Helsinki University of Technology
Dissertations in Computer and Information Science
Espoo 2008 Report D25

STABILITY AND INFERENCE IN DISCRETE DIFFUSION SCALE-SPACES

Ramūnas Girdziušas

Dissertation for the degree of Doctor of Science in Technology to be presented with due permission of the Faculty of Information and Natural Sciences for public examination and debate in Auditorium AS1 at Helsinki University of Technology (Espoo, Finland) on the 29th of February, 2008, at 12 o'clock noon.

Helsinki University of Technology
Faculty of Information and Natural Sciences
Department of Information and Computer Science

Distribution:
Helsinki University of Technology
Department of Information and Computer Science
P.O. Box 5400
FI-02015 TKK
FINLAND
Tel. +358-9-451 3272
Fax +358-9-451 3277
http://www.ics.tkk.fi

ISBN 978-951-22-9206-6 (printed version)
ISBN 978-951-22-9207-3 (electronic version)
ISSN 1459-7020

Multiprint Oy
Espoo 2008

Girdziušas, R. (2008): **Stability and Inference in Discrete Diffusion Scale-Spaces**. Doctoral thesis, Helsinki University of Technology, Dissertations in Computer and Information Science, Report D25, Espoo, Finland.

Keywords: smoothing, scale-space, discrete nonlinear diffusion, majorization, Rose determinant, doubly stochastic matrix, variation diminishing, total positivity, tridiagonal M-matrix, optimal scale selection, marginal likelihood inference, monotonicity, Gaussian models, Faà di Bruno formula.

Abstract

Taking averages of observations is the most basic method to make inferences in the presence of uncertainty. In late 1980's, this simple idea has been extended to the principle of *successively average less where the change is faster*, and applied to the problem of revealing a signal with jump discontinuities in additive noise.

Successive averaging results in a family of signals with progressively decreasing amount of details, which is called the *scale-space* and further conveniently formalized by viewing it as a solution to a certain diffusion-inspired evolutionary partial differential equation (PDE). Such a model is known as the *diffusion scale-space* and it possesses two long-standing problems: (i) *model analysis* which aims at establishing stability and guarantees that averaging does not distort important information, and (ii) *model selection*, such as identification of the optimal scale (diffusion stopping time) given an initial noisy signal and an incomplete model.

This thesis studies both problems in the discrete space and time. Such a setting has been strongly advocated by Lindeberg [1991] and Weickert [1996] among others. The focus of the model analysis part is on necessary and sufficient conditions which guarantee that a discrete diffusion possesses the scale-space property in the sense of sign variation diminishing. Connections with the total variation diminishing and the open problem in a multivariate case are discussed too.

Considering the model selection, the thesis unifies two optimal diffusion stopping principles: (i) the time when the Shannon entropy-based Liapunov function of Sporring and Weickert [1999] reaches its steady state, and (ii) the time when the diffusion outcome has the least correlation with the noise estimate, contributed by Mrázek and Navara [2003]. Both ideas are shown to be particular cases of the marginal likelihood inference. Moreover, the suggested formalism provides first principles behind such criteria, and removes a variety of inconsistencies. It is suggested that the outcome of the diffusion should be interpreted as a certain expectation conditioned on the initial signal of observations instead of being treated as a random sample or probabilities. This removes the need to normalize signals in the approach of Sporring and Weickert [1999], and it also better justifies application of the correlation criterion of Mrázek and Navara [2003].

Throughout this work, the emphasis is given on methods that enable to reduce the problem to that of establishing the positivity of a quadratic form. The necessary and sufficient conditions can then be approached via positivity of matrix minors. A supplementary appendix is provided which summarizes a novel method of evaluating matrix minors. Intuitive examples of difficulties with statistical inference conclude the thesis.

Preface

> *Above all, he [Quesalid] learned the* ars magna *of one of the shamanistic schools of the Northwest Coast: The shaman hides a little tuft of down in a corner of his mouth, and he throws it up, covered with blood, at the proper moment–after having bitten his tongue or made his gums bleed–and solemnly presents it to his patient and the onlookers as the pathological foreign body extracted as a result of his sucking and manipulations.*
>
> *While visiting the neighboring Koskimo Indians, Quesalid attends a curing ceremony of his illustrious colleagues of the other tribe. To his great astonishment he observes a difference in their technique. Instead of spitting out the illness in the form of a "bloody worm" (the concealed down), the Koskimo shamans merely spit a little saliva into their hands, and they dare to claim that this is "the sickness"...*
>
> C. Lévi–Strauss (Anthropologie structurale, 1958, transl. by C. Jacobson and B. Grundfest Schoepf)

This thesis deals with the analysis of the possibility to automate scale-space methods on a computer. It is a summary of work which has been carried out at the Department of Information and Computer Science in the Faculty of Information and Natural Sciences at Helsinki University of Technology under the ComMIT grant 2002–2006. I am very grateful to Prof. Erkki Oja who made this work possible and more concrete. I am indebted to Doc. Jorma Laaksonen who has been my direct instructor for years. I thank both of the reviewers, Prof. Keijo Ruotsalainen and Prof. Samuli Siltanen, for their thorough and constructive comments. A particular credit must be given to my parents, friends and colleagues for their support. The story of two shamans I heard from Doc. Vitalij Nevdacha.

Following a famous phrase, there may exist a certain weakness for mathematics, but no addiction to it. The text is not self-contained, and basic knowledge of matrix theory and statistical inference is assumed from the reader. In order to emphasize contributory aspects, known statements are not proven. The references to the sources with their proofs are given instead. An unnecessary explosion of citations is often prevented by stating a more recent work which contains references to original sources.

In order to ease the reading, acronyms are not much used in the thesis. The term 'Shannon entropy' may refer to 'differential Shannon entropy', 'likelihood' to 'density likelihood', and 'covariance' to 'covariance function', 'covariance matrix', or 'element in the covariance matrix'. When the terms are frequently referred to and the context is clear, the explicit connotations become unnecessary.

Ramūnas Girdziušas

Otaniemi, 28th of January, 2008

Contents

List of Abbreviations

clf.	clarify
e.g.	for example
i.e.	exactly, namely
i.i.d.	independent identically distributed
w.r.t.	with respect to
AWGN	additive white Gaussian noise
LU	lower–upper
PDE	partial differential equation
SNR	signal-to-noise ratio
SR	sign regularity
SSR	strict sign regularity
TP	total positivity
STP	strict total positivity

Frequently Used Symbols

$\mathbf{u} \in \mathbb{R}^n$	column vector of n real-valued elements
$[\mathbf{A}]_{ij}$, a_{ij}	ij-th element of the matrix $\mathbf{A}$
$\lvert\mathbf{A}\rvert$	determinant
$\mathrm{tr}(\mathbf{A})$	trace
$\mathbf{A}^T$	transpose
$\lambda(\mathbf{A})$	eigenvalue
$\mathbf{A} \preceq \mathbf{B}$	$\mathbf{B} - \mathbf{A}$ is positive semi-definite
$\mathbf{x} \prec \mathbf{y}$	$\mathbf{x}$ is majorized by $\mathbf{y}$
$\mathrm{vec}_m(\mathbf{A})$	contraction of the array $\mathbf{A}$ along the dimension m
$\boldsymbol{\alpha}_{1:n}$, $\boldsymbol{\beta}_{1:n}$	sequences of n increasing integers
$\mathbf{A}(\boldsymbol{\alpha}_{1:n}\vert\boldsymbol{\beta}_{1:n})$	submatrix with row and column indices $\boldsymbol{\alpha}_{1:n}$ and $\boldsymbol{\beta}_{1:n}$, resp.
$A(\boldsymbol{\alpha}_{1:n}\vert\boldsymbol{\beta}_{1:n})$	matrix minor
$\lVert\mathbf{x}\rVert^2_{\mathbf{A}}$, $\lVert\mathbf{x}\rVert_p$, $\lVert\cdot\rVert_{L_p}$	quadratic form $\mathbf{x}^T\mathbf{A}\mathbf{x}$, p-norm, L_p-norm
$\gamma_{1:n}$	sequence of the elements $1, 2, \ldots, n$
$\gamma_{1:n}\backslash\{i, j\}$	sequence of elements $1, 2, \ldots, n$ without elements i and j
$\psi_{\mathrm{tv}}(\mathbf{u})$	total variation of vector $\mathbf{u}$
$\psi_s(\mathbf{u})$	sign variation of vector $\mathbf{u}$
$\mathrm{Prob}(E)$	probability of the event E
$p(\mathbf{x})$, $p(\mathbf{x}\vert\mathbf{y})$	same as $p_X(X = \mathbf{x})$, same as $p_{X\vert Y}(X = \mathbf{x}\vert Y = \mathbf{y})$
$\langle g(X)\vert\mathbf{y}\rangle$	expectation of $g(X)$ w.r.t. $p_{X\vert Y}(X = \mathbf{x}\vert Y = \mathbf{y})$ given $\mathbf{y}$
$\mathrm{Cov}(A, B)$	$\langle(A - \langle A\rangle)(B - \langle B\rangle)^T\rangle$, $\mathrm{Cov}(A) \equiv \mathrm{Cov}(A, A)$
$\mathrm{Var}(A)$	$\mathrm{Cov}(A, A)$ when A takes values in $\mathbb{R}$
$X \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	X is normal with the mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$
$H(\mathbf{p})$, $h(X)$	Shannon entropy, differential Shannon entropy
∇	gradient (Cartesian nabla)
$\nabla\cdot$	divergence
$\nabla\nabla^T$	Hesse matrix
$\propto$	proportional to
$\circ$	Hadamard (element-wise) product of matrices
$f(x) = o(x)$	$\lim_{x\to 0} f(x)/x = 0$
$f(x) = O(x)$	$f(x) \leq cx$ for some constant c (asymptotically)
$f(x) \sim g(x)$	$\lim_{x\to\infty} f(x)/g(x) = 1$
$A \Rightarrow B$, $A \Leftrightarrow B$	B holds if A holds, B holds if and only if A holds
$\mathbb{R}_+$, $\mathbb{Z}_+$	positive reals, positive integers

Chapter 1

About the Thesis

1.1 Introduction

This thesis aims at contributing to our understanding on the processing of information. The goal is to discuss the interaction of such concepts as 'stability' and 'incomplete knowledge' by using mathematics. The work allows to better utilize computers in automating model selection.

The basic operative setting is the method of taking (local) averages of observations, called *smoothing*. In this way, we remove unnecessary information (details) and create a more robust representation of measurement, which we can already call as knowledge. When data consists of dense observation sets residing in low-dimensional spaces (time series, images), one can further average the observations iteratively or successively and produce the so-called *scale-space* of signals with progressively lesser details. The smallest amount of useful information can then be attributed to such a member of the scale-space whose values are all constant and equal to the average value of the given signal. Often, a certain optimal scale will exist at which the member of the scale-space reveals some useful properties of the observations. The principle of a scale-space is clarified in Fig. 1.1 as a space of signals created by successively smoothing the observations.

One of the first and most studied scale-spaces can be attributed to the Fourier model of heat flow, which is known as the linear diffusion model in a continuous space and time. Many of its properties have been analyzed in different branches of mathematics and physics. Utilization of diffusion in solving the problems of 'signal in noise' has been studied a lot especially from 1950 to 1970s and can be found under the topic of 'Wiener–Kolmogorov smoothing'.

In the late 1980s, numerous studies appeared on the implementation of the smoothing paradigm 'successively average less where signals change faster'. This type of smoothing revealed the possibility of discontinuity formation and the recovery of discontinuities from corrupted observations. A particular formalism that aims to generalize a linear diffusion model has been independently discussed in a vari-

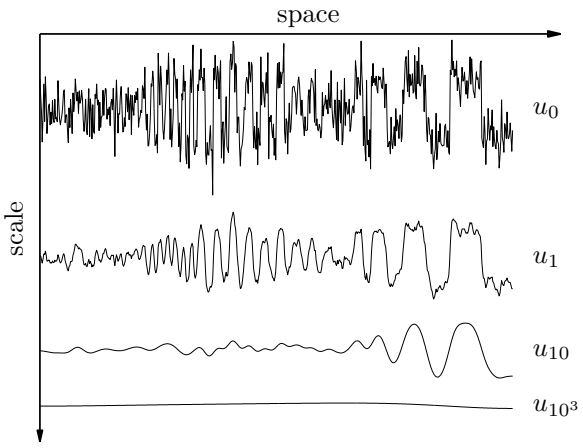


Figure 1.1: The principle of the scale-space is to enrich a given signal u_0 with a family of its 'versions' u_s, $s > 0$. Each signal u_s is assumed to have less structure than u_{s-1}, and a large portion of scale-space analysis is devoted to a better understanding of what 'structure' really means.

ety of fields: (i) image processing, i.e. [Perona and Malik, 1987], [Matsuba, 1988] and [Gidas, 1989], (ii) physics, i.e. [Rosenau et al., 1989] and [Rosenau, 1990, 1992], and (iii) mathematical analysis, see e.g. [Bertsch and Dal Passo, 1989, 1992] and [Barenblatt et al., 1993]. Image processing community considered 'low level vision'. Physicists studied finitely-fast propagating diffusions by unifying hyperbolic systems with their parabolic counterparts. Mathematicians dwelled deeper into the nature of continuity and ill-posedness.

This work will focus on the setting of a finitely discrete space and time, which is very much inspired by the works of Lindeberg [1990] and Weickert [1996]. A central tool in this work is an array of m rows and n columns of real-valued entries, denoted as $\mathbf{A} \in \mathbb{R}^{m \times n}$ and called a matrix. The emphasis will be given on the use of tridiagonal matrices which stand as discrete counterparts of continuous Laplacian operators. There are numerous ways to set the elements of these matrices in order to introduce useful 'nonlinear effects' into the smoothing. A variety of such nonlinearities is summarized in [Teboul et al., 1998]. The work of Keeling [2003] provides comparisons with such a simple smoothing method as the median filter, and D'Almeida [2003] has released a stable numerical implementation of a diffusion which uses a fast LU decomposition of the tridiagonal matrix and additive operator splitting. The latter remains rather little known even among the experts of the numerical analysis of partial differential equations (PDEs).

In this thesis, a discrete scale-space is used as a hypothesis space to make improvements to the methodology of the optimal scale selection. The contributions of this work are related to the model selection problem, which should be here considered as a synonym to model identification, reasoning under uncertainty or information incompleteness, statistical inference, parameter estimation, probabilistic completion and other related concepts.

1.2 Contribution

This work contributes to the optimal scale selection. In general, such a scale may not exist or it might be even better to utilize a variety of scales in order to obtain optimal smoothing of signals. However, this type of remarks concerns with the choice of a specific formalism. One can often reveal a single 'effective scale' when performing model averaging or find out that both formalisms produce meaningful results, but operate with incompatible assumptions.

Instead of summarizing philosophical debates between various schools, the thesis focuses on two important examples which are the works of Sporring and Weickert [1999] and Mrázek and Navara [2003]. It is shown how, on the idea level, a heuristic reasoning in these papers automatically emerges as a byproduct of the marginal likelihood inference. In particular, Propositions 3.6 and 3.7 unify two seemingly different methods and eliminate a variety of inconsistencies. Essential improvements can be mentioned:

1. Rényi's entropies in [Sporring and Weickert, 1999] are known to increase w.r.t. to increasing scale. However, this does not justify them as the optimal scale locators. Models are developed and tested which: (i) support the idea of entropy-based stopping theoretically, and (ii) further bring it to a computational level.

2. It is a nasty habit to confuse probabilities with observations, which takes place in the development of many iterative smoothing algorithms stated in [Carasso, 1999], and the work of Sporring and Weickert [1999] is no exception. The latter introduces a conceptual confusion between stability and inference, and it ignores the fact that probabilities are not physically measurable quantities. As a result, unnecessary preprocessing of signals emerges which fails against intuitive arguments of invariance. This difficulty is eliminated by suggesting the use of differential Shannon entropies whose monotonicity w.r.t. increasing scale is established in Proposition 3.9.

3. Mrázek and Navara [2003] advocate the 'decorrelation' as the optimal selection criterion. It is simple and works well, but it lacks justification. In this case, another source of misunderstandings becomes evident: Process sample paths may be confused with conditional expectations taken over them. The thesis provides explicit probabilistic justification of the principle 'decorrelate the model output with noise' and further generalizes it to non-Gaussian situations in Proposition 3.8.

The basic message is that a discrete diffusion-inspired scale-space can be seen as an incomplete probabilistic statement which provides the knowledge at the level of conditional means. The problem of optimal scale selection is seen as the presence of unknown covariances. The thesis contributes with a nontrivial relationship between the two layers of a statistical hierarchy.

During the probabilistic completion, a variety of technical difficulties are resolved. However, model selection is not possible without model analysis. Stability of the hypothesis space comes first, and probing the plausibility of a particular statement

within such a space should be studied later. Thus, several known theorems of matrix analysis are related in order to establish sufficient conditions for sign variation diminishing and other stability criteria. This problem is reduced to characterization of positive definiteness, which culminates with Proposition 2.13. Computer simulations are provided in Chapter 4 and novel insights can also be found in Appendix A.

1.3 Thesis Organization

Chapter 2 provides a brief introduction to a discrete scale-space model and discusses its stability. Two different approaches are outlined. The first one relies on the majorization theorem and indicates when a discrete diffusion propagator is doubly stochastic. The second one exploits the concept of the total positivity. The principal contribution is Proposition 2.13 which states when the discrete propagator is positive definite, which is essential in establishing the diminishing of a variety of important quantities and the sign variation in particular.

Chapter 3 contains the essential contribution of the thesis. It poses the problem of the optimal scale selection and puts it into the framework of an 'incomplete probabilistic statement'. The concept of incompleteness was especially emphasized by Von Neumann [1955] and Jaynes [2003]. The latter sought for a consistent principle of information conditioning. The approach taken here focuses on the information completion from the level of conditional means to the level of conditional covariances within the family of Gaussian models. Each section of this chapter contributes with a novel insight to a discussed problem. The basic message here is that the use of correlation and entropy should not be viewed as something heuristic and mutually excluding, as both ideas are simplifications of the maximum likelihood inference.

Chapter 4 provides numerical examples with an edge-enhancing nonlinear diffusion. One will find a simple setting of a 'signal in noise' problem, the application of the nonlinear diffusion and a further comparison of temporal evolutions of the optimal stopping criteria. The results confirm the possibility of a successful scale selection.

Chapter 5 states several concluding remarks. It emphasizes the advantages of utilizing Gaussian models and opposes them against other existing frameworks. The importance of energies and variational descriptions is appraised in light of critical remarks existing in the literature.

Appendix A is dedicated to findings which are less relevant to the main contributions of the thesis. Section A.1 discusses novel ways of evaluating matrix determinants, which indirectly characterizes positive definiteness. This could be a starting of a program to investigate the sign variation diminishing in multivariate cases, but it is also of an independent interest as indicated by several applications stated therein. Section A.2 indicates some of the pitfalls that occur in statistical inference. Several existing mistakes, which are sometimes committed even by the great experts in the field, are spotted out and presented with resolutions. The section also falls under the title 'Something what everybody should know, but probably does not'.

1.4 List of Related Publications

This monograph outgrew preliminary ideas communicated to the communities of machine learning and computer vision:

1. M. Aksela, R. Girdziušas, J. Laaksonen, E. Oja, and J. Kangas. Methods for adaptive combination of classifiers with application to recognition of handwritten characters. *Int. Journal of Document Analysis and Recognition*, 6(1):23–41, 2003.

2. R. Girdziušas and J. Laaksonen. Gaussian process regression with fluid hyper-priors. In N. R. Pal, N. Kasabov, and R. K. Mudi et al., editors, *Proc. of 11th Int. Conf. on Neural Information Processing*, volume 3316 of *Lecture Notes in Computer Science*, pages 567–572. Springer, 2004.

3. R. Girdziušas and J. Laaksonen. Gaussian processes of nonlinear diffusion filtering. In *Proc. of Int. Joint Conf. on Neural Networks*, pages 1012–1017. IEEE, 2005.

4. R. Girdziušas and J. Laaksonen. Optimal ratio of Lamé moduli with application to motion of Jupiter storms. In H. Kalviainen, J. Parkkinen, and A. Kaarna, editors, *Proc. of 14th Scandinavian Conference on Image Analysis*, volume 3540 of *Lecture Notes in Computer Science*, pages 1096–1106. Springer, 2005.

5. R. Girdziušas and J. Laaksonen. Jacobi alternative to Bayesian evidence maximization in diffusion filtering. In W. Duch, J. Kacprzyk, and E. Oja et al., editors, *Proc. of Int. Conf. on Artificial Neural Networks*, volume 3697 of *Lecture Notes in Computer Science*, pages 247–252. Springer, 2005.

6. R. Girdziušas and J. Laaksonen. Use of input deformations with Brownian motion filters for discontinuous regression. In S. Singh, M. Singh, C. Apte, and P. Perner, editors, *Proc. of 3rd Int. Conf. on Advances in Pattern Recognition*, volume 3686 of *Lecture Notes in Computer Science*, pages 219–228. Springer, 2005.

7. R. Girdziušas and J. Laaksonen. Optimal stopping and constraints for diffusion models of signals with discontinuities. In J. Gama, R. Camacho, P. Brazdil, A. Jorge, and L. Torgo, editors, *Proc. of 16th European Conference on Machine Learning*, volume 3720 of *Lecture Notes in Artificial Intelligence*, pages 576–583. Springer, 2005.

8. R. Girdziušas and J. Laaksonen. When is a discrete diffusion a scale-space? In A. Shashua D. Metaxas, B. C. Vemuri and H. Shum, editors, *Proc. of 11th IEEE Int. Conf. on Computer Vision*, pages 6. IEEE, 2007.

9. R. Girdziušas and J. Laaksonen. How marginal likelihood inference unifies entropy, correlation and SNR-based stopping in nonlinear diffusion scale-spaces. In I. S. Kweon Y. Yagi, S. B. Kang and H. Zha, editors, *Proc. of 8th Asian Conf. on Computer Vision*, volume 4843 of *Lecture Notes in Computer Science*, pages 811–820. Springer, 2007.

Chapter 2

The Scale-Space Property

A scale-space can be understood as a space of signals with a similar information content but with different 'levels of details'. It consists of: (i) a signal to be analyzed, and (ii) a way to 'smooth' it. The main theme in this chapter is quantification of the 'scale-space property' as discrete variation diminishing, which turns out to be related to other notions of stability.

The basic model in focus is a discrete diffusion, which is defined via a successive multiplication (time-ordered product) of generalized Laplacian matrices. The problem is to establish the conditions which guarantee that multiplying any given vector by such matrices remains stable and possesses the scale-space property, i.e. the diminishing (nonincrease) of sign changes present in a given vector.

The aim is to relate the scale-space property to other notions of stability, and to 'purify' them out of lengthy and complex arguments of continuity. This will also emphasize that many evolutions, irrespectively whether they are linear or nonlinear, are 'smoothing' procedures if the propagator matrices satisfy structural properties such as double stochasticity and (or) total positivity. Thus, it is rather surprising to realize that in a finitely discrete space and discrete time, a nonamplification of global extrema and numerous mathematical caricatures of the second law of thermodynamics are: (i) related to the scale-space property, and (ii) can be modeled via linear algebra, i.e. majorization and total positivity (sign regularity).

The principal contributions of this chapter are Propositions 2.5 and 2.13. They state sufficient (and in some cases necessary) conditions when a discrete diffusion satisfies a variety of stability results, and the scale-space property in particular. Many concepts have their original appearance in continuous settings, and the text draws some parallels with the continuous space. However, the essential statements belong to a finitely discrete space, and there is no ambition to study the possible equivalence of both worlds.

2.1 Introduction to Diffusion Scale-Spaces

A classical linear scale-space consists of an initial signal $u(\mathbf{x}, 0)$ and its convolutions $u(\mathbf{x}, t)$ with the Gaussian density of an increasing variance t, which is called 'scale'. This is known as the Gaussian scale-space which can be derived from the axioms of linearity, separability along dimensions, and invariance w.r.t. spatial shifts, scalings and rotations. Details and references to pioneering works of 1950s can be found in [Weickert et al., 1999].

Gaussian scale-space is also frequently referred as diffusion scale space for the reason that the Gaussian density is a fundamental solution to the parabolic PDE which serves as the basic model of diffusive phenomena. However, the diffusion scale-space is a more general family of models. They can be linear, but non-Gaussian, see e.g. [Lindeberg, 1990] and [Duits, 2005], or nonlinear, such as in edge-preserving diffusions.

One could mention that invariance-based derivations of the Gaussian density were already published by J. Herschel (1850) and J. C. Maxwell (1860). A 'backward' derivation of the Gaussian density from a convolution can be attributed to V. D. Landon (1941). References to these early works can be found in [Jaynes, 2003, p. 200–201].

A single concise definition of the scale-space does not exist. Often, the scale-space analysis focuses either on the 'geometry of smoothing', see e.g. [Kuijper, 2002], or the axiomatic formulations of various smoothing methods, clf. [Duits, 2005]. However, the basic problem revolves around the following question, clf. Lindeberg [1990]:

"*How should one create a multiresolution family of representations with the property that a signal at a coarser level of scale never contains more structure than a signal at a finer level of scale?*"

The 'structure' can be thought as 'information content'. An axiomatic derivation of the Gaussian scale-space is interesting as it is based on weak global constraints of symmetry which lead to a strict local smoothness requirement. However, a common statement that Gaussian scale-space is the only smoothing paradigm that does not distort information content (sometimes mistakenly called assumption-less smoothing) is significantly overrated. In particular, a formal advocation of other linear scale-spaces can be found in [Duits, 2005]. Many nonlinear methods, such as the median filter, clf. [Bangham et al., 1996], have proven to be useful in the presence of outliers or discontinuities.

