (28) to derive a rate for $\|\rho_{r,n}*\rho_{m-r,n}\|^2_{\ell^2(\mathbb{Z})}$. For $m=2$ it holds that

$$\|\rho_{1,n}*\rho_{1,n}\|^2_{\ell^2(\mathbb{Z})} \leq \|\rho_{1,n}\|^4_{\ell^{4/3}(\mathbb{Z})} = \left(\sum_{|k|\leq n-1}|\rho(k)|^{4/3}\right)^3 \lesssim \begin{cases} 1 & : \alpha > \frac{3}{4} \\ n^{3-4\alpha} & : \alpha < \frac{3}{4} \end{cases}$$

and if $m > 2$ we find that

$$\|\rho_{r,n}*\rho_{m-r,n}\|^2_{\ell^2(\mathbb{Z})} \leq \|\rho_{r,n}\|^2_{\ell^2(\mathbb{Z})}\|\rho_{m-r,n}\|^2_{\ell^1(\mathbb{Z})}\,.$$

Moreover,

$$\|\rho_{r,n}\|^2_{\ell^2(\mathbb{Z})} = \sum_{|k|\leq n-1}|\rho(k)|^{2r} \lesssim \begin{cases} 1 & : \alpha > \frac{1}{2r} \\ n^{1-2r\alpha} & : \alpha < \frac{1}{2r}\,, \end{cases} \tag{37}$$

$$\|\rho_{m-r,n}\|^2_{\ell^1(\mathbb{Z})} = \left(\sum_{|k|\leq n-1}|\rho(k)|^{m-r}\right)^2 \lesssim \begin{cases} 1 & : \alpha > \frac{1}{m-r} \\ n^{2-2(m-r)\alpha} & : \alpha < \frac{1}{m-r}\,. \end{cases} \tag{38}$$

Consequently, by changing the roles of $\rho_{r,n}$ and $\rho_{m-r,n}$, one has the estimate

$$\begin{aligned}\|\rho_{r,n}*\rho_{m-r,n}\|^2_{\ell^2(\mathbb{Z})} &\leq \max_{r=1,\dots,[(m-1)/2]}\|\rho_{r,n}\|^2_{\ell^2(\mathbb{Z})}\|\rho_{m-r,n}\|^2_{\ell^1(\mathbb{Z})} \\ &\lesssim \begin{cases} 1 & : \alpha > \frac{1}{2} \\ n^{1-2\alpha} & : \alpha \in \left(\frac{1}{m-1}, \frac{1}{2}\right) \\ n^{3-2m\alpha} & : \alpha < \frac{1}{m-1}\,. \end{cases}\end{aligned}$$

In summary, we arrive at the bounds

$$A_n(2,2,1) \lesssim \begin{cases} n^{-1} & : \alpha > \frac{3}{4} \\ n^{2-4\alpha} & : \alpha < \frac{3}{4}, \end{cases} \tag{39}$$

$$A_n(m,m,r) \lesssim \begin{cases} n^{-1} & : \alpha > \frac{1}{2} \\ n^{-2\alpha} & : \alpha \in \left(\frac{1}{m-1}, \frac{1}{2}\right) \\ n^{2-2m\alpha} & : \alpha < \frac{1}{m-1}. \end{cases} \tag{40}$$

Now, one can plug-in the estimates into the right hand side of (15) and obtain by defining $A_n := \left(\max_{r=1,\dots,[(m-1)/2]} A_n(m,m,r) + A_n(m,m+\gamma,m)\right)^{1/2}$ that

$$\begin{aligned}
&\sum_{p=m}^{\infty} p \sum_{r=1}^{p-1} (r-1)! \binom{p-1}{r-1}^2 \sqrt{(2(p-r))!} \left\| \tfrac{f_{p,n}}{\sigma_n} \otimes_r \tfrac{f_{p,n}}{\sigma_n} \right\|_{2(p-r)} \\
&\quad + \sum_{\substack{p,q=m \\ p\neq q}}^{\infty} p \sum_{r=1}^{\min(p,q)} (r-1)! \binom{p-1}{r-1}\binom{q-1}{r-1} \sqrt{(p+q-2r)!} \left\| \tfrac{f_{p,n}}{\sigma_n} \otimes_r \tfrac{f_{q,n}}{\sigma_n} \right\|_{p+q-2r} \\
&\le \frac{A_n}{\sigma_n^2} \Bigg\{ \sum_{p=m}^{\infty} c_p^2\, p \sum_{r=1}^{p-1} (r-1)! \binom{p-1}{r-1}^2 \sqrt{(2(p-r))!} \\
&\quad + \sum_{\substack{p,q=m \\ p\neq q}}^{\infty} |c_p||c_q|\, p \sum_{r=1}^{\min(p,q)} (r-1)! \binom{p-1}{r-1}\binom{q-1}{r-1} \sqrt{(p+q-2r)!} \Bigg\}.
\end{aligned} \tag{41}$$

The claim follows upon proving that the two sums in the brackets converge. Put

$$B_1(p) := \sum_{r=1}^{p-1} (r-1)! \binom{p-1}{r-1}^2 \sqrt{(2(p-r))!}, \tag{42}$$

note that due to (30), the first sum converges provided that $B_1(p) \lesssim (p-1)!\, 4^p \lesssim (p-1)!\,(2+\varepsilon)^{2p}$. However, using the Cauchy-Schwarz inequality, Vandermonde's identity for binomial coefficients and Stirling's formula, we see that

$$\begin{aligned}
\frac{B_1(p)}{(p-1)!} &= \sum_{r=1}^{p-1} \binom{p-1}{r-1}\binom{2(p-r)}{p-r}^{1/2} \le \left(\sum_{r=1}^{p-1} \binom{p-1}{r-1}^2\right)^{1/2} \left(\sum_{r=1}^{p-1} \binom{2(p-r)}{p-r}\right)^{1/2} \\
&= \left(\binom{2(p-1)}{p-1} - 1\right)^{1/2} \left(\sum_{r=1}^{p-1} \binom{2r}{r}\right)^{1/2} \\
&\le (p-1)^{1/2} \binom{2(p-1)}{p-1} \sim \frac{4^{p-1}}{\sqrt{\pi}}.
\end{aligned}$$

