Generic Conditions for Forecast Dominance[*]

Fabian Krüger[†] Johanna F. Ziegel[‡]

December 13, 2019

Abstract

Recent studies have analyzed whether one forecast method dominates another under a class of consistent scoring functions. While the existing literature focuses on empirical tests of forecast dominance, little is known about the theoretical conditions under which one forecast dominates another. To address this question, we derive a new characterization of dominance among forecasts of the mean functional. We present various scenarios under which dominance occurs. Unlike existing results, our results allow for the case that the forecasts' underlying information sets are not nested, and allow for uncalibrated forecasts that suffer, e.g., from model misspecification or parameter estimation error. We illustrate the empirical relevance of our results via data examples from finance and economics.

Key words: loss function, model comparison, prediction

[*] We thank Werner Ehm, Tilmann Gneiting, Malte Knüppel, Sebastian Lerch, Melanie Schienle and seminar and conference participants at the Heidelberg Institute for Theoretical Studies, Statistische Woche (Linz, 2018), the University of Cologne and QFFE (Marseille, 2019) for helpful discussions, and Yoon-Jae Whang for making program code related to Linton et al. (2005) available via his website. Johanna F. Ziegel gratefully acknowledges financial support from the Swiss National Science Foundation.

[†] Karlsruhe Institute of Technology, Department of Economics, Blücherstraße 17, 76185 Karlsruhe, Germany. Email: `fabian.krueger@kit.edu`

[‡] University of Bern, Institute of Mathematical Statistics and Actuarial Science, Alpeneggstrasse 22, 3012 Bern, Switzerland. Email: `johanna.ziegel@stat.unibe.ch`.

1 Introduction

Forecasts of a random variable Y (such as the inflation rate, a financial volatility measure, or the sale price of a house) play an important role in economics. Recent technological advances have contributed to an ever increasing array of data sources and forecasting techniques, which necessitates statistically principled comparisons of forecast quality. Here we focus on the typical task of predicting the mean of Y. It is well known that squared error loss sets the incentive to correctly forecast the mean, conditional on a certain information set. This basic insight underlies the use of squared error for estimating regression models. However, Savage (1971) shows that there are infinitely many other scoring (or loss) functions that are also consistent with the goal of forecasting the mean. Consider, for example, the task of modeling and forecasting the mean of a binary variable $Y \in \{0, 1\}$, which is simply the probability that $Y = 1$. In this case, squared error is often referred to as the 'Brier score' (following Brier, 1950). While squared error can be used to construct consistent parameter estimators in regression models and to evaluate probability forecasts out-of-sample, there is a continuum of other scoring functions that can be used as well (see e.g. Buja et al., 2005). The Bernoulli log likelihood function, which corresponds to maximum likelihood estimation, is arguably the most popular of these choices. In the general case where Y is not restricted to be binary, squared error continues to be a popular scoring function, and can be motivated as the (negative) log likelihood function of a Gaussian density with known variance. Log likelihood functions corresponding to other single-parameter families (such as Poisson or Exponential) can be employed as well; Table 1 below provides examples.

The non-uniqueness of consistent scoring functions is challenging, in that rankings of two forecast methods by average scores may depend on the specific function used for out-of-sample evaluation. Ehm et al. (2016), Ehm and Krüger (2018), Yen and Yen (2018), Ziegel et al. (2018) and Barendse and Patton (2019) therefore propose graphical tools and hypothesis tests to analyze the robustness of empirical forecast rankings. In their terminology, one forecast method dominates another if it performs better in terms of every consistent scoring function.

Adopting a theoretical perspective, Holzmann and Eulert (2014) show that a correctly specified forecast method dominates a competitor that is based on a smaller (nested) information set. However, forecasts based on diverse and thus non-nested information sets play a major role in applications, and are often encouraged by designers of forecast surveys and contests. For example, the European Central Bank's 'Survey of Professional Forecasters' features private and public-sector, financial and non-financial institutions from all over Europe (European Central Bank, 2018). Patton (2018) demonstrates that non-nested information sets may lead to lack of forecast dominance, i.e., to forecast rankings that fail to be robust across consistent scoring functions. This issue has been tackled for probability forecasts of a binary variable (DeGroot and Fienberg, 1983; Krzysztofowicz and Long, 1990), but results for more general situations are available only under specific assumptions. Furthermore, all existing theoretical results assume that the forecasts under comparison specify the correct expectation of Y, given some information set. As illustrated by Patton (2018), this assumption is often violated in applications, which may lead to non-robust forecast rankings.

The present paper sheds new light on the theoretical conditions under which forecast dominance occurs. An understanding of these conditions is useful to interpret empirical results of (non-)robust forecast rankings, and to identify desiderata of forecasting methods that may inspire improvements of existing methods. Unlike previous studies, we derive conditions that allow for non-nested information sets. Furthermore, we allow for various types of forecast imperfections resulting, amongst others, from model misspecification and parameter estimation error (if forecasts are generated by statistical methods) or cognitive biases (if forecasts are judgmental, generated by humans). These phenomena are ubiquitous in practice but have not been tackled by the existing theoretical literature on forecast dominance.

The paper is structured as follows. Section 2 presents our main technical result, a new characterization of dominance among mean forecasts. We then discuss alternative sets of assumptions that yield natural conditions for dominance. Section 3 considers the case of auto-calibrated forecasts, which means that the forecast matches the conditional expectation of Y, given the forecast itself. Under this condition, which allows for non-nested information sets, the forecast which is more variable in the sense of convex order (see e.g. Shaked and Shanthikumar, 2007; Levy, 2016) dominates the other. This result generalizes the result of Holzmann and Eulert (2014) mentioned above, and thus provides weaker sufficient conditions for forecast dominance. Section 4 drops the auto-calibration assumption, but instead requires joint normality of each forecast with the predictand. Alternatively, Section 5 assumes that both forecasts are based on the same information set $\mathcal{F}$, but yield imperfect approximations of the conditional expectation of the predictand given $\mathcal{F}$. Our results in Sections 4 and 5 demonstrate that there can well be dominance relations among two uncalibrated (i.e., not auto-calibrated) forecasts. In Section 6, we illustrate our theoretical results via data examples from finance and economics. Section 7 concludes with a discussion of the results and open problems. All proofs are deferred to the appendix. An online appendix presents additional analytical examples and details on hypothesis testing in our data examples.

2 A Characterization of Forecast Dominance

Savage (1971) considers scoring functions of the form

$$S(x, y) = \phi(y) - \phi(x) - \phi'(x)\ (y - x), \tag{1}$$

where $x \in \mathbb{R}$ is a forecast, $y \in \mathbb{R}$ is a realization, and ϕ is a convex function with subgradient ϕ'. Here, a scoring function assigns a negatively oriented penalty, such that a smaller value of S corresponds to a better forecast. Functions of the form given in (1) are *consistent* for the mean (Gneiting, 2011): If Y has cumulative distribution function (CDF) F, then

$$\mathbb{E}\left(S(m(F), Y)\right) \leq \mathbb{E}\left(S(x, Y)\right), \quad \text{for any } x \in \mathbb{R}. \tag{2}$$

Here $m(F) = \int x \, \mathrm{d}F(x)$ is the mean of F (which we always assume to exist and be finite), and $\mathbb{E}$ denotes expectation. Equation (2) states that a forecaster minimizes their expected score when stating the mean of Y as their forecast. The scoring function S is *strictly consistent* if equality in (2) implies $x = m(F)$. Strict consistency corresponds to a strictly convex function ϕ in (1). Under some additional assumptions (see Gneiting, 2011, Theorem 7), the scoring functions given at (1) are the only consistent scoring functions for the mean. Note that the additive term $\phi(y)$ in (1) is included to enforce the convention that $S(y, y) = 0$. However, the term does not depend on x, and is hence irrelevant in terms of optimal forecasting. Table 1, which is a modified version of Yen and Yen (2018, Table 1), presents examples of strictly consistent scoring functions for the mean.

		Range	Range	
$S(x, y)$	$\phi(z)$	of X	of Y	Comment(s)
$(y - x)^2$	z^2	$\mathbb{R}$	$\mathbb{R}$	squared error
$-y \log x - (1 - y)\log(1 - x)^*$	$z \log z + (1 - z)\log(1 - z)$	$(0, 1)$	$[0, 1]$	negative log likelihood of Bernoulli dist.
$\log x + \frac{y}{x} - 1^*$	$-\log z$	$(0, \infty)$	$[0, \infty)$	negative log likelihood of exponential dist.; equal to QLIKE loss (Patton, 2011)
$-y \log x + x^*$	$z \log z - z$	$(0, \infty)$	$[0, \infty)$	negative log likelihood of Poisson dist.

Table 1: Examples of strictly consistent scoring functions for the mean. Each example is characterized by a strictly convex function $\phi(z)$. Scoring functions marked by an asterisk (*) differ from Equation (1) by subtracting $\phi(y)$. This transformation ensures that the scoring function is well-defined over the entire range of Y. Rankings of any two forecasts x_1, x_2 remain unchanged, and strict consistency of the scoring function is preserved.

Consider two generic forecasters (or forecasting methods) A and B who issue forecasts X_A and X_B of the mean of Y. We treat these forecasts as random variables and consider their joint distribution with Y, the random variable to be predicted. We assume throughout that X_A, X_B and Y are integrable. The random variables are defined on the probability space $(\Omega, \mathcal{A}, \mathbb{Q})$ whereby the point forecasts X_A, X_B are measurable with respect to information sets $\mathcal{A}_A, \mathcal{A}_B \subseteq \mathcal{A}$; see Ehm et al. (2016, Section 3.1) for a detailed discussion. This setup includes the case of a binary predictand $Y \in \{0,1\}$, in which the mean forecasts X_A, X_B quote the probability that $Y = 1$, conditional on their respective information sets. We emphasize that the setup is consistent with the case that $Y \equiv Y_t$ is a time series and $X_j \equiv X_{tj}, j \in \{A, B\}$ are associated forecasts. The only requirement is that the joint distribution of the forecasts and the predictand is strictly stationary, such that the objects that we use in the following (notably expectations and CDFs) are well defined and do not depend on time. See Strähl and Ziegel (2017, Definition 2.2) for a formal probability space setup involving time series of forecasts and realizations, and Example 3.3 for an illustration. The following notion of forecast dominance is central to this paper.