When considering nonlinear models, one starting point is to assume that 'no more structure' is equivalent to: (i) non-amplification of global and local extrema, and (ii) non-creation of local extrema. Both properties should hold as the scale increases. This requirement is sometimes referred as the principle of *causality*, and it could also be associated with the concept of 'irreversibility'. The non-amplification of global extrema is known as the *extremum principle*. In the case of nonlinear diffusions, this property has been established by means of the Gauss–Green theorem independently by Weickert [1996] and Kawohl and Kutev [1998]. The non-amplification and non-creation of local extrema is called the property of

monotonicity. The work of Harten [1983] has shown that monotonicity implies the L_1-norm contraction, and that monotonicity is equivalent to the total variation diminishing.

In this chapter, the phrase 'no more structure' will be understood in the sense of nonincreasing total or sign variation in a finitely discrete space of a single dimension. Before filling certain gaps in this topic, it is first good to introduce the basic model.

One could begin with a 'dynamical law' of the Gaussian density, which is usually demonstrated via Fourier analysis, see e.g. [Kolmogorov and Fomin, 1957]. This requires a careful choice of function spaces and working with the field of complex numbers. Simpler and more formal derivations can be given by using matrix derivatives.

Let us specify a shortcut $\|\mathbf{a}\|_{\mathbf{B}}^2 \equiv \mathbf{a}^T\mathbf{B}\mathbf{a}$ and define the Gaussian density as:

$$p(\mathbf{x}) = |2\pi\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}\|\mathbf{x}-\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}^{-1}}^2\right), \quad \mathbf{x} \in \mathbb{R}^n. \tag{2.1}$$

Furthermore, let us take the derivatives $\partial p(\mathbf{x})/\partial[\boldsymbol{\Sigma}]_{ij}$ and gather the results into a matrix $\partial p(\mathbf{x})/\partial\boldsymbol{\Sigma}$:

$$\frac{\partial p(\mathbf{x})}{\partial \boldsymbol{\Sigma}} = -\frac{1}{2}p(\mathbf{x})|2\pi\boldsymbol{\Sigma}|^{-1}\frac{\partial|2\pi\boldsymbol{\Sigma}|}{\partial\boldsymbol{\Sigma}} - \frac{1}{2}p(\mathbf{x})\frac{\partial}{\partial\boldsymbol{\Sigma}}\left(\|\mathbf{x}-\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}^{-1}}^2\right). \tag{2.2}$$

The derivative of the determinant can be found in [Anderson, 1958, p. 347]:

$$\frac{\partial|2\pi\boldsymbol{\Sigma}|}{\partial\boldsymbol{\Sigma}} = (2\pi)^n|\boldsymbol{\Sigma}|(2\boldsymbol{\Sigma}^{-1} - \mathbf{I}\circ\boldsymbol{\Sigma}^{-1}), \tag{2.3}$$

where $\circ$ denotes the Hadamard (element-wise) product. The derivative of the second term can also be borrowed from [Anderson, 1958, p. 349]:

$$\begin{aligned}\frac{\partial}{\partial\boldsymbol{\Sigma}}\left(\|\mathbf{x}-\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}^{-1}}^2\right) = -2\boldsymbol{\Sigma}^{-T}(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-T} + \\ +\mathbf{I}\circ\boldsymbol{\Sigma}^{-T}(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-T}.\end{aligned} \tag{2.4}$$

Knowing that $\boldsymbol{\Sigma}^{-T} = \boldsymbol{\Sigma}^{-1}$, the gradient of the Gaussian density is given by:

$$\nabla p(\mathbf{x}) = -p(\mathbf{x})\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}). \tag{2.5}$$

The Hesse matrix then becomes:

$$\nabla\nabla^T p(\mathbf{x}) = -\nabla p(\mathbf{x})(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-T} - p(\mathbf{x})\nabla\left((\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-T}\right) \tag{2.6}$$

$$= p(\mathbf{x})\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-T} - p(\mathbf{x})\boldsymbol{\Sigma}^{-T}. \tag{2.7}$$

Verifying Eqs. (2.2)–(2.4) with Eq. (2.7) produces the identity:

$$\frac{\partial p(\mathbf{x})}{\partial\boldsymbol{\Sigma}} = \nabla\nabla^T p(\mathbf{x}) - \frac{\mathbf{I}}{2}\circ\nabla\nabla^T p(\mathbf{x}). \tag{2.8}$$

Following Jaynes [2003], a univariate case of this beautiful result can be attributed to V. D. Landon (1941). Plackett [1954] was apparently the first one to state it in

the matrix form, and Price [1958] generalized it to arbitrary functions and higher orders of the derivatives.

If the elements of the covariance matrix $\mathbf{\Sigma}$ depend on the variable $t \in \mathbb{R}$, then the equivalence $\sum_{i,j=1}^{n} A_{ij}B_{ij} = \mathrm{tr}(\mathbf{A}^T\mathbf{B}) = \mathrm{tr}(\mathbf{B}\mathbf{A}^T)$ yields:

$$\frac{\partial p(\mathbf{x})}{\partial t} = \mathrm{tr}\left(\frac{\partial \mathbf{\Sigma}}{\partial t}\left(\nabla\nabla^T p(\mathbf{x}) - \frac{\mathbf{I}}{2} \circ \nabla\nabla^T p(\mathbf{x})\right)\right). \tag{2.9}$$

A classical diffusion PDE can be obtained by setting $\partial\mathbf{\Sigma}/\partial t = \mathbf{I}$:

$$\frac{\partial p(\mathbf{x})}{\partial t} = \frac{1}{2}\,\mathrm{tr}\;\nabla\nabla^T p(\mathbf{x}) \tag{2.10}$$

$$= \frac{1}{2}\nabla^T\nabla p(\mathbf{x}). \tag{2.11}$$

Let us finally convolve the Gaussian density with any bounded function:

$$g(\mathbf{x}) = \int_{\mathbb{R}^n} f(\boldsymbol{\mu})p(\mathbf{x} - \boldsymbol{\mu})\,\mathbf{d}\boldsymbol{\mu} \tag{2.12}$$

The result is not a Gaussian density, and not even 'density', but it nevertheless makes Eqs. (2.8) and (2.9) valid. The integral does not affect the derivatives w.r.t. the elements of $\mathbf{\Sigma}$ and $\mathbf{x}$.

If the elements of $\mathbf{\Sigma}$ depend on t in such a way that the Gaussian density becomes the delta distribution as $t \to 0$, the convolution at $t = 0$ then yields the function $f(\mathbf{x})$. Therefore, one of the solutions to Eq. (2.9) with an initial condition $p(\mathbf{x}) = f(\mathbf{x})$ is given by Eq. (2.12).

Despite its historical significance, a continuous space and time setting is not the only one to approach diffusion scale-spaces. For example, relevant studies of diffusions on an infinite lattice can be found in [Lindeberg, 1990] and [Ignat, 2005], but there are more such settings. Depending on whether the space or time is discrete or continuous, and whether the domain is finite or infinite, one can obtain at least sixteen different settings.

Thus, limiting oneself to a certain community seems to be inevitable, and the preference will here be given to a finitely discrete setting. Such a limitation results in a loss of understanding of continuity, which, in a narrow sense, can be understood as the stability w.r.t. a 'decreasing grid size' when the elements of the grid become infinitesimal quantities. Many nonlinear diffusions are known which develop jump discontinuities in the initially continuous signals, see e.g. [Perona and Malik, 1987], [Rosenau et al., 1989], [Bertsch and Dal Passo, 1989], Bertsch and Dal Passo [1992] and [Barenblatt et al., 1993]. The use of mathematical techniques related to continuous spaces is essential for thorough understanding of the formation of discontinuities.

However, the problems of stability and statistical inference exist independently of the question 'what happens in an infinite limit'. In a finitely discrete case, matrix theory allows to unify seemingly different notions of stability and view them as consequences of simple postulates. Notably, numerous discrete counterparts of nonlinear diffusions become 'ill-posed' only in an 'infinite limit', see e.g. [Catté et al., 1992].

The focus of this work is on two particular evolutions of a finite-dimensional vector $\mathbf{u}_t \in \mathbb{R}^n$ in a discrete time $t = 0, \tau, \ldots$ with $\tau > 0$ being a known constant.

Definition 2.1 (Discrete diffusion scale-spaces)**.** *Let the matrix $\mathbf{F} \in \mathbb{R}^{n\times(n+1)}$ be a bidiagonal forward-difference operator:*

$$\mathbf{F} = \begin{pmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & \ddots & \\ & & & -1 & 1 \end{pmatrix}. \tag{2.13}$$

Let the matrix $\mathbf{D} \in \mathbb{R}^{(n+1)\times(n+1)}$ be diagonal with the elements $d_{ii} \equiv b_i(\mathbf{u}_t)$:

$$\mathbf{D}(\mathbf{u}) = \begin{pmatrix} b_1(\mathbf{u}) & & & \\ & b_2(\mathbf{u}) & & \\ & & \ddots & \\ & & & b_{n+1}(\mathbf{u}) \end{pmatrix}. \tag{2.14}$$

These elements can be seen as the coordinates of some $\mathbf{b}(\mathbf{u}_t) \in \mathbb{R}^n \mapsto \mathbb{R}^{n+1}$.

Furthermore, let the generalized Laplacian matrix $\mathbf{B}(\mathbf{u})$ be:

$$\mathbf{B}(\mathbf{u}) \equiv \mathbf{F}\mathbf{D}(\mathbf{u})\mathbf{F}^T, \tag{2.15}$$

*The first diffusion scale-space, referred as **semi-implicit**, is given by:*

$$\mathbf{u}_{t+\tau} = \big(\mathbf{I} - \tau\mathbf{B}(\mathbf{u}_t)\big)^{-1}\mathbf{u}_t, \tag{2.16}$$

*where each element $b_{ij}(\mathbf{u}_t) : \mathbb{R}^n \mapsto \mathbb{R}$ of the matrix $\mathbf{B}(\mathbf{u}_t) \in \mathbb{R}^{n\times n}$ is a function of $\mathbf{u}_t$. The second evolution, which could be called **explicit**, is defined as:*

$$\mathbf{u}_{t+\tau} = \big(\mathbf{I} + \tau\mathbf{B}(\mathbf{u}_t)\big)\mathbf{u}_t\,. \tag{2.17}$$

This equation retains the first two terms in the geometric series expansion of the matrix in Eq. (2.16). The convergence holds if the spectral radius $\rho(\mathbf{B}(\mathbf{u}_t)) < 1$. The matrices $\big(\mathbf{I} - \tau\mathbf{B}(\mathbf{u}_t)\big)^{-1}$ and $\mathbf{I} + \tau\mathbf{B}(\mathbf{u}_t)$ will sometimes be referred as single-step diffusion propagators, or simply propagators. The explicit dependence on $\mathbf{u}_t$ will frequently be dropped out in order to reduce the number of parentheses in equations.

Throughout the work, the constraint $b_1 = b_{n+1} = 0$ will be imposed and referred as the von Neumann boundary conditions. The exceptions will be found in Sections 4.3 and 4.4, where numerical simulations will be performed with the Dirichlet conditions $b_1 = -1$ and $b_{n+1} = 1$, and the Robin conditions $b_1 = -1$ and $b_{n+1} = 0$.

It is assumed that the matrix $\mathbf{I} - \tau\mathbf{B}(\mathbf{u}_t)$ is nonsingular, which guarantees the existence of the semi-implicit propagator. In order to put this statement to a more computational level, it is good to emphasize the concept of the matrix minor and two essential facts about. The first one will further quantify the existence of inverses in the discrete space and time, whereas the latter will be subsequently applied to the stability analysis of the diffusion scale-spaces.

Let the determinant $|\mathbf{A}|$ be a function $f(\mathbf{A}) : \mathbb{R}^{n\times n} \mapsto \mathbb{R}$ which is multilinear and skew-symmetric w.r.t. the columns of $\mathbf{A}$, and satisfies the normalization condition $|\mathbf{A}| = 1$ when $\mathbf{A} = \mathbf{I}$ is the identity matrix, see e.g. [Gantmacher, 1959]. Cullen and Gale [1965] prefer the axiom $|\mathbf{AB}| = |\mathbf{A}|\,|\mathbf{B}|$ to a skew-symmetric multilinearity, but their normalization condition becomes more complicated.

Definition 2.2 (Matrix minor)**.** *Let $\boldsymbol{\alpha}$ be a strictly increasing sequence of k integers chosen from $\{1, 2, \ldots, n\}$, e.g.*

$$1 \leq \alpha_1 < \alpha_2 < \ldots < \alpha_k \leq n. \tag{2.18}$$

Following Ando [1987], let $Q_{k,n}$ denote the totality of such sequences. There are $\binom{n}{k}$ of them in $Q_{k,n}$. Given the set $\boldsymbol{\alpha} \in Q_{k,n}$, its complementary set $\boldsymbol{\alpha}^c \equiv \{1, 2, \ldots, n\}\backslash\boldsymbol{\alpha}$. Here and in what follows $\backslash$ denotes a usual set minus, but $\boldsymbol{\alpha}^c$ must be strictly increasing. A particular sequence will be denoted as $\boldsymbol{\alpha}_{1:n} = \{\alpha_1, \ldots, \alpha_n\}$. The shortcuts $\boldsymbol{\gamma}_n \equiv \{1, 2, \ldots, n\}$, $\boldsymbol{\gamma}_{i:j} \equiv \{i, i+1, \ldots, j\}$ $(i < j)$, and $Q_{n,n} \equiv Q_n$ will frequently be employed. The k-th order minor of the matrix $\mathbf{A}$ is the determinant

$$A(\boldsymbol{\alpha}|\boldsymbol{\beta}) \equiv \begin{vmatrix} a_{\alpha_1\beta_1} & a_{\alpha_1\beta_2} & \ldots & a_{\alpha_1\beta_k} \\ a_{\alpha_2\beta_1} & a_{\alpha_2\beta_2} & \ldots & a_{\alpha_2\beta_k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{\alpha_k\beta_1} & a_{\alpha_k\beta_2} & \ldots & a_{\alpha_k\beta_k} \end{vmatrix}, \quad \boldsymbol{\alpha}, \boldsymbol{\beta} \in Q_{k,n}. \tag{2.19}$$

The matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ has $\binom{n}{k}^2$ minors of order k and $\sum_{k=1}^{n}\binom{n}{k}^2 = \binom{2n}{n}$ minors in total. The minors $A(\boldsymbol{\alpha}|\boldsymbol{\alpha})$ are called principal. There are $\binom{n}{k}$ principal minors of order k and $\sum_{k=1}^{n}\binom{n}{k} = 2^n$ in total. The quantities $A(\boldsymbol{\gamma}_k|\boldsymbol{\gamma}_k)$ are called successive (ascending) principal minors. There are only n of them.

The primary importance of matrix minors lies in their ability to characterize: (i) existence and uniqueness of LU-decomposition, i.e. the ability to solve linear systems via Gaussian elimination, and (ii) positivity of quadratic forms.

Proposition 2.1 (Existence of LU-decomposition, see e.g. [Gantmacher, 1959])**.** *A matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ can be represented as the product of a lower triangular matrix and an upper triangular matrix if and only if none of the successive principal minors are equal to zero.*

Thus, the existence of the semi-implicit propagator is established provided none of the successive principal minors of the propagator are equal to zero. One can characterize the elements b_i which ensure that Eqs. (2.16) and (2.17) produce a scale-space in the sense that the number of sign changes in a univariate signal at discrete space locations does not increase w.r.t. an increasing scale. The focus will be given on how the problem reduces to the characterization of positive definiteness when employing known results of the matrix theory related to double stochasticity and total positivity. Here it is useful to remind the reader of the following characterization.

Proposition 2.2 (Criterion of positive definiteness, see e.g. Gantmacher and Krein [2002])**.** *Let $\mathbf{x}^T\mathbf{A}\mathbf{x} = \sum_{ij} a_{ij}x_ix_j$ be called a quadratic form, where a_{ij} is the ij-th element of the matrix $\mathbf{A}$, which will often be also denoted as $[\mathbf{A}]_{ij}$. The matrix is positive definite (semidefinite), i.e. $\mathbf{x}^T\mathbf{A}\mathbf{x} > 0$ (≥ 0) for all $\mathbf{x} \neq \mathbf{0}$, if and only if its successive principal minors are positive (all principal minors are nonnegative).*

This question whether diffusion is a scale-space has been asked before, but no definite answer exist even in a linear case, see e.g. [Hummel and Moniot, 1989]. Partial studies appear numerously in a variety of contexts. Glashoff and Kreth [1980] and Horváth [2000] study restrictions on the time stepping constant τ in linear cases. Harten [1983] discovered the equivalence between the total variation diminishing and monotonicity, and further designed numerical schemes for nonlinear conservation laws. Lindeberg [1990] has applied the results of total positivity to the Laplacian matrices which arise from linear models. Weickert and Benhamouda [1997] and Mrázek et al. [2005] has finally focused on discrete nonlinear cases, but the difference between the necessity and sufficiency is not investigated much.

2.2 Total Variation and Double Stochasticity

Discrete variation diminishing can be seen as a part of the Liapunov stability which deals with the diminishing of other functions such as extrema of signals and negative entropies, see e.g. Weickert and Benhamouda [1997]. Liapunov stability can be established as the consequence of majorization.

Definition 2.3 (Total variation diminishing)**.** *The total variation* $\psi_{\mathrm{tv}}(\mathbf{u}) : \mathbb{R}^n \mapsto \mathbb{R}$ *is*

$$\psi_{\mathrm{tv}}(\mathbf{u}) \equiv \sum_{i=1}^{n} |u_{i+1} - u_i|, \;\; u_{n+1} = u_1 \,. \tag{2.20}$$

The matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ *is said to be total variation diminishing if the inequality* $\psi_{\mathrm{tv}}(\mathbf{Au}) \leq \psi_{\mathrm{tv}}(\mathbf{u})$ *holds for any* $\mathbf{u} \in \mathbb{R}^n$*.*

Discrete total variation diminishing is a particular case of the p-norm diminishing:

$$\|\mathbf{u}\|_p \equiv \left(\sum_{i=1}^{n} |u_i|^p \right)^{1/p} . \tag{2.21}$$

This can be seen by introducing the permutation matrix

$$\mathbf{P} = \begin{pmatrix} 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 1 \\ 1 & 0 & \dots & 0 \end{pmatrix}, \tag{2.22}$$

and rewriting Eq. (2.20) as the 1-norm:

$$\psi_{\mathrm{tv}}(\mathbf{u}) = \|\mathbf{Tu}\|_1, \;\; \mathbf{T} = \mathbf{P} - \mathbf{I} \,. \tag{2.23}$$

In turn, 1-norm diminishing can be seen as a consequence of double stochasticity.

Definition 2.4 (Double stochasticity)**.** *The matrix* $\mathbf{A} \in \mathbb{R}^{n\times n}$ *is row-stochastic (or column-stochastic) if it is nonnegative, i.e.* $a_{ij} \geq 0$ *for each pair of indices* (i, j) *and* $\sum_{j=1}^{n} a_{ij} = 1$ *(or* $\sum_{i=1}^{n} a_{ij} = 1$*). A row and column-stochastic matrix is called doubly-stochastic.*

Proposition 2.3 (Double stochasticity of $\mathbf{I}+\tau\mathbf{B}$)**.** *The matrix* $\mathbf{I}+\tau\mathbf{B}$ *is doubly-stochastic if and only if* $0 \leq b_2 \leq \tau^{-1}$, $0 \leq b_{i-1} + b_i \leq \tau^{-1}$, *for* $i = 3, \ldots, n$ *and* $0 \leq b_n \leq \tau^{-1}$. *It is doubly-stochastic if* $0 \leq b_i \leq (2\tau)^{-1}$ *for all* $i = 2, \ldots, n$.

Proof. Expanding Eq. (2.15) yields:

$$\mathbf{B} = -\begin{pmatrix} b_2 & -b_2 & & & \\ -b_2 & b_2+b_3 & -b_3 & & \\ & \ddots & \ddots & \ddots & \\ & & -b_{n-1} & b_{n-1}+b_n & -b_n \\ & & & -b_n & b_n \end{pmatrix}. \tag{2.24}$$

The proof becomes a trivial application of Definition 2.4. □

The case of sufficiency can be established to Eq. (2.16) by using the property of M-matrix.

Proposition 2.4 (M-matrix definition and its properties, Berman and Plemmons [1979])**.** *Let* $\mathbf{A} \in \mathbb{R}^{n\times n}$ *be nonsingular and Z-matrix, where the latter states that the off-diagonal elements are nonpositive. The following statements are equivalent:*

1. $\mathbf{A}$ *is an M-matrix.*
2. $\mathbf{A}$ *is positive definite, i.e.* $\mathbf{u}^T\mathbf{A}\mathbf{u} > 0$ *for any* $\mathbf{u} \neq \mathbf{0}$.
3. $\mathbf{A}^{-1}$ *is nonnegative, i.e. each element is nonnegative.*
4. $\mathbf{A}$ *has all positive diagonal elements, and there exists a positive diagonal* $\mathbf{G} \in \mathbb{R}^{n\times n}$ *such that* $\mathbf{AG}$ *is strictly diagonally-dominant, i.e.* $a_{ii}g_i > \sum_{j\neq i} |a_{ij}|g_j$ *for* $i = 1, \ldots, n$.
5. $\mathbf{A}$ *is sign-preserving, i.e.* $[\mathbf{A}\mathbf{u}]_i[\mathbf{u}]_i > 0$ *for* $[\mathbf{u}]_i \neq 0$.

Proposition 2.5 (Double stochasticity of $(\mathbf{I}-\tau\mathbf{B})^{-1}$)**.** *The matrix* $(\mathbf{I}-\tau\mathbf{B})^{-1}$ *is doubly-stochastic if* $b_i \geq 0$ *for* $i = 2, \ldots, n$.

Proof. Firstly, we borrow the argument in [Marshall and Olkin, 1979, p. 48]. For the vector $\mathbf{1} \in \mathbb{R}^n$ whose all elements are equal to unity, $(\mathbf{I}-\tau\mathbf{B})\mathbf{1} = \mathbf{1}$. This implies that if $(\mathbf{I}-\tau\mathbf{B})$ is nonsingular, $(\mathbf{I}-\tau\mathbf{B})^{-1}\mathbf{1} = \mathbf{1}$. Therefore, each row of $(\mathbf{I}-\tau\mathbf{B})^{-1}$ sums to unity. *Secondly*, we resort to the fundamental result of Gantmacher and Krein [2002], which states that $(\mathbf{I}-\tau\mathbf{B})^{-1}$ is Green's matrix which is symmetric when $\mathbf{B}$ is symmetric. Thus, each row of $(\mathbf{I}-\tau\mathbf{B})^{-1}$ equals to its column, and therefore, each column sums to unity. *Thirdly*, the nonnegativity $(\mathbf{I}-\tau\mathbf{B})^{-1} \geq 0$ can be established by requiring that $\mathbf{I}-\tau\mathbf{B}$ should be an M-matrix. Clearly, the nonnegativity $b_i \geq 0$ for $i = 2, \ldots, n$ enforces the Z-matrix property. *The fact that this nonnegativity is also sufficient to guarantee the positive definiteness of* $\mathbf{I}-\tau\mathbf{B}$ *will be proven in Proposition 2.13.* □

The necessity remains unclear. At least one element of b_i, $(i = 2, \ldots, n)$ should be positive. For if all the elements are nonpositive, then $\mathbf{I}-\tau\mathbf{B}$ is doubly-stochastic,

and each row and column of $(\mathbf{I} - \tau\mathbf{B})^{-1}$ sums to unity. However, except the trivial case when $\mathbf{B} = \mathbf{0}$, the matrix $(\mathbf{I} - \tau\mathbf{B})^{-1}$ cannot be nonnegative because only permutation matrices and their inverses can be doubly-stochastic together, clf. Marshall and Olkin [1979].

Definition 2.5 (Majorization, Marshall and Olkin [1979])**.** *Let*

$$x_{[1]} \geq x_{[2]} \geq \ldots \geq x_{[n]} \tag{2.25}$$

denote the components of $\mathbf{x} \in \mathbb{R}^n$ *arranged in a nonincreasing order. If for* $k = 1, 2, \ldots, n-1$*,*

$$\sum_{i=1}^{k} x_{[i]} \leq \sum_{i=1}^{k} y_{[i]}, \quad \text{and} \quad \sum_{i=1}^{n} x_{[i]} = \sum_{i=1}^{n} y_{[i]}, \tag{2.26}$$

the vector $\mathbf{x}$ *is said to be majorized by* $\mathbf{y}$*, which is denoted as* $\mathbf{x} \prec \mathbf{y}$*.*

Proposition 2.6 (Double stochasticity and majorization, Marshall and Olkin [1979])**.** $\mathbf{A} \in \mathbb{R}^{n \times n}$ *is doubly stochastic if and only if* $\mathbf{Ax} \prec \mathbf{x}$ *for all* $\mathbf{x} \in \mathbb{R}^n$*.*

The proof with pointers to generalizations can be found in [Marshall and Olkin, 1979, p. 20]. It becomes clear that successive multiplications by a doubly stochastic matrix, which can be seen as a nonlinear evolution in a discrete time, do not increase all the cumulative sums of nonincreasing rearrangement of any vector $\mathbf{x} \in \mathbb{R}^n$. This is a powerful result with many consequences.