To show that the sum in (41) converges, it is sufficient to prove that

$$\sum_{\substack{p,q=m \\ p\neq q}}^{\infty} (c_p^2 + c_q^2)\, p \sum_{r=1}^{\min(p,q)} (r-1)! \binom{p-1}{r-1}\binom{q-1}{r-1} \sqrt{(p+q-2r)!} < \infty, \tag{43}$$

thanks to the inequality $ab \le a^2 + b^2$, valid for all $a, b \in \mathbb{R}$. To this end, observe that for $p \ge m + 1$, again by Stirling's formula,

$$\begin{aligned}
&\sum_{q=m}^{p-1} \sum_{r=1}^{q} (r-1)! \binom{p-1}{r-1} \binom{q-1}{r-1} \sqrt{(p+q-2r)!} \\
&\le (p-1)! \sum_{q=m}^{p-1} \sum_{r=1}^{q} \binom{q-1}{r-1} \binom{2(p-r)}{p-r}^{1/2} \\
&\le (p-1)! \sum_{q=m}^{p-1} \binom{2(q-1)}{q-1}^{1/2} q^{1/2} \binom{2(p-1)}{p-1}^{1/2} \\
&\le (p-1)!\, (p-1)^{3/2} \binom{2(p-1)}{p-1} \\
&\sim (p-1)!\, (p-1) \frac{4^{p-1}}{\sqrt{\pi}} \\
&\lesssim (p-1)! \frac{(2+\varepsilon)^{2p-2}}{\sqrt{\pi}},
\end{aligned}$$

for any $\varepsilon > 0$, which implies (43) in view of (30). Thus, one can formally apply Theorem 3.1 such that there is a constant $C \in (0, \infty)$ only depending on the class Φ and g (or more specifically on the sequence $(c_q)_{q \in \mathbb{N}}$) such that

$$d_\Phi\left(\frac{F_n}{\sigma_n}, Z\right) \le C\, \frac{A_n}{\sigma_n^2}\,.$$

The proof is complete. □

Remark 4.3. If $c_{2q} \neq 0$ for some $q \ge 1$, it follows immediately that $\|f_{2q,n}\|^2 \to c_{2q}^2 \sum_{k \in \mathbb{Z}} \rho^{2q}(k) > 0$, as $n \to \infty$, and hence $\sigma^2 > 0$. As a consequence, we see that in this situation, the variance of F_n has no influence on the rates in Theorem 4.1.

Remark 4.4. The chaotic gap γ of the function g is visible in Theorem 4.1 only in the case $\gamma = 1$. As our proof shows, this is just a coincidence, since for $\gamma = 2$ the terms involving the chaotic gap are of the same size as the other leading terms and for $\gamma > 2$ become even subdominant in our situation. Consequently, for $\gamma \ge 2$ one gets exactly the same rates as for $\gamma = \infty$, and the rates in Theorem 4.1 coincide with the rates given in [3, Proposition 2.15].

Let us finally in this section consider the set-up of Theorem 4.1 in the special case that g has Hermite rank $m = 1$ (which has been excluded in the statement). Clearly, if g is linear (so $m = 1$ and $\gamma = \infty$), F_n is already centred Gaussian and $F_n/\sqrt{\mathrm{var}(F_n)}$ coincides in distribution with the standard Gaussian random variable Z. However, an inspection of the proof of Theorem 4.1 shows that for non-linear g with $m = 1$ and arbitrary chaotic gap $1 \le \gamma < \infty$ the leading term in (15) is $\|f_{1,n} \otimes_1 f_{1+\gamma,n}\|$, which finally yields a rate of convergence of order $n^{-1/2}$ as long as $\alpha > 1 = 1/m$.

4.2 Power variations of the fractional Brownian motion and processes from the Cauchy class

We build on the example we have seen in the previous section and let again $X = (X_k)_{k \in \mathbb{Z}}$ be a one-dimensional centred and stationary Gaussian process with unit variance and with covariance function ρ such that the assumption (21) is satisfied. Moreover, we assume again that for any $n \geq 1$ the vector $(X_1, \ldots, X_n)$ is jointly Gaussian with a non-degenerate covariance matrix. As function g we take now

$$(44) \qquad g(x) := |x|^p - \mu_p\,, \quad p > 0\,, \qquad \text{where} \qquad \mu_p = \mathbb{E}|X_1|^p\,.$$

In this situation, our random variable F_n defined by (22) becomes a so-called (centred) power variation of the Gaussian process X. The asymptotic behaviour of these functionals has attracted considerable interest in probability theory and mathematical statistics, see [2, 3], for example, as well as the references cited therein.

Now, we notice that unless in the special case $p = 2$ of the quadratic variation, the Hermite expansion of the function g at (44) is not finite. Moreover, since g is an even function it holds that the Hermite rank of g is $m = 2$ and, if $p \neq 2$, that the chaotic gap is $\gamma = 2$. Consequently, part (b) of Theorem 4.1 applies. (Note also that $\mathbb{E}|g(X_1)|^2 = \mu_{2p} < \infty$ and $\mathbb{E}|g((2+\varepsilon)X_1)|^2 < \infty$.)

Instead of repeating the bounds, let us consider a more concrete situation. We assume that the Gaussian process $X = (X_k)_{k\in\mathbb{Z}}$ is obtained from the increments of a fractional Brownian motion $B^H = (B^H_t)_{t\in\mathbb{R}}$ with Hurst parameter $H \in (0, 1)$. Let us recall that this means that B^H is a centred Gaussian process in continuous time with covariance function given by

$$\mathbb{E}[B^H_s B^H_t] = \frac{1}{2}\big(|s|^{2H} + |t|^{2H} - |s-t|^{2H}\big)\,, \qquad s, t \in \mathbb{R},$$

and that $X_k = B^H_{k+1} - B^H_k$ for all $k \in \mathbb{Z}$. Then X has covariance function

$$\rho(j) = \frac{1}{2}\big(|j+1|^{2H} + |j-1|^{2H} - 2|j|^{2H}\big)$$

and (21) is satisfied with $\alpha = 2 - 2H$. Moreover, since $(X_1, \ldots, X_n)$ possesses a non-degenerate covariance function,

$$F_n = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \{|X_k|^p - \mu_p\}$$

has a density and moreover, it is known that $\mathrm{var}(F_n)$ converges, as $n \to \infty$, to a positive and finite constant; see Remark 4.3. Thus, the following result is a direct consequence of Theorem 4.1.

Corollary 4.5. *Let d_Φ be one of the probability distances d_{TV}, d_K, d_W or d_{bW}. Then*

$$d_\Phi\left(\frac{F_n}{\sqrt{\mathrm{var}(F_n)}}\right) \lesssim \begin{cases} n^{-1/2} & : H \in \big(0, \frac{5}{8}\big) \\ n^{4H-3} & : H \in \big(\frac{5}{8}, \frac{3}{4}\big)\,. \end{cases}$$

To the best of our knowledge, the previous corollary is the only known result showing that for general power variations with $p > 0$ the speed of convergence in the central limit theorem is universal. By this we mean that the rate we obtain for general $p > 0$ coincides with the known rate for the quadratic variation functional, where $p = 2$. This should also be compared with the discussion in Remark 4.4 and especially with the result in [3, Proposition 2.15].

Remark 4.6. We emphasize that for $H > 3/4$ the fluctuations of F_n are no more Gaussian, while for the boundary case $H = 3/4$ one can derive a logarithmic rate of convergence. However and as already discussed in Remark 4.2, we will not pursue this direction in the present paper.