Definition 2.1 (Forecast dominance)**.** Forecast A *dominates* forecast B if

$$\mathbb{E}\left(S(X_A, Y)\right) \leq \mathbb{E}\left(S(X_B, Y)\right)$$

for every function S of the form given in (1).

The preceding definition implies that the better performance of A compared to B is robust across all consistent scoring functions S. Theorem 1b and Corollary 1b of Ehm et al. (2016) imply that forecast dominance holds if and only if $\mathbb{E}\left(S_\theta(X_A, Y)\right) \leq \mathbb{E}\left(S_\theta(X_B, Y)\right)$ for all $\theta \in \mathbb{R}$, where

$$S_\theta(x, y) = \frac{1}{2}(\theta - y)\mathbf{1}_{(x>\theta)} \tag{3}$$

is the so-called elementary score for the mean indexed by the parameter $\theta \in \mathbb{R}$, up to a term that does not depend on x and is thus irrelevant in terms of forecast rankings (see Lemma A.3 for details). Building upon the elementary score, we next present a novel characterization of forecast dominance.

Theorem 2.1. *Let A and B be forecasts for the mean. Then A dominates B if and only if*

$\psi_A(\theta) \geq \psi_B(\theta)$ *for all* $\theta \in \mathbb{R}$, *where*

$$\psi_j(\theta) = \frac{1}{2}\int_\theta^\infty \mathbb{P}(X_j > w)\,\mathrm{d}w + \frac{1}{2}\mathbb{E}\left((\mathbb{E}\left(Y \middle| X_j\right) - X_j)\mathbf{1}_{(X_j>\theta)}\right) \quad \textit{for } j \in \{A, B\}.$$

The function $\psi_j(\theta)$ appearing in Theorem 2.1 is the expected value of the random variable $-S_\theta(X_j, Y)$, where $S_\theta(x, y)$ has been defined at (3). Theorem A.4 in Appendix A is a more general version of Theorem 2.1, covering forecast dominance for expectiles at level $\tau \in (0, 1)$. Expectiles are an asymmetric generalization of the mean which is the expectile at level $\tau = 1/2$ (Newey and Powell, 1987). While the representation of forecast dominance in Ehm et al. (2016) is an important prerequisite for our Theorems A.4 and 2.1, our derivation of an analytical expression for the expected score is novel, and is crucial in order to establish forecast dominance (or lack thereof) in theoretical scenarios. The two summands of the function ψ_j separate the influence of the variability (first summand) and the calibration (second summand) of the forecast. Roughly speaking, calibration refers to the statistical compatibility of forecasts and observations; see Section 3 for details. Variability of a forecast may or may not be desirable depending on the calibration properties; see Theorem 3.1 and Proposition 4.1. In the remainder of this paper, we derive various interpretable scenarios under which the technical condition of Theorem 2.1 is satisfied.

3 Auto-Calibrated Forecasts

Definition 3.1 (Auto-calibration). X is an *auto-calibrated* forecast of Y if $\mathbb{E}\left(Y \middle| X\right) = X$ almost surely.

The definition implies that the forecast X of Y can be used 'as is', without any need to perform bias correction. The prefix 'auto' indicates that X is an optimal forecast relative to the information set $\sigma(X)$ generated by X itself. Patton (2018, Proposition 2) also considered this notion of auto-calibration in the context of forecast dominance. In the literature on forecasting binary probabilities, which are mean forecasts and thus nested in the current setting, the same notion is often simply called 'calibration', see e.g. Ranjan and Gneiting (2010, Section 2.1). Furthermore, the definition coincides with the null hypothesis of the popular Mincer and Zarnowitz (1969, henceforth MZ) regression, given by

$$Y = \alpha + \beta X + \text{error}; \tag{4}$$

the null hypothesis $(\alpha, \beta) = (0, 1)$ corresponds to X being an auto-calibrated forecast of Y. Auto-calibration relates to the joint distribution of the forecast X_j and the realization Y. Below we make use of the concept of convex order that compares univariate distributions.

Definition 3.2 (Convex order)**.** A random variable Z_1 is *greater* than Z_2 in *convex order* if $\mathbb{E}\left(\phi(Z_1)\right) \geq \mathbb{E}\left(\phi(Z_2)\right)$, for all convex functions ϕ such that the expectations exist.

By Strassen's (1965) theorem, Z_1 is greater than Z_2 in convex order if and only if there are random variables Z_1', Z_2' on a joint probability space such that $Z_1' \sim Z_1$, $Z_2' \sim Z_2$ and $\mathbb{E}\left(Z_1' \middle| Z_2'\right) = Z_2'$. Here, $\sim$ denotes equality in distribution. If Z_1 is greater than Z_2 in convex order then $\mathbb{V}\left(Z_1\right) \geq \mathbb{V}\left(Z_2\right)$, where $\mathbb{V}$ denotes variance. The converse is generally false; however, in the special case that Z_1 and Z_2 are both Gaussian with the same mean, $\mathbb{V}\left(Z_1\right) > \mathbb{V}\left(Z_2\right)$ implies that Z_1 is greater in convex order than Z_2.

If Z_1 is greater than Z_2 in convex order, then $-Z_2$ second-order stochastically dominates $-Z_1$. (A random variable V second-order stochastically dominates another random variable W if $\mathbb{E}\left(u(V)\right) \geq \mathbb{E}\left(u(W)\right)$ for all non-decreasing and concave functions u; see Levy (2016, Section 3.6). Note that this definition is weaker than convex order since the latter involves both increasing and decreasing functions ϕ.) Furthermore, writing $Z_1' = Z_2' + \varepsilon$ with $\varepsilon = Z_1' - Z_2'$, we obtain $\mathbb{E}\left(\varepsilon \middle| Z_2'\right) = 0$. In the economic literature, Z_1 is sometimes referred to as being equal in distribution to 'Z_2 plus noise' (Rothschild and Stiglitz, 1970; Machina and Pratt, 1997). The term 'noise' for ε suggests that the variation in Z_1 is undesirable. Indeed, if $-Z_1$ and $-Z_2$ represent two investments with stochastic monetary payoffs, then every risk-averse decision maker with concave utility function will prefer $-Z_2$ to $-Z_1$. We avoid the 'noise' terminology since the negative connotation of the term is not justified in the present context; by contrast, the following result indicates that being more volatile is highly desirable in the context of auto-calibrated mean forecasts.

Theorem 3.1. *Assume that A and B are both auto-calibrated mean forecasts. Then, A dominates B if and only if X_A is greater than X_B in convex order.*

According to Theorem 3.1, it is desirable for a forecast to be large in convex order: Given the assumption that forecasts are auto-calibrated, being large in convex order implies that the forecast is more variable and is based on a 'larger' information set $\mathcal{A}_j$. Without the assumption of auto-calibration, a forecast could be more variable simply because of erratic variation (see Sections 4 and 5 below). In the case that Y is binary and X_A, X_B are discretely distributed with finite support, Theorem 3.1 coincides with DeGroot and Fienberg (1983, Theorem 1). However, Theorem 3.1 is much more widely applicable since it imposes no assumptions on the distributions of Y, X_A and X_B.

Example 3.1. Let $Y = Z_1 + Z_2 + Z_3 + Z_4$, where $\{Z_k\}_{k=1}^4$ are independent and identically distributed random variables with mean zero. The distribution may be non-Gaussian, may involve skewness and excess kurtosis, or could be discrete. Now let $X_A = Z_1 + Z_2$ and $X_B = Z_3$, such that both A and B are auto-calibrated for Y, and X_A is greater than X_B in convex order. By Theorem 3.1, A dominates B. This setup includes the example of Ehm et al. (2016, p. 557) where Z_k are all standard normal and dominance is established via calculations that exploit normality.

Example 3.2. Suppose that X_A and X_B are both auto-calibrated and normally distributed. If $\mathbb{V}(X_A) > \mathbb{V}(X_B)$, then normality implies that X_A is greater than X_B in convex order, so that A dominates B by Theorem 3.1. This example generalizes Patton (2018, Proposition 2) since it is based on slightly weaker assumptions and establishes dominance under all consistent scoring functions instead of a subclass called exponential Bregman loss.

Proposition 3.2. *For $j = A, B$, let $X_j = \mathbb{E}\left(Y \middle| \mathcal{F}_j\right)$, where $\mathcal{F}_B \subset \mathcal{F}_A$. Then X_A and X_B are both auto-calibrated and X_A is greater than X_B in convex order.*

Examples 3.1 and 3.2 both feature non-nested information sets. Proposition 3.2 establishes that two forecasts with nested information sets satisfy the auto-calibration and convex order conditions that underlie Theorem 3.1. The latter then states that X_A dominates X_B, as would be expected given that X_A has access to a larger information set and both forecasts are correctly specified. The result of Holzmann and Eulert (2014, final line of Corollary 2) uses the same setup as Proposition 3.2 above, and is thus a special case of Theorem 3.1. Hence, Theorem 3.1 provides sufficient conditions for forecast dominance that are weaker than the ones by Holzmann and Eulert. However, the result of Holzmann and Eulert applies to general functionals, whereas we focus on the mean functional. The following example concerns forecasts made at different points in time, which is an important special case of nested information sets in practice.

Example 3.3. Let $Y_t = a\ Y_{t-1} + \varepsilon_t$, where $|a| < 1$ and ε_t is independent and identically distributed with mean zero and variance σ^2, and let $\mathcal{F}_t$ be the information set generated by observations until time t. Suppose $X_{tA} = \mathbb{E}\left(Y_t \middle| \mathcal{F}_{t-1}\right) = a\ Y_{t-1}$ and $X_{tB} = \mathbb{E}\left(Y_t \middle| \mathcal{F}_{t-h}\right) = a^h\ Y_{t-h}$ for some $h \in \{2, 3, \ldots\}$. Then Y_t, X_{tA} and X_{tB} are all strictly stationary time series, and $\mathcal{F}_{t-h} \subset \mathcal{F}_{t-1}$. Proposition 3.2 thus implies that both forecasts are auto-calibrated, and that X_{tA} is greater than X_{tB} in convex order. Hence, the variance of X_{tA} exceeds that of X_{tB}, which also follows from Corollary 2 of Patton and Timmermann (2012).

Finally, the following corollary describes a simple implication of Theorem 3.1 that is closely related to empirical practice in econometrics.