Definition 2.6 (Schur-convex function, Marshall and Olkin [1979])**.** *A function* $\psi : \mathbb{R}^n \mapsto \mathbb{R}$ *is called Schur-convex (isotone) if* $\mathbf{x} \prec \mathbf{y}$ *on* $\mathbb{R}^n \Rightarrow \psi(\mathbf{x}) \leq \psi(\mathbf{y})$*.*

The necessary and sufficient conditions of Schur-convexity, its preservation under myriads of transformations and important particular cases are summarized by Marshall and Olkin [1979]. In our case, it is vital to realize that any Schur-convex function automatically becomes the Liapunov function in a discrete time, and several specific examples deserve a brief discussion.

Proposition 2.7 (Examples of Schur-convex functions, Marshall and Olkin [1979])**.** *The following functions are Schur-convex:*

1. $f(\mathbf{u}) = \left(\sum_{i=1}^{n} |u_i|^p\right)^{1/p}$, $\mathbf{u} \in \mathbb{R}^n$, $p \geq 1$ *including* $p \to \infty$*.*
2. $f(\mathbf{u}) = \max_{1 \leq i_1 < \ldots < i_k \leq n} \left(|u_{i_1}| + \ldots + |u_{i_k}|\right)$, $\mathbf{u} \in \mathbb{R}^n$*.*
3. $f(\mathbf{u}) = (\alpha - 1)^{-1} \ln\left(\sum_{i=1}^{n} u_i^{\alpha}\right)$, $\mathbf{u} \in \mathbb{R}^+$, $\alpha \geq 0$*.*

The p-norm is a particular case of the Schur-convex function known as a symmetric gauge function. The special case $p \to \infty$ establishes the extremum principle: Global extrema do not amplify. This result has been established in [Weickert and Benhamouda, 1997].

The second property states the diminishing of the mixed $p = \infty/1$ norm, which can be seen as one generalization of the extremum principle in a discrete space and time. The third case shows that negative Rényi's entropies do not increase in a

discrete time. A special case $\alpha = 1$ corresponds to the Shannon entropy and can be seen as a mathematical caricature of the second law of thermodynamics: entropy increases in time. Sporring and Weickert [1999] speculate that Rényi's entropies can be useful in optimal scale selection when models are conditioned on data. This question is discussed in Chapter 3, where a more advantageous utilization of entropies is suggested.

In summary, in the case of a discrete diffusion with the propagator $(\mathbf{I} - \tau\mathbf{B})^{-1}$, the positivity of the elements $b_i \geq 0$, for $i = 2, \ldots, n$, guarantees diminishing of p-norms, total variation and negative Rényi's entropies. In essence, many instances of the Liapunov stability take place simultaneously.

2.3 Sign Variation and Total Positivity

Sign variation turns out to be related to the total variation, but it can also be approached via the concept of the total positivity. This is advantageous as both sufficient and necessary conditions can be established.

Definition 2.7 (Sign variation diminishing)**.** *Let a sign variation* $\psi_s(\mathbf{u}) : \mathbb{R}^n \mapsto \mathbb{R}$ *count the number of sign changes along the elements of* $\mathbf{u} \in \mathbb{R}^n$ *in the order* $i = 1, 2, \ldots, n$*:*

$$\psi_s(\mathbf{u}) = \frac{1}{2}\sum_{i=1}^{n} |\operatorname{sgn}(u_{i+1}) - \operatorname{sgn}(u_i)|, \quad u_{n+1} = u_n. \tag{2.27}$$

It is assumed that zero elements are discarded before employing Eq. (2.27)*. The matrix* $\mathbf{A} \in \mathbb{R}^{n\times n}$ *is said to be sign variation diminishing if the inequality* $\psi_s(\mathbf{Au}) \leq \psi_s(\mathbf{u})$ *holds for any* $\mathbf{u} \in \mathbb{R}^n$*.*

Definition 2.8 (Sign regularity and total positivity, Karlin [1968])**.** *A matrix* $\mathbf{A} \in \mathbb{R}^{m\times n}$ *is called sign-regular of order* r *and abbreviated as* SR_r *(strictly sign regular,* SSR_r*) if there exists a sequence of numbers* ϵ_p *each either* $+1$ *or* -1 *such that for* $p = 1, 2, \ldots, r \leq \min(m, n)$*,* $\epsilon_p A(\boldsymbol{\alpha}|\boldsymbol{\beta}) \geq 0$*,* (> 0)*,* $\boldsymbol{\alpha}, \boldsymbol{\beta} \in Q_{p,n}$*. If all* $\epsilon_p \geq 0$ (> 0)*, then the matrix is said to be totally-positive (strictly totally positive), which is denoted as* TP_r *(*STP_r*).*

Sign regularity is known to be a necessary and sufficient condition for a matrix to be sign variation diminishing. This deserves a brief historical appraisal. Pinkus [1996] explains how a vague connection between the signature of the matrix minors and sign variation diminishing has already been foreseen by Kellogg [1918]. The first systematic investigation of totally positive matrices can be found in [Gantmacher and Krein, 2002] which is a thoroughly revised edition of the famous work. The pioneering formulation of sign variation diminishing is often attributed to the ideas of I. J. Schoenberg, T. Motzkin and A. Whitney, see [Hirschman and Widder, 1955, p. 97–103] for the references and a revised proof which is lengthy as it is based on the elementary operations on matrix columns (rows). A more direct proof which operates directly with minors was stated by Karlin [1968]. A complete departure from the continuous space has been finally emphasized by Ando [1987] and several new results related to discrete cases can also be found in [Gantmacher and Krein,

2002]. There exist a third proof based on the bidiagonal decompositions, albeit in a special case of square totally positive matrices. This is a suggestion of Fallat [2001] realized rigorously by Dopico and Koev [2007].

The following statement is more restrictive, but also more relevant to the context of this thesis. It is borrowed from the work of Gantmacher and Krein [2002] which has also established the theory of tridiagonal (Jacobi) matrices by stating, in particular, their explicit inverses and Sturmian theory of eigenvalues.

Proposition 2.8 (Sign variation diminishing, Gantmacher and Krein [2002])**.** *For* $\mathbf{A} \in \mathbb{R}^{n\times n}$ *and any* $\mathbf{u} \in \mathbb{R}^n$*, the sign variation diminishing holds, i.e.* $\psi_s(\mathbf{Au}) \leq \psi_s(\mathbf{u})$*, and the vectors* $\mathbf{u}$ *and* $\mathbf{Au}$ *oscillate in the same way in the case of equality, if and only if* $\mathbf{A} \in \mathrm{TP}_n$*.*

It is generally impossible to check the total positivity directly. There are simply too many minors to consider. However, a tridiagonal M-matrix property reduces the problem to evaluating only n minors! The following two propositions reduce the problem of total positivity to the requirement of positive definiteness.

Proposition 2.9 (Total positivity of a tridiagonal matrix, Gantmacher and Krein [2002])**.** *Let* $\mathbf{A} \in \mathbb{R}^{n\times n}$ *be nonsingular and tridiagonal. The matrix is totally-positive, i.e.* $\mathbf{A} \in \mathrm{TP}_n$ *if and only if its off-diagonal elements are nonnegative and the successive principal minors are positive.*

Proposition 2.10 (Totally positive, tridiagonal and M-matrices, Markham [1972], Lewin [1980] and Peña [1995])**.** *Let* $\mathbf{A} \in \mathbb{R}^{n\times n}$ *be nonsingular. Any two statements listed below imply the third one:*

- $\mathbf{A}$ *is an M-matrix.*
- $\mathbf{A}^{-1} \in \mathrm{TP}_n$*.*
- $\mathbf{A}$ *is tridiagonal.*

Proposition 2.10 indicates a remarkable property of a nonsingular tridiagonal M-matrix: Not only its inverse is positive, but it is also totally positive, and, moreover, given a tridiagonal matrix, the M-matrix property is both, necessary and sufficient to guarantee inverse total positivity.

It deserves to be mentioned that the sign and total variation were connected by Karlin [1968] via the 'Banach indicatrix':

$$\psi_{\mathrm{tv}}(\mathbf{u}) = \int_{\mathbb{R}} \psi_s(\mathbf{u} - \rho\mathbf{1}) d\rho, \tag{2.28}$$

where $\mathbf{1}$ is a vector of unity elements. Given that the matrix $\mathbf{A}$ is sign variation diminishing, it will also be total variation diminishing if, in addition, each row sums to unity. This can be seen from $\mathbf{Au} - \rho\mathbf{1} = \mathbf{A}(\mathbf{u} - \rho\mathbf{1})$ and diminishing of the positive integrand in Eq. (2.28).

2.4 Positive Definiteness

What remains to be shown are the conditions when the matrices $\mathbf{I}+\tau\mathbf{B}$ and $(\mathbf{I}-\tau\mathbf{B})^{-1}$ are positive definite. As the eigenvalues are difficult to obtain, the easiest way to characterize positive definiteness seems to be the equivalent requirement of the positivity of successive principal minors.

The determinant of a positive definite tridiagonal matrix can be computed via the LU-decomposition in $O(n)$ multiplications, but that does not provide an easy access to the signature of the minors, it only gives a verification procedure for a particular matrix. However, analytical computations can be performed by utilizing the determinant of Rose [1969]:

$$\begin{vmatrix} b_1+b_2 & -b_2 & & & \\ -b_2 & b_2+b_3 & -b_3 & & \\ & \ddots & \ddots & \ddots & \\ & & -b_n & b_n+b_{n+1} \end{vmatrix} = \sum_{i=1}^{n+1} \prod_{\substack{j=1 \\ j\neq i}}^{n+1} b_j \,. \tag{2.29}$$

The easiest way to arrive at this expression is to employ a well-known two-term recurrence for the principal minors of the tridiagonal matrix, see e.g. [Kawasaki, 2001]:

$$A(\boldsymbol{\gamma}_k|\boldsymbol{\gamma}_k) = a_{ii}A(\boldsymbol{\gamma}_{k-1}|\boldsymbol{\gamma}_{k-1}) - a_{i-1,i}a_{i,i-1}A(\boldsymbol{\gamma}_{k-2}|\boldsymbol{\gamma}_{k-2}), \tag{2.30}$$

where $A(\boldsymbol{\gamma}_{-1}|\boldsymbol{\gamma}_{-1}) = 0$ and $A(\boldsymbol{\gamma}_0|\boldsymbol{\gamma}_0) = 1$.

Explicit computation of the determinant for $k = 1, 2, 3$ reveals the pattern of the expression in Eq. (2.29). The original derivation of Rose [1969] applies the Cauchy–Binet theorem to Eq. (2.15). Notice that the determinant equals to zero whenever any two elements b_i and b_j $(i \neq j)$ equal to zero. Thus, it should be clear that Laplacian will be a singular matrix if the von Neumann boundary conditions $b_1 = b_{n+1} = 0$ are imposed.

Here we can contribute a geometric insight to Eq. (2.29), which is shown in Figs. 2.1a,b and stated below.

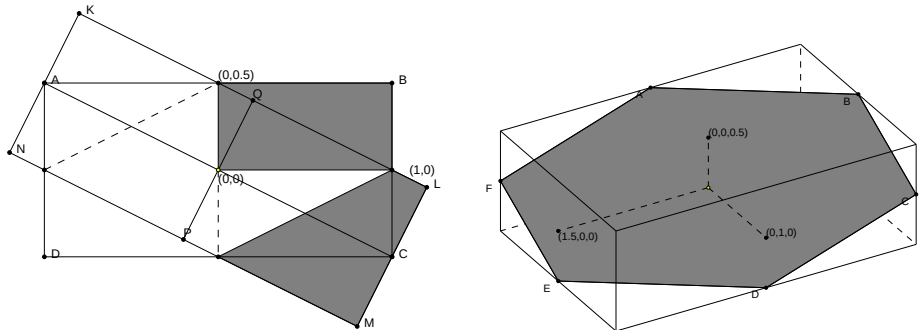
Proposition 2.11 (Geometry of Rose's determinant)**.** *Rose's determinant is a squared volume of an n-dimensional parallelepiped defined as an intersection:*

$$\mathcal{P}_n \equiv \begin{cases} b_k^{-1/2} x_k \in [-\tfrac{1}{2}, \tfrac{1}{2}], & k = 1, \ldots, n+1, \\ \sum_{k=1}^{n+1} b_k^{-1/2} x_k = 0. \end{cases} \tag{2.31}$$

In particular, when $n = 1$, Rose's determinant is a squared length of the edge, and it becomes the squared area of the convex hexagon when $n = 2$.

Proof. It follows from Fig. 2.1a that Vol(ABCD) = Vol(KLMN), where Vol(ABCD) here means the length of the edge, the area of the polygon, or generally 'hypervolume'. Clearly,

$$\mathrm{Vol(AC)} = \frac{\mathrm{Vol(ABCD)}}{\mathrm{Vol(PQ)}} \,. \tag{2.32}$$



(a) (b)

Figure 2.1: Geometric meaning of Rose's determinant: (a) the determinant is a squared length of the edge AC when $n = 1$, which equals to 5 when $b_1 = 1$ and $b_2 = 4$, and (b) the determinant becomes the squared area of the convex hexagon when $n = 2$, which is 49 when $b_1 = 1$, $b_2 = 4$, and $b_3 = 9$.

The square of the quantity Vol(ABCD) is Grammian, clf. e.g. Gantmacher [1959]:

$$\mathrm{Vol(ABCD)}^2 = \begin{vmatrix} \mathbf{v}_1^T\mathbf{v}_1 & \mathbf{v}_1^T\mathbf{v}_2 & \dots & \mathbf{v}_1^T\mathbf{v}_{n+1} \\ \mathbf{v}_2^T\mathbf{v}_1 & \mathbf{v}_2^T\mathbf{v}_2 & \dots & \mathbf{v}_2^T\mathbf{v}_{n+1} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{v}_{n+1}^T\mathbf{v}_1 & \mathbf{v}_{n+1}^T\mathbf{v}_2 & \dots & \mathbf{v}_{n+1}^T\mathbf{v}_{n+1} \end{vmatrix}, \tag{2.33}$$

where $\mathbf{v}_k = (0, 0, \dots, b_k^{1/2}, \dots, 0)^T \in \mathbb{R}^n$. Therefore,

$$\mathrm{Vol(ABCD)}^2 = \prod_{k=1}^{n+1} b_k \,. \tag{2.34}$$

The quantity Vol(PQ) can be evaluated by means of the identity:

$$d = \frac{|\mathbf{w}^T\mathbf{x}_0 + c|}{\|\mathbf{w}\|_2} \,, \tag{2.35}$$

where d is the shortest distance between the point $\mathbf{x}_0 \in \mathbb{R}^n$ to the hyperplane $\mathbf{w}^T\mathbf{x} + c = 0$. In our case, the space is $\mathbb{R}^{n+1}$, the hyperplane is given by the second line of Eq. (2.31) and from the case with $n = 2$ one infers that the distance between the points P and Q equals twice the shortest distance between the point $(b_1^{1/2}/2, 0)$ and the hyperplane. In general, $(b_1^{1/2}/2, 0)$ will be any point whose position is determined by either of the vectors $v_k/2$:

$$\mathrm{Vol(PQ)}^2 = \frac{\left(2(0 + 0 + \ldots + b_k^{-1/2} b_k^{1/2}/2 + 0 + \ldots + 0)\right)^2}{\sum_{k=1}^{n+1} (b_k^{-1/2})^2} = \frac{1}{\sum_{k=1}^{n+1} b_k^{-1}} \,. \tag{2.36}$$

Substituting Eqs. (2.34) and (2.36) to Eq. (2.32) produces Rose's determinant:

$$\mathrm{Vol(AC)}^2 = \left(\sum_{k=1}^{n+1} b_k^{-1}\right) \prod_{k=1}^{n+1} b_k = \sum_{i=1}^{n+1} \prod_{\substack{j=1 \\ j\neq i}}^{n+1} b_j \,. \tag{2.37}$$

□

Proposition 2.12 (Determinant of the matrix $\mathbf{I}-\tau\mathbf{B}$)**.**

$$|\mathbf{I}-\tau\mathbf{B}| = 1+\sum_{k=1}^{n}\tau^k \sum_{\boldsymbol{\omega}\in Q_{k,n}} \prod_{s=1}^{r} \sum_{i=\nu_{s-1}}^{\nu_s+1} \prod_{\substack{j=\nu_s-1\\ j\neq i}}^{\nu_s+1} b_j\,, \tag{2.38}$$

with $\nu_0=1$ and $\nu_r=k$. The integer s denotes the s-th group of the connected indices in $\boldsymbol{\omega}\in Q_{k,n}$.

Proof. For any matrix $\mathbf{A}\in\mathbb{C}^{n\times n}$,

$$|\mathbf{I}+\mathbf{A}| = 1+\sum_{k=1}^{n}\sum_{\boldsymbol{\omega}\in Q_{k,n}} A(\boldsymbol{\omega}|\boldsymbol{\omega})\,, \tag{2.39}$$

clf. Gantmacher and Krein [2002]. This demands explicit expressions for the minors $B(\boldsymbol{\omega}|\boldsymbol{\omega})$. The way to derive them is best seen by example, say the case of $\boldsymbol{\omega}\in Q_{5,20}$, and a particular sequence from this set such as $\boldsymbol{\omega}=\{3,5,6,8,9\}$. The corresponding minor is:

$$B(\boldsymbol{\omega}|\boldsymbol{\omega}) = \begin{pmatrix} b_3+b_4 & 0 & 0 & 0 & 0\\ 0 & b_5+b_6 & -b_6 & 0 & 0\\ 0 & -b_6 & b_6+b_7 & 0 & 0\\ 0 & 0 & 0 & b_8+b_9 & -b_9\\ 0 & 0 & 0 & -b_9 & b_9+b_{10} \end{pmatrix}. \tag{2.40}$$

The connected indices form 'isles' which are particular cases of Rose's determinant. The determinant of a block-diagonal matrix splits into a product of the block-determinants. In general, one must divide any sequence $\boldsymbol{\omega}\in Q_{k,n}$ into r groups of connected indices:

$$\boldsymbol{\omega} = \{\underbrace{\omega_1,\ldots,\omega_{\nu_1}}_{\text{1st group}},\ \underbrace{\omega_{\nu_1+1},\ldots,\omega_{\nu_2}}_{\text{2nd group}},\ \ldots,\underbrace{\omega_{\nu_{r-1}+1},\ldots,\omega_k}_{r\text{-th group}}\}\,. \tag{2.41}$$

Each group has at least one integer, otherwise one does not form a group. The result then follows immediately by applying Eq. (2.29). □

Proposition 2.12 can now be directly applied not only to characterize variation diminishing of the matrix-vector products with $(\mathbf{I}-\tau\mathbf{B})^{-1}$ and $\mathbf{I}+\tau\mathbf{B}$, but also to complete the proof of Proposition 2.5.

Proposition 2.13 (Positive definiteness of the matrices $(\mathbf{I}-\tau\mathbf{B})^{-1}$ and $\mathbf{I}+\tau\mathbf{B}$)**.** *The matrix $(\mathbf{I}-\tau\mathbf{B})^{-1}$ is positive definite if and only if*

$$1+\sum_{k=1}^{p}\tau^k \sum_{\boldsymbol{\omega}\in Q_{k,p}} \prod_{s=1}^{r} \sum_{i=\nu_{s-1}}^{\nu_s+1} \prod_{\substack{j=\nu_s-1\\ j\neq i}}^{\nu_s+1} b_j > 0, \tag{2.42}$$

for all $p=1,\ldots,n$, where the notations were explained previously in the proof of Proposition 2.12. The matrix $\mathbf{I}+\tau\mathbf{B}$ is positive definite if and only if

$$1+\sum_{k=1}^{p}\tau^k \sum_{\boldsymbol{\omega}\in Q_{k,p}} \prod_{s=1}^{r} \sum_{i=\nu_{s-1}}^{\nu_s+1} \prod_{\substack{j=\nu_s-1\\ j\neq i}}^{\nu_s+1} (-b_j) > 0, \tag{2.43}$$

for all $p=1,\ldots,n$.

Proof. A direct application of Proposition 2.12. □

Notice that the first statement is satisfied if $b_i > 0$ for $i = 2, \ldots, n$, which completes the proof of Proposition 2.5. The second statement is satisfied if $0 \leq b_2 \leq \tau^{-1}$, $0 \leq b_{i-1} + b_i \leq \tau^{-1}$, for $i = 3, \ldots, n$ and $0 \leq b_n \leq \tau^{-1}$. This follows from the fourth property of Proposition 2.4 where one may set $\mathbf{G} = \mathbf{I}$. This has been proven already in Bourgin [1939].

Proposition 2.13 reveals an interesting paradox. The positivity $b_i > 0$ for $i = 2, \ldots, n$ is sufficient to guarantee sign variation diminishing of the map $(\mathbf{I} - \tau\mathbf{B})^{-1}$. Naturally, the same is expected to hold for the map $\mathbf{I} + \tau\mathbf{B}$ when τ is small because it is close to $(\mathbf{I} - \tau\mathbf{B})^{-1}$ in the sense of any matrix norm. However, the map $\mathbf{I} + \tau\mathbf{B}$ is sign variation diminishing only if a nontrivial constraint, given by Eq. (2.43), is satisfied.

As a simple example, consider the case $n = 1$ with the Laplacian

$$\mathbf{B} = \begin{pmatrix} -b_2 & b_2 \\ b_2 & -b_2 \end{pmatrix} . \tag{2.44}$$

Given a fixed $\tau > 0$, a positive definite $(\mathbf{I} - \tau\mathbf{B})^{-1}$ requires that $b_2 \geq 0$. However, the matrix $\mathbf{I} + \tau\mathbf{B}$ is positive definite if and only if $0 \leq b_2 < 0.5\tau^{-1}$. Notice also that $\mathbf{I} + \tau\mathbf{B}$ is doubly-stochastic if and only if $0 \leq b_2 < \tau^{-1}$, which confirms the conclusion in Glashoff and Kreth [1980] that sign stability is more stringent than a more conventional Liapunov stability.

2.5 Multivariate Diffusion

Characterizing variation diminishing in a multivariate setting is a much more difficult problem, but at least it can be posed explicitly. Consider a d-dimensional array $\mathbf{U}$ with n_m points in the m-th dimension of the array and $n = n_1 \cdots n_d$ number of points in total. Let $\mathrm{vec}_m(\mathbf{U}_t) \in \mathbb{R}^n$ for $m = 1, \ldots, d$ be a vector which is obtained by dividing the array into lexicographically arranged vectors along the m-th dimension and stacking them together into a single-column vector. For example, when $d = 2$, the column vectors $\mathrm{vec}_1(\mathbf{U})$ and $\mathrm{vec}_2(\mathbf{U})$ are the conventional $\mathrm{vec}(\mathbf{U})$ and $\mathrm{vec}(\mathbf{U}^T)$ operators, respectively.

In order to write the Laplacian matrix in a multivariate case, it is good to introduce the permutation matrices $\mathbf{P}_{ij} \in \mathbb{R}^{n \times n}$ for $i, j = 1, \ldots, d$ such that $\mathbf{P}_{ij}\,\mathrm{vec}_j(\mathbf{U}) = \mathrm{vec}_i(\mathbf{U})$, and $\mathbf{P}_{ii} = \mathbf{I}$ with $\mathbf{I}$ being the identity matrix. Let us choose any integer r out of $1, 2, \ldots, d$. Given the vector $\mathrm{vec}_r(\mathbf{U})$, the Laplacian in the coordinate space of this vector becomes:

$$\mathbf{B} = \sum_{m=1}^{d} \mathbf{P}_{rm}\mathbf{B}_m\mathbf{P}_{mr} . \tag{2.45}$$

The matrix $\mathbf{B}_m$ is block-diagonal in the basis $\mathrm{vec}_m(\mathbf{U})$. However, the sum does no longer possess this property: Nontrivial sparsity patterns emerge.

It is clear that characterizing matrices $(\mathbf{I}-\tau\mathbf{B})^{-1}$ and $\mathbf{I}+\tau\mathbf{B}$) becomes now considerably more difficult. An interesting trick in avoiding a non-tridiagonal sparsity is to apply the additive operator splitting, see e.g. [Fischer and Modersitzki, 2002], which deserves some insights. The idea is to apply the first order expansion in the von Neumann series:

$$
\begin{aligned}
(\mathbf{I}-\tau\mathbf{B})^{-1}\,\mathrm{vec}_r(\mathbf{U}) &\approx (\mathbf{I}+\tau\mathbf{B})\,\mathrm{vec}_r(\mathbf{U}) && (2.46)\\
&= \left(\mathbf{I}+\tau\sum_{m=1}^{d}\mathbf{P}_{rm}\mathbf{B}_m\mathbf{P}_{mr}\right)\mathrm{vec}_r(\mathbf{U}) && (2.47)\\
&= \frac{1}{d}\sum_{m=1}^{d}(\mathbf{I}+d\,\tau\mathbf{P}_{rm}\mathbf{B}_m\mathbf{P}_{mr})\,\mathrm{vec}_r(\mathbf{U}) && (2.48)\\
&\approx \frac{1}{d}\sum_{m=1}^{d}(\mathbf{I}-d\,\tau\mathbf{P}_{rm}\mathbf{B}_m\mathbf{P}_{mr})^{-1}\,\mathrm{vec}_r(\mathbf{U}) && (2.49)\\
&= \frac{1}{d}\sum_{m=1}^{d}\mathbf{P}_{rm}\,(\mathbf{I}-d\,\tau\mathbf{B}_m)^{-1}\,\mathrm{vec}_m(\mathbf{U}). && (2.50)
\end{aligned}
$$

The approximation in Eq. (2.49) is rather unusual, but it stabilizes the expression w.r.t. $\tau\to\infty$. The last equality is the most notable as it reduces the overall matrix-vector product to the products with tridiagonal inverses! Another way of thinking is to first write Eq. (2.46) for each $r=1,2,\ldots,d$. After obtaining the redundant set of equations, the trick is then to replace the m-th term $\mathbf{P}_{rm}\mathbf{A}_m\mathbf{P}_{mr}$ with $\mathbf{A}_r$. The final approximation, given by Eq. (2.50), is obtained by summing d such solutions re-permuted back to any fixed basis $\mathrm{vec}_r(\mathbf{U})$.