Another flexible and prominent class of random processes $X = (X_k)_{k\in\mathbb{Z}}$ to which our theory applies are the members of the so-called Cauchy class. These processes are centred Gaussian and their covariance function is given by

$$\rho(j) = (1 + |j|^{\beta})^{-\alpha/\beta}, \qquad j \in \mathbb{Z},$$

where the parameters α and β have to satisfy $\alpha > 0$ and $\beta \in (0, 2]$, see [2]. It is clear that these processes satisfy the assumptions made in the previous secion and that Theorem 4.1 can be applied. This shows that, if $\alpha > 1/2$, the normalized power variations F_n with parameter $p > 0$ satisfy the quantitative central limit theorem

$$d_\Phi\left(\frac{F_n}{\sqrt{\mathrm{var}(F_n)}}\right) \lesssim \begin{cases} n^{-1/2} & : \alpha > \frac{3}{4} \\ n^{1-2\alpha} & : \alpha \in \left(\frac{1}{2}, \frac{3}{4}\right), \end{cases}$$

where d_Φ can be any of the probability distances d_{TV}, d_K, d_W or d_{bW}, adding thereby to the limit theorems developed in [2]. Again, the rate of convergence is universal and does not depend on the choice of the power p.

4.3 Functionals of Gaussian subordinated processes in continuous time revisited

Finally, we apply our methods to functionals of Gaussian subordinated processes in continuous time. More precisely, we revisit the example from Section 6 in [11] and show that the rate of convergence there can be improved by our methods, confirming thereby the conjecture made in Remark 6.2 in [11]. That is, we consider a centred Gaussian process $X = (X_t)_{t\in\mathbb{R}}$ in continuous time with stationary increments. For example, X could be a (two-sided) fractional Brownian motion. The covariance function of X is defined by $\rho(u - v) := \mathbb{E}[(X_{u+1} - X_u)(X_{v+1} - X_v)]$, $u, v \in \mathbb{R}$. It is clear that X might be considered as a suitable isonormal Gaussian process, see Example 2.1.5 in [8].
Let $g : \mathbb{R} \longrightarrow \mathbb{R}$ be a non-constant and measurable function and fix two real numbers $a < b$. For any $T > 0$ define the functional

$$F_T = \frac{1}{\sqrt{T}} \int_{aT}^{bT} \{g(X_{u+1} - X_u) - \mathbb{E}[g(Z)]\}\, \mathrm{d}u, \tag{45}$$

where $Z \sim \mathcal{N}(0, 1)$ denotes a standard Gaussian random variable. To present the next result, we adopt the $\lesssim$-notation for sequences to the continuous case. In particular, we shall write $a(T) \lesssim b(T)$ for two functions $a, b : (0, \infty) \longrightarrow \mathbb{R}$ if the quotient $a(T)/b(T)$ stays bounded for all T. For simplicity, we restrict ourself to the case of the Wasserstein distance and do not investigate under which conditions the functionals F_T posess a density.

Proposition 4.7. *Assume that $\rho(0) = 1$ and that $\int_{\mathbb{R}} |\rho(u)|\, \mathrm{d}u < \infty$. Further, suppose that $g : \mathbb{R} \longrightarrow \mathbb{R}$ is a non constant and measurable function, for which $\mathbb{E}|g(X_1)|^2 < \infty$ and*

$\mathbb{E}|g((2+\varepsilon)X_1)|^2 < \infty$ *for some* $\varepsilon > 0$*. Then, as* $T \to \infty$*, one has that*

$$d_W\left(\frac{F_T}{\sqrt{\mathrm{var}(F_T)}}, Z\right) \lesssim \frac{T^{-1/2}}{\mathrm{var}(F_T)}, \tag{46}$$

Moreover, if g *is symmetric, then*

$$d_W\left(\frac{F_T}{\sqrt{\mathrm{var}(F_T)}}, Z\right) \lesssim T^{-1/2}.$$

Proof. The proof is almost literally the same as that for Theorem 4.1: the sums there have to be replaced by integrals and estimated using the integrability assumption of the covariance function, which corresponds to the case $\alpha > 1$ in the proof of Theorem 4.1. All combinatorial considerations remain unchanged. We also remark that according to Proposition 6.3 in [11] the symmetry of the function g implies the asymptotic variance $\sigma^2 := \mathrm{var}(F_T)$ exists in $(0, \infty)$, see also Remark 4.3. This leads to the second bound. We leave the details to the reader. □

We would like to emphasize that the central limit theorem for F_T in [11], which is based on an application of the second-order Poincaré inequality on the Wiener space, only works under considerably stronger smoothness assumptions on the function g. In particular, g has to be twice continuously differentiable. In this sense, Proposition 4.7 improves and extends the result of Section 6 in [11].

5 A multivariate extension

The purpose of this final section is to provide a multivariate extension of Theorem 3.1. To this end, we measure the distance between two d-dimensional ($d \geq 2$) random vectors $\mathbf{X}$ and $\mathbf{Y}$ by the multivariate Wasserstein distance

$$d_{mW}(\mathbf{X}, \mathbf{Y}) := \sup \left|\mathbb{E}[\varphi(\mathbf{X})] - \mathbb{E}[\varphi(\mathbf{Y})]\right|,$$

where the supremum is running over all Lipschitz functions $\varphi : \mathbb{R}^d \to \mathbb{R}$ with Lipschitz constant less than or equal to 1. The following result can be seen as the natural multi-dimensional generalization of Theorem 3.1.

Theorem 5.1. *Fix* $d \geq 2$ *and let* $C = (C_{ij})_{i,j=1}^d$ *be a positive definite* $d \times d$ *matrix. Suppose that* $\mathbf{F} = (F_1, \ldots, F_d)$ *is a centred* d*-dimensional random vector with covariance matrix* C*, and such that* $F_i \in L^2(\Omega)$ *and* $F_i \in \mathbb{D}^{1,4}$ *for all* $i \in \{1, \ldots, d\}$*. Further, let* $\mathbf{Z} \sim \mathcal{N}(0, C)$ *be a centred Gaussian random vector with covariance matrix* C *and denote for each* $i \in \{1, \ldots, d\}$ *and* $q \geq 0$ *by* $h_{q,i} \in \mathfrak{H}^{\odot q}$ *the kernels of the chaotic decomposition of* F_i*. Then*

$$\begin{aligned} d_{mW}(\mathbf{F}, \mathbf{Z}) \leq c \sum_{i,j=1}^{d} \sum_{p=1}^{\infty} p \sum_{r=1}^{p-1} (r-1)! \binom{p-1}{r-1}^2 \sqrt{(2(p-r))!}\, \|h_{p,i} \otimes_r h_{p,j}\|_{\mathfrak{H}^{\otimes 2(p-r)}} \\ + c \sum_{i,j=1}^{d} \sum_{\substack{p,q=1 \\ p \neq q}}^{\infty} p \sum_{r=1}^{\min(p,q)} (r-1)! \binom{p-1}{r-1}\binom{q-1}{r-1} \sqrt{(p+q-2r)!}\, \|h_{p,i} \otimes_r h_{q,j}\|_{\mathfrak{H}^{\otimes p+q-2r}}, \end{aligned}$$

where $c = \sqrt{d}\,\|C^{-1}\|_{\mathrm{op}}\|C\|_{\mathrm{op}}^{1/2}$ *and* $\|\cdot\|_{\mathrm{op}}$ *indicates the operator norm of the argument matrix.*