Corollary. *Consider MZ regressions as in Equation (4), conducted separately for forecast $j \in \{A, B\}$. Suppose that A and B satisfy the conditions of Theorem 3.1. Then in population, the MZ regression for A attains a higher R^2 than the one for B.*

This relates to the empirical literature on forecasting financial volatility, where R^2s of MZ regressions are commonly used to assess forecasting ability of alternative methods (e.g. Andersen et al., 2003, Tables III.A and III.B). See Section 6.1 for an empirical illustration.

4 Forecast Dominance under Normality

Auto-calibration essentially rules out uninformative variation ('noise') in a forecast that may result from an overfitted statistical model, for example.

Example 4.1. Let $Y = X_A + \varepsilon$, where X_A and ε are independently standard normal. Suppose forecaster A quotes X_A as a mean forecast for Y, and forecaster B quotes $X_B = X_A + \zeta$, where $\zeta \sim \mathcal{N}(0, \sigma_\zeta^2)$, independently of X_A and ε. One obtains easily that $\mathbb{E}\left(Y \middle| X_B\right) = X_B/(1+\sigma_\zeta^2)$, which implies that forecast B is uncalibrated.

In Example 4.1, intuition suggests that A is a better forecast than B since the latter simply adds the noise term ζ on top of the former. Theorem 3.1 cannot be used to derive this statement since B is uncalibrated. In this section and in Section 5, we dispense with the auto-calibration assumption. In order to arrive at interpretable conditions, we investigate the scenario in which the forecast $X_j, j \in \{A, B\}$ and the realization Y follow a bivariate normal distribution, such that

$$\begin{pmatrix} X_j \\ Y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_j \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_j^2 & \rho_{Yj}\ \sigma_j\sigma_Y \\ \rho_{Yj}\ \sigma_j\sigma_Y & \sigma_Y^2 \end{pmatrix} \right), \tag{5}$$

where $\rho_{Yj} \in [-1, 1]$ is the correlation between X_j and Y. The Gaussian setup is similar to Satopää et al. (2016) who motivate joint normality of forecasts and realizations from a situation in which forecasters observe small bits ('particles') of the information that generates the predictand; see their Section 3.2. Forecast dominance does not depend on the dependence structure between the forecasts. Hence Equation (5) refers to the pair $(X_j, Y)'$ only; the joint distribution of $(X_A, X_B)'$ is left unspecified, and may be non-Gaussian. The distribution in

(5) is an unconditional one, and does not specify the dependence (or independence) across forecast instances. See Example A.1 in the Online Appendix for a stationary time series illustration that fits into the Gaussian framework.

We assume that $\mu_Y = \mu_A = \mu_B$, which means that forecasts A and B correctly assess the unconditional mean of Y. This simplifies our analysis but does not seem restrictive in most applications. The setup in Equation (5) allows for a wide range of scenarios in terms of forecast accuracy. In particular, the correlation parameter ρ_{Yj} may be positive or negative, and there is no prespecified relation between the variance parameters σ_j and σ_Y. This modeling approach hence is capable of describing the behavior of imperfect forecasts.

Proposition 4.1. *Assume that for $j \in \{A, B\}$ the distribution of (X_j, Y) is bivariate normal as in Equation (5). Then*

$$\begin{aligned}\mathbb{E}(S_\theta(X_B, Y)) - \mathbb{E}(S_\theta(X_A, Y)) = \frac{\sigma_Y}{2}\left\{\rho_{YA}\,\varphi\left(\frac{\theta-\mu_Y}{\sigma_A}\right) - \rho_{YB}\,\varphi\left(\frac{\theta-\mu_Y}{\sigma_B}\right)\right\} \\ + \frac{(\theta-\mu_Y)}{2}\left\{\Phi\left(\frac{\theta-\mu_Y}{\sigma_A}\right) - \Phi\left(\frac{\theta-\mu_Y}{\sigma_B}\right)\right\}, \quad (6)\end{aligned}$$

where $S_\theta(x, y)$ is the elementary score function defined at (3), and φ and Φ are the probability density and CDF of a standard normal distribution, respectively.

By Ehm et al. (2016, Theorem 1b and Corollary 1b), A dominates B if the left hand side of (6) is non-negative for all $\theta \in \mathbb{R}$. The expression in (6) yields several sets of sufficient conditions for forecast dominance, where we use the notation $\beta_j = \rho_{Yj}\ \sigma_Y/\sigma_j$ to denote the population slope coefficient in a MZ regression of Y on X_j as in Equation (4). The condition $\beta_j = 1$ is necessary and sufficient for auto-calibration.

Case 1 Let $\sigma_A \geq \sigma_B$, and assume that $\beta_B \leq 1 \leq \beta_A$. Then A dominates B.

Case 2 Let $\sigma_A \leq \sigma_B$.

Case 2a Assume that $0 \leq \beta_A, \beta_B \leq 1$. If $\beta_A\sigma_A^2 \geq \beta_B\sigma_B^2$, then A dominates B.

Case 2b If $\beta_B \leq 0 \leq \beta_A$, then A dominates B.

Case 3 Suppose that $\beta_A \sigma_A = \beta_B \sigma_B$, and that either $\beta_A, \beta_B > 1$ or $\beta_A, \beta_B < 1$. Then the forecast j for which $|\beta_j - 1|$ is smaller dominates the other.

Case 4 If $\sigma_A = \sigma_B$, the forecast j for which β_j is higher dominates the other.

Justification of these claims is given in the Appendix. For two auto-calibrated forecasts ($\beta_A = \beta_B = 1$), Case 1 implies that the one with higher variance is dominant, which echoes the statement of Theorem 3.1. (Since both forecasts are Gaussian with the same mean, having higher variance is the same as being greater in convex order.) However, Case 1 does not require auto-calibration. It implies that there may be dominance relations among two uncalibrated forecasts, or dominance of an auto-calibrated forecast over an uncalibrated competitor, or vice versa. Case 2a describes a situation in which A has lower variance than B, but at the same time has higher covariance with Y. This suggests that A has a more favorable signal-to-noise ratio than B, explaining dominance of A over B. In Case 2b, B is a particularly poor forecast, featuring high variance and negative correlation with Y. Case 3 describes situations in which both forecasts have the same correlation with Y, and both are uncalibrated. In these situations, the forecast that comes closer to being auto-calibrated is dominant. Finally, Case 4 describes a simple condition for dominance if both forecasts have the same variance.

Proposition 4.1 yields a simple necessary condition for forecast dominance: For A to dominate B, it must hold that $\rho_{YA} \geq \rho_{YB}$. (This can be seen by evaluating the expected score difference in Proposition 4.1 at $\theta = \mu_Y$.) If the forecast parameters satisfy this necessary condition but can not be classified into one of the four cases presented above, it is unclear whether a dominance relation exists. In this situation, one can use the result of Proposition 4.1 for an informal numerical check of dominance; see Example A.2 in the Online Appendix for an illustration.

A major implication of the Gaussian case is that auto-calibration – which underlies Section

3, as well as all of the previous literature – is not generally required to establish forecast dominance. In particular, there may well be dominance relations among forecasts generated from mis-specified statistical models; see Section 5.

5 Forecasts based on a Common Information Set

The results in Section 4 do not require auto-calibration, but require joint Gaussianity of forecasts and realizations. In this section, we present a result that requires neither auto-calibration nor Gaussianity, but assumes that both forecasts can be represented as $\mathbb{E}\left(Y\middle|\mathcal{F}\right)$ plus noise, where the information set $\mathcal{F}$ is common across forecasting methods. The forecast methods can be viewed as different ways of exploiting $\mathcal{F}$, based on statistical models using alternative estimation algorithms or functional form assumptions, for example.

Theorem 5.1. *Let $\mathcal{F} \subset \mathcal{A}$ be a σ-algebra, and let*

$$Y = \mathbb{E}\left(Y\middle|\mathcal{F}\right) + \varepsilon, \quad X_j = \mathbb{E}\left(Y\middle|\mathcal{F}\right) + \eta_j, \quad j \in \{A, B\},$$

where $\mathbb{E}\left(\varepsilon\middle|\mathcal{F}\right) = 0$, and η_j is conditionally independent of ε given $\mathcal{F}$. Assume that, conditionally on $\mathcal{F}$, the distributions of η_A and η_B are both symmetric around zero and are such that $|\eta_A|$ is smaller than $|\eta_B|$ with respect to first order stochastic dominance. Then A dominates B.

Conditional independence of η_j and ε says that, given the information $\mathcal{F}$, η_j must not contain information about ε. This requirement seems natural given our interpretation of η_j as a modeling error. The assumptions about η_A and η_B imply that the former is less variable (Shaked and Shanthikumar, 2007, Section 3.D). In the special case that $\eta_j|\mathcal{F} \sim \mathcal{N}(0, \sigma_j^2)$, the condition is satisfied if $\sigma_A^2 < \sigma_B^2$. The assumption that $\mathbb{E}\left(\eta_j\middle|\mathcal{F}\right) = 0$ implies that modeling errors are unsystematic, which seems plausible in the context of overfitted statistical models, for example. The theorem nests the case that $\eta_j = 0$ almost surely for one model $j \in \{A, B\}$.

In contrast to Theorem 3.1 and Proposition 4.1, the conditions of Theorem 5.1 are not directly testable for empirical data. However, we present testable implications.

Proposition 5.2. *Under the conditions of Theorem 5.1, the following statements hold:*

(a) $\mathbb{E}(X_A) = \mathbb{E}(X_B) = \mathbb{E}(Y)$.

(b) $\mathbb{C}ov(X_j, Y) \leq \mathbb{V}(X_j)$ *for* $j \in \{A, B\}$, *that is, both forecasts attain a slope coeffient* $\beta_j \leq 1$ *in MZ regressions.*

(c) $\mathbb{E}\left(X_B^{2k}\right) \geq \mathbb{E}\left(X_A^{2k}\right)$ *for all* $k \in \mathbb{N}$.

Theorem 5.1 has implications for out-of-sample prediction in linear models.