Here one may see that the univariate analysis applies up to each term

$$(\mathbf{I}-d\,\tau\mathbf{B}_m)^{-1}\,\mathrm{vec}_m(\mathbf{U}) \tag{2.51}$$

in Eq. (2.50). Univariate conditions simply guarantee that every such a vector for $m=1,\ldots,d$ does not increase the number of sign changes present in $\mathrm{vec}_m(\mathbf{U})$. The results about extrema and entropy functions apply as well. However, the subsequent sum and permutation does no longer preserve the stability properties of each component. The trick with the additive operator splitting is stated here because establishing the total positivity and other relevant properties of the sum of the matrices in Eq. (2.49) could be easier than working with the matrix $(\mathbf{I}-\tau\mathbf{B})^{-1}$ directly. It probably deserves to be mentioned that additive splitting is little known. Fischer and Modersitzki [2002] state it in a general multivariate case, but provide verifications only in a bivariate case.

This chapter can be concluded by providing a certain parallel with the analysis in the continuous space and time. The interest lies in the possibility to show that sampling the constraint for the extremum principle in a continuous space and time may yield a different requirement compared to the previously discussed ones. Precision with functional spaces is beyond the scope here. We do not solve any nonlinear equations in a continuous space and time, and there is no equivalence between the models in different modalities of the space and time.

Consider the following PDE:

$$\frac{\partial u(\mathbf{x},t)}{\partial t} = \nabla \cdot \mathbf{v}\left(\nabla u(\mathbf{x},t)\right) , \tag{2.52}$$

where $u(\mathbf{x},t)$ is defined on the domain $\Omega \subset \mathbb{R}^n \times \mathbb{R}_+$. In what follows, all the necessary quantities are assumed to be sufficiently measurable and continuous enough to validate the equations (this is not the key point here).

Proposition 2.14 (Extremum principle, Weickert [1996], Kawohl and Kutev [1998]). *Let $\|u(t)\|_{L_p}$ denote the $L^p(\Omega)$ norm. For any $p \in [1,\infty)$, and $t > 0$,*

$$\|u(t)\|_{L_p} \leq \|u(0)\|_{L_p} \quad \text{if} \quad \nabla u(\mathbf{x},t) \cdot \mathbf{v}(\nabla u(\mathbf{x},t)) > 0. \tag{2.53}$$

Proof. We restate the proof of Kawohl and Kutev [1998] because it makes rather transparent the fact that the proof does not depend whether the flux $\mathbf{v}(\nabla u(\mathbf{x},t))$ is linear or nonlinear. In essence, this is, *mutatis mutandis*, derivation of Beckenbach and Bellman [1965, p. 149] and dates back to the work of Ladyzhenskaja et al. [1968, p. 429]. Paradoxically, extensions of the same technique to a more complicated problem of 'gradient estimates' appear earlier in [Kawohl, 1988].

One relies on the Gauss–Green theorem:

$$\int_\Omega \nabla \cdot \mathbf{g}\, \mathbf{dx} = \int_{\partial\Omega} \mathbf{g} \cdot \mathbf{n}\, \mathbf{dx}\,, \tag{2.54}$$

where $\mathbf{g} : \Omega \mapsto \mathbb{R}^d$, $\mathbf{n}$ is outward normal to the boundary $\partial\Omega$. Here and in what follows the obvious dependence on the arguments $\mathbf{x}$ and t is suppressed for the sake of brevity. The substitution of $\mathbf{g} = f\,\mathbf{v}(\nabla u)$ into Eq. (2.54) provides an auxiliary result:

$$\int_\Omega f \nabla \cdot \mathbf{v}(\nabla u)\, \mathbf{dx} = \int_{\partial\Omega} f \mathbf{v}(\nabla u) \cdot \mathbf{n}\, \mathbf{dx} - \int_\Omega \nabla f \cdot \mathbf{v}(\nabla u)\, \mathbf{dx}. \tag{2.55}$$

The proof is first reduced to establishing a nondecreasing p-th power of the L_p-norm:

$$\frac{\partial}{\partial t}\|u\|_{L_p} = \frac{1}{p}\|u\|_{L_p}^{1-p} \frac{\partial}{\partial t} \int_\Omega |u|^p\, \mathbf{dx}\,. \tag{2.56}$$

A further use of Eq. (2.55) with $f = |u|^{p-2}u$ produces the following result:

$$\begin{aligned}
\frac{\partial}{\partial t} \int_\Omega |u|^p \mathbf{dx} &= p \int_\Omega |u|^{p-1} \operatorname{sgn}(u) \frac{\partial u}{\partial t}\, \mathbf{dx}\,, && (2.57)\\
&= p \int_\Omega |u|^{p-2} u\, \nabla \cdot \mathbf{v}(\nabla u)\, \mathbf{dx} && (2.58)\\
&= -p \int_\Omega \nabla \left(|u|^{p-2}u\right) \cdot \mathbf{v}(\nabla u)\, \mathbf{dx} && (2.59)\\
&= -p(p-1) \int_\Omega |u|^{p-2} \nabla u \cdot \mathbf{v}(\nabla u)\, \mathbf{dx}\,. && (2.60)
\end{aligned}$$

Substituting Eq. (2.60) into Eq. (2.56) establishes the nonincrease of the L_p norm when the 'if' condition in Eq. (2.53) is satisfied. As $p \to \infty$, the L_p-norm becomes the supremum norm which reveals the maximum value of $|u(x,t)|$, i.e. the global extremum. □

Assume now a particular form of the flux:

$$v(\nabla u(\mathbf{x},t)) = \varphi(\|\nabla u(\mathbf{x},t)\|)\nabla u(\mathbf{x},t), \quad \varphi : \mathbb{R} \mapsto \mathbb{R}. \tag{2.61}$$

Such an equation is common in the analysis of the edge-preserving diffusions, clf. [Perona and Malik, 1987].

Furthermore, let us limit ourselves to a univariate case $\Omega = [0,1]$ and consider a formal transition towards a discrete space and time by employing the following replacements:

$$u(x_i,t) \leftrightarrow [\mathbf{u}_t]_i, \quad \nabla u(x_i,t) \leftrightarrow \frac{1}{n}[\mathbf{F}\mathbf{u}]_i, \quad \int_{x\in[0,1]} (\cdot)\, dx \leftrightarrow \frac{1}{n}\sum_{i=1}^{n+1}, \tag{2.62}$$

where $\mathbf{F}$ is the forward-difference matrix, clf. Eq. (2.13). The function φ can now be seen as the map $\mathbb{R}^n \mapsto \mathbb{R}$:

$$\varphi\left(|\nabla u(x_i,t)|\right) \leftrightarrow \varphi\left(\frac{1}{n}|[\mathbf{F}\mathbf{u}_t]_i|\right) \equiv b_i(\mathbf{u}_t). \tag{2.63}$$

Eq. (2.60) would now tell us that the p-norm of $\mathbf{u}_t$ does not increase if each $b_i(\mathbf{u}_t) \geq 0$. This would imply that the generalized Laplacian $\mathbf{B}$ should be negative semi-definite, clf. Eqs. (2.15) and (2.13). In other words, the adoption of the condition on the L_p-norm diminishing to a discrete case requires that $-\mathbf{B}$ should be positive semi-definite. This is more restrictive than the positive definiteness of $\mathbf{I} - \tau\mathbf{B}$, and it also differs from the requirement of a positive definite $\mathbf{I} + \tau\mathbf{B}$. Positive semi-definiteness of the negative Laplacian will become more important in Chapter 3.

Chapter 3

Selection of Optimal Scale

Once stability is discussed, one may proceed towards automating the identification (selection) of the model from data. Model selection will be viewed here as a problem of statistical inference. The latter puts an emphasis on consistent predictions as opposed to stability, computational efficiency or invariance laws.

The theory of inference closest to this work is the classical maximum likelihood estimation, see e.g. Anderson [1958], equipped with warnings of Jaynes [2003]. All the philosophical aspects are delegated to discussions advocating the probability system of Kolmogorov [1956] and the model selection of Jaynes [2003].

A discrete diffusion can be viewed as an information processing device capable of revealing a signal in noise. If an initial condition is set to the signal of observations, then an empirical evidence suggests that there is an optimal stopping time when the diffused quantity is closest to an ideal signal assumed to exist in the noisy observations. The need for optimal stopping criteria is summarized by Carasso [1999]:

"*Attentive viewing of a computer screen for quite long periods of time may be necessary, and, because changes from one iteration to the next are usually imperceptible, locating the optimal point at which to terminate the process becomes highly elusive.*"

The principal contribution of this chapter is the postulation of Gaussian models which resolve several difficulties with the present optimal scale selection. Among the main improvements one could list: (i) the removal of normalization requirements present in the entropies of Sporring and Weickert [1999], (ii) achievement that in a linear case the entropies depend on the variance of noise and do not depend on actual signals, (iii) explicit formulation of the Bayesian assumptions behind the principle 'orthogonalize the model output to noise', which will here be referred as the decorrelation of Mrázek and Navara [2003].

3.1 Setting the Problem

An attempt will be made to unify two taste-specific but practical approaches which, at the first glance, seem to be very different: (i) the correlation criterion which is simple and stated as 'near optimal' by Mrázek and Navara [2003] and also analyzed by Gilboa et al. [2006], and (ii) the idea of the starting time of the stationary entropy change analyzed by Sporring and Weickert [1999].

It is first useful to recall Proposition 2.7 which states that the multiplication of any vector by a doubly stochastic matrix diminishes the value of the negative entropy, i.e. $H(\mathbf{Au}) \geq H(\mathbf{u})$, where

$$H(\mathbf{w}) \equiv -\sum_{i=1}^{n} w_i \ln w_i, \quad w_i \in \mathbb{R}_+. \tag{3.1}$$

Due to a tradition, the quantity $H(\mathbf{w})$ will be called the entropy of Shannon [1948], albeit it is well-known that his work is not the first investigation of the concept. The Shannon entropy appears in earlier editions of the work of Von Neumann [1955], and the research can be traced back to the masterpiece of L. Boltzmann (1877) whose English translation can be found in [Roux, 2002].

If a discrete diffusion is a successive multiplication by the doubly stochastic matrices, its steady state tends to a signal whose values equal to a constant. This matches well with the following property of the Shannon entropy, clf. [Marshall and Olkin, 1979]:

$$H(1, 0, \ldots, 0) \leq H(\mathbf{w}) \leq H(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}), \tag{3.2}$$

where the components of the vector $\mathbf{w}$ sum to unity. This a simple mathematical caricature of the second law of thermodynamics which states that heat flows from hot to cold, which increases the entropy.

The quantity in Eq. (3.2) may serve as an inverse measure of information content, in a spectral sense. The 'delta-peak' is richer than a constant signal in terms of the frequency content. According to this view, a linear diffusion reduces the information content of a signal, which conforms to the 'scale-space' idea where the diffusion time acts as a characteristic length-scale of a signal variability.

The spectral notion of information content is useful, but is not general enough:

1. A nonlinear diffusion may develop jump discontinuities, but otherwise reduce the variability of a signal. A spectral viewpoint becomes inconvenient.

2. The function in Eq. (3.2) applies to positive signals. It is unclear how to preprocess any real signal without adjusting its information content. Eq. (3.2) is already noninvariant w.r.t. the addition of a constant to a signal.

3. The entropy depends directly on the actual signal, but intuition suggests that the amount of 'uncertainty' should depend only on certain characteristics of noise present in the signal.

Clearly, the use of the Shannon entropy in the above-outlined context is suboptimal, but its misuses are numerous. One frequently applies an *ad hoc* normalization to an observed image in order to obtain a bivariate density function. Such approaches do exist because they often report a nontrivial mathematical statement, and demonstrate its usefulness with an experiment. This includes a variety of rather famous methods, such as those reported in [Meinel, 1986a], [Meinel, 1986b] and [Carasso, 1999]. One of the reasons of their success is that probabilistic diversity measures can be useful as 'smoothing energies'. However, there are better, more systematic, ways to utilize probabilities.

Use of entropies with scale-spaces is motivated by aesthetic and historical reasons. In essence, Sporring and Weickert [1999] attempt to import entropy production principles of Prigogine [1967] to image analysis. Furthermore, they put forward the statement which is presumably based on unreported experiments:

"*This correspondence has focused on the maximal entropy change by scale to estimate the size of image structures. The minimal change by scale, however, indicates especially stable scales with respect to evolution time. We expect these scales to be good candidates for stopping times in nonlinear diffusion scale-spaces.*"

The idea is interesting as it relates the Liapunov stability to the second law of thermodynamics and the maximum entropy inference. However, mixing the concept of 'observation' with the 'probability density' raises unnecessary questions, e.g.: Is Liapunov stability supposed to replace model selection? Is there any best way to preprocess a given image, so that when viewed as a scalar-valued function of the spatial coordinates, it would become a probability density? A current status of the entropy-based stopping suggested by Sporring and Weickert [1999] remains summarized in [Mrázek and Navara, 2003]:

"*However, as the entropy can be stable on whole intervals, it may be difficult to decide on a single stopping instant from that interval; we are unaware of their idea being brought into practice in the field of image restoration.*"

The approach of Mrázek and Navara [2003] is very simple. The authors view the difference between the diffusion outcome and the initial signal of observations as noise and suggest to optimally stop the diffusion at the time which gives the smallest correlation between the diffusion outcome and the noise estimate. However, they indiscriminate between the concepts of 'the sample of a random process (chain)' and the 'conditional expectation under assumption of a given random process (chain)'. Naturally, as in the case with the work of Sporring and Weickert [1999], the methodology of Mrázek and Navara [2003] lacks explicit probabilistic model too. It is rather evident that most of a critique directed against the entropy-based stopping applies to the use of correlation as well. Moreover, the remark on the 'entropic stability' in [Mrázek and Navara, 2003] pertains to rare cases in which the correlation might have very shallow minimum as well, or no minima at all, as indicated in [Gilboa et al., 2006].

However, neither entropy nor correlation-based stopping should be excluded by the developments related to robust statistics. Entropy and correlation are fundamental concepts, and one can always find many ways to make improvements in special cases. It is just wrong to advocate particular statistics as universally

applicable scale selection methods. We suggest a unification which allows to: (i) avoid unnecessary preprocessing of signals, (ii) arrive at a more general criterion, which merges both ideas into a single equation and further clarifies their probabilistic assumptions, and (iii) view the optimal diffusion stopping as an example where Bayesian arguments simplify the likelihood inference, not vice versa, as is commonly practiced.

3.2 Constructing Joint Probability Densities

A diffusion scale-space can be seen as the statement with the conditional expectation over some unknown probability measure. Knowledge of such a measure allows to employ numerous methods of a statistical inference, and the likelihood method is one of the simplest. The problem, however, is to first build such a density.

Let the scale-space be given by the following propagator:

$$\mathbf{u}_m = \mathbf{P}_\theta^{-1}\mathbf{u}_0 \equiv \prod_{i=0}^{m-1} \left(\mathbf{I} - \tau\mathbf{B}(\mathbf{u}_i)\right)^{-1} \mathbf{u}_0, \tag{3.3}$$

where the Laplacian matrices $\mathbf{B}(\mathbf{u}_i)$ have been discussed in Chapter 2. The subscript θ in the nonsingular symmetric propagator $\mathbf{P}_\theta \in \mathbb{R}^{n\times n}$ marks its dependence on the parameters, such as m, and the goal is to set the 'first principles' which will latter enable to apply the statistical inference in the parameter identification. Conceptually, from a statistical inference viewpoint, identifying the dynamical model is the same as 'fitting a line' to observations, and let us further simplify the notation so that $\mathbf{y} \equiv \mathbf{u}_0$ is the known vector of observations, and $\mathbf{u} \equiv \mathbf{u}_m$ is the model output which is a synonym to predictions or smoothing outcome.

Our first assumption is the existence of the discrete variational criterion which allows to formally relate the model output $\mathbf{u}$ with the observations $\mathbf{y}$. Eq. (3.3) suggests defining the cost function $\frac{1}{2}\|\mathbf{u}\|^2_{\mathbf{P}_\theta} - \mathbf{u}^T\mathbf{y}$. The optimality conditions of the first-order then lead to Eq. (3.3). However, it will prove to be useful to state a more general form of the energy:

$$E_\theta(\mathbf{u},\mathbf{y}) = \frac{1}{2}\begin{pmatrix}\mathbf{u}\\ \mathbf{y}\end{pmatrix}^T \begin{pmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{C} & \mathbf{D}\end{pmatrix}\begin{pmatrix}\mathbf{u}\\ \mathbf{y}\end{pmatrix}. \tag{3.4}$$

Assuming that the matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ is nonsingular, the first-order optimality leads to the following relationship:

$$\mathbf{u}^* = \arg\min_{\mathbf{u}} E_\theta(\mathbf{u},\mathbf{y}) \tag{3.5}$$

$$= (\mathbf{A}+\mathbf{A}^T)^{-1}(\mathbf{B}+\mathbf{C}^T)\mathbf{y}. \tag{3.6}$$

At this point, one is free in choosing the matrices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ so that Eq. (3.6) becomes formally equivalent to Eq. (3.3). A note of caution is in place. The Laplacians $\mathbf{B}(\mathbf{u})$ depend on $\mathbf{u}$ in a nonlinear way. The energy minimization here is not used to solve any nonlinear equations. Given the scale-space of signals $\mathbf{u}_0, \mathbf{u}_1, \ldots, \mathbf{u}_m$ we simply construct the propagator matrix and the discrete diffusion becomes linear *a posteriori*. The energy merely suggests an interpretation,

which latter allows to deduce the criteria for the selection of parameters θ. In a linear case, the optimality conditions of the first order produce a formal relationship between the two known quantities, i.e. the vectors $\mathbf{u}^* \equiv \mathbf{u}_m$ and $\mathbf{y}$. In what follows, the asterisk is dropped out to simplify the notations.

The second assumption is to consider a meaningful Gaussian density

$$p(\mathbf{u}, \mathbf{y}|\theta) = |2\pi\mathbf{\Sigma}|^{-1} \exp\left(-E_\theta(\mathbf{u}, \mathbf{y})\right), \tag{3.7}$$

which further constrains the quadratic form. The matrix

$$\mathbf{\Sigma} \equiv \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}^{-1} \tag{3.8}$$

must now be symmetric. Here 'meaningful' does not necessarily imply that $E_\theta(\mathbf{u}, \mathbf{y})$ should be finite and positive, which requires a positive definite $\mathbf{\Sigma}$. The transition from a set of linear equations to the Gaussian density can be useful even if the latter is not well-defined as subsequent changes can be made to remove singularities.

We shall now construct a particular model $\mathcal{H}_1$ for $\mathbf{\Sigma}$, which will turn out to be singular, and discuss a particular improvement, denoted as $\mathcal{H}_2$.

Definition 3.1 (Gaussian model)**.** *Assume that the values of a random variable Z are the elements of the joint vector of the model outputs $\mathbf{u} \in \mathbb{R}^n$ and the observations $\mathbf{y} \in \mathbb{R}^n$. Let Z be distributed according to $N(\mathbf{0}, \mathbf{\Sigma})$ with the covariance matrix*

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{uu} & \mathbf{\Sigma}_{uy} \\ \mathbf{\Sigma}_{uy}^T & \mathbf{\Sigma}_{yy} \end{pmatrix}. \tag{3.9}$$

Here $\mathbf{\Sigma}_{ab} \equiv \mathrm{Cov}(A, B)$, where A and B are vector-valued and

$$\mathrm{Cov}(A, B) \equiv \left\langle (A - \langle A \rangle)(B - \langle B \rangle)^T \right\rangle. \tag{3.10}$$

It is always implied that the elements of the covariance matrix in Eq. (3.9) *depend on model parameters $\boldsymbol{\theta}$.*

Proposition 3.1 (Gaussian conditioning, clf. Anderson [1958])**.**

$$\langle U|\mathbf{y}, \boldsymbol{\theta}\rangle = \mathbf{\Sigma}_{uy}(\boldsymbol{\theta})\mathbf{\Sigma}_{yy}^{-1}(\boldsymbol{\theta})\mathbf{y}, \tag{3.11}$$

$$2\ln p(\mathbf{y}|\boldsymbol{\theta}) = -\|\mathbf{y}\|^2_{\mathbf{\Sigma}_{yy}^{-1}(\boldsymbol{\theta})} - \ln|2\pi\mathbf{\Sigma}_{yy}(\boldsymbol{\theta})|, \tag{3.12}$$

$$\mathrm{Cov}\,(U|\mathbf{y}, \boldsymbol{\theta}) = \mathbf{\Sigma}_{uu}(\boldsymbol{\theta}) - \mathbf{\Sigma}_{uy}(\boldsymbol{\theta})\mathbf{\Sigma}_{yy}^{-1}(\boldsymbol{\theta})\mathbf{\Sigma}_{uy}^T(\boldsymbol{\theta}). \tag{3.13}$$

This proposition summarizes three important quantities of a statistical conditioning: (i) conditional mean which determines 'model output' or 'predictions', (ii) marginal likelihood, which is useful in choosing the model parameters and will be discussed in Section 3.3, and (iii) conditional covariance, which can sometimes be associated with uncertainty of predictions. Notice that the vector $\mathbf{y}$ does not enter the right hand side of Eq. (3.13) explicitly.

Proposition 3.2 (Additive uncorrelated Gaussian noise)**.** *If the observation $Y = U + N$, where N is uncorrelated with U, then*

$$\mathbf{\Sigma}_{yy} = \mathbf{\Sigma}_{uu} + \mathbf{\Sigma}_{nn}. \tag{3.14}$$

Proof. By substituting the sum to Eq. (3.10), one can verify that

$$\boldsymbol{\Sigma}_{u+n} = \boldsymbol{\Sigma}_{uu} + 2\boldsymbol{\Sigma}_{un} + \boldsymbol{\Sigma}_{nn}. \tag{3.15}$$

If the signals are uncorrelated, $\langle UN^T \rangle = \langle U \rangle \langle N^T \rangle$ and

$$\begin{aligned} \boldsymbol{\Sigma}_{un} &= \left\langle (U - \langle U \rangle)(N - \langle N \rangle)^T \right\rangle & (3.16) \\ &= \langle UN^T \rangle - \langle U \rangle \langle N^T \rangle = \mathbf{0}. & (3.17) \end{aligned}$$

□

Proposition 3.3 (Variance-reducing conditioning)**.** *A further restriction* $\boldsymbol{\Sigma}_{uu} = \boldsymbol{\Sigma}_{uy} = \boldsymbol{\Sigma}_{uy}^T$ *guarantees that conditioning reduces variance:*

$$\begin{aligned} \operatorname{Cov}(U|\mathbf{y}, \boldsymbol{\theta}) &= \boldsymbol{\Sigma}_{uu} - \boldsymbol{\Sigma}_{uu}(\boldsymbol{\Sigma}_{uu} + \boldsymbol{\Sigma}_{nn})^{-1}\boldsymbol{\Sigma}_{uu} & (3.18) \\ &= (\boldsymbol{\Sigma}_{uu}^{-1} + \boldsymbol{\Sigma}_{nn}^{-1})^{-1} \preceq \boldsymbol{\Sigma}_{uu}\,. & (3.19) \end{aligned}$$

The operator $\mathbf{A} \preceq \mathbf{B}$ *should be understood in the sense that* $\mathbf{B} - \mathbf{A}$ *is positive semidefinite. The dependence of the right hand sides on* $\boldsymbol{\theta}$ *is suppressed for the sake of brevity.*

Proposition 3.4 (Model $\mathcal{H}_1$)**.** *Assumptions of the additive uncorrelated Gaussian noise and variance-reducing conditioning restrict the Gaussian model in Definition 3.1 with the conditions* $\boldsymbol{\Sigma}_{uu} = \boldsymbol{\Sigma}_{uy} = \boldsymbol{\Sigma}_{yu}$ *and* $\boldsymbol{\Sigma}_{yy} = \boldsymbol{\Sigma}_{uu} + \boldsymbol{\Sigma}_{nn}$*:*

$$\boldsymbol{\Sigma}_{\mathcal{H}_1} = \begin{pmatrix} \boldsymbol{\Sigma}_{uu} & \boldsymbol{\Sigma}_{uu} \\ \boldsymbol{\Sigma}_{uu} & \boldsymbol{\Sigma}_{uu} + \boldsymbol{\Sigma}_{nn} \end{pmatrix}. \tag{3.20}$$

The joint probability density factors into

$$\begin{aligned} p(\mathbf{u}|\boldsymbol{\theta}, \mathcal{H}_1) &= N(\mathbf{0}, \boldsymbol{\Sigma}_{uu}), & p(\mathbf{y}|\boldsymbol{\theta}, \mathcal{H}_1) &= N(\mathbf{0}, \boldsymbol{\Sigma}_{uu} + \boldsymbol{\Sigma}_{nn}), & (3.21) \\ p(\mathbf{y}|\boldsymbol{\theta}, \mathbf{u}, \mathcal{H}_1) &= N(\mathbf{u}, \boldsymbol{\Sigma}_{nn}), & p(\mathbf{u}|\mathbf{y}, \boldsymbol{\theta}, \mathcal{H}_1) &= N(\boldsymbol{\mu}_{u|y}, \boldsymbol{\Sigma}_{u|y}), & (3.22) \end{aligned}$$

where the conditional mean vector $\boldsymbol{\mu}_{u|y}$ *and the covariance matrix* $\boldsymbol{\Sigma}_{u|y}$ *are shortcuts to* $\langle U|\mathbf{y}, \boldsymbol{\theta}, \mathcal{H}_1 \rangle$ *and* $\operatorname{Cov}(U|\mathbf{y}, \boldsymbol{\theta}, \mathcal{H}_1)$*, respectively.*

Proof. The exponential factor in the joint Gaussian density (total energy) is:

$$E_{\mathcal{H}_1}(\mathbf{u}, \mathbf{y}) = \frac{1}{2} \begin{pmatrix} \mathbf{u} \\ \mathbf{y} \end{pmatrix}^T \begin{pmatrix} \boldsymbol{\Sigma}_{uu} & \boldsymbol{\Sigma}_{uu} \\ \boldsymbol{\Sigma}_{uu} & \boldsymbol{\Sigma}_{uu} + \boldsymbol{\Sigma}_{nn} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{u} \\ \mathbf{y} \end{pmatrix}. \tag{3.23}$$

Application of Schur complements and the Sherman–Morrison–Woodbury identity to the matrix in Eq. (3.20) results in:

$$\begin{pmatrix} \boldsymbol{\Sigma}_{uu} & \boldsymbol{\Sigma}_{uu} \\ \boldsymbol{\Sigma}_{uu} & \boldsymbol{\Sigma}_{uu} + \boldsymbol{\Sigma}_{nn} \end{pmatrix}^{-1} = \begin{pmatrix} \boldsymbol{\Sigma}_{nn}^{-1} + \boldsymbol{\Sigma}_{uu}^{-1} & -\boldsymbol{\Sigma}_{nn}^{-1} \\ -\boldsymbol{\Sigma}_{nn}^{-1} & \boldsymbol{\Sigma}_{nn}^{-1} \end{pmatrix}. \tag{3.24}$$

Separating variables $\mathbf{u}$ and $\mathbf{y}$ establishes the result. □

Substitution of Eq. (3.24) to Eq. (3.23) yields Eq. (3.25). One could note that the total energy can be expressed as the 'regularization energy':

$$E_{\mathcal{H}_1}(\mathbf{u}, \mathbf{y}) = \frac{1}{2} \left(\|\mathbf{y} - \mathbf{u}\|^2_{\boldsymbol{\Sigma}_{nn}^{-1}} + \|\mathbf{u}\|^2_{\boldsymbol{\Sigma}_{uu}^{-1}} \right). \tag{3.25}$$

This gives the extremal property of the conditional expectation, i.e. $\langle U|\mathbf{y},\mathcal{H}_1\rangle = \arg\min_{\mathbf{u}} E_{\mathcal{H}_1}(\mathbf{u},\mathbf{y})$. This is so because the minimum of the total energy in Eq. (3.25) is at $\boldsymbol{\Sigma}_{uu}(\boldsymbol{\Sigma}_{uu}+\boldsymbol{\Sigma}_{nn})^{-1}\mathbf{y}$, which coincides with Eq. (3.11).