Proof. As in the proof of Theorem 3.1 we assume without loss of generality that $\mathfrak{H} = L^2(A)$ for some σ-finite non-atomic measure space $(A, \mathcal{A}, \mu)$, write $\|\cdot\|_q$ instead of $\|\cdot\|_{\mathfrak{H}^{\otimes q}}$ and $f_{q,i}$ for $h_{q,i}$, $i \in \{1, \ldots, d\}$.
From Theorem 6.1.1 in [8] we have that

$$d_{mW}(\mathbf{F}, \mathbf{Z}) \le \sqrt{d}\,\|C^{-1}\|_{\mathrm{op}}\|C\|_{\mathrm{op}}^{1/2}\Big(\sum_{i,j=1}^{d} \mathbb{E}\Big[\Big(C_{ij} - \int_A (D_x F_i)(-D_x L^{-1} F_j)\,\mu(\mathrm{d}x)\Big)^2\Big]\Big)^{\frac{1}{2}}.$$

The covariance representation (7) as well as the definitions (11) and (13) of D and L^{-1}, respectively, imply that the expectation is bounded by

$$\mathbb{E}\Big[\Big(\int_A \sum_{p=1}^{\infty} pI_{p-1}(f_{i,p}(x,\cdot)) \sum_{q=1}^{\infty} I_{q-1}(f_{q,j}(x,\cdot))\,\mu(\mathrm{d}x) - \sum_{n=1}^{\infty} n!\langle f_{n,i}, f_{n,j}\rangle_n\Big)^2\Big]$$

and hence

$$\begin{aligned} d_{mW}(\mathbf{F}, \mathbf{Z}) &\le c \sum_{i,j=1}^{d} \sum_{p,q=1}^{\infty} p\Big(\mathbb{E}\Big[\Big(\int_A I_{p-1}(f_{p,i}(x,\cdot)) I_{q-1}(f_{q,j}(x,\cdot))\,\mu(\mathrm{d}x) \\ &\qquad - \mathbb{E}\int_A I_{p-1}(f_{p,i}(x,\cdot)) I_{q-1}(f_{q,j}(x,\cdot))\,\mu(\mathrm{d}x)\Big)^2\Big]\Big)^{1/2} \\ &= c \sum_{i,j=1}^{d} \sum_{p,q=1}^{\infty} p\Big(\mathrm{var}\Big(\int_A I_{p-1}(f_{p,i}(x,\cdot)) I_{q-1}(f_{q,j}(x,\cdot))\,\mu(\mathrm{d}x)\Big)\Big)^{1/2} \end{aligned}$$

The variance in the last expression equals $T_1(F, G) - T_2(F, G)^2$ with

$$\begin{aligned} T_1(F, G) = \int_A \int_A \mathbb{E}[&I_{p-1}(f_{p,i}(x,\cdot)) I_{q-1}(f_{q,j}(x,\cdot)) \\ &\times I_{p-1}(f_{p,i}(y,\cdot)) I_{q-1}(f_{q,j}(y,\cdot))]\,\mu(\mathrm{d}y)\mu(\mathrm{d}x) \end{aligned}$$

and

$$T_2(F, G) = \mathbb{E}\int_A I_{p-1}(f_{p,i}(x,\cdot)) I_{q-1}(f_{q,j}(x,\cdot))\,\mu(\mathrm{d}x).$$

Arguing as in the proof of Theorem 3.1, we see that

$$T_1(F, G) = \sum_{r=1}^{\min(p,q)} ((r-1)!)^2 \binom{p-1}{r-1}^2 \binom{q-1}{r-1}^2 (p+q-2r)!\|f_{p,i}\widetilde{\otimes}_r f_{q,j}\|_{p+q-2r}^2$$

we well as

$$T_2(F, G) = \mathbf{1}(p = q)(p-1)!\langle f_{p,i}, f_{p,j}\rangle_p.$$

As a consequence, we find that

$$
\begin{aligned}
&T_1(F,G) - T_2(F,G)^2 \\
&= \mathbf{1}(p=q) \sum_{r=1}^{p-1} ((r-1)!)^2 \binom{p-1}{r-1}^4 (2(p-r))! \| f_{p,i} \widetilde{\otimes}_r f_{p,j} \|^2_{2(p-r)} \\
&\quad + \mathbf{1}(p \neq q) \sum_{r=1}^{\min(p,q)} ((r-1)!)^2 \binom{p-1}{r-1}^2 \binom{q-1}{r-1}^2 (p+q-2r)! \| f_{p,i} \widetilde{\otimes}_r f_{q,j} \|^2_{p+q-2r}
\end{aligned}
$$

and finally

$$
\begin{aligned}
d_{mW}(\mathbf{F},\mathbf{Z}) \le c &\sum_{i,j=1}^{d} \sum_{p=1}^{\infty} p \sum_{r=1}^{p-1} (r-1)! \binom{p-1}{r-1}^2 \sqrt{(2(p-r))!} \| f_{p,i} \otimes_r f_{p,j} \|_{2(p-r)} \\
&+ c \sum_{i,j=1}^{d} \sum_{\substack{p,q=1 \\ p \neq q}}^{\infty} p \sum_{r=1}^{\min(p,q)} (r-1)! \binom{p-1}{r-1} \binom{q-1}{r-1} \sqrt{(p+q-2r)!} \| f_{p,i} \otimes_r f_{q,j} \|_{p+q-2r} .
\end{aligned}
$$

This completes the proof. □

6 Appendix

Let $\bar{\mathcal{B}}(\mathbb{R}^n)$ be the completion of the Borel σ-field $\mathcal{B}(\mathbb{R}^n)$ with respect to the n-dimensional Lebesgue measure $\lambda^n : \bar{\mathcal{B}}(\mathbb{R}^n) \to [0,\infty]$. We call a $\bar{\mathcal{B}}(\mathbb{R})$-measurable function $f : \mathbb{R}^n \to \mathbb{R}^n$ 0-measure-preserving if $\lambda^n(B) = 0$ for all $B \in \bar{\mathcal{B}}(\mathbb{R}^n)$ implies that $\lambda^n(f^{-1}(B)) = 0$.