Example 5.1. Let

$$Y = Z'\beta + \varepsilon,$$

where Z is a p-dimensional vector of regressors, and ε is an error term satisfying $\mathbb{E}\left(\varepsilon \middle| Z\right) = 0$. Suppose that forecast $j \in \{A, B\}$ is based on some estimator for β, obtained from training data $\{Y_i, Z_i\}_{i=1}^n$. We seek to make predictions for a new observation $Y_0 = Z_0'\beta + \varepsilon_0$, where Z_0 and ε_0 are independent of the training data. We have that $X_j = Z_0'\hat{\beta}_j^n = Z_0'\beta + Z_0'\ (\hat{\beta}_j^n - \beta)$, where $\hat{\beta}_j^n$ is the estimator underlying forecast j, and $\eta_j = Z_0'\ (\hat{\beta}_j^n - \beta)$ represents the approximation error of forecast j. Setting $\mathcal{F} = \sigma(Z_0)$, we can apply Theorem 5.1. By assumption, $\hat{\beta}_j^n - \beta$ (which is generated from training data) is independent of ε_0, such that η_j is conditionally independent of ε_0 given $\mathcal{F}$. For large training samples, it is natural to assume multivariate normality of $\hat{\beta}_j^n - \beta$ for $j \in \{A, B\}$ with mean zero and covariance matrix Σ_j. Under this assumption, dominance of A over B occurs if $a'\Sigma_A a \leq a'\Sigma_B a$, for all $a \in \mathbb{R}^k$, which is equivalent to $(\Sigma_B - \Sigma_A)$ being positive semi-definite. This is the standard notion of A being a more precise estimator of β (Lehmann and Casella, 1998, Equation 4.4).

6 Data Examples

6.1 Forecasting the volatility of financial asset returns

Following Andersen et al. (2003), a large literature is concerned with modeling and forecasting realized measures of asset return volatility. Here we consider forecasting $\log \mathrm{RK}_t$, where RK_t is a realized kernel estimate (Barndorff-Nielsen et al., 2008) for the Dow Jones Industrial Average on day t. The two forecast specifications we compare are of the form

$$\widehat{\log \mathrm{RK}}_t = \hat{\beta}_0 + \hat{\beta}_1 Z_{t-1} + \hat{\beta}_2 \sum_{l=1}^{5} Z_{t-l} + \hat{\beta}_3 \sum_{l=1}^{22} Z_{t-l},$$

where $\{Z_t\}_t$ is a sequence of predictor variables. This functional form follows Corsi (2009), and provides a simple way of capturing the temporal persistence in $\log \mathrm{RK}_t$ that is typical of financial volatilities. For forecast A, Z_t corresponds to the daily logarithmic value of the VIX index, an implied volatility index computed from financial options. For forecast B, Z_t corresponds to the logarithmic value of the absolute index return on day t. We estimate both specifications using ordinary least squares, based on a rolling window of 1000 observations. Data on the realized kernel measure and daily returns are from the Oxford-Man Realized library at `https://realized.oxford-man.ox.ac.uk/`; data on the VIX are from the FRED database of the Federal Reserve Bank of St. Louis (`https://fred.stlouisfed.org/series/VIXCLS`). The sample obtained from merging both data sources covers daily observations from January 4, 2000 to May 10, 2018. The initial part of the sample is reserved for estimating the model. We evaluate forecasts for an out-of-sample period ranging from February 13, 2004 to May 10, 2018 (3580 observations).

To illustrate the conditions for Theorem 3.1 empirically, we first consider MZ regressions for both forecasts, based on the out-of-sample period. For forecast A (based on VIX), we obtain the estimate

$$\begin{array}{ccccc} Y_t = & 0.029 & + & 1.010 & X_{tA} + \text{error}; \\ & [0.030] & & [0.022] & \end{array}$$

the R^2 of the regression is 64%, and standard errors that are robust to autocorrelation and heteroscedasticity are reported in brackets. The standard errors are computed using the function NeweyWest from the R package sandwich (Zeileis, 2004), which implements the Newey and West (1987, 1994) variance estimator. For forecast B (based on absolute returns), we obtain

$$\begin{array}{ccccc} Y_t = & 0.015 & + & 1.003 & X_{tB} + \text{error}, \\ & [0.051] & & [0.046] & \end{array}$$

with an R^2 of 48.2%. In both regressions, a Wald test of the hypothesis of auto-calibration (corresponding to an intercept of zero and a slope of one) cannot be rejected at conventional significance levels.

To assess the convex order condition empirically, let F_j denote the CDF of forecast $j \in \{A, B\}$. Then A is greater than B in convex order if and only if

$$\int_{-\infty}^{x} F_A(z)\ dz - \int_{-\infty}^{x} F_B(z)\ dz \geq 0 \tag{7}$$

for every $x \in \mathbb{R}$, and equality holds in the limit as $x \to \infty$ (see the proof of Theorem 3.1 in Appendix B). Figure 2 plots the empirical CDFs of both forecasts. Visual inspection suggests that the integral condition in Equation 7 is plausible in the current example. In order to provide a more formal assessment, we use the subsampling based test by Linton et al. (2005) to investigate the hypothesis that one distribution is smaller than another in convex order. (Linton et al. (2005) test for second order stochastic dominance (SOSD). Under the assumption of auto-calibration, both forecasts have the same expected value, so that SOSD and convex order coincide, except for a differential sign convention.) We abbreviate the hypothesis of interest as 'A is CO-smaller than B' in the following discussion. Since the test depends on a tuning parameter (the size b of the subsamples) that is hard to select in

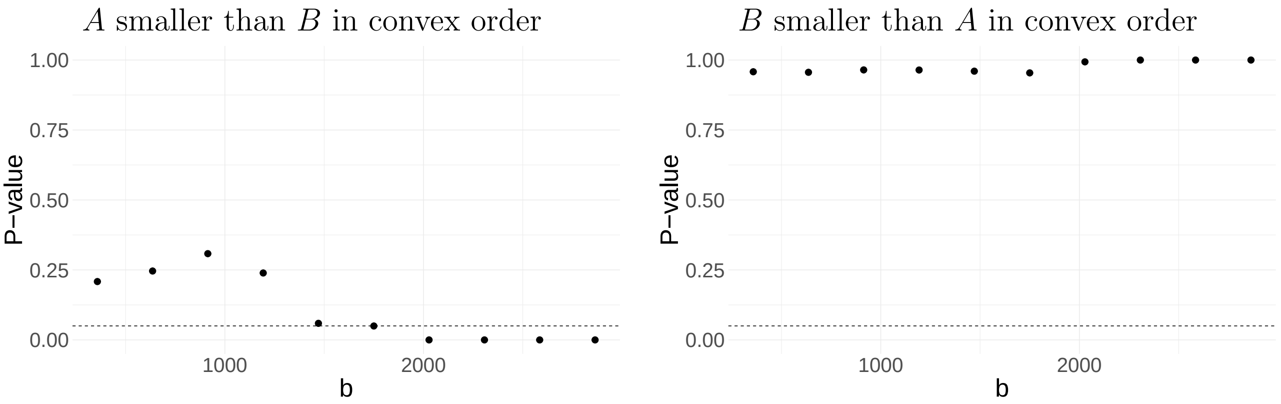

Figure 1: Subsampling based p-values of the test by Linton et al. (2005) plotted against the subsample size parameter b. The dashed horizontal line marks a p-value of five percent.

practice, Linton et al. (2005, Section 5.2) suggest to plot the test's p-value against b, and select b from within a range over which p is stable; see Online Appendix B.1 for details. Figure 1 shows the test results. The hypothesis that A is CO-smaller than B is rejected at the five percent levels for a range of $b \geq 2000$ over which the p-values are stable. By contrast, the right panel of Figure 1 shows no evidence against the hypothesis that B is CO-smaller than A, with large p-values for all values of b. In summary, the test thus reinforces the impression that a convex ordering (with A being greater than B) is plausible in the present example.

Hence both conditions of Theorem 3.1 seem plausible, and forecast A appears to be more informative than forecast B. Thus, we expect A to dominate B. In order to test dominance empirically, we use the bootstrap-based test by Ziegel et al. (2018) which we modify to cover the class of Bregman scoring functions at (1), instead of the class of scoring functions related to Expected Shortfall that is used by Ziegel et al. (2018). Following their implementation, we use a stationary bootstrap with block length drawn from a geometric distribution with mean $1.36\ n^{-1/3}$, where n is the size of the forecast evaluation sample. We use $10,000$ bootstrap iterations; see Online Appendix B.2 for further details. In line with the implication of Theorem 3.1, the hypothesis that A dominates B is not rejected by the test, with a bootstrap p-value of one. In contrast, the hypothesis that B dominates A is rejected with a bootstrap

p-value below one percent.

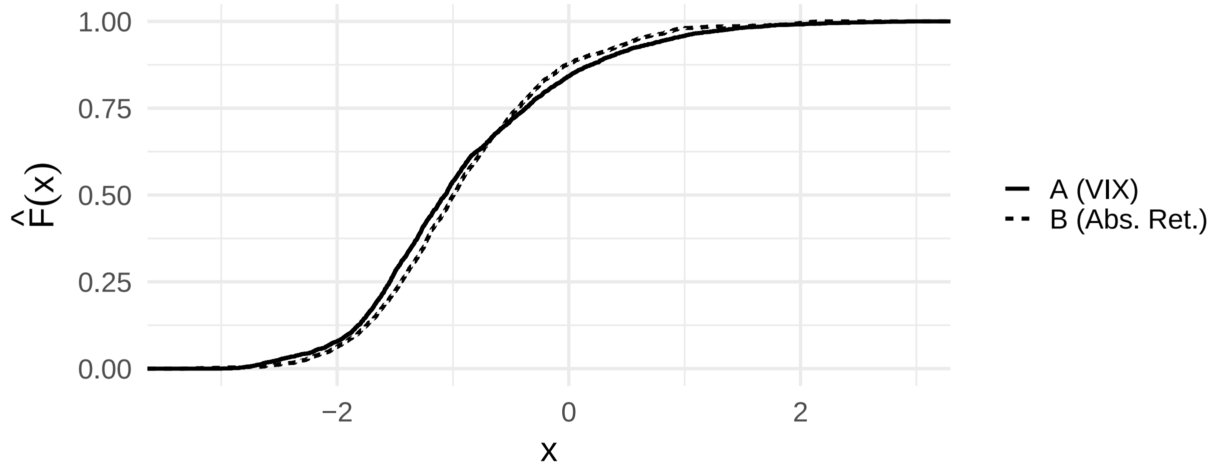

Figure 2: Volatility example: Empirical CDFs of both forecasts.