The joint covariance matrix is now reduced to the matrix with two independent quantities: (i) the signal covariance $\boldsymbol{\Sigma}_{uu}$ and the noise covariance $\boldsymbol{\Sigma}_{nn}$. We can now produce a stochastic picture of Eq. (3.3) by relating it to the conditional expectation:

$$\mathbf{P}_\theta^{-1} \equiv \boldsymbol{\Sigma}_{uu}(\boldsymbol{\Sigma}_{uu}+\boldsymbol{\Sigma}_{nn})^{-1}. \tag{3.26}$$

There still remain too many degrees of freedom in constructing the probability density, but let us further assume that the noise is white, i.e. $\boldsymbol{\Sigma}_{nn} \equiv \theta_0$. Then,

$$\boldsymbol{\Sigma}_{uu}^{-1} = \frac{1}{\theta_0}(\mathbf{P}_\theta - \mathbf{I}). \tag{3.27}$$

The concentration matrix $\boldsymbol{\Sigma}_{uu}^{-1}$ will be nonsingular if and only if the propagator has no eigenvalues qual to unity. When the Laplacians $\mathbf{B}(\mathbf{u})$ are singular, this is the case indeed. Thus, the terms such as $p(\mathbf{y}|\mathcal{H}_1)$ have no meaning.

However, the model can be modified. Instead of adding the signal variable with the covariance matrix $\boldsymbol{\Sigma}_{uu}$ to the uncorrelated noise variable with the covariance matrix $\boldsymbol{\Sigma}_{nn}$, it will prove useful to add the signal with $\boldsymbol{\Sigma}_{nn}^{-1}$ to an uncorrelated noise with $\boldsymbol{\Sigma}_{uu}^{-1}$. This trick motivates the model $\mathcal{H}_2$.

Proposition 3.5 (Model $\mathcal{H}_2$)**.** *Assume that the joint variable* $Z = (U,Y) \sim N(\mathbf{0},\boldsymbol{\Sigma})$ *with*

$$\boldsymbol{\Sigma}_{\mathcal{H}_2} = \begin{pmatrix} \boldsymbol{\Sigma}_{nn}^{-1} & \boldsymbol{\Sigma}_{nn}^{-1} \\ \boldsymbol{\Sigma}_{nn}^{-1} & \boldsymbol{\Sigma}_{nn}^{-1}+\boldsymbol{\Sigma}_{uu}^{-1} \end{pmatrix}. \tag{3.28}$$

The joint probability density factors into

$$p(\mathbf{u}|\boldsymbol{\theta},\mathcal{H}_2) = N(\mathbf{0},\boldsymbol{\Sigma}_{nn}^{-1}), \qquad p(\mathbf{y}|\boldsymbol{\theta},\mathcal{H}_2) = N(\mathbf{0},\boldsymbol{\Sigma}_{uu}^{-1}+\boldsymbol{\Sigma}_{nn}^{-1}), \tag{3.29}$$

$$p(\mathbf{y}|\boldsymbol{\theta},\mathbf{u},\mathcal{H}_2) = N(\mathbf{u},\boldsymbol{\Sigma}_{uu}^{-1}), \qquad p(\mathbf{u}|\mathbf{y},\boldsymbol{\theta},\mathcal{H}_2) = N(\boldsymbol{\mu}_{u|y},\boldsymbol{\Sigma}_{nn}). \tag{3.30}$$

Proof. The total energy of the model $\mathcal{H}_2$ is:

$$E_{\mathcal{H}_2}(\mathbf{u},\mathbf{y}) = \frac{1}{2}\begin{pmatrix}\mathbf{u}\\ \mathbf{y}\end{pmatrix}^T \begin{pmatrix} \boldsymbol{\Sigma}_{nn}^{-1} & \boldsymbol{\Sigma}_{nn}^{-1} \\ \boldsymbol{\Sigma}_{nn}^{-1} & \boldsymbol{\Sigma}_{nn}^{-1}+\boldsymbol{\Sigma}_{uu}^{-1} \end{pmatrix}^{-1} \begin{pmatrix}\mathbf{u}\\ \mathbf{y}\end{pmatrix}. \tag{3.31}$$

Application of Schur complements with the Sherman–Morrison–Woodbury identity yields:

$$\begin{pmatrix} \boldsymbol{\Sigma}_{nn}^{-1} & \boldsymbol{\Sigma}_{nn}^{-1} \\ \boldsymbol{\Sigma}_{nn}^{-1} & \boldsymbol{\Sigma}_{nn}^{-1}+\boldsymbol{\Sigma}_{uu}^{-1} \end{pmatrix}^{-1} = \begin{pmatrix} \boldsymbol{\Sigma}_{uu}+\boldsymbol{\Sigma}_{nn} & -\boldsymbol{\Sigma}_{uu} \\ -\boldsymbol{\Sigma}_{uu} & \boldsymbol{\Sigma}_{uu} \end{pmatrix}. \tag{3.32}$$

Separating the quantities $\mathbf{u}$ and $\mathbf{y}$ produces the required decompositions. □

The regularization alternative to $E_{\mathcal{H}_2}(\mathbf{u},\mathbf{y})$ becomes:

$$E_{\mathcal{H}_2}(\mathbf{u},\mathbf{y}) = \frac{1}{2}\left(\|\mathbf{y}-\mathbf{u}\|^2_{\boldsymbol{\Sigma}_{uu}} + \|\mathbf{u}\|^2_{\boldsymbol{\Sigma}_{nn}}\right). \tag{3.33}$$

A minimum of the total energy in Eq. (3.33) is at $\boldsymbol{\Sigma}_{nn}^{-1}(\boldsymbol{\Sigma}_{nn}^{-1}+\boldsymbol{\Sigma}_{uu}^{-1})^{-1}\mathbf{y}$, which coincides with Eq. (3.11) applied to the model $\mathcal{H}_2$. This establishes the extremal

property of the conditional expectation, which is $\langle U|\mathbf{y},\mathcal{H}_2\rangle = \arg\min_{\mathbf{u}} E_{\mathcal{H}_2}(\mathbf{u},\mathbf{y})$. Notice that $\langle U|\mathbf{y},\mathcal{H}_1\rangle = \langle U|\mathbf{y},\mathcal{H}_2\rangle$, but $E_{\mathcal{H}_1}(\mathbf{u},\mathbf{y}) \neq E_{\mathcal{H}_2}(\mathbf{u},\mathbf{y})$.

The problem with the existence of the likelihood $p(\mathbf{y}|\mathcal{H}_2)$ is eliminated because the density $N(\mathbf{0}, \boldsymbol{\Sigma}_{uu}^{-1} + \boldsymbol{\Sigma}_{nn}^{-1})$ is finite and non-degenerate even when $\boldsymbol{\Sigma}_{uu}^{-1}$ is singular. This comes at the cost that the noise covariance is no longer white.

3.3 Marginal Likelihoods of the Models

The meaning of Eq. (3.12) allows to unify a variety of model selection ideas which need not be derived by heuristic means. It will be shown that the unnormalized correlation and entropy are particular cases of the marginal likelihood inference. Historically, the concept of likelihood was considered already by C. F. Gauss in 1809 and named so by R. A. Fisher around 1922, clf. [Plackett, 1966]. The marginal likelihood is often attributed to Robbins [1983], who has used it much earlier in the statistical inference setting known as *empirical Bayes*. Modern uses of this concept can be traced back to the work of Lejeune and Faulkenberry [1982] and the comment given by Leonard [1982].

Given the parameters $\boldsymbol{\theta}$ with the assumptions pertaining to either of the models $\mathcal{H}_i$, where $i = \{1, 2\}$, let the joint probability density of observations $\mathbf{y}$ be $p(\mathbf{y}|\boldsymbol{\theta},\mathcal{H}_i)$. When viewed as a function of the model parameters $\boldsymbol{\theta}$, the quantity is called *marginal likelihood*. The word *likelihood* signifies that this function is not a probability density function, and the marginal likelihood can be opposed to the likelihood $p(\mathbf{y}|\mathbf{u},\boldsymbol{\theta},\mathcal{H}_i)$.

When the dependence on the argument $\mathbf{y}$ is emphasized, the probability with the density $p(\mathbf{y}|\mathcal{H}_i)$ is often called Bayesian evidence. Bayesian philosophers advise to make inferences from data by using only the probabilities of quantities conditioned on data, such as those obtainable from the density $p(\mathbf{u},\boldsymbol{\theta}|\mathbf{y},\mathcal{H}_i)$, clf. the famous statement: 'We do not analyze unseen data sets'. This means that marginalizing probability density over the observation variables is strictly forbidden, but one is free to integrate over parameters or even multiple hypothesis spaces.

Another, rather minor, Bayesian requirement is to postulate the probability densities for the parameters $\boldsymbol{\theta}$ too. However, in the applications to scale-spaces it is easier to treat the quantities $\mathbf{u}$ and $\boldsymbol{\theta}$ differently. If desired, probability densities for the parameters $\boldsymbol{\theta}$ could be incorporated by adding additional energy terms to the likelihood or resorting to fiducial inference, see e.g. [Salomé, 1998].

Proposition 3.6 (Marginal likelihood $p(\mathbf{y}|\boldsymbol{\theta},\mathcal{H}_1)$)**.** *Assume a white covariance* $\boldsymbol{\Sigma}_{nn} = \theta_0\mathbf{I}$ *for some* $\theta_0 > 0$*. The marginal likelihood* $p(\mathbf{y}|\boldsymbol{\theta},\mathcal{H}_1)$ *in Eq.* (3.21) *decomposes into:*

$$-2\ln p(\mathbf{y}|\boldsymbol{\theta},\mathcal{H}_1) = \frac{1}{\theta_0}\big(\|\mathbf{y}-\boldsymbol{\mu}_{u|y}\|^2 + (\mathbf{y}-\boldsymbol{\mu}_{u|y})^T\boldsymbol{\mu}_{u|y}\big) + \ln|2\pi(\boldsymbol{\Sigma}_{uu}+\theta_0\mathbf{I})| \ . \quad (3.34)$$

Proof. The term $p(\mathbf{y}|\boldsymbol{\theta},\mathcal{H}_1)$, given by Eq. (3.21), can be written in an expanded form:

$$-2\ln p(\mathbf{y}|\boldsymbol{\theta},\mathcal{H}_1) = \|\mathbf{y}\|^2_{(\boldsymbol{\Sigma}_{uu}+\theta_0\mathbf{I})^{-1}} + \ln|2\pi(\boldsymbol{\Sigma}_{uu}+\theta_0\mathbf{I})| \ . \quad (3.35)$$

The first term splits into the squared Euclidean distance between the model prediction and observations and the orthogonality of 'noise' to the model output:

$$\|\mathbf{y}\|^{-1}_{\boldsymbol{\Sigma}_{uu}+\theta_0\mathbf{I}} = \mathbf{y}^T\boldsymbol{\Sigma}^{-1}_{uu}\boldsymbol{\mu}_{u|y} \qquad [\text{use } \boldsymbol{\mu}_{u|y} \equiv \boldsymbol{\Sigma}_{uu}(\boldsymbol{\Sigma}_{uu}+\theta_0\mathbf{I})^{-1}\mathbf{y}] \tag{3.36}$$

$$= \theta_0^{-1}\mathbf{y}^T(\mathbf{y}-\boldsymbol{\mu}_{u|y}) \quad [\text{rearrange } \boldsymbol{\Sigma}^{-1}_{uu}\boldsymbol{\mu}_{u|y} = \theta_0^{-1}(\mathbf{y}-\boldsymbol{\mu}_{u|y})] \tag{3.37}$$

$$= \theta_0^{-1}(\|\mathbf{y}-\boldsymbol{\mu}_{u|y}\|^2 + (\mathbf{y}-\boldsymbol{\mu}_{u|y})^T\boldsymbol{\mu}_{u|y})\,, \tag{3.38}$$

where Eq. (3.38) follows from its predecessor by adding and subtracting $-\mathbf{y}^T\boldsymbol{\mu}_{u|y} + \boldsymbol{\mu}^T_{u|y}\boldsymbol{\mu}_{u|y}$. □

When splitting $\boldsymbol{\Sigma}_{uu}+\theta_0\mathbf{I} = \theta_0\boldsymbol{\Sigma}_{uu}(\theta_0^{-1}\mathbf{I}+\boldsymbol{\Sigma}^{-1}_{uu})$ and using $\ln|2\pi\boldsymbol{\Sigma}_x| = 2h(X)-n$, the term $\ln|2\pi\boldsymbol{\Sigma}_{uu}+\theta_0\mathbf{I}|$ can be written in two different ways, which results in an interesting identity:

$$2h(Y|\boldsymbol{\theta},\mathcal{H}_1) = 2h(U|\boldsymbol{\theta},\mathcal{H}_1) - \ln\left|\frac{\mathrm{Cov}(U|\boldsymbol{\theta},\mathbf{y},\mathcal{H}_1)}{\theta_0}\right|. \tag{3.39}$$

Therefore, minimizing $\ln|2\pi\boldsymbol{\Sigma}_{uu}+\theta_0\mathbf{I}|$ reduces the uncertainty of the prior density $p(\mathbf{u}|\boldsymbol{\theta},\mathcal{H}_1)$ and maximizes the generalized signal-to-noise ratio (SNR).

Another way to understand the determinant term in Eq. (3.34) is to apply the Sherman–Morrison–Woodbury identity and perform splitting into:

$$\ln|2\pi(\boldsymbol{\Sigma}_{uu}+\theta_0\mathbf{I})| = n\ln 2\pi\theta_0 - \ln\left|\mathbf{I}-\boldsymbol{\Sigma}_{uu}(\boldsymbol{\Sigma}_{uu}+\theta_0\mathbf{I})^{-1}\right|. \tag{3.40}$$

Here it is useful to further apply the identity $\ln|\mathbf{A}| = \mathrm{tr}\ln\mathbf{A}$. Furthermore, let us utilize the Taylor series approximation of the matrix logarithm up to the first order. The term $\mathrm{tr}\left(\boldsymbol{\Sigma}_{uu}(\boldsymbol{\Sigma}_{uu}+\theta_0\mathbf{I})^{-1}\right)$ can then be seen as the number of good parameters. The second term in Eq. (3.40) resembles the number of noisy degrees of freedom and when subtracted from the first term, it measures the number of effective model parameters. However, the total energy of the system penalizes the model complexity too.

Proposition 3.7 (Marginal likelihood $p(\mathbf{y}|\boldsymbol{\theta},\mathcal{H}_2)$)**.** *Assume a white covariance* $\boldsymbol{\Sigma}_{nn} = \theta_0\mathbf{I}$ *for some* $\theta_0 > 0$*. The marginal likelihood* $p(\mathbf{y}|\boldsymbol{\theta},\mathcal{H}_2)$ *given by Eq.* (3.29) *decomposes into*

$$-2\ln p(\mathbf{y}|\boldsymbol{\theta},\mathcal{H}_2) = \theta_0\big(\|\boldsymbol{\mu}_{u|y}\|^2 + (\mathbf{y}-\boldsymbol{\mu}_{u|y})^T\boldsymbol{\mu}_{u|y}\big) + \ln\left|2\pi(\boldsymbol{\Sigma}^{-1}_{uu}+\theta_0^{-1}\mathbf{I})\right|. \tag{3.41}$$

Proof. The explicit form of the marginal likelihood $p(\mathbf{y}|\boldsymbol{\theta},\mathcal{H}_2)$ in Eq. (3.29) reads:

$$-2\ln p(\mathbf{y}|\boldsymbol{\theta},\mathcal{H}_2) = \|\mathbf{y}\|^2_{(\boldsymbol{\Sigma}^{-1}_{uu}+\theta_0^{-1}\mathbf{I})^{-1}} + \ln\left|2\pi(\boldsymbol{\Sigma}^{-1}_{uu}+\theta_0^{-1}\mathbf{I})\right|. \tag{3.42}$$

Contrary to the model $\mathcal{H}_1$, the matrix $\boldsymbol{\Sigma}^{-1}_{uu}$ is now allowed to be singular. A simple identity $\mathbf{y}^T(\boldsymbol{\Sigma}^{-1}_{uu}+\theta_0^{-1}\mathbf{I})^{-1}\mathbf{y} = \theta_0\mathbf{y}^T\boldsymbol{\Sigma}_{uu}(\boldsymbol{\Sigma}_{uu}+\theta_0\mathbf{I})^{-1}\mathbf{y}$ rewrites the first term in Eq. (3.42):

$$\|\mathbf{y}\|^2_{(\boldsymbol{\Sigma}^{-1}_{uu}+\theta_0^{-1}\mathbf{I})^{-1}} = \theta_0\mathbf{y}^T\boldsymbol{\mu}_{u|y} \tag{3.43}$$

$$= \theta_0((\mathbf{y}-\boldsymbol{\mu}_{u|y})^T\boldsymbol{\mu}_{u|y} + \|\boldsymbol{\mu}_{u|y}\|^2)\,. \tag{3.44}$$

□

One could note that the both models are related. For example, comparison of Eqs. (3.38) and (3.44) reveals that

$$E_{\mathcal{H}_2}(\boldsymbol{\mu}_{u|y}, \mathbf{y}) = \theta_0^2 E_{\mathcal{H}_1}(\boldsymbol{\mu}_{u|y}, \mathbf{y}). \tag{3.45}$$

By noticing that $\mathrm{Cov}(Y|\boldsymbol{\theta}, \mathcal{H}_2) = (\mathrm{Cov}(U|\boldsymbol{\theta}, \mathbf{y}, \mathcal{H}_1))^{-1}$, one may also discover that

$$h(Y|\boldsymbol{\theta}, \mathcal{H}_2) = n \ln(2\pi e) - h(U|\mathbf{y}, \boldsymbol{\theta}, \mathcal{H}_1). \tag{3.46}$$

It is also possible to replace the inner-products with the determinants via the identity $1 + \mathbf{u}^T\mathbf{v} = |\mathbf{I} + \mathbf{u}\mathbf{v}^T|$, but the reverse route seems to be more practical. The latter is discussed in Section 3.5.

3.4 Decorrelation Priors of the Models

Mrázek and Navara [2003] have noticed that when the observations $\mathbf{y}$ present an additive corruption with noise, and $\mathbf{u}$ is the outcome of the diffusion scale space, the 'correlation' between the vector $\mathbf{y}-\mathbf{u}$ and $\mathbf{u}$ evolves in time so that the location of its minimum yields a practically approved values of the scale.

Propositions 3.6 and 3.7 indicate that minimizing the total energy of the system, making the model output orthogonal to noise, maximizing the differential entropy and generalized SNR need not be seen as heuristic ideas, but rather particular principles stemming from the marginal likelihood framework. Let us clarify the meaning of 'decorrelation'.

Definition 3.2 (Correlation, clf. [Anderson, 1958])**.** *The correlation $\varrho \in [-1, 1]$ between the two random variables X and Y is:*

$$\varrho = \frac{\mathrm{tr}\,\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{tr}\,\mathrm{Cov}(X, X)}\sqrt{\mathrm{tr}\,\mathrm{Cov}(Y, Y)}}, \tag{3.47}$$

It is assumed that both variables take values in $\mathbb{R}^n$.

Mrázek and Navara [2003] treat the diffusion outcome as if it were merely a stochastic sample. The authors do not define the density $p(\mathbf{y}, \mathbf{u}|\boldsymbol{\theta})$, but simply substitute the definitions of the covariance functions in terms of expectations and replace the latter ones with a simple averaging, i.e.

$$\mathrm{tr}\,\mathrm{Cov}(X, Y) \to \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}), \tag{3.48}$$

where $\bar{x} = n^{-1}\sum_{i=1}^{n} x_i$ and $\bar{y}$ is defined similarly.

Practice indicates that normalizing the criterion in Eq. (3.47) does not significantly affect the optima, and in this case, the 'decorrelation' reduces to:

$$t^* = \arg\min_t (\mathbf{y} - \boldsymbol{\mu}_{u|y})^T \boldsymbol{\mu}_{u|y}, \tag{3.49}$$

where it is implicitly assumed that $\boldsymbol{\mu}_{u|y}$ depends on the time t.

Treating the diffusion outcome in the same way as observations $\mathbf{y}$ is inconsistent from a methodological viewpoint:

1. The output of the diffusion scale space $\mathbf{u}$ can be seen as the result of computing the conditional expectation $\boldsymbol{\mu}_{u|y}$. This is not a vector of observations, and should not be treated as a stochastic realization (sample of a discrete chain). It is a part of our hypothesis space about $\mathbf{y}$, and we seek for probabilistic assignments to particular statements within a hypothesis space.

2. Justification of any statistics without the explicit noise model is always weak.

The criterion in Eq. (3.49) can also be further reconciled from a Bayesian perspective. The removal of the terms in the marginal likelihood can be seen as the introduction of their 'anti-terms'. This can be interpreted as multiplying the marginal likelihood with the density density $p(\boldsymbol{\theta}|\mathcal{H}_1)$ *a priori*. This picture is not entirely Bayesian because the densities depend on data. However, they generalize the situation in terms of the density functions which do not necessarily have to be Gaussian.