Lemma 6.1. *Let $g : \mathbb{R} \to \mathbb{R}$ be a $\bar{\mathcal{B}}(\mathbb{R})$-measurable function. For fixed $n \ge 1$ define the measurable function $g_n : \mathbb{R}^n \to \mathbb{R}^n$ by applying g to each coordinate, that is, $g_n(x_1,\dots,x_n) := (g(x_1),\dots,g(x_n))$, $(x_1,\dots,x_n) \in \mathbb{R}^n$. Then g is 0-measure-preserving if and only if g_n is 0-measure-preserving.*

Proof. If $n = 1$, then there is nothing to show. So, we let $n > 1$ and assume that g is 0-measure-preserving. Now, the assertion follows by induction on n. For this reason, it is sufficient to restrict to the case that $n = 2$. For any $B \in \bar{\mathcal{B}}(\mathbb{R}^2)$ and any $x \in \mathbb{R}$ we write $B_x := \{y \in \mathbb{R} : (x,y) \in B\}$, which is an element of $\bar{\mathcal{B}}(\mathbb{R})$.
Choose a $B \in \bar{\mathcal{B}}(\mathbb{R}^2)$ with $\lambda^2(B) = 0$. Then

$$
\lambda^2(B) = \int_{\mathbb{R}} \lambda^1(B_x)\, \lambda^1(\mathrm{d}x) = 0 .
$$

As a consequence, the set $N := \{x \in \mathbb{R} : \lambda^1(B_x) > 0\}$ satisfies $\lambda^1(N) = 0$. Now, let $A := g_2^{-1}(B)$. Then, for each $y \in \mathbb{R}$, one has that

$$
\begin{aligned}
A_y &= \{z \in \mathbb{R} : (y,z) \in A\} = \{z \in \mathbb{R} : (g(y), g(z)) \in B\} \\
&= \{z \in \mathbb{R} : g(z) \in B_{g(y)}\} = g^{-1}(B_{g(y)}) .
\end{aligned}
$$

Due to our assumptions on g it holds that $\{y \in \mathbb{R} : \lambda^1(A_y) > 0\} = \{y \in \mathbb{R} : \lambda^1(B_{g(y)}) > 0\} = g^{-1}(N)$, which is a set of Lebesgue measure 0. Now, the claim follows upon observing that

$$
\lambda^2(A) = \int_{\mathbb{R}} \lambda^1(A_y)\, \lambda^1(\mathrm{d}y) = \int_{g^{-1}(N)} \lambda^1(A_y)\, \lambda^1(\mathrm{d}y) = 0 .
$$

On the other hand, if g_n is 0-measure-preserving, then g is also 0-measure-preserving by considering sets of the form $B_1 \times \cdots \times B_n$ for $B_i \in \bar{\mathcal{B}}(\mathbb{R})$ where $\lambda^1(B_1) = \cdots = \lambda^1(B_n) = 0$. □

Lemma 6.2. *Let μ be a measure on $\bar{\mathcal{B}}(\mathbb{R}^n)$, which is equivalent to the Lebesgue measure λ^n and let $f : \mathbb{R}^n \to \mathbb{R}^n$ be measurable function. Then $\mu \circ f^{-1}$ is absolutely continuous with respect to λ^n if and only if f is* 0*-measure-preserving.*

Proof. The proof is standard. □

Lemma 6.3. *If a random vector $(X_1, \dots, X_n)$ has a density with respect to λ^n for some $n \in \mathbb{N}$, then also $X_1 + \cdots + X_n$ has a density with respect to λ^1.*

Proof. The claim follows by means of the transformation theorem. □

Acknowledgement

We would like to thank Johanna F. Ziegel and Anja Mühlemann for inspiring discussions. TF acknowledges funding by the Swiss National Science Foundation (SNF) via grant 152609.

References

[1] J.-M. Azaïs, F. Dalmao, and J. R. León, *CLT for the zeros of classical random trigonometric polynomials*, Ann. Inst. H. Poincaré Probab. Statist., 52 (2016), pp. 804–820.

[2] O. E. Barndorff-Nielsen, J. M. Corcuera, and M. Podolskij, *Power variation for Gaussian processes with stationary increments*, Stochastic Process. Appl., 119 (2009), pp. 1845–1865.

[3] H. Biermé, A. Bonami, and J. R. León, *Central limit theorems and quadratic variations in terms of spectral density*, Electron. J. Probab., 16 (2011), pp. 362–395.

[4] A. Estrade and J. R. León, *A central limit theorem for the Euler characteristic of a Gaussian excursion set*, to appear in Ann. Probab., (2016+).

[5] D. Hug, G. Last, and M. Schulte, *Second-order properties and central limit theorems for geometric functionals of Boolean models*, Ann. Appl. Probab., 26 (2016), pp. 73–135.

[6] S. Janson, *Gaussian Hilbert Spaces*, vol. 129 of Cambridge Tracts in Mathematics, Cambridge University Press, Cambridge, 1997.

[7] I. Nourdin and G. Peccati, *Stein's method on Wiener chaos*, Probab. Theory Related Fields, 145 (2009), pp. 75–118.

[8] I. Nourdin and G. Peccati, *Normal Approximations with Malliavin Calculus - From Stein's Method to Universality*, Cambridge University Press, Cambridge, new. ed., 2012.

[9] I. Nourdin and G. Peccati, *The optimal fourth moment theorem*, Proc. Amer. Math. Soc., 143 (2015), pp. 3123–3133.

[10] I. Nourdin, G. Peccati, and M. Podolskij, *Quantitative Breuer-Major theorems*, Stochastic Process. Appl., 121 (2011), pp. 793–812.

[11] I. Nourdin, G. Peccati, and G. Reinert, *Second order Poincaré inequalities and CLTs on Wiener space*, J. Funct. Anal., 257 (2009), pp. 593–609.

[12] D. Nualart, *The Malliavin Calculus and Related Topics*, Probability and its Applications (New York), Springer-Verlag, Berlin, second ed., 2006.

[13] G. Peccati and M. S. Taqqu, *Wiener Chaos: Moments, Cumulants and Diagrams*, vol. 1 of Bocconi & Springer Series, Springer, Milan; Bocconi University Press, Milan, 2011.

10. Discussion

The two research articles Fissler and Thäle (2016a) and Fissler and Thäle (2016b) can be regarded as complementary in many aspects: Whereas the first one considers qualitative Gamma approximations of Poisson functionals inside a fixed chaos, the second one studies quantitative normal approximations of Gaussian functionals with a possibly infinite chaos representation. The dichotomy also applies to future research plans based on the respective articles.