6.2 Forecasting US inflation

We illustrate the results of the normally distributed case from Section 4 with inflation forecasts from the Survey of Professional Forecasters (SPF), a widely used survey of macroeconomic experts. We compare the survey against two simple forecasting schemes: A random walk forecast (RW) that states the latest realization available to SPF participants, and a rolling mean forecast (RM) considering the four latest available observations (Atkeson and Ohanian, 2001). Given their simplicity, these methods act as minimal benchmarks for more sophisticated competitors, and are routinely included in practical forecast comparisons (see e.g. Faust and Wright, 2013, Section 2.5). Our analysis is based on real-time data published by the Federal Reserve Bank of Philadelphia at `https://www.philadelphiafed.org/research-and-data/real-time-center`. We focus on inflation as measured by the GDP deflator; the relevant series codes are PGDP (SPF forecasts) and P (realizations). We compare the forecasts against the second vintages of the realizations data. We further center the forecasts and realizations at zero in order to enforce the common mean assumption made in Section 4 ($\mu_Y = \mu_A = \mu_B$); however, our results are very similar if we omit this centering step.

We first assess the assumption that forecasts X_{tj} and realizations Y_t follow a bivariate normal distribution. To this end, we implement the test by Lobato and Velasco (2004) for the null hypothesis that a univariate stationary time series is unconditionally Gaussian. The test is appealing in that it is free of tuning parameters. We apply the test to the forecasts X_{tj}, the outcome Y_t and the forecast errors $Y_t - X_{tj}$, all of which are normally distributed if X_{tj} and Y_t are jointly normal. Repeating this procedure for three different forecast methods j (SPF, random walk and rolling mean) and at five forecast horizons (ranging from zero to four quarters ahead), we obtain p-values above 20% in all but one case. These results indicate that there is little evidence against pairwise bivariate normality of forecasts and realizations. Analogous tests for other macroeconomic variables (GDP growth and consumer price inflation) yielded clear rejections of normality, which is why we do not consider these variables here.

As a simple summary measure of forecast performance, Table 2 presents the methods' mean squared error (MSE) at various forecast horizons. The SPF attains the smallest MSE among the three methods, with the rolling mean method performing similarly well at some horizons. The random walk method attains the largest MSE at all horizons. In order to assess the plausibility of various dominance scenarios (see below Proposition 4.1), Table 2 presents some relevant statistics related to the covariance matrix of $(X_{tj}, Y_t)'$. We check whether these statistics match any of the scenarios under which dominance may occur. Consider, for example, the comparison of SPF versus RW at horizon $h = 0$ in the first column of Table 2. The SPF forecasts have a smaller empirical standard deviation than the random walk forecasts ($\sigma_{SPF} = 0.916 < 1.156 = \sigma_{RW}$). At the same time, the SPF's MZ regression coefficient ($\beta_{SPF} = 0.903$) exceeds that of the random walk ($\beta_{RW} = 0.471$). These findings indicate that the SPF forecasts have a better signal-to-noise ratio than the random walk. Indeed, the point estimates satisfy the conditions of Case 2a in Section 4, with the SPF taking the role of the dominant forecast A.

The left panel of Table 3 summarizes the outcomes of similar comparisons for all forecast horizons h. This analysis is based on the empirical point estimates, and can hence be thought of as calibrating the theoretical results of Section 4 to empirical data. The table reports a '✓' entry whenever the parameters in Table 2 belong to one of the sufficient conditions for dominance presented in Section 4 (Case 1-4). The SPF forecasts are dominant in six instances, all of which satisfy the conditions of Case 2a. These findings hence indicate that the SPF forecasts tend to contain less noise and more signal than the simple time series methods. Furthermore, according to the parameter estimates, the RM forecast dominates the RW forecast at the three shortest horizons, with the parameters again belonging to Case 2a in each case.

The right panel of Table 3 reports bootstrap p-values for various possible dominance relations. The bootstrap implementation is analogous to the one in Section 6.1. The bootstrap is nonparametric, contrasting the Gaussian setup of the theory in Section 4. In comparing the left and right panels of Table 3, one can see a fairly close correspondence between the theoretical implications and the empirical test results. In particular, instances where theory predicts dominance (symbol ✓ in left panel) correspond to high bootstrap p-values in the right panel, such that there is no evidence against dominance. Cases where theory rules out dominance (symbol X in left panel) tend to go along with low bootstrap p-values in the right panel, corresponding to evidence against dominance.

The preceding analysis shows that our theoretical results under normality can inform empirical forecast comparisons. In addition, the comparisons between the two simple time series methods (RW and RM) are also in line with the theoretical conditions of Theorem 5.1: First, both methods are based on the same information set generated by observations up until time t. Second, the theorem's testable implications in Proposition 5.2 all seem plausible here; compare the coefficients β_j, σ_j and $\mathbb{E}(X_j^4)$ reported in Table 2. The theorem then predicts dominance of RM over RW. As shown in Table 3, this conclusion is broadly in line with empirical nonparametric bootstrap tests.

h	0	1	2	3	4
MSE_{SPF}	0.665	0.778	0.862	0.920	1.001
MSE_{RW}	1.412	1.535	1.497	1.348	1.538
MSE_{RM}	0.886	0.917	0.991	1.103	1.201
σ_Y	1.160	1.160	1.160	1.160	1.160
σ_{SPF}	0.916	0.917	0.967	1.008	1.012
σ_{RW}	1.156	1.161	1.176	1.213	1.221
σ_{RM}	0.924	0.935	0.950	0.971	0.987
β_{SPF}	0.903	0.834	0.755	0.706	0.665
β_{RW}	0.471	0.425	0.441	0.496	0.432
β_{RM}	0.766	0.741	0.692	0.624	0.570
$\mathbb{E}(X^4_{SPF})$	2.088	2.220	2.488	3.172	2.813
$\mathbb{E}(X^4_{RW})$	5.393	5.383	5.515	6.506	6.534
$\mathbb{E}(X^4_{RM})$	1.895	1.932	2.028	2.234	2.354

Table 2: Sample estimates for the US inflation data. h indicates the forecast horizon (in quarters); the sample period is 1984:Q1 to 2018:Q2. For forecast method $j \in \{SPF, RW, RM\}$, MSE_j denotes the mean squared error, σ_j denotes the standard deviation, β_j denotes the slope coefficient from a regression of realized inflation on the forecast, and $\mathbb{E}(X_j^4)$ is the fourth moment of the forecast. σ_Y is the standard deviation of the realized inflation rates.

	Theory implications					Bootstrap p-values				
h	0	1	2	3	4	0	1	2	3	4
SPF $\succ^?_{fd}$ RW	✓	✓	✓	?	✓	1.000	1.000	1.000	0.968	0.798
RW $\succ^?_{fd}$ SPF	X	X	X	X	X	0.031	0.043	0.050	0.030	0.056
SPF $\succ^?_{fd}$ RM	✓	✓	?	?	?	0.911	0.650	0.723	0.441	0.736
RM $\succ^?_{fd}$ SPF	X	X	X	X	X	0.435	0.255	0.160	0.291	0.225
RW $\succ^?_{fd}$ RM	X	X	X	X	X	0.032	0.014	0.027	0.473	0.298
RM $\succ^?_{fd}$ RW	✓	✓	✓	?	?	1.000	1.000	0.999	0.760	0.919

Table 3: The notation 'A $\succ^?_{fd}$ B' denotes the possibility that A dominates B. h indicates the forecast horizon. Left panel: ✓ means that one of the sufficient conditions for dominance is satisfied. X means that the necessary condition is not satisfied. ? means that the necessary condition (but none of the sufficient conditions) is satisfied. Right panel: Bootstrap p-values of nonparametric forecast dominance test.

7 Discussion

Patton (2018) identifies three reasons why forecast dominance may not hold in practice: Non-nested information sets, misspecification, and estimation error. Motivated by this assessment, the present paper provides a theoretical analysis of forecast dominance that relates to each of these situations. Under the assumption that forecasts are auto-calibrated, our results in Section 3 provide a novel characterization of the role played by information sets that may or may not be nested. Misspecification and estimation error are likely to lead to uncalibrated forecasts for which no analytical results are available in the existing literature on forecast dominance. Our results in Sections 4 and 5 cover this case in detail, based on two distinct sets of assumptions that allow us to arrive at interpretable conditions.

Conceptually, our results indicate that the notion of forecast dominance may be less strong than suggested by Patton (2018), Nolde and Ziegel (2017, Section 2.3), and others. In particular, there can be dominance relations among two forecasts that are both highly imperfect. From a more technical perspective, an interesting question is whether similar conditions for forecast dominance can be derived for functionals other than the mean. As starting points of the analysis, our Theorem A.4 specifies conditions for dominance for the expectile functional (which includes the mean as a special case), and we treat quantiles in Online Appendix C. An open challenge are full distributional forecasts.

References

Andersen, T. G., Bollerslev, T., Diebold, F. X., and Labys, P. (2003). Modeling and forecasting realized volatility. *Econometrica*, 71:579–625.

Atkeson, A. and Ohanian, L. E. (2001). Are Phillips curves useful for forecasting inflation? *Federal Reserve Bank of Minneapolis Quarterly Review*, 25:2–11.

Barendse, S. and Patton, A. (2019). Comparing predictive accuracy in the presence of a loss function shape parameter. Working Paper, Duke University, November 2019.

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., and Shephard, N. (2008). Designing realized kernels to measure the ex post variation of equity prices in the presence of noise. *Econometrica*, 76:1481–1536.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78:1–3.

Buja, A., Stuetzle, W., and Shen, Y. (2005). Loss functions for binary class probability estimation and classification: Structure and applications. Working Paper, University of Washington, November 2005.

Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, 7:174–196.

DeGroot, M. H. and Fienberg, S. E. (1983). The comparison and evaluation of forecasters. *The Statistician*, 32:12–22.

Ehm, W., Gneiting, T., Jordan, A., and Krüger, F. (2016). Of quantiles and expectiles: Consistent scoring functions, Choquet representations and forecast rankings (with discussion and rejoinder). *Journal of the Royal Statistical Society: Series B*, 78:505–562.

Ehm, W. and Krüger, F. (2018). Forecast dominance testing via sign randomization. *Electronic Journal of Statistics*, 12:3758–3793.

European Central Bank (2018). ECB survey of professional forecasters (documentation). Available at `https://www.ecb.europa.eu/stats/ecb_surveys/survey_of_professional_forecasters/html/index.en.html`, accessed: September 17, 2018.

Faust, J. and Wright, J. H. (2013). Forecasting inflation. In *Handbook of Economic Forecasting*, volume 2, pages 2–56. Elsevier, Amsterdam.

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106:746–762.

Holzmann, H. and Eulert, M. (2014). The role of the information set for forecasting–with applications to risk management. *Annals of Applied Statistics*, 8:595–621.