Proposition 3.8 (Decorrelating priors)**.** *Minimizing the 'unnormalized correlation'* $(\mathbf{y}-\boldsymbol{\mu}_{u|y})^T\boldsymbol{\mu}_{u|y}$ *is equivalent to the maximization of the marginal likelihood with the models* $\mathcal{H}_1$ *and* $\mathcal{H}_2$ *when they are supplied with the following priors:*

$$p(\boldsymbol{\theta}|\mathcal{H}_1) \propto \theta_0^{-n/2} \exp\left(-\ln p(\mathbf{y}|\boldsymbol{\mu}_{u|y},\boldsymbol{\theta},\mathcal{H}_1) + \langle \ln p(Y|\boldsymbol{\theta},\mathcal{H}_1)\rangle\right), \tag{3.50}$$

$$p(\boldsymbol{\theta}|\mathcal{H}_2) \propto \theta_0^{+n/2} \exp\left(-\ln p(\boldsymbol{\mu}_{u|y}|\boldsymbol{\theta},\mathcal{H}_2) + \langle \ln p(Y|\boldsymbol{\theta},\mathcal{H}_2)\rangle\right). \tag{3.51}$$

Proof. When $\mathbf{u}_\theta \equiv \boldsymbol{\mu}_{u|y}$, in order to isolate the term $(\mathbf{y}-\mathbf{u}_\theta)^T\mathbf{u}_\theta$ in Eq. (3.34), one has to multiply the marginal likelihood with

$$p(\boldsymbol{\theta}|\mathcal{H}_1) \propto \exp\left(\frac{1}{2\theta_0}\|\mathbf{y}-\boldsymbol{\mu}_{u|y}\|^2 + \frac{1}{2}\ln|2\pi(\boldsymbol{\Sigma}_{uu}+\theta_0\mathbf{I})|\right). \tag{3.52}$$

The first term can be generalized by writing $p(\mathbf{y}|\boldsymbol{\mu}_{u|y},\boldsymbol{\theta},\mathcal{H}_1) = N(\boldsymbol{\mu}_{u|y},\theta_0\mathbf{I})$, clf. Eq. (3.22) with $\boldsymbol{\Sigma}_{nn} = \theta_0\mathbf{I}$, in an expanded form and taking logarithms of the both sides. This results in

$$\frac{1}{2\theta_0}\|\mathbf{y}-\boldsymbol{\mu}_{u|y}\|^2 = -\frac{n}{2}\ln(2\pi\theta_0) - \ln p(\mathbf{y}|\boldsymbol{\mu}_{u|y},\mathcal{H}_1). \tag{3.53}$$

Up to a constant, the second term in Eq. (3.52) is the entropy $h(Y|\boldsymbol{\theta},\mathcal{H}_1)$, and when ignoring the terms that do not depend on $\boldsymbol{\theta}$, this leads to Eq. (3.50).

When changing the marginal likelihood of the model $\mathcal{H}_2$ to the unnormalized correlation, Eq. (3.41) suggests the prior

$$p(\boldsymbol{\theta}|\mathcal{H}_2) \propto \exp\left(\frac{1}{2\theta_0^{-1}}\|\boldsymbol{\mu}_{u|y}\|^2 + \frac{1}{2}\ln\left|2\pi(\boldsymbol{\Sigma}_{uu}^{-1}+\theta_0^{-1}\mathbf{I})\right|\right). \tag{3.54}$$

Consulting Eq. (3.29) with $\boldsymbol{\Sigma}_{nn} = \theta_0^{-1}$ shows that $p(\boldsymbol{\mu}_{u|y}|\mathcal{H}_2) = N(\boldsymbol{\mu}_{u|y},\theta_0^{-1}\mathbf{I})$, which allows to rewrite the first term in Eq. (3.54):

$$\frac{1}{2\theta_0^{-1}}\|\boldsymbol{\mu}_{u|y}\|^2 = -\frac{n}{2}\ln(2\pi\theta_0^{-1}) - \ln p(|\boldsymbol{\mu}_{u|y}|\mathcal{H}_2). \tag{3.55}$$

Again noticing that the second term up to a constant is the entropy $h(Y,\boldsymbol{\theta},\mathcal{H}_2)$, and ignoring the terms that do not depend on $\boldsymbol{\theta}$, reveals Eq. (3.51). □

If one uses $\langle \ln p(Y|\boldsymbol{\mu}_{u|y}, \boldsymbol{\theta}, \mathcal{H}_1)\rangle$ instead of $\ln p(\mathbf{y}|\boldsymbol{\mu}_{u|y}, \boldsymbol{\theta}, \mathcal{H}_1)$, the prior $p(\boldsymbol{\theta}|\mathcal{H}_1)$ becomes a uniform prior because the term $\theta_0^{-n/2}$ can then be conveniently introduced to the exponent as the Gaussian entropy $h(X|\theta_0)$. The exponent disappears on the basis of the identity $h(A) = h(B) + h(B|A)$. Likewise, $p(\boldsymbol{\theta}|\mathcal{H}_2)$ becomes a uniform prior when using $\langle \ln p(U|\boldsymbol{\theta}, \mathcal{H}_2)\rangle$ instead of $\ln p(\boldsymbol{\mu}_{u|y}|\boldsymbol{\theta}, \mathcal{H}_2)$ in Eq. (3.51).

If for some reason the terms in the exponents sum to zero, the priors $p(\boldsymbol{\theta}|\mathcal{H}_1) \propto \theta_0^{-n/2}$ and $p(\boldsymbol{\theta}|\mathcal{H}_2) \propto \theta_0^{n/2}$ become Jeffreys' priors for multinomial densities with the parameters θ_0 and θ_0^{-1}, respectively, clf. [Plackett, 1966].

3.5 Better Utilization of Shannon Entropy

Up to irrelevant constants, the logarithms of the determinants of the covariance matrices in Eq. (3.34) and Eq. (3.41) are differential Shannon entropies of the corresponding Gaussian densities. Optimizing these entropies alone can be seen as making an adjustment to the maximization of the marginal likelihood. For example, it is clear from Eq. (3.34) that the first term gives preference to small stopping times. It is therefore expected that the entropy aims to counterbalance this effect if the marginal likelihood does possess the maximum. Thus, the idea of Sporring and Weickert [1999] now gets a proper justification. In summary:

1. Optimizing the entropy with an early stopping can be seen as part of the marginal likelihood maximization.
2. Rényi entropies suggested in [Sporring and Weickert, 1999] should be replaced with the differential Shannon entropies defined over properly introduced random variables, not signals. This change does not solve all the problems related to an 'invariant inference', but at least the need to map signals into positive quantities disappears. Moreover, differential entropies depend only on the covariance (curvature) in the Gaussian case, which, to some extent, allows to maintain invariance w.r.t. a constant component of the signal.
3. Differential Shannon entropies do not depend on the actual signal in a linear case, only on the variance of noise. This is important as the optimal stopping time determines the amount of smoothing which, at least in a linear diffusion case, seems to depend on the level of noise, but not on the actual signal in noise.

Proposition 3.9 (Monotonicity of entropies in a discrete time)**.** *Let the propagator be time-homogeneous, i.e.* $\mathbf{P} = (\mathbf{I} - \mathbf{B})^m$*. The entropy is nondecreasing, i.e.*

$$h(Y|m+1, \mathcal{H}_1) \geq h(Y|m, \mathcal{H}_1), \tag{3.56}$$

provided that the matrix $-\mathbf{B}$ *is positive definite. If the propagator is nonlinear, the following inequality is true:*

$$h(Y|m+1, \mathcal{H}_2) \geq h(Y|m, \mathcal{H}_2), \tag{3.57}$$

provided that each matrix $-\mathbf{B}(\mathbf{u}_m)$ *is positive semidefinite for every* $m \in \mathbb{Z}_+ \cup 0$.

Proof. The time-behavior of the entropy $h(Y|m,\mathcal{H}_1)$ is determined by the term $\ln|2\pi(\boldsymbol{\Sigma}_{uu}+\theta_0\mathbf{I})|$ which, up to irrelevant constants, equals to $-\ln|\mathbf{I}-\mathbf{P}^{-1}|$. Let the eigenvalues $\lambda(-\mathbf{B})$ be denoted as λ_i for $i=1,\ldots,n$. The Taylor series expansion leads to:

$$-\ln|\mathbf{I}-\mathbf{P}_t^{-1}| = -\sum_{i=1}^{n}\ln\left(1-(1+\lambda_i)^{-m}\right) = -\sum_{i=1}^{n}(1+\lambda_i)^{-m} + \text{h.o.t.}\,, \quad (3.58)$$

which follows from $\ln(1-x) = -\sum_{k=1}^{\infty}x^k/k$. Clearly, if the matrix $-\mathbf{B}$ is positive definite, then each $\lambda_i > 0$ and the entropy increases w.r.t. m. If we further assume that the largest term, i.e. $(1+\lambda_{\min})^{-(t+1)}$ with $\lambda_{\min} > 0$, is dominating, the decay of the negative entropy will be exponential in time.

It follows from Eq. (3.41) that, up to irrelevant constants, the entropy $h(Y|\boldsymbol{\theta},\mathcal{H}_2)$ is determined by:

$$\ln|\mathbf{P}| = \sum_{i=1}^{n}\ln(1+\lambda_i)^m = m\sum_{i=1}^{n}\ln(1+\lambda_i). \quad (3.59)$$

Therefore, the entropy $h(Y|\boldsymbol{\theta},\mathcal{H}_2)$ grows linearly in time, and $\mathbf{B}$ is allowed to be singular. The nondecrease of $h(Y|\boldsymbol{\theta},\mathcal{H}_2)$ can be established for a general propagator:

$$\ln|\mathbf{P}| = \sum_{k=0}^{m}\ln|\mathbf{I}-\mathbf{B}(\mathbf{u}_k)| = \sum_{k=0}^{m}\sum_{i=1}^{n}\ln(1+\lambda_i(k))\,. \quad (3.60)$$

Here each eigenvalue $\lambda_i(k) \geq 0$ comes from the set $\lambda(-\mathbf{B}(\mathbf{u}_k))$ and is now time-dependent. The positivity of the eigenvalues guarantees that the term $\ln|\mathbf{P}|$ is nondecreasing, which proves the inequality in Eq. (3.57). □

The smallest eigenvalue $\lambda_{\min}$ determines the nondecrease of the entropy $h(Y|\boldsymbol{\theta},\mathcal{H}_1)$ in the homogeneous case, whereas it is the maximal eigenvalue $\lambda_{\max}$ which affects the entropy $h(Y|\boldsymbol{\theta},\mathcal{H}_2)$. Convenient bounds follow from the Schur theorem discussed in [Marshall and Olkin, 1979]. It states that the eigenvalues of a Hermitian matrix majorize its diagonal elements. As a special case, the following inequalities are true:

$$\lambda_{\min} \leq \min_{i\in\{1,2,\ldots,n\}}(-b_{ii}), \quad \lambda_{\max} \geq \max_{i\in\{1,2,\ldots,n\}}(-b_{ii})\,, \quad (3.61)$$

where b_{ii} are the diagonal elements of $\mathbf{B}$, and they are typically negative.

The ideology of Sporring and Weickert [1999] now gets a proper justification. Utilization of the differential entropy first establishes it as a model complexity measure, and then proves that it is indeed a Liapunov function, whereas the original connection with model selection is rather weak. In the model of Sporring and Weickert [1999], the signal is assumed to be normalized in order to satisfy the constraints of the probability density, which can be written such as $\mathbf{u}_m = \langle\delta(U-\mathbf{u})|\mathbf{y},\boldsymbol{\theta},\mathcal{H}\rangle$. However, the observations must be preprocessed in order to validate this density, and the diffusions are restricted to positive evolutions. In this thesis, $\mathbf{u}_m \equiv \boldsymbol{\mu}_{u|y} \equiv \langle U|\mathbf{y},\boldsymbol{\theta},\mathcal{H}_{1(2)}\rangle$ and $\mathbf{y}$ does not have to be preprocessed.

An insight to the differential Shannon entropies can be stated. The key observation is that the logarithms of determinants can be expressed solely in terms of inner

products. The entropies, as well marginal likelihoods, depend then only on the conditional mean operator. This is relevant to simplifications based on the operator splitting, which would allow one to avoid the nontridiagonal matrices in evaluating the criteria.

Let us borrow well-known identities from [Barry and Pace, 1999]:

$$\ln|\mathbf{I}-\mathbf{A}| = \operatorname{tr}(\ln(\mathbf{I}-\mathbf{A})) = -\sum_{k=1}^{\infty}\frac{\operatorname{tr}(\mathbf{A}^k)}{k}\,, \tag{3.62}$$

This expansion is valid if the spectral radius of the matrix $\mathbf{A}\in\mathbb{R}^{n\times n}$ does not exceed unity.

The computation of the trace can be performed probabilistically:

$$\frac{1}{n}\operatorname{tr}(\mathbf{A}) = \left\langle \frac{X^T\mathbf{A}X}{X^TX}\right\rangle, \quad X\sim N(\mathbf{0},\mathbf{I}). \tag{3.63}$$

Skilling [1993] states that when n is large, e.g. $n=O(10^6)$, even a single point can be useful in the approximate evaluation of trace. This is not very surprising, clf. Section A.2.

The logarithmic term in Eq. (3.35) can be rewritten as:

$$\ln|2\pi(\boldsymbol{\Sigma}_{uu}+\theta_0\mathbf{I})| = n\ln(2\pi\theta_0)+\ln|\mathbf{P}|-\ln|\mathbf{P}-\mathbf{I}|\,, \tag{3.64}$$

$$= n\ln(2\pi\theta_0)-\ln|\mathbf{I}-\mathbf{P}^{-1}|\,. \tag{3.65}$$

The application of Eqs. (3.62) and (3.63) results in:

$$\ln|2\pi(\boldsymbol{\Sigma}_{uu}+\theta_0\mathbf{I})| = n\ln(2\pi\theta_0)+\sum_{k=1}^{\infty}\frac{\operatorname{tr}(\mathbf{P}^{-k})}{k}, \tag{3.66}$$

$$= n\ln(2\pi\theta_0)+n\sum_{k=1}^{\infty}\frac{1}{k}\left\langle\frac{X^T\mathbf{P}^{-k}X}{X^TX}\right\rangle. \tag{3.67}$$

The product matrix-vector product $\mathbf{P}^{-k}X$ is a k times iterated conditional expectation. For example, when $k=3$,

$$\mathbf{P}^{-k}X \equiv \langle U|\,\langle U|\langle U|X,\boldsymbol{\theta},\mathcal{H}_1\rangle,\boldsymbol{\theta},\mathcal{H}_1\rangle\,,\boldsymbol{\theta},\mathcal{H}_1\rangle\,. \tag{3.68}$$

Similar expressions can be provided for the entropy in the model $\mathcal{H}_2$. Eq. (3.27) simplifies the logdet term in Eq. (3.42):

$$\ln|2\pi(\boldsymbol{\Sigma}_{uu}^{-1}+\theta_0^{-1}\mathbf{I})| = n\ln 2\pi\theta_0^{-1}+\ln|\mathbf{P}|\,. \tag{3.69}$$

Writing $\mathbf{P}=\mathbf{I}-(\mathbf{I}-\mathbf{P})$ allows to apply Eq. (3.62), which further restricts the eigenvalues of $\mathbf{P}$ to the interval $(0,2)$. This results in an expression with powers $(\mathbf{I}-\mathbf{P})^k$ which can be expanded by using the binomial theorem:

$$(\mathbf{I}-\mathbf{P})^k = \sum_{m=0}^{k}\frac{k!}{m!(k-m)!}\mathbf{P}^m. \tag{3.70}$$

Disentanglement of traces in Eq. (3.63) leads to:

$$\ln\left|2\pi(\mathbf{\Sigma}_{uu}^{-1}+\theta_0^{-1}\mathbf{I})\right| = n\ln(2\pi\theta_0^{-1}) + \sum_{k=1}^{\infty}\sum_{m=0}^{k}\frac{(k-1)!}{m!(k-m)!}\operatorname{tr}(\mathbf{P}^k), \tag{3.71}$$

$$= n\ln(2\pi\theta_0^{-1}) + n\sum_{k=1}^{\infty}\frac{1}{k}\sum_{m=0}^{k}\frac{(k-1)!}{m!(k-m)!}\left\langle\frac{X^T\mathbf{P}^kX}{X^TX}\right\rangle. \tag{3.72}$$

The product $\mathbf{P}^kX$ is now a successive application of the 'inverse conditional expectation'. Both Eqs. (3.67) and Eqs. (3.72) can be seen as novel definitions of the differential Shannon entropy.

Inner-products also allow to extend the applicability of the marginal likelihood concept with infinite-dimensional systems. For example, one could substitute the inner-product representations of the entropies to marginal likelihoods and then multiply both sides of the equations with '$\partial\Omega/|\Omega| = 1/n$'. A further use of the mean value theorem and replacement of the inner-products with the Hilbert product is one available research program. This aspect goes beyond the scope of this work.

Chapter 4

Computational Analysis

Computational simulations will provide answers to the questions that are otherwise difficult to obtain theoretically. The scale-space will be further narrowed down to discrete edge-preserving diffusions, see e.g. [Weickert et al., 1998].

Section 4.1 explains a synthetic problem that will be analyzed in detail. This is a classical problem of revealing a rectangular pulse corrupted with blur and additive noise.

Section 4.2 studies the difference between a linear diffusion in a discrete and continuous modes of the space and time. It is shown that for achieving a linear dependence of the maximal error on the decreasing size of the smallest spatial element, one needs to perform more iterations than usually expected. A linear decrease of the maximal error w.r.t. the size of the time step τ also requires increasing the size of the grid.

Section 4.3 emphasizes the difference between the von Neumann and Dirichlet boundary conditions when the model is extended beyond the conditional mean to the level of covariances. The covariance matrix can be singular and nonunique and at the same time the conditional expectation and the marginal likelihood may yield a unique well-defined model output.

Section 4.4 further shows how boundary conditions affect the time evolution of a nonlinear diffusion. Several useful-to-know difficulties with nonlinear diffusions are emphasized with two particular examples. The first one shows a slow-down of a diffusion near edges, which indicates that a nonlinear diffusion may take very long simulation times. The second example restates a well-known stair-casing phenomenon, clf. [Esedoglu, 2001, Lassas and Siltanen, 2004].

Section 4.5 provides an example of a successful stopping of a nonlinear diffusion when the noise variance is very large and a linear diffusion is hopeless.

4.1 Basic Setting

It is helpful to work with the simplest possible problem: Smoothing of a binary signal in the additive Gaussian noise. Despite that this does not show all capacity of the scale-spaces, a great advantage of such a setting is its transparency.

Consider a linear univariate diffusion PDE:

$$\begin{cases} \frac{\partial u}{\partial t} = c\frac{\partial^2 u}{\partial x^2} & \text{on } \Omega = [0,1], \\ \frac{\partial u}{\partial x} = 0 & \text{on } \partial\Omega. \end{cases} \tag{4.1}$$

Here $c = \text{const}$. The Fourier analysis helps to develop a semi-analytical solution to this equation as an expansion in Fourier series. For this reason, let us define an initial signal $u(x,0) \equiv u_0 \equiv y$ according to

$$u(x,0) = \frac{a_0}{2} + \sum_{k=1}^{r} a_k \cos(2\pi k x), \tag{4.2}$$

The exact solution to Eq. (4.1) introduces an exponential damping to Eq. (4.2), clf. [Berg and McGregor, 1966]:

$$u(x,t) = \frac{a_0}{2} + \sum_{k=1}^{r} a_k e^{-(2\pi k)^2 ct} \cos(2\pi k x). \tag{4.3}$$

For the sake of simplicity and a further analysis of a nonlinear diffusion it is good to choose an initial signal as the Fourier series of a rectangular pulse of width 0.5 centered in the interval $x \in [0,1]$ as shown in Fig. 4.1. The Fourier coefficients are easy to compute and are given by

$$a_0 = 1,\ a_k = -\frac{2}{\pi k} \sin\frac{\pi k}{2},\ k = 1, 2, \ldots, r. \tag{4.4}$$

Eqs. (4.3) and (4.4) allow to make sure that the discrete model approximates its continuous counterpart in a linear case because an explicit solution is readily available.

It will be assumed that an ideal signal is a rectangular pulse and the observations $u_0 \equiv y$ include two distortions: (i) blur of edges, which is imposed by using a fixed number $r = 20$ of Fourier components, and (ii) ringing near the edges, introduced by the Wilbraham–Gibbs phenomenon. The latter has two meanings: (i) non-disappearance of oscillatory jumps at the discontinuity even when the number of Fourier components is infinite, and (ii) loss of (exponential) approximation accuracy w.r.t. the number of Fourier components away from discontinuity. Notably, the Wilbraham–Gibbs phenomenon takes place when using the least squares projections even with discontinuous basis, e.g. [Foster and Richards, 1991], and one might suspect its appearance even in the use of the random projections of Haupt and Nowak [2006].

In addition to the distortion, the observations will be contaminated with a zero-mean additive white Gaussian noise (AWGN) and the main challenge will be to recover the rectangular pulse from a sample of its noisy observations.

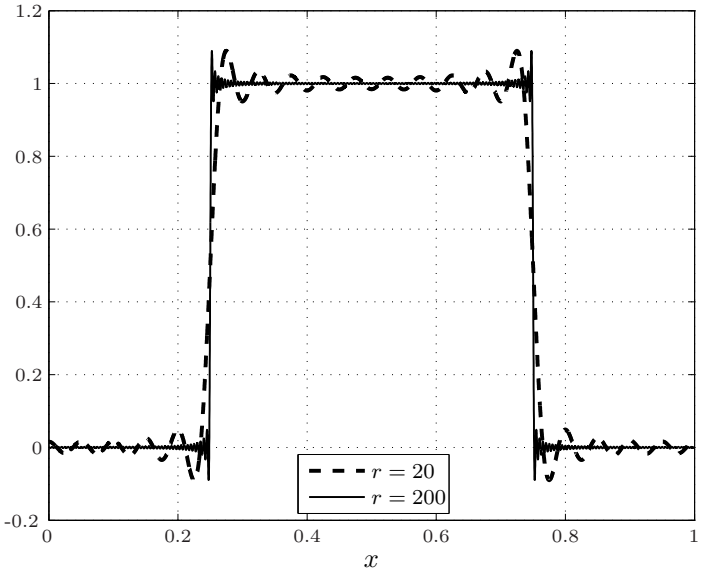


Figure 4.1: A signal specified according to Eqs. (4.3) and (4.4) with $r = 20$ and $r = 200$. The discrepancies caused by the Wilbraham–Gibbs phenomenon are at the locations of jump discontinuities: $x = 0.25$ and $x = 0.75$.

It is well known that both discrete diffusions of Definition 2.1 can be seen as approximate solutions to Eq. (4.1) and the convergence is achieved when the size of the matrix $\mathbf{B}$ tends to infinity and $\tau \to 0$, see e.g. [Babuška et al., 1966]. Linear models, however, cannot achieve a recovery of steep changes in a signal, and therefore, we shall apply discrete diffusions with the following nonlinearity:

$$b_i(\mathbf{u}) \equiv 1 - \exp\left(-\frac{\nu}{(|[\mathbf{Fu}]_i|/\lambda)^s}\right), \quad i = 2, 3, \ldots, n\,. \tag{4.5}$$

Here the matrix $\mathbf{F}$ is the bidiagonal forward-difference operator given by Eq. (2.13). The effect of the nonlinearity can be understood by treating the elements $b_i(\mathbf{u})$ as the values of the function $b(w)$ of a single scalar-valued variable $w \equiv |[\mathbf{Fu}]_i|$. This variable possesses positive values whose magnitude indicates how large the difference between the elements $[\mathbf{u}]_i$ and $[\mathbf{u}]_{i+1}$ is. If this difference is small, w is close to zero and $b(w)$ is close to unity. If the difference is large, $b(w)$ is close to zero. The overall effect is best seen by referring to Eq. (2.17). Clearly, the rows with small elements $b_i(\mathbf{u}_t)$ do not affect the corresponding coordinates of $\mathbf{u}_{t+\tau}$ much. On the other hand, larger values of $b_i(\mathbf{u}_t)$ enforce the averaging of the nearest neighbors. Thus, Eq. (4.5) introduces a spatially non-uniform smoothing: Average less where the change is faster.

A particular value $\nu \geq 0$ can be set up by borrowing some arguments from the analysis in a continuous space and time, but one should know that discrete nonlinear diffusions do not, generally, approximate their continuous counterparts. For example, D'Almeida [2003] chooses the parameter ν in such a way that the derivative of the 'flux' $wb(w)$ changes its sign at $w = \lambda$. The threshold $\lambda \geq 0$ of a

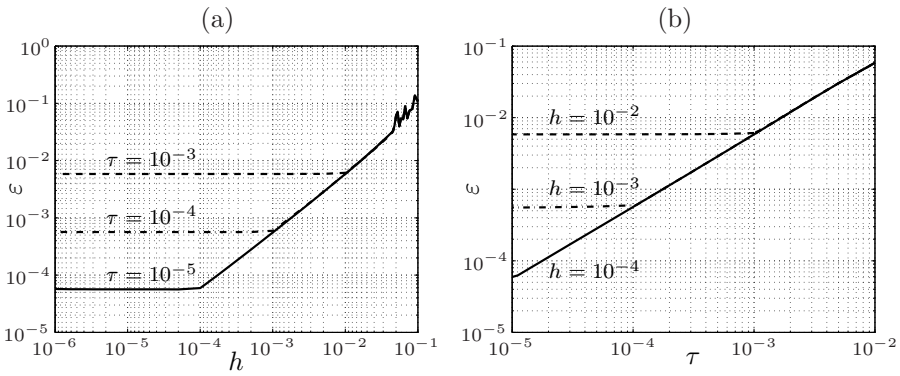


Figure 4.2: Maximal absolute deviation ε vs. the spatial grid resolution h and the time step size τ. Despite the accuracy of the discrete projection of the second derivative operator is $o(h)$, the overall accuracy of the diffusion outcome is $O(h)$ at best. This can be seen in the case (a). Moreover, linear time stepping accuracy $o(\tau)$ holds only until certain resolution thresholds, as can be seen in (b).

'discrete gradient', and the sharpness of nonlinearity $s = 2, 4, \ldots$, need to be set up experimentally. The use of Eq. (4.5) in the context of an optimal diffusion stopping is suggested in [Mrázek and Navara, 2003].