As already discussed in the original article Fissler and Thäle (2016a), the conditions of the results are quite restrictive in view of the order of the integrals and the sign conditions. In the new light of the corresponding erratum, one has to impose even more conditions to ensure the equivalence stated in the four moments theorem (Theorem 3.5). In Remark 5.8, we called for new methods to establish a four moments theorem for Poisson integrals with a centered Gamma limit for higher orders. In fact, Döbler and Peccati (2017, Theorem 1.6) established a more general and satisfying result, being able to treat arbitrary orders (and also odd orders!) and to dispense with the sign condition on the kernels. However, even with their refined methods, their result still does not yield a four moments theorem unless one imposes some more conditions on the contractions of the kernels; see Döbler and Peccati (2017, Remark 1.7). This corresponds to our additional assumptions of (b') in the erratum. In summary, we do not see a promising ansatz how to tackle the discussed problem under less restrictive assumptions.

On the contrary, the article Fissler and Thäle (2016b) establishes a method how to derive quantitative central limit theorems for situations where it seemed impossible until now. As we have already discussed in the article, we plan to apply the technique also to other Gaussian functionals, such as the number of zeros of a random trigonometric polynomial (Azaïs et al., 2016), or the Euler characteristic of a Gaussian excursion set (Estrade and León, 2016). Further promising objects to study with this method are non-stationary Gaussian processes such as Volterra processes discussed in Nourdin et al. (2016).

Bibliography

C. Acerbi. Spectral measures of risk: A coherent representation of subjective risk aversion. *Journal of Banking & Finance*, 26:1505–1518, 2002.

C. Acerbi and B. Székely. Backtesting Expected Shortfall. *Risk Magazine*, 2014.

C. Acerbi and B. Székely. General properties of backtestable statistics. In preparation, 2017.

C. D. Aliprantis and K. C. Border. *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer, Berlin Heidelberg New York, 3rd edition, 2006.

D. E. Allen, M. McAleer, R. J. Powell, and A. K. Singh. Down-Side Risk Metrics as Portfolio Diversification Strategies across the Global Financial Crisis. *Journal of Risk and Financial Management*, 9(2):6, 2016.

Y. Armenti, S. Crepey, S. Drapeau, and A. Papapantoleon. Multivariate Shortfall Risk Allocation and Systemic Risk. *Preprint*, 2016. URL https://arxiv.org/abs/1507.05351.

P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath. Coherent measures of risk. *Math. Finance*, 9:203–228, 1999.

A. V. Asimit, V. Bignozzi, K. C. Cheung, J. Hu, and E.-S. Kim. Robust and Pareto Optimality of Insurance Contract. *Prepint*, 2016. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2834079.

M. Avriel, E. Diewert, S. Schaible, and I. Zang. *Generalized Concavity*. SIAM, Philadelphia, 2010.

J.-M. Azaïs, F. Dalmao, and J. León. CLT for the zeros of classical random trigonometric polynomials. *Ann. Inst. H. Poincaré Probab. Statist.*, 52(2):804–820, 05 2016.

A. Banerjee, X. Guo, and H. Wang. On the Optimality of Conditional Expectation as a Bregman Predictor. *IEEE Trans. Inform. Theory*, 51:2664–2669, 2005.

S. Bayer and T. Dimitriadis. A Joint (Value-at-Risk, Expected Shortfall) Regression Framework with Financial Applications. In preparation, 2017.

F. Bellini and V. Bignozzi. On elicitable risk measures. *Quant. Finance*, 15(5):725–733, 2015.

A. Bourguin and G. Peccati. The Malliavin–Stein Method on the Poisson Space. In G. Peccati and M. Reitzner, editors, *Stochastic Analysis for Poisson Point Processes: Malliavin Calculus, Wiener-Itô Chaos Expansions and Stochastic Geometry*, Bocconi & Springer Series 7, chapter 6, pages 185–228. Springer International Publishing Switzerland, 2016.

M. Burzoni, I. Peri, and C. M. Ruffo. On the properties of the Lambda value at risk: robustness, elicitability and consistency. *Preprint*, 2016. URL https://arxiv.org/abs/1603.09491.

A. Caponnetto. A note on the role of squared loss in regression. Technical report, Massachusetts Institute of Technology, 2005. URL http://cbcl.mit.edu/projects/cbcl/publications/ps/caponnetto-squareloss-6-05.pdf.

L. H. Y. Chen, L. Goldstein, and Q.-M. Shao. *Normal Approximation by Stein's Method*. Probability and Its Applications. Springer Berlin Heidelberg, 2011.

S. Clift, N. Costanzino, and M. Curran. Empirical Performance of Backtesting Methods for Expected Shortfall. *Preprint*, 2016. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2618345.

R. Cont, R. Deguest, and G. Scandolo. Robustness and sensitivity analysis of risk measurement procedures. *Quant. Finance*, 10:593–606, 2010.

J. Corbetta and I. Peri. A new approach to backtesting and risk model selection. *Preprint*, 2016. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2796253.

J.-P. Crouzeix and J. A. Ferland. Criteria for quasi-convexity and pseudo-convexity: Relationships and comparisons. *Mathematical Programming*, 23(1):193–205, 1982.

J. Daníelsson, P. Embrechts, C. Goodhart, C. Keating, F. Muennich, O. Renault, and H. S. Shin. An Academic Response to Basel II. Special paper no. 130, Financial Markets Group, London School of Economics, 2001.

M. H. A. Davis. Verification of internal risk measure estimates. *Statistics & Risk Modeling*, 33(3–4):67–93, 2016.

A. P. Dawid. Present Position and Potential Developments: Some Personal Views: Statistical Theory: The Prequential Approach. *Journal of the Royal Statistical Society. Series A (General)*, 147(2):278–292, 1984.

F. X. Diebold and R. S. Mariano. Comparing predictive accuracy. *Journal of Business and Economic Statistics*, 13:253–263, 1995.

W. E. Diewert, M. Avriel, and I. Zang. Nine kinds of quasiconcavity and concavity. *Journal of Economic Theory*, 25(3):397–420, 1981.

C. Diks, V. Panchenko, and D. van Dijk. Likelihood-based scoring rules for comparing density forecasts in tails. *J. Econometrics*, 163(2):215–230, 2011.

C. Döbler and G. Peccati. The fourth moment theorem on the Poisson space. *Preprint*, 2017. URL https://arxiv.org/abs/1701.03120.

B. Efron. Regression percentiles using asymmetric squared error loss. *Statist. Sinica*, 1 (1):93–125, 1991.

W. Ehm, T. Gneiting, A. Jordan, and F. Krüger. Of quantiles and expectiles: consistent scoring functions, Choquet representations and forecast rankings. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(3):505–562, 2016.

P. Embrechts and M. Hofert. Statistics and Quantitative Risk Management for Banking and Insurance. *Ann. Rev. Stat. Appl.*, 1, 2014.

P. Embrechts, H. Liu, and R. Wang. Quantile-Based Risk Sharing. *Preprint*, 2016. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2744142.