Krzysztofowicz, R. and Long, D. (1990). Fusion of detection probabilities and comparison of multisensor systems. *IEEE Transactions on Systems, Man, and Cybernetics*, 20:665–677.

Lehmann, E. L. and Casella, G. (1998). *Theory of Point Estimation.* Springer, New York, 2 edition.

Levy, H. (2016). *Stochastic Dominance: Investment Decision Making Under Uncertainty.* Springer, New York, 3 edition.

Linton, O., Maasoumi, E., and Whang, Y.-J. (2005). Consistent testing for stochastic dominance under general sampling schemes. *Review of Economic Studies*, 72:735–765.

Lobato, I. N. and Velasco, C. (2004). A simple test of normality for time series. *Econometric Theory*, 20:671–689.

Machina, M. and Pratt, J. (1997). Increasing risk: Some direct constructions. *Journal of Risk and Uncertainty*, 14:103–127.

Mincer, J. A. and Zarnowitz, V. (1969). The evaluation of economic forecasts. In Mincer, J. A., editor, *Economic Forecasts and Expectations: Analysis of Forecasting Behavior and Performance*, pages 3–46. Columbia University Press, New York.

Müller, A. and Rüschendorf, L. (2001). On the optimal stopping values induced by general dependence structures. *Journal of Applied Probability*, 38:672–684.

Newey, W. K. and Powell, J. L. (1987). Asymmetric least squares estimation and testing. *Econometrica*, 55:819–847.

Newey, W. K. and West, K. D. (1987). A simple, positive semi-definite, heteroscedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55:703–708.

Newey, W. K. and West, K. D. (1994). Automatic lag selection in covariance matrix estimation. *Review of Economic Studies*, 61:631–653.

Nolde, N. and Ziegel, J. F. (2017). Elicitability and backtesting: Perspectives for banking regulation (with discussion and rejoinder). *Annals of Applied Statistics*, 11:1833–1874.

Patton, A. J. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics*, 160:246–256.

Patton, A. J. (2018). Comparing possibly misspecified forecasts. *Journal of Business & Economic Statistics*. Forthcoming.

Patton, A. J. and Timmermann, A. (2012). Forecast rationality tests based on multi-horizon bounds. *Journal of Business & Economic Statistics*, 30:1–17.

Ranjan, R. and Gneiting, T. (2010). Combining probability forecasts. *Journal of the Royal Statistical Society. Series B*, 72:71–91.

Rothschild, M. and Stiglitz, J. E. (1970). Increasing risk: I. A definition. *Journal of Economic Theory*, 2:225 – 243.

Satopää, V. A., Pemantle, R., and Ungar, L. H. (2016). Modeling probability forecasts via information diversity. *Journal of the American Statistical Association*, 111:1623–1633.

Savage, L. J. (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66:783–801.

Shaked, M. and Shanthikumar, J. G. (2007). *Stochastic Orders*. Springer, New York.

Strähl, C. and Ziegel, J. F. (2017). Cross-calibration of probabilistic forecasts. *Electronic Journal of Statistics*, 11:608–639.

Strassen, V. (1965). The existence of probability measures with given marginals. *Annals of Mathematical Statistics*, 36:423–439.

Yen, T.-J. and Yen, Y.-M. (2018). Testing forecast accuracy of expectiles and quantiles with the extremal consistent loss functions. Working Paper, National Chengchi University, July 2018.

Zeileis, A. (2004). Econometric computing with HC and HAC covariance matrix estimators. *Journal of Statistical Software*, 11:1–17.

Ziegel, J. F., Krüger, F., Jordan, A., and Fasciati, F. (2018). Robust forecast evaluation of Expected Shortfall. *Journal of Financial Econometrics.* Forthcoming.

Appendix

A Result for Dominance of Expectile Forecasts

We state and prove a more general version of Theorem 2.1. We consider the expectile functional of Y at level $\tau \in (0,1)$ (Newey and Powell, 1987). The expectile is the unique value t that satisfies

$$(1-\tau)\int_{(-\infty,t]}(t-y)\,\mathrm{d}F(y) = \tau\int_{[t,\infty)}(y-t)\,\mathrm{d}F(y),$$

where $F(y)$ is the CDF of Y. The mean functional is obtained as a special case for $\tau = 1/2$. As shown by Gneiting (2011), the class of consistent scoring functions for the expectile at level τ is given by

$$S(x,y) = |\mathbf{1}_{(y<x)} - \tau| \;\left(\phi(y) - \phi(x) - \phi'(x)\;(y-x)\right), \tag{8}$$

where ϕ is a convex function with subgradient ϕ'. The relevant class for the mean (see Equation 1) emerges for $\tau = 1/2$. Analogously to Definition 2.1, we then have the following definition of forecast dominance for expectiles.

Definition A.1 (Forecast dominance for expectiles)**.** Forecast A *dominates* forecast B if

$$\mathbb{E}\left(S(X_A, Y)\right) \leq \mathbb{E}\left(S(X_B, Y)\right)$$

for every function S of the form given in (8).

Lemma A.1. *For any Borel set $A \subset \mathbb{R}$,*

$$\begin{aligned}
\mathbb{E}\left((X-Y)_+ \mathbf{1}_A(X)\right) &= \int_{-\infty}^{\infty} \mathbb{P}(Y < w, X > w, X \in A)\,\mathrm{d}w,\\
\mathbb{E}\left((Y-X)_+ \mathbf{1}_A(X)\right) &= \int_{-\infty}^{\infty} \mathbb{P}(Y \geq w, X \leq w, X \in A)\,\mathrm{d}w.
\end{aligned}$$

Proof. By Fubini's theorem, we obtain

$$\begin{aligned}
\mathbb{E}\left((X-Y)_+ \mathbf{1}_A(X)\right) &= \int_{\mathbb{R}} \int_{\mathbb{R}} (x-y)_+ \mathbf{1}_A(x)\,\mathrm{d}F(y|X=x)\,\mathrm{d}G(x)\\
&= \int_{\mathbb{R}} \mathbf{1}_A(x) \int_{(-\infty,x]} (x-y)\,\mathrm{d}F(y|X=x)\,\mathrm{d}G(x)\\
&= \int_{\mathbb{R}} \mathbf{1}_A(x) \int_{(-\infty,x]} \int_{y}^{x} \mathrm{d}w\,\mathrm{d}F(y|X=x)\,\mathrm{d}G(x)\\
&= \int_{\mathbb{R}} \int_{\mathbb{R}} \int_{-\infty}^{\infty} \mathbf{1}_{(-\infty,w)}(y)\mathbf{1}_{(w,\infty)}(x)\mathbf{1}_A(x)\,\mathrm{d}w\,\mathrm{d}F(y|X=x)\,\mathrm{d}G(x)\\
&= \int_{-\infty}^{\infty} \int_{(w,\infty)} \mathbf{1}_A(x) \int_{(-\infty,w)} \mathrm{d}F(y|X=x)\,\mathrm{d}G(x)\,\mathrm{d}w\\
&= \int_{-\infty}^{\infty} \mathbb{E}\left(\mathbf{1}_{(w,\infty)}(X)\mathbf{1}_{(-\infty,w)}(Y)\mathbf{1}_A(X)\right)\,\mathrm{d}w
\end{aligned}$$

where $F(\cdot|X=x)$ denotes the conditional CDF of Y given $X = x$, and G denotes the CDF of X. The proof of the second equality is analogous. □

Lemma A.2. *Let X, Z be two random variables such that $\mathbb{E}(XZ)$ exists and is finite. Then,*

$$\mathbb{E}(XZ) = \int_0^\infty \int_0^\infty \big(H(x,z) - F(x) - G(z) + 1\big)\,\mathrm{d}x\,\mathrm{d}z + \int_{-\infty}^0 \int_{-\infty}^0 H(x,z)\,\mathrm{d}x\,\mathrm{d}z \\ + \int_{-\infty}^0 \int_0^\infty \big(H(x,z) - G(z)\big)\,\mathrm{d}x\,\mathrm{d}z + \int_0^\infty \int_{-\infty}^0 \big(H(x,z) - F(x)\big)\,\mathrm{d}x\,\mathrm{d}z, \qquad (9)$$

where $H(x,z) = \mathbb{P}(X \le x, Z \le z)$, $F(x) = \mathbb{P}(X \le x)$, $G(z) = \mathbb{P}(Z \le z)$ are the joint and marginal CDFs of (X,Z), X and Z, respectively.

Proof. For a random variable Y, we can write

$$Y_+ = \int_0^\infty (1 - \mathbf{1}_{[Y,\infty)}(x))\,\mathrm{d}x, \quad Y_- = \int_{-\infty}^0 \mathbf{1}_{[Y,\infty)}(x)\,\mathrm{d}x,$$

where $Y_+ = \max\{Y, 0\}$, $Y_- = \max\{-Y, 0\}$ are the positive and the negative part of Y, respectively. Therefore,

$$(XZ)_+ = X_+Z_+ + X_-Z_- = \int_0^\infty \int_0^\infty (1 - \mathbf{1}_{[X,\infty)}(x))(1 - \mathbf{1}_{[Z,\infty)}(z))\,\mathrm{d}x\,\mathrm{d}z \\ + \int_{-\infty}^0 \int_{-\infty}^0 \mathbf{1}_{[X,\infty)}(x)\mathbf{1}_{[Z,\infty)}(z)\,\mathrm{d}x\,\mathrm{d}z. \quad (10)$$

Taking the expectation in (10) and using Fubini's theorem, we obtain

$$\mathbb{E}((XZ)_+) = \int_0^\infty \int_0^\infty H(x,z) - F(x) - G(z) + 1\,\mathrm{d}x\,\mathrm{d}z + \int_{-\infty}^0 \int_{-\infty}^0 H(x,z)\,\mathrm{d}x\,\mathrm{d}z,$$

and, similarly, with $(XZ)_- = X_+Z_- + X_-Z_+$,

$$\mathbb{E}((XZ)_-) = \int_{-\infty}^0 \int_0^\infty G(z) - H(x,z)\,\mathrm{d}x\,\mathrm{d}z + \int_0^\infty \int_{-\infty}^0 F(x) - H(x,z)\,\mathrm{d}x\,\mathrm{d}z. \qquad \square$$

Lemma A.3. *The elementary scoring function for expectiles in Ehm et al. (2016, Equation*

12) is identical to the function

$$S_\theta(x, y) = |\mathbf{1}_{(y<\theta)} - \tau|(\theta - y)\mathbf{1}_{(x>\theta)}, \tag{11}$$

up to a difference of $\tau\ (y-\theta)_+$ which does not depend on x.