4.2 Choosing Parameter τ

It is well-known that a discrete diffusion with a semi-implicit propagator, clf. Definition 2.1, serves as a stable approximation to a linear continuous diffusion. The error is known to depend linearly on the stepping size τ. Let us now introduce another quantity, the size of the smallest grid element $h = n^{-1}$, which is also an inverse of the size of the Laplacian $\mathbf{B}$. This section will emphasize the fact that the accuracy of the semi-implicit diffusion is sublinear w.r.t. either h or τ when one of them is fixed.

The following study was performed with a linear diffusion of a noiseless signal with $c = 1$, $t = 10^{-2}$, and a variety of the time steps τ and the spatial sizes $h = 1/n$, as summarized in Fig. 4.2. The maximal absolute deviation ε of the solution w.r.t. the ideal signal was chosen as the error measure. In a way, this choice might be too pessimistic, but the worst case analysis allows to better reveal the effects that could be hidden when using average errors. One could also note that when c is constant, 'semi-implicit' becomes equivalent to 'implicit'. The semi-implicit iteration works with any unbounded $\tau > 0$. However, despite that computational requirements to perform a single iteration grow linearly with the size of the problem, one needs more iterations to achieve a linear accuracy $O(h)$. Fig. 4.2a shows that when the problem requires $n = 1/h > 10^4$ grid elements, an intuitively small time step $\tau = 10^{-3}$ yields $\varepsilon = O(10^{-2})$ error. In that case, one has to decrease the time step down to $\tau = 10^{-5}$ to attain $\varepsilon = O(10^{-4})$ error. This requires 100 times longer simulation. Fig. 4.2b shows the same effect w.r.t. the decreasing time constant τ.

In order to obtain $O(\tau)$-accurate maximal error norm, one needs to increase the size of the discrete Laplacian, i.e. decrease the spatial constant h.

Therefore, despite the fact that the semi-implicit iteration is often stated as $O(h)$ accurate w.r.t. the grid resolution h and $O(\tau)$ accurate w.r.t. the time step τ, the overall accuracy in the sense of the maximum error norm has certain barriers: simply decreasing either h or τ when another is kept fixed is not enough.

In summary, the choice of the parameter h is typically dictated by the available resolution of the observation-gathering device. However, the stepping constant τ is a free parameter and its choice is up to the one who applies diffusion models. In the case of a linear diffusion, it is possible to rely on the measure of how closely the discrete model follows its continuous counterpart. In the presence of nonlinearities, the convergence is no longer available, unless one deals with particular nonlinearities and introduces careful regularizations as in [Catté et al., 1992].

Weickert et al. [1998] studied numerically the implications of choosing the value of τ to the preservation of a rotational invariance. Clearly, this is just one aspect of the question: How many iterations should take place? This difficulty can also be related to the choice of a minimal change of the scale in a scale-space. Presently, it must be determined experimentally.

4.3 Singularity of Covariance Matrix $\mathbf{\Sigma}_{uu}$

Before approaching a nonlinear diffusion, it is good to study the impact of the singularity of the Laplacian $\mathbf{B}$ to the singularity of the covariance matrix $\mathbf{\Sigma}_{uu}$. Gaussian conditioning allows singular covariance matrices, but the marginal likelihood criterion is well-defined in a singular case only for the model $\mathcal{H}_2$.

A single iteration of a discrete linear diffusion can be given a probabilistic interpretation by using Eqs. (3.27). Linearity implies the possibility to interpret a single diffusion step as the Gaussian smoothing (regression analysis).

The relationship between the propagator and covariance matrices can be restated:

$$\mathbf{\Sigma}_{uu}^{-1} = \frac{1}{\theta_0}(\mathbf{P}_\theta - \mathbf{I}), \quad \mathbf{P}_\theta \equiv (\mathbf{I} - \tau\mathbf{B})^{-1}. \tag{4.6}$$

It is interesting to inquire what the elements of the propagator $\mathbf{P}$ and the covariance matrix $\mathbf{\Sigma}_{uu}$ look like and how they are affected by the boundary conditions. For this reason, let us introduce the following functions, which are uniformly sampled in $x \in [0, 1]$:

$$k(x_i, x_j) \equiv [\mathbf{\Sigma}_{uu}]_{ij}\,, \tag{4.7}$$

$$g(x_i, x_j) \equiv [\mathbf{P}_\theta]_{ij}\,. \tag{4.8}$$

The covariance functions $k(x_i, x_j)$ can also be referred as kernels, and $g(x_i, x_j)$ are also known as Green's functions. Fig. 4.3 indicates the difference between the kernels read from the matrix $\mathbf{\Sigma}_{uu}$ and discrete Green's functions in the case of a

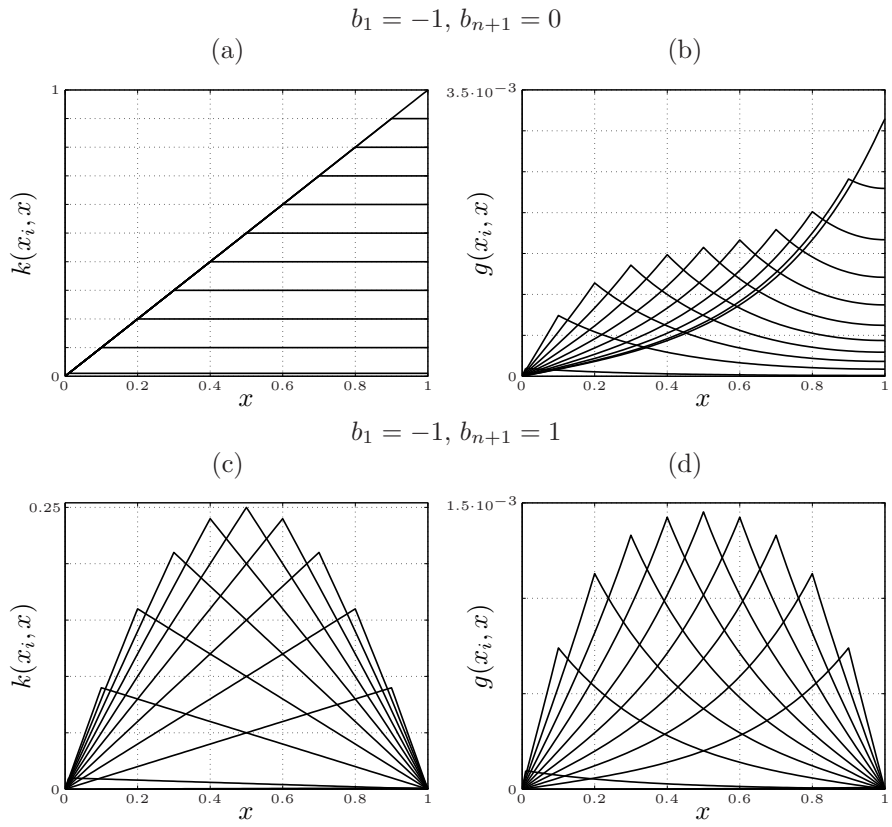


Figure 4.3: Kernel and Green's functions on $\Omega = [0, 1]$: (a),(b) Brownian motion, (c),(d) Brownian bridge. The Green's functions are shown in the case of strong smoothing with $h = 10^{-3}$ and $\tau = 10^{-1}$. They get narrower if τ is smaller. In both cases the relationship between the kernel function and Green's function is unique.

linear diffusion iteration with $\theta_0 = \tau$. Two boundary conditions are considered: (i) the Robin conditions $b_1 = -1$ and $b_{n+1} = 0$, which correspond to Fig. 4.3a,b, and (ii) the Dirichlet conditions $b_1 = -1$ and $b_{n+1} = 1$, which correspond to Fig. 4.3c,d.

The interested reader may check that a single step of a linear diffusion with the boundary conditions $b_1 = -1$ and $b_{n+1} = 0$ corresponds to the Gaussian smoothing with the assumption of Brownian motion, i.e. $k(x_i, x_j) = \min(x_i, x_j)$. The conditions $b_1 = -1$ and $b_{n+1} = 1$ implement the assumption of Brownian bridge, i.e. $k(x_i, x_j) = \min(x_i, x_j) - x_i x_j$. We do not study continuous space-scales in this thesis, but certain constraints are inspired from the diffusions in a continuous world, which deserves to be mentioned. It is, however, essential that the covariance and Green matrices are positive definite and there is a unique relationship between them, given that the value of θ_0 is known.

If one imposes the von Neumann boundary conditions, i.e. $b_1 = b_{n+1} = 0$, the Laplacian $\mathbf{B}$ becomes singular, clf. remarks given below Eq. (2.30). The covariance

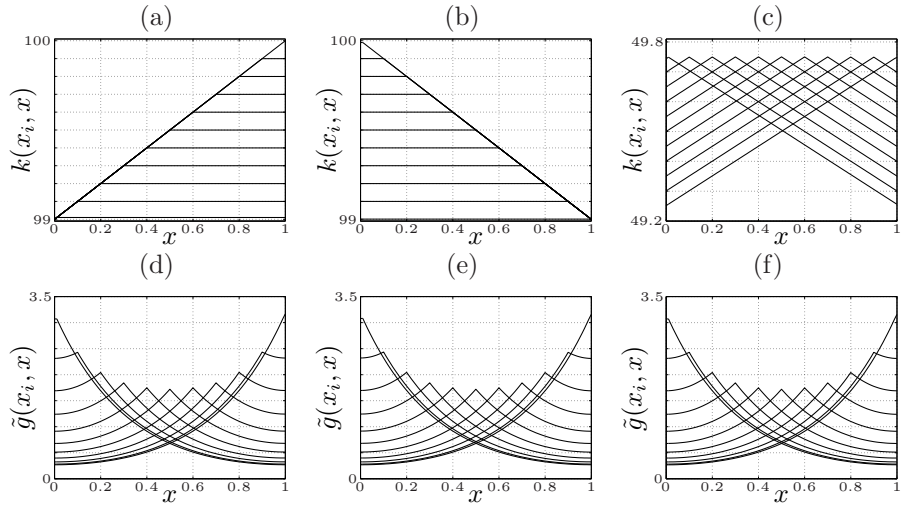


Figure 4.4: Kernels $k(x_i, x)$ and Green's functions $\tilde{g}(x_i, x) = 10^3 \cdot g(x_i, x)$ in the case of the Robin boundary conditions which approach the von Neumann boundary conditions as $\epsilon \to \infty$: (a),(d) $\mathcal{A}_1(\epsilon)$, (b),(e) $\mathcal{A}_2(\epsilon)$, and (c),(f) $\mathcal{A}_3(\epsilon)$. Here $\epsilon = 10^{-2}$. Up to the error of $O(10^{-3})$ all the Green's functions are the same, yet the covariance functions are different.

matrix $\mathbf{\Sigma}_{uu}$ becomes singular too because the propagator has the eigenvalue equal to unity. A single iteration of a linear diffusion has no unique equivalent of the Gaussian smoothing with an AWGN assumption.

Fig. 4.4 shows that a slight perturbation of the von Neumann boundary conditions determines a unique kernel for Green's function. Here we consider three different boundary conditions. The first one is given by:

$$\mathcal{A}_1(\epsilon): \;\; b_1 = h\epsilon(1 - \epsilon - h\epsilon)^{-1}, \;\; b_{n+1} = 0. \tag{4.9}$$

These conditions approach the von Neumann conditions as $\epsilon \to 0$. A specific expression for the element b_1 also ensures that if $\epsilon = 1$, then $b_1 = -1$. Thus, the Robin condition would be 'Dirichlet–von Neumann' when $\epsilon = 1$.

The second conditions bears the same idea, but reverses the conditions for the indices $i = 1$ and $i = n + 1$:

$$\mathcal{A}_2(\epsilon): \;\; b_1 = 0, \;\; b_{n+1} = h\epsilon(1 - \epsilon + h\epsilon)^{-1}. \tag{4.10}$$

The Robin boundary conditions would become 'von Neumann–Dirichlet' for $\epsilon = 1$.

The third condition is made to be symmetrical:

$$\mathcal{A}_3(\epsilon): \;\; b_1 = h\epsilon(1 - \epsilon - h\epsilon)^{-1}, \;\; b_{n+1} = h\epsilon(1 - \epsilon + h\epsilon)^{-1}. \tag{4.11}$$

The Robin boundary conditions are 'Dirichlet–Dirichlet' when $\epsilon = 1$.

The von Neumann boundary conditions, i.e. $\epsilon = 0$, result in a singular kernel matrix, but Green's matrix remains positive definite. Slightly perturbed boundary

conditions yield positive definite kernel matrices with unique Green's matrices. The maximal difference between Green's functions is only $O(10^{-6})$ and, as can be seen in Figs. 4.4d–f, they are nearly indistinguishable. Therefore, it does not matter which of the covariance functions of Fig. 4.4a,b,c is used as an assumption of the Gaussian process in AWGN, the conditioning always gives approximately the same answer.

4.4 Nonlinear Phenomena

This section reveals several peculiarities of a univariate discrete diffusion with the nonlinearity given by Eq. (4.5).

Fig. 4.5 compares the outcomes of linear and nonlinear diffusions with three boundary conditions and at particular time instants. The values of the parameters $\tau = 10^{-3}$ and $h = 10^{-2}$ were chosen in order to maintain convenient computation times while achieving a reasonably small $O(10^{-2})$ maximal error, clf. Fig. 4.2. There is quite a variety of nonlinear diffusions that one can generate by imposing different values of the gradient norm threshold λ and the sharpness of nonlinearity s in Eq. (4.5). Here they are chosen so that a nonlinear diffusion is close to linear. In this way, one can better see both the effect of boundary conditions and the edge preservation.

An essential observation is that the boundary conditions are less significant in a nonlinear case: Blocking of a diffusion across the edge propagates globally. This becomes even more prominent when the sharpness of nonlinearity s is further increased, clf. Fig. 4.6. In this case, the diffusion slow-down near edges appears immediately in a few iterations and the diffusion converges to a nontrivial steady state independently of the boundary conditions. This result indicates that the nonlinear diffusivity cannot be too sharp, otherwise the diffusion does not take place.

Another example can be given to show how a diffusion slow-down near edges produces undesired 'outliers' in the solution. Figs. 4.7a,b show 500 observations of the blurred rectangular pulse which are further contaminated with AWGN of moderately large variance. The gradient threshold $\lambda = 28.3709666$ and the sharpness $s = 8$ are the same as in the experiment summarized in Fig. 4.6. The nonlinear diffusion almost perfectly recovers the edges, but the peculiarities of time evolution are notable. After $t = 100$ iterations the outcome of the nonlinear diffusion produces two outliers close to the edge at $x = 0.75$. This is caused by the fact that the relevant elements $b_i(\mathbf{u}_t)$ are very small in the vicinity of the edges. This particular parameter setting and noise realization dramatically slow down the diffusion in the vicinity of the right edge: It takes 700 iterations more to achieve convergence to the true signal.

Contrary to the previous example of Fig. 4.6, this setting with Gaussian noise is not so sensitive to the value of the gradient threshold λ, but changing the parameters of the time stepping requires adjusting the threshold value. A faster advancement in global time units (but not in the computational speed) can be achieved with $\tau = 0.01$, $t = 2000$ iterations and $\lambda = 35$. The steady state of a constant signal

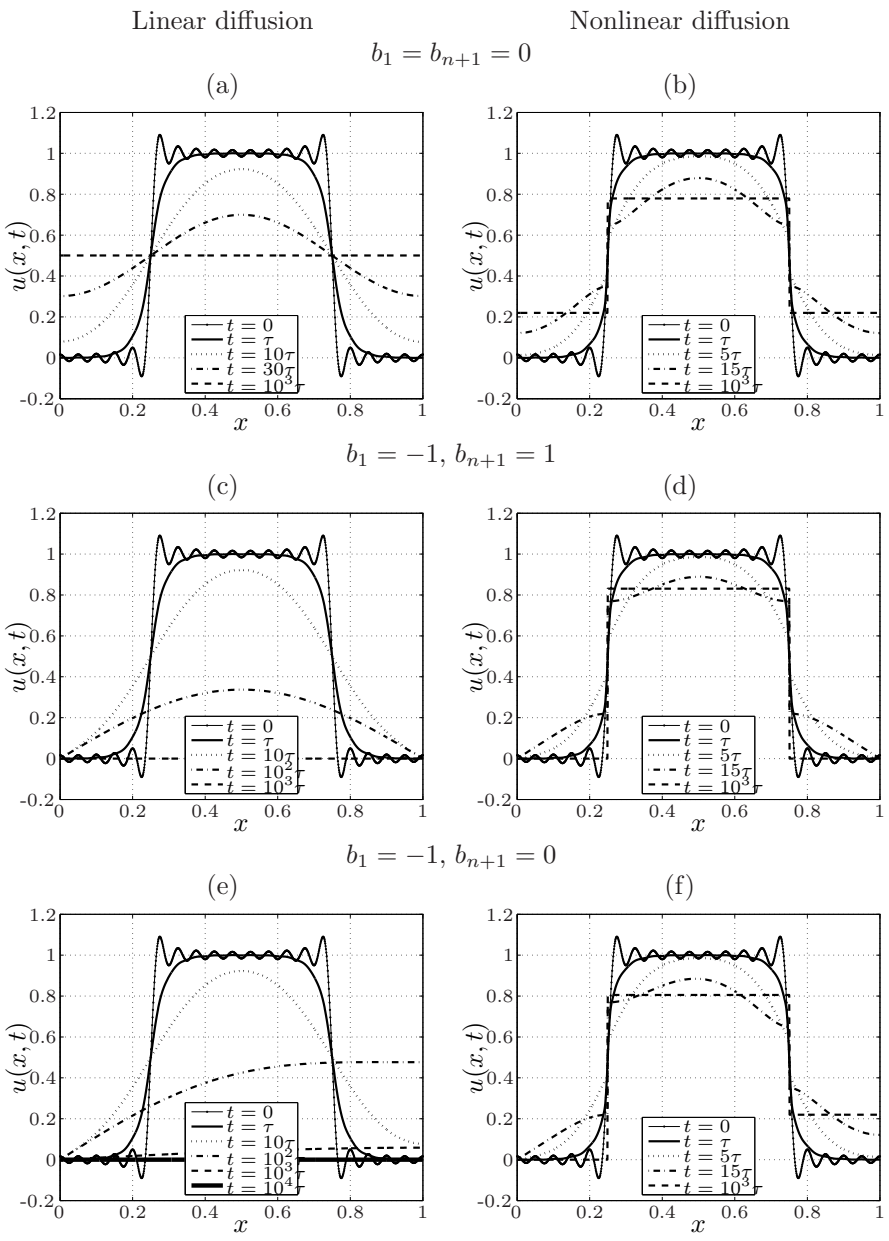


Figure 4.5: Example of linear and nonlinear diffusion filtering: (a),(b) von Neumann, (c),(d) Dirichlet and (d),(e) Robin boundary conditions. Here $\tau = 10^{-3}$ and the nonlinear diffusion is simulated with the gradient norm threshold $\lambda = 28.3709666$ and the sharpness of nonlinearity $s = 8$,clf. Eq. (4.5). The gradient threshold value is critical: $\lambda = 28.3709667$ puts the diffusion into a linear regime. Maintaining a nonlinear diffusion close to its linear counterpart allows to better visualize the dynamics and the asymmetry caused by the boundary conditions.

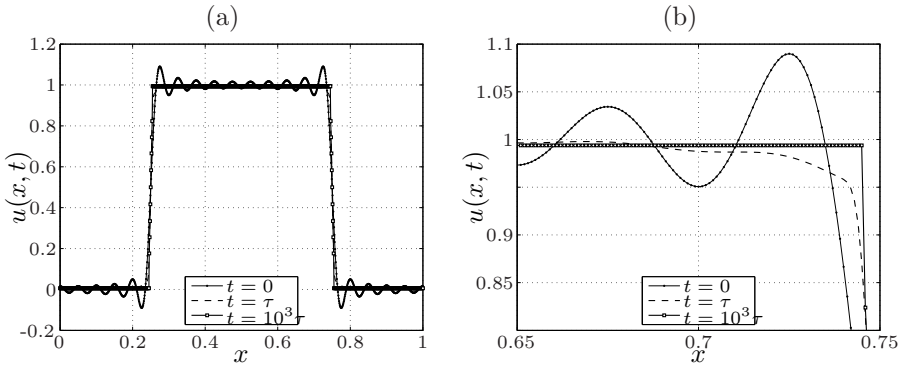


Figure 4.6: The result of nonlinear diffusion filtering (a) and its details (b) when the sharpness of the nonlinearity is increased up to $s = 20$ in Eq. (4.5). This edge-preserving outcome does not depend on the boundary conditions. No smoothing of the gradient was applied.

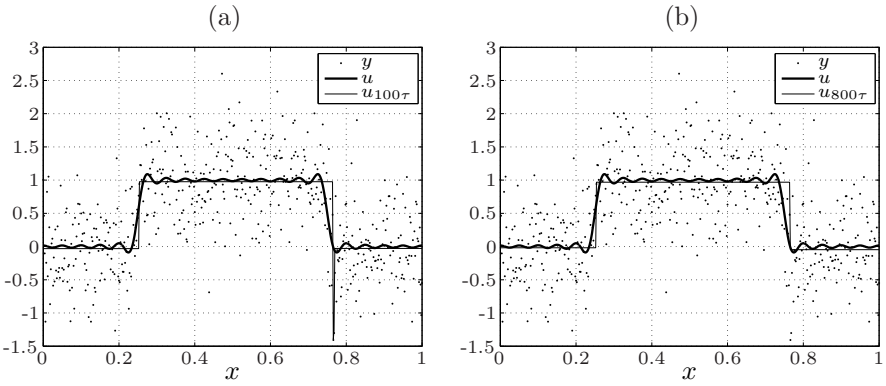


Figure 4.7: A long convergence time to the steady state when $\theta_0 = 0.25$, $\tau = 0.1$, $m = 8$, $\lambda = 28.3709666$. The stopping time (a) $t = 100\tau$ and (b) $t = 800\tau$. It takes 700 iterations more just to smooth two points of the signal correctly. In both cases the sharpness of nonlinearity $s = 8$. No smoothing of the gradient was applied.

emerges after about $t = 10^5$ iterations in this setting.

Finally, it is important to emphasize that a qualitatively different behavior emerges when the size of the matrix $\mathbf{B}$ (a spatial grid size) changes. Figs. 4.8a,b show that in addition to a nontrivial steady state, the appearance of jumps depends on the gradient threshold λ and the spatial grid size n. Significant differences can be seen already when doubling the values, i.e. the parameters vary as $\lambda = 1, 2$ and $n = 500, 1000$. This phenomenon is known as the stair-casing, see e.g. [Esedoglu, 2001, Lassas and Siltanen, 2004]. The effect is shown here to warn the reader who wants to view the nonlinear diffusion as a 'black-box' filter.

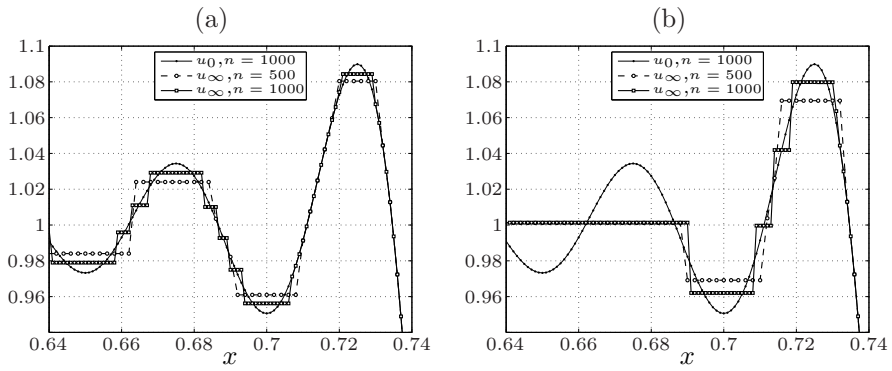


Figure 4.8: The stair-casing phenomenon when (a) $\lambda = 1$ and (b) $\lambda = 2$. Choosing a different size of a grid results in a qualitatively different diffusion.

4.5 Simulations With Large-Variance AWGN

This section demonstrates the ability of an optimally stopped diffusion to recover the edges of a blurred signal which is completely hidden in noise. This setting could be opposed to conventional simulations which utilize the AWGN with 'toy-variance', see e.g. [Osher and Fedkiw, 2003, Haupt and Nowak, 2006]. A relevant real-world example of smoothing nuclear magnetic resonance measurements of the rock strata can be found in [Ruanaidh and Fitzgerald, 1996]. That is a rare example of smoothing real data with jump discontinuities, but it is still a toy problem: Outliers are clearly visible and the noise variance is very small.

A synthetic problem is indicated in Fig. 4.9 where a true signal, whose range is $[0, 1]$, is hardly visible in the noisy values scattered in $[-30, 30]$. During the simulation, n, the number of observations, is one million elements.

After preliminary experiments, the following five stopping criteria were chosen: (i) the marginal likelihood $p(\mathbf{y}|t, \mathcal{H}_2)$, (ii) the entropy $h(Y|t, \mathcal{H}_2)$, (iii) the 'orthogonality' term $(\mathbf{y} - \boldsymbol{\mu}_{u|y})^T \boldsymbol{\mu}_{u|y}$, (iv) the correlation of Mrázek and Navara [2003], and (v) the mean absolute error between the true signal (a rectangular pulse) and the diffusion outcome. The entropy criteria of Sporring and Weickert [1999] have a merit of simplicity, but they have been discarded here for the reason that they require a proper normalization, which is inconvenient as it introduces additional degrees of freedom into the problem. The mean of the absolute error values was chosen as a true measure because the mean of the squared errors has a rather shallow minimum.