S. Emmer, M. Kratz, and D. Tasche. What is the best risk measure in practice? A comparison of standard risk measures. *Journal of Risk (The)*, 8:31–60, 2015.

A. Estrade and J. R. León. A central limit theorem for the Euler characteristic of a Gaussian excursion set. *Ann. Probab.*, 44(6):3849–3878, 11 2016.

Z. Feinstein, B. Rudloff, and S. Weber. Measures of Systemic Risk. *Prerpint*, 2016. URL https://arxiv.org/abs/1502.07961.

T. Fissler and C. Thäle. A four moments theorem for Gamma limits on a Poisson chaos. *ALEA, Lat. Am. J. Probab. Math. Stat.*, 13(1):163–192, 2016a.

T. Fissler and C. Thäle. A new quantitative central limit theorem on the Wiener space with applications to Gaussian processes. *Preprint*, 2016b. URL https://arxiv.org/abs/1610.01456.

T. Fissler and J. F. Ziegel. Higher order elicitability and Osband's principle. *Preprint*, 2015. URL http://arxiv.org/abs/1503.08123.

T. Fissler and J. F. Ziegel. Higher order elicitability and Osband's principle. *Ann. Statist.*, 44(4):1680–1707, 2016.

T. Fissler, J. F. Ziegel, and T. Gneiting. Expected shortfall is jointly elicitable with value-at-risk: implications for backtesting. *Risk Magazine*, pages 58–61, January 2016.

D. Friedman. Effective scoring rules for probabilistic forecasts. *Management Science*, 29 (4):447–454, 1983. URL http://www.jstor.org/stable/2630866.

R. Frongillo and I. Kash. Vector-Valued Property Elicitation. *JMLR: Workshop and Conference Proceedings*, 40:1–18, 2015a.

R. Frongillo and I. Kash. On Elicitation Complexity. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 3258–3266. Curran Associates, Inc., 2015b. URL http://papers.nips.cc/paper/5832-on-elicitation-complexity.pdf.

R. Frongillo and I. Kash. On elicitation complexity and conditional elicitation. *Preprint*, 2015c. URL https://arxiv.org/abs/1506.07212.

R. Frongillo, I. Kash, and S. Becker. Open Problem: Property Elicitation and Elicitation Complexity. *JMLR: Workshop and Conference Proceedings*, 49:1–4, 2016.

R. E. Gaunt, A. Pickett, and G. Reinert. Chi-square approximation by Stein's method with application to Pearson's statistic. *Preprint*, 2016. URL https://arxiv.org/abs/1507.01707v3.

R. Giacomini and H. White. Tests of conditional predictive ability. *Econometrica*, 74(6): 1545–1578, 2006.

T. Gneiting. Editorial: Probabilistic forecasting. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 171(2):319–321, 2008.

T. Gneiting. Making and Evaluating Point Forecasts. *J. Amer. Statist. Assoc.*, 106: 746–762, 2011.

T. Gneiting and A. Raftery. Strictly Proper Scoring Rules, Prediction, and Estimation. *J. Amer. Statist. Assoc.*, 102:359–378, 2007.

T. Gneiting and R. Ranjan. Comparing Density Forecasts Using Threshold- and Quantile-Weighted Scoring Rules. *Journal of Business & Economic Statistics*, 29(3):411–422, 07 2011.

T. Gneiting and R. Ranjan. Combining predictive distributions. *Electron. J. Statist.*, pages 1747–1782, 2013.

T. Gneiting, F. Balabdaoui, and A. E. Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society Series B*, 69:243–268, 2007.

H. J. Greenberg and W. P. Pierskalla. A Review of Quasi-Convex Functions. *Operations Research*, 19(7):1553–1570, 1971. URL http://www.jstor.org/stable/169178.

Y. Grushka-Cockayne, K. C. Lichtendahl, V. R. R. Jose, and R. L. Winkler. Quantile Evaluation, Sensitivity to Bracketing, and Sharing Business Payoffs. *Preprint*, 2016. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2628322.

C. Heinrich. The mode functional is not elicitable. *Biometrika*, 101(1):245–251, 2014.

Herodotus and A. D. Godley. *Herodotus I*, volume 5. London: W. Heinemann; New York: G.P. Putnam, 1920. URL http://www.archive.org/details/herodotus05hero.

H. Holzmann and M. Eulert. The role of the information set for forecasting – with applications to risk management. *Annals of Applied Statistics*, 8:79–83, 2014.

H. Holzmann and B. Klar. Weighted scoring rules and hypothesis testing. *Prepint*, 2016. URL https://arxiv.org/abs/1611.07345.

P. J. Huber. Robust Estimation of a Location Parameter. *Ann. Math. Statist.*, pages 73–101, 1964.

P. J. Huber and E. M. Ronchetti. *Robust Statistics*. John Wiley & Sons, Inc., Hoboken, New Jersey, second edition edition, 2009.

D. Hug and M. Reitzner. Introduction to Stochastic Geometry. In G. Peccati and M. Reitzner, editors, *Stochastic Analysis for Poisson Point Processes: Malliavin Calculus, Wiener-Itô Chaos Expansions and Stochastic Geometry*, Bocconi & Springer Series 7, chapter 5, pages 145–184. Springer International Publishing Switzerland, 2016.

E. Jakobsons and S. Vanduffel. Dependence uncertainty bounds for the expectile of a portfolio. *Risks*, 3(4):599, 2015.

E. Jouini, M. Meddeb, and N. Touzi. Vector-valued coherent risk measures. *Finance Stochast.*, 8:531–552, 2004.

C. Kascha and F. Ravazzolo. Combining inflation density forecasts. *Journal of Forecasting*, 29(1-2):231–250, 2010.

R. Koenker. *Quantile Regression*. Cambridge University Press, Cambridge, 2005.

S. Kou and X. Peng. On the measurement of economic tail risk. *Operations Research*, 2015. To appear.

R. Lachièze-Rey and M. Reitzner. *U*-Statistics in Stochastic Geometry. In G. Peccati and M. Reitzner, editors, *Stochastic Analysis for Poisson Point Processes: Malliavin Calculus, Wiener-Itô Chaos Expansions and Stochastic Geometry*, Bocconi & Springer Series 7, chapter 7, pages 229–254. Springer International Publishing Switzerland, 2016.

N. Lambert. Elicitation and Evaluation of Statistical Functionals. *Preprint*, 2013. URL http://web.stanford.edu/~nlambert/papers/elicitation.pdf.

N. Lambert, D. M. Pennock, and Y. Shoham. Eliciting properties of probability distributions. In *Proceedings of the 9th ACM Conference on Electronic Commerce*, pages 129–138, Chicago, Il, USA, 2008. ACM.