Proof. Adjusting the notation in Equation (12) of Ehm et al. (2016) (using the symbol τ instead of α for the expectile level), we have

$$S_\theta(x, y) = |\mathbf{1}_{(y<x)} - \tau|\ \left\{(y-\theta)_+ - (x-\theta)_+ - (y-x)\ \mathbf{1}_{(\theta<x)}\right\}.$$

Since $|\mathbf{1}_{(y<x)} - \tau| = \mathbf{1}_{(y<x)}(1-2\tau) + \tau$ and $(z)_+ = z\mathbf{1}_{(z>0)}$, the score can be rewritten as

$$S_\theta(x, y) = \left[\mathbf{1}_{(y<x)}(1-2\tau) + \tau\right](\theta - y)\left[\mathbf{1}_{(x>\theta)} - \mathbf{1}_{(y>\theta)}\right].$$

Subtracting the term $\tau(y-\theta)\mathbf{1}_{(y>\theta)}$ (which does not depend on x) and rearranging, we get

$$\begin{aligned} S_\theta(x, y) &= \left[\mathbf{1}_{(y<x)}(1-2\tau)\right](y-\theta)\left(\mathbf{1}_{(y>\theta)} - \mathbf{1}_{(x>\theta)}\right) + \tau(\theta-y)\mathbf{1}_{(x>\theta)} \\ &= \begin{cases} \left[\mathbf{1}_{(y<x)}(1-2\tau) + \tau\right](\theta-y)\mathbf{1}_{(x>\theta)} & y \le \theta \\ \left[\mathbf{1}_{(y<x)}(1-2\tau)\right](y-\theta)\left(1 - \mathbf{1}_{(x>\theta)}\right) + \tau(\theta-y)\mathbf{1}_{(x>\theta)} & y > \theta \end{cases} \\ &= \begin{cases} (1-\tau)(\theta-y)\mathbf{1}_{(x>\theta)} & y \le \theta \\ \tau(\theta-y)\mathbf{1}_{(x>\theta)} & y > \theta \end{cases} \\ &= |\mathbf{1}_{(y\le\theta)} - \tau|(\theta-y)\mathbf{1}_{(x>\theta)} = |\mathbf{1}_{(y<\theta)} - \tau|(\theta-y)\mathbf{1}_{(x>\theta)}. \end{aligned}$$

□

Theorem A.4. *Let A and B be forecasts for the τ-expectile. Then A dominates B if and* *only if $\psi_A(\theta) \geq \psi_B(\theta)$, for all $\theta \in \mathbb{R}$, where*

$$\begin{aligned}\psi_j(\theta) = \int_\theta^\infty & \tau\mathbb{P}(X_j > w, Y > w) + (1-\tau)\mathbb{P}(X_j > w, Y \leq w)\,\mathrm{d}w \\ & + \tau\mathbb{E}\left((Y - X_j)_+\mathbf{1}_{(X_j>\theta)}\right) - (1-\tau)\mathbb{E}\left((X_j - Y)_+\mathbf{1}_{(X_j>\theta)}\right), \quad \textit{for } j \in \{A, B\}.\end{aligned}$$

Proof. By Ehm et al. (2016, Corollary 1b), A dominates B if and only if $\mathbb{E}\left(S_\theta(X_B, Y)\right) \geq \mathbb{E}\left(S_\theta(X_A, Y)\right)$ for all $\theta \in \mathbb{R}$, where S_θ is given at (11), see Lemma A.3. Note that $S_\theta(X_j, Y)$ is integrable if Y is integrable, $j \in \{A, B\}$. We apply Lemma A.2 to the random variables $\mathbf{1}_{(X_j>\theta)}$ and $|\mathbf{1}_{(Y<\theta)} - \tau|(\theta - Y)$. We have $F(x) = \mathbb{P}(\mathbf{1}_{(X_j>\theta)} \leq x) = \mathbf{1}_{(x\geq 1)} + \mathbf{1}_{(x\in[0,1))}\mathbb{P}(X_j \leq \theta)$,

$$\begin{aligned}G(z) &= \mathbb{P}(|\mathbf{1}_{(Y<\theta)} - \tau|(\theta - Y) \leq z) \\ &= \mathbb{P}((1-\tau)(\theta - Y) \leq z, Y < \theta) + \mathbb{P}(\tau(\theta - Y) \leq z, Y \geq \theta) \\ &= \mathbf{1}_{(z>0)}\mathbb{P}(Y \geq \theta - z/(1-\tau)) + \mathbf{1}_{(z\leq 0)}\mathbb{P}(Y \geq \theta - z/\tau), \\ H(x,z) &= \mathbb{P}(\mathbf{1}_{(X_j>\theta)} \leq x, |\mathbf{1}_{(Y<\theta)} - \tau|(\theta - Y) \leq z) \\ &= \mathbf{1}_{(x\geq 1)}G(z) + \mathbf{1}_{(x\in[0,1),z>0)}\mathbb{P}(X_j \leq \theta, Y \geq \theta - z/(1-\tau)) \\ &\quad + \mathbf{1}_{(x\in[0,1),z\leq 0)}\mathbb{P}(X_j \leq \theta, Y \geq \theta - z/\tau).\end{aligned}$$

Therefore, the first integral on the right hand side of (9) is

$$\begin{aligned}&\int_0^\infty \int_0^\infty H(x,z) - F(x) - G(z) + 1\,\mathrm{d}x\,\mathrm{d}z \\ &= \int_0^\infty \int_0^1 \mathbb{P}(X_j \leq \theta, Y \geq \theta - z/(1-\tau)) - \mathbb{P}(X_j \leq \theta)) - \mathbb{P}(Y \geq \theta - z/(1-\tau)) + 1\,\mathrm{d}x\,\mathrm{d}z \\ &= \int_0^\infty \mathbb{P}(X_j > \theta, Y < \theta - z/(1-\tau))\,\mathrm{d}z.\end{aligned}$$

Similarly, we can compute the third integal on the right hand side of (9) to obtain

$$\int_{-\infty}^0 \int_0^\infty H(x,z) - G(z)\,\mathrm{d}x\,\mathrm{d}z = -\int_{-\infty}^0 \mathbb{P}(X_j > \theta, Y \geq \theta - z/\tau)\,\mathrm{d}z.$$

The second and the fourth integral on the right hand side of (9) are zero because $H(x,z)$ and $F(x)$ are zero for $x<0$. Using a change of variables, we obtain

$$\begin{aligned}\psi_j(\theta) &= -\mathbb{E}\left(S_\theta(X_j,Y)\right)\\ &= \tau\int_\theta^\infty \mathbb{P}(X_j>\theta, Y\geq w)\,\mathrm{d}w - (1-\tau)\int_{-\infty}^\theta \mathbb{P}(X_j>\theta, Y<w)\,\mathrm{d}w.\end{aligned}$$

We can rewrite this as

$$\begin{aligned}\psi_j(\theta) = &\int_\theta^\infty \tau\mathbb{P}(X_j>w, Y\geq w) + (1-\tau)\mathbb{P}(X_j>w,Y<w)\,\mathrm{d}w\\ &+\tau\int_\theta^\infty \mathbb{P}(w\geq X_j>\theta, Y\geq w)\,\mathrm{d}w\\ &-(1-\tau)\left(\int_\theta^\infty \mathbb{P}(X_j>w,Y<w)\,\mathrm{d}w + \int_{-\infty}^\theta \mathbb{P}(X_j>\theta,Y<w)\,\mathrm{d}w\right)\\ =&\int_\theta^\infty \tau\mathbb{P}(X_j>w, Y\geq w) + (1-\tau)\mathbb{P}(X_j>w,Y<w)\,\mathrm{d}w\\ &+\tau\int_{-\infty}^\infty \mathbb{P}(X_j\leq w, Y\geq w, X_j>\theta)\,\mathrm{d}w\\ &-(1-\tau)\int_{-\infty}^\infty \mathbb{P}(X_j>w,Y<w,X_j>\theta)\,\mathrm{d}w\\ =&\int_\theta^\infty \tau\mathbb{P}(X_j>w, Y\geq w) + (1-\tau)\mathbb{P}(X_j>w,Y<w)\,\mathrm{d}w\\ &+\tau\mathbb{E}\left((Y-X_j)_+\mathbf{1}_{(X_j>\theta)}\right) - (1-\tau)\mathbb{E}\left((X_j-Y)_+\mathbf{1}_{(X_j>\theta)}\right),\end{aligned}$$

where the second equality holds because $\mathbb{P}(w\geq X_j>\theta, Y\geq w)=0$ for $w<\theta$ and $\mathbb{P}(X_j>w,Y<w,X_j>\theta)=\mathbb{P}(X_j>w,Y<w)$ for $w\geq\theta$, $\mathbb{P}(X_j>w,Y<w,X_j>\theta)=\mathbb{P}(Y<w,X_j>\theta)$ for $w<\theta$. The last equality follows from Lemma A.1 with $A=(\theta,\infty)$. □

B Proofs and Technical Details

Proof of Theorem 2.1. The result follows from Theorem A.4 with $\tau=1/2$ because $\mathbb{P}(X_j>w,Y\geq w)+\mathbb{P}(X_j>w,Y<w)=\mathbb{P}(X_j>w)$ and $\mathbb{E}\left(((Y-X_j)_+-(X_j-Y)_+)\mathbf{1}_{(X_j>\theta)}\right)$

$= \mathbb{E}\left((Y - X_j)\mathbf{1}_{(X_j>\theta)}\right) = \mathbb{E}((\mathbb{E}\left(Y\middle|X_j\right) - X_j)\mathbf{1}_{(X_j>\theta)})$, where the second equality uses the law of iterated expectations. □

Proof of Theorem 3.1. Under auto-calibration, $\mathbb{E}\left(Y\middle|X_j\right) = X_j$ holds almost surely. In view of Theorem 2.1, Theorem 3.1 then follows from Müller and Rüschendorf (2001, Corollary 4.1) which shows that X_A is greater than X_B in convex order if and only if $\int_a^\infty \mathbb{P}(X_A > t)\,\mathrm{d}t \geq \int_a^\infty \mathbb{P}(X_B > t)\,\mathrm{d}t$ for all $a \in \mathbb{R}$, and

$$\lim_{a\to-\infty}\left(\int_a^\infty \mathbb{P}(X_A > t)\,\mathrm{d}t - \int_a^\infty \mathbb{P}(X_B > t)\,\mathrm{d}t\right) = 0. \tag{12}$$

To see why (12) holds, note that

$$\lim_{a\to-\infty}\left(\int_a^\infty \mathbb{P}(X_A > t)\,\mathrm{d}t - \int_a^\infty \mathbb{P}(X_B > t)\,\mathrm{d}t\right) = \mathbb{E}\left(X_A\right) - \mathbb{E}\left(X_B\right) = \mathbb{E}\left(Y\right) - \mathbb{E}\left(Y\right) = 0,$$

where the first equality follows from Müller and Rüschendorf (2001, Proposition 4.1.(a)(iii)), and the second equality follows from auto-calibration. □

Proof of Proposition 3.2. Auto-calibration of X_j holds because $\sigma(X_j) \subseteq \mathcal{F}_j$ and $\mathbb{E}\left(Y\middle|X_j\right) = \mathbb{E}\left(\mathbb{E}\left(Y\middle|\mathcal{F}_j\right)\middle|X_j\right) = \mathbb{E}\left(X_j\middle|X_j\right) = X_j$, where the first equality uses the tower property of conditional expectation. To show that X_A is greater than X_B in convex order, note that $\mathbb{E}\left(X_A\middle|X_B\right) = \mathbb{E}\left(\mathbb{E}\left(Y\middle|\mathcal{F}_A\right)\middle|X_B\right) = \mathbb{E}\left(Y\middle|X_B\right) = X_B$, where the second equality again uses the tower property, together with the fact that $\sigma(X_B) \subset \mathcal{F}_A$. Strassen's 1965 characterization mentioned in Section 2 thus implies that X_A is greater than X_B in convex order. □

Proof of Corollary at the end of Section 3. Due to auto-calibration, $\mathrm{Cov}(X_j, Y) = \mathbb{V}\left(X_j\right)$ for $j \in \{A, B\}$, where Cov denotes covariance. The convex order condition implies that $\mathbb{V}\left(X_A\right) \geq \mathbb{V}\left(X_B\right)$, and hence that $\mathrm{Cor}(X_A, Y) = \sqrt{R_A^2} \geq \mathrm{Cor}(X_B, Y) = \sqrt{R_B^2}$, where Cor denotes correlation and R_j^2 is the R^2 from the Mincer-Zarnowitz regression for forecast j. □

Proof of Proposition 4.1. Suppose that (X_j, Y) follow a bivariate normal distribution. We compute $\psi_j(\theta)$ defined in Theorem 2.1 for $j \in \{A, B\}$. We have that $\mathbb{E}\left(Y\middle|X_j\right) = \mu_Y +$

$\rho_{Yj}(\sigma_Y/\sigma_j)(X_j-\mu_j)$, and hence

$$\begin{aligned}\mathbb{E}\left((\mathbb{E}\left(Y\middle|X_j\right)-X_j)\mathbf{1}_{(X_j>\theta)}\right)&=\mathbb{E}\left(\left(\mu_Y+\rho_{Yj}\frac{\sigma_Y}{\sigma_j}(X_j-\mu_j)-X_j\right)\mathbf{1}_{(X_j>\theta)}\right)\\&=\left(\mu_Y-\theta-\rho_{Yj}\frac{\sigma_Y}{\sigma_j}(\mu_j-\theta)\right)\left(1-\Phi\left(\frac{\theta-\mu_j}{\sigma_j}\right)\right)+\left(\rho_{Yj}\frac{\sigma_Y}{\sigma_j}-1\right)\sigma_j\Psi\left(\frac{\theta-\mu_j}{\sigma_j}\right),\end{aligned}$$

where we define for $\theta\in\mathbb{R}$, $\Psi(\theta)=\int_\theta^\infty 1-\Phi(w)\,\mathrm{d}w$. Then,

$$\begin{aligned}\psi_j(\theta)&=\frac{\sigma_j}{2}\Psi\left(\frac{\theta-\mu_j}{\sigma_j}\right)+\frac{1}{2}\mathbb{E}\left((\mathbb{E}\left(Y\middle|X_j\right)-X_j)\mathbf{1}_{(X_j>\theta)}\right)\\&=\frac{1}{2}\left(\mu_Y-\theta-\rho_{Yj}\frac{\sigma_Y}{\sigma_j}(\mu_j-\theta)\right)\left(1-\Phi\left(\frac{\theta-\mu_j}{\sigma_j}\right)\right)+\frac{\rho_{Yj}\sigma_Y}{2}\Psi\left(\frac{\theta-\mu_j}{\sigma_j}\right).\end{aligned}\tag{13}$$

Using the assumption that $\mu_A=\mu_B=\mu_Y$ and the fact that $\Psi(\theta)=\varphi(\theta)-\theta\ (1-\Phi(\theta))$, Equation (13) yields that

$$2\ \psi_j(\theta)=\rho_{Yj}\sigma_Y\ \varphi\left(\frac{\theta-\mu_Y}{\sigma_j}\right)-(\theta-\mu_Y)\ \left(1-\Phi\left(\frac{\theta-\mu_Y}{\sigma_j}\right)\right).\tag{14}$$

□

Notes on Cases 1 to 4. Case 1 holds because, for each $\theta\in\mathbb{R}$, we have $2\psi_A(\theta)+(\theta-\mu_Y)\geq\sigma_A\ \varphi((\theta-\mu_Y)/\sigma_A)+(\theta-\mu_Y)\Phi((\theta-\mu_Y)/\sigma_A)\geq\sigma_B\varphi((\theta-\mu_Y)/\sigma_B)+(\theta-\mu_Y)\Phi((\theta-\mu_Y)/\sigma_B)\geq 2\psi_B(\theta)+(\theta-\mu_Y)$, where $\psi_j(\theta)$ has been defined at (14). Case 2a can be shown by reparametrizing $\sigma_{Yj}=\sigma_Y\sigma_j\rho_{Yj}$, and differentiating $2\psi_j(\theta)$ with respect to σ_j. Case 3 can be shown by differentiating $2\psi_j(\theta)$ with respect to σ_j. Cases 2b and 4 are immediate. □

Proof of Theorem 5.1. Denote the CDF of η_j, conditional on $\mathcal{F}$, by $F_j^{\mathcal{F}}$, for $j\in\{A,B\}$. By Shaked and Shanthikumar (2007, Theorem 3.D.1), the assumptions of Theorem 5.1 imply that

$$F_A^{\mathcal{F}}(z)-F_B^{\mathcal{F}}(z)\begin{cases}\geq 0, & \text{for } z\geq 0,\\ \leq 0, & \text{for } z\leq 0.\end{cases}\tag{15}$$

By Ehm et al. (2016, Corollary 1b), A dominates B if and only if $\mathbb{E}\left(S_\theta(X_B,Y)\right)\geq\mathbb{E}\left(S_\theta(X_A,Y)\right)$

for all $\theta \in \mathbb{R}$, where S_θ is given at (3). The random variable $S_\theta(X_j, Y)$ is integrable if Y is integrable. Define $W = \mathbb{E}\left(Y \middle| \mathcal{F}\right)$, and let $\theta \in \mathbb{R}$. Then,

$$\begin{aligned} 2\,\mathbb{E}\left(S_\theta(X_j, Y)\right) &= \mathbb{E}\left(\mathbf{1}_{(\theta < X_j)}(\theta - Y)\right) = \mathbb{E}\left(\mathbb{E}\left(\mathbf{1}_{(\theta < W + \eta_j)}(\theta - W - \varepsilon)\middle|\mathcal{F}\right)\right) \\ &= \mathbb{E}\left(\mathbb{E}\left(\mathbf{1}_{(\theta - W < \eta_j)}\middle|\mathcal{F}\right)\mathbb{E}\left((\theta - W - \varepsilon)\middle|\mathcal{F}\right)\right) = \mathbb{E}\left((1 - F_j^{\mathcal{F}}(\theta - W))(\theta - W)\right). \end{aligned}$$

Hence, (15) implies

$$\mathbb{E}\left(S_\theta(X_B, Y)\right) - \mathbb{E}\left(S_\theta(X_A, Y)\right) = \frac{1}{2}\mathbb{E}\left(\left(F_A^{\mathcal{F}}(\theta - W) - F_B^{\mathcal{F}}(\theta - W)\right)(\theta - W)\right) \geq 0. \qquad \square$$

Proof of Proposition 5.2. Parts (a) and (b) are immediate given the setup of Theorem 5.1. Regarding (c), we have the following inequality for any strictly increasing function ϕ:

$$\begin{aligned} \mathbb{E}\left(\phi(|\eta_B|)\middle|\mathcal{F}\right) &= \int_0^\infty \mathbb{P}(\phi(|\eta_B|) > w|\mathcal{F})\ dw \\ &= \int_0^\infty \left(\mathbb{P}(\eta_B < -\phi^{-1}(w)|\mathcal{F}) + \mathbb{P}(\eta_B > \phi^{-1}(w)|\mathcal{F})\right)\ dw \\ &\geq \int_0^\infty \left(\mathbb{P}(\eta_A < -\phi^{-1}(w)|\mathcal{F}) + \mathbb{P}(\eta_A > \phi^{-1}(w)|\mathcal{F})\right)\ dw = \mathbb{E}\left(\phi(|\eta_A|)|\mathcal{F}\right), \end{aligned} \tag{16}$$

where the inequality follows from (15) in the proof of Theorem 5.1.

Now let $W = \mathbb{E}\left(Y \middle| \mathcal{F}\right)$, such that $(X_j)^{2k} = (W + \eta_j)^{2k}$. For terms of the form $W^c \eta_j^d$, with c and d being odd integers, it holds that $\mathbb{E}\left(W^c \eta_j^d\right) = \mathbb{E}\left(W^c \mathbb{E}\left(\eta_j^d \middle| \mathcal{F}\right)\right) = 0$, where the last equality follows from symmetry of $F_j^{\mathcal{F}}$ around zero. For terms of the form $W^c \eta_j^d$, with c and d being even integers, it holds that $\mathbb{E}\left(W^c \eta_B^d\right) = \mathbb{E}\left(W^c\ \mathbb{E}\left(\eta_B^d \middle| \mathcal{F}\right)\right) \geq \mathbb{E}\left(W^c\ \mathbb{E}\left(\eta_A^d \middle| \mathcal{F}\right)\right)$, where the inequality follows from (16). Part (b) of Theorem 5.1 then follows from the binomial theorem. □