G. Last. Stochastic Analysis for Poisson Point Processes. In G. Peccati and M. Reitzner, editors, *Stochastic Analysis for Poisson Point Processes: Malliavin Calculus, Wiener-Itô Chaos Expansions and Stochastic Geometry*, Bocconi & Springer Series 7, chapter 1, pages 1–36. Springer International Publishing Switzerland, 2016.

S. Lerch, T. L. Thorarinsdottir, F. Ravazzolo, and T. Gneiting. Forecaster's Dilemma: Extreme Events and Forecast Evaluation. *Preprint*, 2015. URL https://arxiv.org/abs/1512.09244.

C. Ley, G. Reinert, and Y. Swan. Stein's method for comparison of univariate distributions. *Probab. Surveys*, 14:1–52, 2017.

F. Liu and R. Wang. A Theory for Measures of Tail Risk. *Preprint*, 2016. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2841909.

M. Luk. *Stein's method for the Gamma distribution and related statistical applications*. PhD thesis, University of Southern California, Los Angeles, USA, 1994.

V. Maume-Deschamps, D. Rullière, and K. Said. Multivariate extensions of expectiles as risk measures. *Preprint*, 2016. URL https://arxiv.org/abs/1609.07637.

A. H. Murphy and H. Daan. Forecast Evaluation. In A. H. Murphy and R. W. Katz, editors, *Probability, Statistics and Decision Making in the Atmospheric Sciences*, pages 379–437. Westview Press, Boulder, Colorado, 1985.

R. F. Nau. Should Scoring Rules Be 'Effective'? *Management Science*, 31(5):527–535, 1985.

W. K. Newey and J. L. Powell. Asymmetric Least Squares Estimation and Testing. *Econometrica*, 55:819–847, 1987.

N. Nolde and J. F. Ziegel. Elicitability and backtesting. *Preprint*, 2016. URL https://arxiv.org/abs/1608.05498.

I. Nourdin and G. Peccati. Stein's method on Wiener chaos. *Probability Theory and Related Fields*, 145(1):75–118, 2009.

I. Nourdin and G. Peccati. *Normal Approximations with Malliavin Calculus – From Stein's Method to Universality*. Cambridge University Press, Cambridge, new. edition, 2012.

I. Nourdin, D. Nualart, and R. Zintout. Multivariate central limit theorems for averages of fractional Volterra processes and applications to parameter estimation. *Statistical Inference for Stochastic Processes*, 19(2):219–234, 2016.

D. Nualart. *The Malliavin Calculus and Related Topics*. Springer, Berlin, Heidelberg, 2006.

K. H. Osband. *Providing Incentives for Better Cost Forecasting*. PhD thesis, University of California, Berkeley, 1985.

A. J. Patton. Data-based ranking of realised volatility estimators. *J. Econometrics*, 161 (2):284–303, 2011.

A. J. Patton. Evaluating and comparing possibly misspecified forecasts. *Preprint*, 2015. URL http://public.econ.duke.edu/~ap172/.

G. Peccati and M. S. Taqqu. *Wiener chaos: moments, cumulants and diagrams*, volume 1 of *Bocconi & Springer Series*. Springer, Milan; Bocconi University Press, Milan, 2011. A survey with computer implementation, Supplementary material available online.

G. Peccati and C. Thäle. Gamma limits and U-statistics on the Poisson space. *ALEA Lat. Am. J. Probab. Math. Stat.*, 10(1):525–560, 2013.

G. Peccati, J. L. Solé, M. S. Taqqu, and F. Utzet. Stein's method and Normal approximation of Poisson functionals. *The Annals of Probability*, 38(2):443–478, March 2010.

A. Pickett. *Rates of convergence of χ^2 approximations via Stein's method*. PhD thesis, Oxford, 2004.

M. Pitera and T. Schmidt. Unbiased estimation of risk. *Preprint*, 2016. URL https://arxiv.org/abs/1603.02615.

S. Reichelstein and K. Osband. Incentives in government contracts. *Journal of Public Economics*, 24(2):257–270, 7 1984.

M. Saerens. Building cost functions minimizing to some summary statistics. *IEEE Transactions on Neural Networks*, 11(6):1263–1271, Nov 2000.

L. J. Savage. Elicitation of Personal Probabilities and Expectations. *J. Amer. Statist. Assoc.*, 66:783–801, 1971.

S. Schaible and W.T. Ziemba, editors. *Generalized concavity in optimization and economics*. Number Bd. 1. Academic Press, 1981.

M. Schulte and C. Thäle. Poisson Point Process Convergence and Extreme Values in Stochastic Geometry. In G. Peccati and M. Reitzner, editors, *Stochastic Analysis for Poisson Point Processes: Malliavin Calculus, Wiener-Itô Chaos Expansions and Stochastic Geometry*, Bocconi & Springer Series 7, chapter 8, pages 255–294. Springer International Publishing Switzerland, 2016.

J. Scott and G. Marshall. *A Dictionary of Sociology A Dictionary of Sociology*. Oxford University Press, 3 rev. edition, 2009.

I. Steinwart. How to compare different loss functions and their risks. *Constructive Approximation*, 26(2):225–287, 2007.

I. Steinwart, C. Pasin, R. Williamson, and S. Zhang. Elicitation and Identification of Properties. *JMLR: Workshop and Conference Proceedings*, 35:1–45, 2014.

C. Strähl and J. F. Ziegel. Cross-calibration of probabilistic forecasts. *Preprint*, 2015. URL https://arxiv.org/abs/1505.05314.

W. Thomson. Eliciting production possibilities from a well-informed manager. *Journal of Economic Theory*, 20(3):360–380, 6 1979.

S. Weber. Distribution-Invariant Risk Measures, Information, and Dynamic Consistency. *Math. Finance*, 16:419–441, 2006.

J. F. Ziegel. Coherence and elicitability. *Mathematical Finance*, 26(4):901–918, 2016.

T. Zwingmann and H. Holzmann. Asymptotics for the expected shortfall. *Preprint*, 2016. URL https://arxiv.org/abs/1611.07222.

Erklärung

gemäss Art. 28 Abs. 2 RSL 05

Name/Vorname: Fissler, Tobias

Matrikelnummer: 13-123-823

Studiengang: Statistik (Dissertation)

Titel der Arbeit: On Higher Order Elicitability
and Some Limit Theorems on the Poisson and Wiener Space

Leiterin der Arbeit: Prof. Dr. Johanna Fasciati-Ziegel

Ich erkläre hiermit, dass ich diese Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen benutzt habe. Alle Stellen, die wörtlich oder sinngemäss aus Quellen entnommen wurden, habe ich als solche gekennzeichnet. Mir ist bekannt, dass andernfalls der Senat gemäss Artikel 36 Absatz 1 Buchstabe r des Gesetzes vom 5. September 1996 über die Universität zum Entzug des auf Grund dieser Arbeit verliehenen Titels berechtigt ist. Ich gewähre hiermit Einsicht in diese Arbeit.

Ort/Datum

Unterschrift