The result of the nonlinear diffusion stopped at $t = 10\tau$ is indicated in Figs. 4.10a,b. The locations of the recovered edges are not exactly at $x = 0.25$ and $x = 0.75$, and the levels of the recovered signal do not change from 0 to 1 and vice versa as it should happen in the ideal algorithm. Nevertheless, the result rather well approximates the true signal considering that the noise attains values which are approximately 30 times larger than the true signal. Recovering signals in a very large AWGN is possible only with very dense observation sets, and the nonlinear

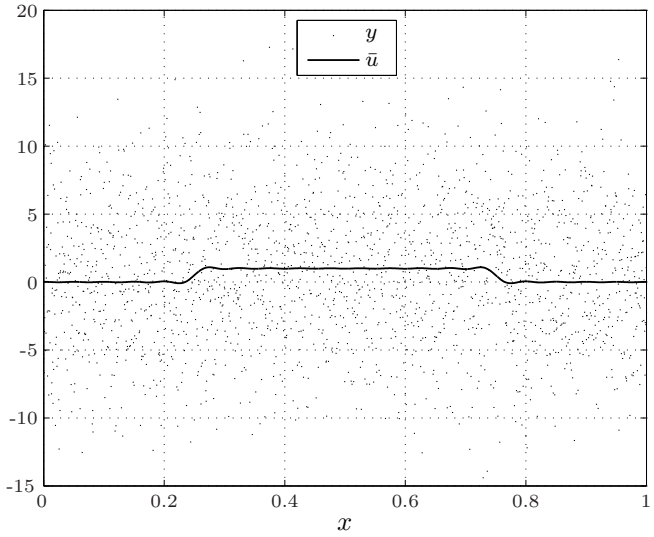


Figure 4.9: A signal specified according to Eqs. (4.3) and (4.4) with $r = 20$ and denoted here as $\bar{u}$, is further contaminated with an AWGN whose variance $\theta_0 = 25$. The goal is to recover a rectangular pulse without loosing the location of its edges. Here only a sample of 2000 noisy observations out of the set of $n = 10^6$ elements is visualized. The actual range of the observations is $[-30, 30]$.

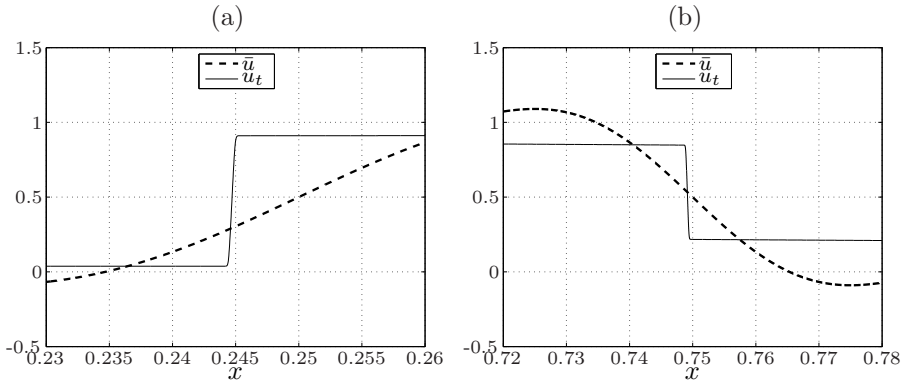


Figure 4.10: Nonlinear diffusion with $n = 10^6$, $\theta_0 = 25$, $\tau = 0.025$, $m = 8$, $\lambda = 200$, $\sigma = 1000$, $t = 10\tau$: (a) restored left front and (b) restored right front. Here $\bar{u}$ denotes a noiseless blurred rectangular pulse obtained by using its Fourier approximation with $r = 20$ components, and u_t is its recovery from the observations of $\bar{u}$ in a very large-variance noise.

diffusion is an especially efficient device to apply as it demands linear computational resources w.r.t. the size of the grid. The selection of the optimal stopping time can thus be automated. However, the choice of the diffusivity function and time stepping parameters requires manual tuning.

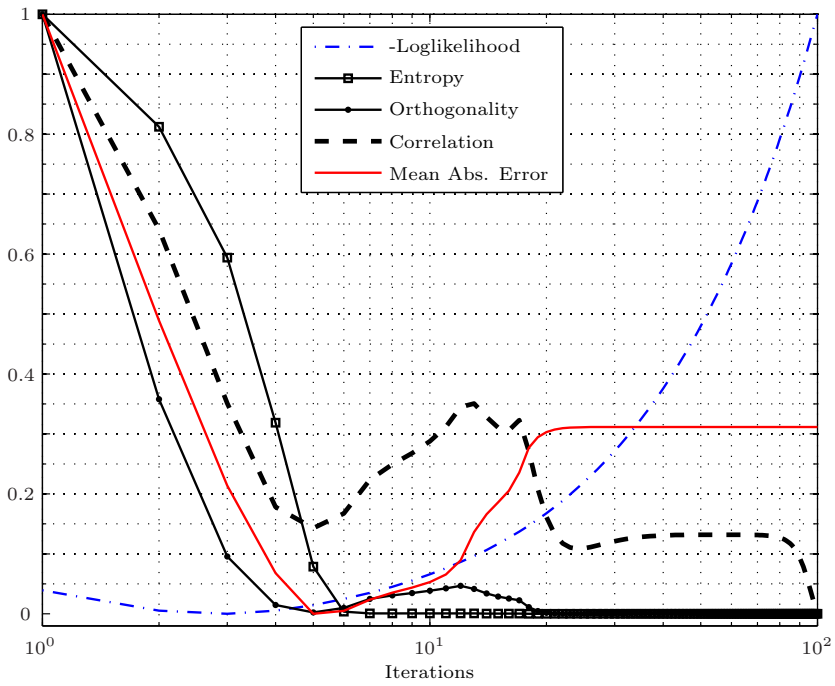


Figure 4.11: Time evolution of the criteria for optimal stopping.

Fig. 4.11 summarizes the evolution of the mean absolute error and the four applied criteria that are useful in determining the optimal stopping time. All criteria are normalized by subtracting their minimal values and dividing them by their range. According to the least mean absolute error criterion, the optimal stopping appears at $t = 5\tau$, but a visual inspection of signals would reveal that the outcome is not very sensitive to the range $t = 4\tau \dots 10\tau$. The maximum likelihood criterion with $t = 3\tau$ underestimates the stopping time, but its simplifications are helpful indeed. The unnormalized correlation estimator, which is simply an inner product $(\mathbf{y} - \mathbf{u})^T \mathbf{u}$, evolves similarly to the correlation used by Mrázek and Navara [2003]. The first distinct minima of both criteria is located at $t = 5\tau$. The entropy criterion reaches the steady state one iteration later, i.e. at $t = 6\tau$.

The evolutionary patterns of all the stopping criteria seem to remain stable during a variety of simulations and produce approximately equivalent behavior of the criteria w.r.t. their optimal stopping times. It is rather notable that the correlation criteria possess multiple minima, and one chooses the minimum with the smallest time value because it is typically the most pronounced. The temporal evolution of the entropy, as a matter of fact, is at least as useful indicator of the optimal stopping (scale) as the correlation criteria.

Chapter 5

Conclusions

In this thesis, it has been shown how the discrete diffusion scale-space can be seen as an incomplete probabilistic statement. The idea of the optimal entropy-based stopping, proposed by [Sporring and Weickert, 1999], and the strategy of decorrelating noise with the model output, suggested by [Mrázek and Navara, 2003], have been unified, and methodologically improved.

One could emphasize that despite the existing general guidelines of statistical inference, see e.g. [Cramér, 1945], [Anderson, 1958] or [Jaynes, 2003], it is always a great challenge to choose a right formalism and design models that achieve both correct predictions and simplicity, i.e. they pass Occam's razor. Of the two, simplicity is more subjective and extremely immune to axiomatic approaches. Let us emphasize the role of variational (energy) representations and Gaussian model space explored in this work, and defend the philosophy of the undertaken path against the opposing arguments.

The use of Gaussian models should not be seen as a statistical limitation. It follows from the desire to keep the inference transparent. This choice expresses the only fact that one prefers to work with symmetric positive definite matrices. The principle of information conditioning then reduces to the application of Schur complements. This is not an assumption, but a choice of convenient mathematical objects, namely, matrices. Probabilistic completion then proceeds from the postulate of the conditional mean towards the determination of the conditional covariance. The thesis has aimed to demonstrate that such a transition is by no means trivial and allows to discover a variety of model selection criteria.

This work has advocated the idea that there is no need, in principle, to consider probabilities explicitly. One seldom knows them *a priori*, can never measure them, and they are definitely very hard to estimate – if their estimation in high-dimensional spaces ever makes any sense. In brief, they represent a certain idealization of knowledge. Total failures of applying 'powerful' inference methods with an inadequately postulated probability space are becoming numerous. The levels of means and covariances already account for uncertainties, where probabilities enter equations implicitly through expectations.

From the model selection perspective, there is no difference between linear or nonlinear, spatiotemporal or completely homogeneous propagators. These modalities represent technical details which must be considered when establishing the stability of the model. They also have an impact on dealing with the non-singularity and positive definiteness of the covariance matrices in Gaussian models. However, the essential difficulty lies in reducing 'information incompleteness'. It is supposed to meaningfully decrease when one moves away from the statement with the conditional expectation towards the descriptions with the conditional covariance, likelihood, entropies, marginal and conditional densities, and finally arrives at an ultimate Utopia with a completely specified joint probability.

According to the strategy presented in Section 3.2, one must first relate the observations with the model outcome. *A posteriori*, when all the signals in the scale-space are known, we are free to remove the dependence of the propagator on the signal. All the matrix entries which have been dependent on the signal at the previous time instant, are simply known.

Examples of the probabilistic models with nonlinear energies can be found in Fessler et al. [2000] and [Champagnat and Idier, 2004]. One first writes a variational energy for the problem and later interprets it as a negative exponent of a hypothetical probability density. The Euler–Lagrange equation of the variational functional can be viewed as a nonlinear transformation which relates the observation variable Y, postulated as a Gaussian random variable, to the diffusion outcome variable U. Fessler et al. [2000] obtain analytical expression for the density of $U|\mathbf{y}$, but it is 'exact' only w.r.t. the postulated nonlinear relation and Gaussianity of Y. The question is whether one should think of the non-Gaussianity of the probability density function, which is a significant mathematical complication, as a necessary tool in a better understanding of the scale-spaces. Gaussian densities are simpler.

Here one could emphasize that there are many ways to complete the scale-space model, and simplicity is probably the only guiding principle. The first problem emerges already at the level of variational interpretations. For example, Newton equations of motion can be seen as Euler–Lagrange equations of many possible Lagrangians. There are many 'least action' principles, and an extension of knowledge to probabilities (stochastic quantization) introduces further degrees of freedom.

However, this does not seem obvious for those who are used to certain invariance principles and always deal with a particular energy (Lagrangian). Moreover, historically, the use of Green's theorem often presented a reformulation with a weaker requirement on the continuity classes of the solutions. Thus, variational formulations are often thought as more 'fundamental'. In order to further stimulate a discussion, consider the following remark of Finlayson and Scriven [1967]:

"Apart from self-adjoint, linear systems, which are comparatively rare, there is no practical need for variational formalism. When approximate solutions are in order the applied scientist and engineer are better advised to turn immediately to direct approximation methods for their problems, rather than search for or try to understand quasi-variational formulations and restricted variational principles."

According to [Finlayson and Scriven, 1967], the true 'energy formulation' should allow one to reduce the continuity requirements. This has an impact on experi-

ments as taking derivatives of measurements always increases errors.

Finlayson and Scriven [1967] were ignorant about the fact that quasi-variational formulations may induce probabilistic models. This is not very surprising because they discussed the variational calculus in the fluid dynamics setting. Contrary to the scale-space models, the 'fluid propagator' is either much more complex, or is simply not explicitly available because of the presence of a more complicated dynamics and constraints. The identifiability of properties such as incompressibility is in jeopardy as the latter does not directly relate to 'smoothing' in the sense of removing additive noise, whereas the scale does.

Seeking for a consistent probability density that meaningfully improves an initially incomplete knowledge of the problem helps to remove the proliferation of seemingly unrelated arguments and an explosion of model selection criteria. This goal is about building shorter, more systematic descriptions of smoothing, which should fit well with the axiomatic nature of research on scale-spaces. The thesis concludes that marginal likelihood inference allows to achieve this goal in the optimal scale selection.

Bibliography

T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley, 1958.

T. Ando. Totally positive matrices. *Lin. Alg. Appl.*, 90:165–219, 1987.

I. Babuška, M. Práger, and E. Vitásek. *Numerical Processes in Differential Equations*. Wiley, 1966.

S. Banerjee and A. E. Gelfand. On smoothness properties of spatial process. *J. of Multiv. Anal.*, 84:85–100, 2003.

J. A. Bangham, P. D. Ling, and R. Harvey. Scale-space from nonlinear filters. *IEEE Trans. PAMI*, 18(5):520–528, 1996.

G. I. Barenblatt, M. Bertsch, R. D. Passo, and M. Ughi. A degenerate pseudoparabolic regularization of a nonlinear forward–backward heat equation arising in the theory of heat and mass exchange in stably stratified turbulent shear flow. *SIAM J. Math. Anal.*, 24(6):1414–1439, 1993.

R. P. Barry and R. K. Pace. Monte Carlo estimates of the log determinant of large sparse matrices. *Lin. Alg. Appl.*, 289:41–54, 1999.

E. F. Beckenbach and R. Bellman. *Inequalities*. Springer–Verlag, 1965.

P. W. Berg and J. L. McGregor. *Elementary Partial Differential Equations*. Holden–Day, 1966.

A. Berman and R. J. Plemmons. *Nonnegative Matrices in the Mathematical Sciences*. Academic, 1979.

M. Bertsch and R. Dal Passo. A parabolic equation with mean-curvature type operator. In *Nonlinear Diffusion Equations and Their Equilibrium States III*, pages 89–97. Birkhäuser, 1989.

M. Bertsch and R. Dal Passo. Hyperbolic phenomena in a strongly degenerate parabolic equation. *Arch. Rat. Mech. Anal.*, 117:349–387, 1992.

S. N. Biswas, K. Datta, R. P. Saxena, P. K. Srivastava, and V. S. Varma. The Hill determinant: an application to the anharmonic oscillator. *Phys. Rev. D*, 4(12): 3617–3620, 1971.

H. W. Bode and C. E. Shannon. A simplified derivation of linear least square smoothing and prediction theory. *Proc. IRE*, 38(4):417–425, 1950.

D. G. Bourgin. The diagram method for determinant expansions. *Am. Math. Month.*, 43(6):344–347, 1936.

D. G. Bourgin. Positive determinants. *Am. Math. Month.*, 46(4):225–226, 1939.

A. S. Carasso. Linear and nonlinear image deblurring: A documented study. *SIAM J. Numer. Anal.*, 36(6):1659–1689, 1999.

F. Catté, P. L. Lions, J. M. Morel, and T. Coll. Image selective smoothing and edge detection by nonlinear diffusion. *SIAM J. Num. Anal.*, 29(1):182–193, 1992.

F. Champagnat and J. Idier. A connection between half-quadratic criteria and EM algorithms. *IEEE Sig. Proc. Lett.*, 11(9):709–712, 2004.

C. A. Charalambides. *Enumerative Combinatorics*. Chapman and Hall/CRC, 2002.

G. Chatterjee. Negative integral powers of a bidiagonal matrix. *Math. Comp.*, 28 (127):713–714, 1974.

M. T. Chu. Symbolic calculation of the trace of the power of a tridiagonal matrix. *Computing*, 35(3–4):257–268, 1985.

R. W. Cox and R. J. Chapman. *Problems and Solutions*, chapter A Toeplitz Matrix Eigenproblem, pages 245–249. SIAM Review, Vol. 37, No. 2, 1995.

H. Cramér. *Mathematical Methods of Statistics*. Almqvist and Wiksells, 1945.

C. G. Cullen and K. J. Gale. A functional definition of the determinant. *Am. Math. Month.*, 72(4):403–406, 1965.

L. D. Landau and E. M. Lifshitz. *Statistical Physics*. Pergamon Press, 1st edition, 1958.

F. D'Almeida. Nonlinear diffusion toolbox. MATLAB Central, 2003.

C. R. Dietrich. Modality of the restricted likelihood for spatial Gaussian random fields. *Biometrika*, 78(4):833–839, 1991.

F. M. Dopico and P. Koev. Bidiagonal decompositions of oscillating systems of vectors. Submitted to Lin. Alg. Appl., 2007.

R. Duits. *Perceptual Organization in Image Analysis: A Mathematical Approach Based on Scale, Orientation and Curvature*. Ph. D. thesis, Technische Universiteit Eindhoven, 2005.

S. Esedoglu. An analysis of the Perona–Malik scheme. *Comm. Pure Appl. Math*, 54(1442–1487), 2001.

S. M. Fallat. Bidiagonal factorizations of totally nonnegative matrices. *Am. Math. Month.*, 108(8):697–712, 2001.

J. A. Fessler, H. Erdoğan, and W. B. Wu. Exact distribution of edge-preserving MAP estimators for linear signal models with Gaussian measurement noise. *IEEE Trans. Im. Proc.*, 9(6):1049–1055, 2000.

B. A. Finlayson and L. E. Scriven. On the search for variational principles. *Int. J. Heat Mass Transfer*, 10:799–821, 1967.

B. Fischer and J. Modersitzki. *Inverse Problems, Image Analysis, and Medical Imaging*, volume 313 of *AMS Contemporary Mathematics*, chapter Fast Diffusion Registration, pages 117–129. 2002.

J. Foster and F. B. Richards. The Gibbs phenomenon for piecewise-linear approximation. *Am. Math. Month.*, 98(1):47–49, 1991.

M. Foster. An application of the Wiener–Kolmogorov smoothing theory to matrix inversion. *J. of SIAM*, 9(3):387–392, 1961.

F. R. Gantmacher. *The Theory of Matrices*, volume 1,2. Chelsea, 1959.

F. R. Gantmacher and M. G. Krein. *Oscillation Matrices and Kernels and Small Vibrations of Mechanical Systems*. AMS, revised edition edition, 2002.

I. M. Gelfand and S. V. Fomin. *Calculus of Variations*. Prentice-Hall, 1963.

I. M. Gelfand and A. M. Yaglom. Integration in functional spaces and its applications in quantum physics. *J. Math. Phys.*, 1(1):48–69, 1960.

B. Gidas. A renormalization group approach to image processing problems. *IEEE Trans. PAMI*, 11(2):164–180, 1989.

G. Gilboa, N. Sochen, and Y. Y. Zeevi. Estimation of optimal PDE-based denoising in the SNR sense. *IEEE Trans. Im. Proc.*, 15(8):2269–2280, 2006.

K. Glashoff and H. Kreth. Sign-stability in difference schemes for parabolic initinal-boundary value problems. *Numerische Mathematik*, 35(3):343–354, 1980.

A. Harten. High resolution schemes for hyperbolic conservation laws. *J. Comp. Phys.*, 49:357–393, 1983.

J. Haupt and R. Nowak. Signal reconstruction from noisy random projections. *IEEE Trans. Inf. Th.*, 52(9):4036–4048, 2006.

I. I. Hirschman and D. V. Widder. *The Convolution Transform*. Princeton Univ. Press, 1955.

R. Horváth. *Qualitative Properties of Numerical Solutions of the Heat Conduction Equation*. Ph. d. thesis, Univ. of West Hungary, 2000.

R. Hummel and R. Moniot. Reconstructions from zero crossings in scale space. *IEEE Trans. ASSP*, 37(12):2111–2130, 1989.

L. I. Ignat. Qualitative properties of a numerical scheme for the heat equation. In *Proc. Eur. Conf. Numer. Math. Adv. Appl.*, pages 593–600. Springer–Verlag, 2005.

C. Itzykson and J. B. Zuber. *Quantum Field Theory*. Mcgraw-Hill, 1980.

E. T. Jaynes. *Probability Theory : The Logic of Science*. Cambridge Univ. Press, 2003.

W. P. Johnson. The curious history of Faà di Bruno's formula. *Am. Math. Month.*, 109(3):217–234, 2002.

S. Karlin. *Total Positivity*, volume 1. Stanford Univ. Press, 1968.

H. Kawasaki. A conjugate points theory for a nonlinear programming problem. *SIAM J. Control Optim.*, 40(1):54–63, 2001.

B. Kawohl. Some qualitative properties of nonlinear partial differential equations. In *Nonlinear Diffusion Equations and Their Equilibrium States II*, pages 19–31. Springer–Verlag, 1988.

B. Kawohl and N. Kutev. Maximum and comparison principle for one-dimensional anisotropic diffusion. *Mathematische Annalen*, 311:107–123, 1998.

S. L. Keeling. Total variation based convex filters for medical imaging. *Appl. Math. Comp.*, 139:101–119, 2003.

O. D. Kellogg. Orthogonal function sets arising from integral equations. *Am. J. Math.*, 40(2):145–154, 1918.

A. N. Kolmogorov. *Foundations of the Theory of Probability*. Chelsea, 2nd (1st in 1950, original in german (1933)) edition, 1956.

A. N. Kolmogorov and S. V. Fomin. *Elements of the Theory of Functions and Functional Analysis*. Graylock Pr, 1957.

A. Kuijper. *The deep structure of Gaussian scale space images*. Ph. D. thesis, Universiteit Utrecht, 2002.

O. A. Ladyzhenskaja, V. A. Solonnikov, and N. N. Uralceva. *Linear and Quasi-linear Equations of Parabolic Type*. AMS, 1968.

M. Lassas and S. Siltanen. Can one use total variation prior for edge-preserving Bayesian inversion? *Inverse Problems*, 20:1537–1563, October 2004.

M. Lejeune and G. D. Faulkenberry. A simple predictive density function. *J. Am. Stat. Assoc.*, 77(379):654–657, 1982.

T. Leonard. A simple predictive density function. *J. Am. Stat. Assoc.*, 77(379): 657–658, 1982.

M. Lewin. Totally non-negative, M- and Jacobi matrices. *SIAM J. Algebraic and Discrete Methods*, 1(4):419–421, 1980.

T. Lindeberg. Scale-space for discrete signals. *IEEE Trans. PAMI*, 12(3):234–254, 1990.

T. Lindeberg. *Discrete Scale-Space Theory and the Scale-Space Primal Sketch*. Ph. d. Thesis, Royal Inst. of Techn., Stockholm, 1991.

J. R. Magnus and H. Neudecker. *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Wiley, 2nd edition, 1999.

T. L. Markham. Nonnegative matrices whose inverses are M-matrices. *Proc. Am. Math. Soc.*, 36(2):326–330, 1972.

A. W. Marshall and I. Olkin. *Inequalities: Theory of Majorization and Its Applications*. Academic Press, 1979.

M. Maruyama, F. Girosi, and T. Poggio. A connection between GRBF and MLP. Tech. Rep. AI-1291, MIT, 1992.

I. Matsuba. Renormalization group approach to hierarchical image analysis. In *Proc. ICASSP*, volume 2, pages 1044–1047, 1988.

E. S. Meinel. Origins of linear and nonlinear recursive restoration algorithms. *J. Opt. Soc. Am. A*, 3(6):787–799, 1986a.

E. S. Meinel. Maximum-entropy image restoration: Lagrange and recursive techniques. *J. Opt. Soc. Am. A*, 5(1):25–29, 1986b.

T. P. Minka. Deriving quadrature rules from Gaussian processes. Tech. rep., Carnegie Mellon University, 2000.

P. Mrázek and M. Navara. Selection of optimal stopping time for nonlinear diffusion filtering. *Int. Journal of Computer Vision*, 52(2):189–203, 2003.

P. Mrázek, J. Weickert, and G. Steidl. Diffusion-inspired shrinkage functions and stability results for wavelet denoising. *Int. Journal of Computer Vision*, 64(2–3): 171–186, 2005.

S. Osher and R. Fedkiw. *Level Set Methods and Dynamic Implicit Surfaces*, volume 153 of *Applied Mathematical Sciences*. Springer, 2003.

J. M. Peña. M-matrices whose inverses are totally positive. *Lin. Alg. Appl.*, 221: 189–193, 1995.

P. Perona and J. Malik. Scale space and edge detection using anisotropic diffusion. In *IEEE Workshop on Comp. Vis.*, pages 16–22, 1987.

A. Pinkus. *Total Positivity and its Applications*, chapter Spectral Properties of Totally Positive Kernels and Matrices, pages 477–511. Kluwer Academic Publishers, 1996.

R. L. Plackett. A reduction formula for normal multivariate integrals. *Biometrika*, 41(3,4):351–360, 1954.

R. L. Plackett. Current trends in statistical inference. *J. R. Stat. Soc. A*, 129(2): 249–267, 1966.

R. Price. A useful theorem for nonlinear devices having Gaussian inputs. *IRE Trans. Inf. Th.*, 4(2):69–72, 1958. See also 10(2):171, 1964.

I. Prigogine. *Introduction to thermodynamics of irreversible processes*. Interscience, 3rd edition, 1967.

H. Robbins. Some thoughts on empirical Bayes estimation. *Ann. Stat.*, 11(3): 713–723, 1983.

D. J. Rose. An algorithm for solving a special class of tridiagonal systems of linear equations. *Comm. ACM*, 12(4):234–236, 1969.

P. Rosenau. Free-energy functionals at the high-gradient limit. *Phys. Rev. A*, 41 (4):2227–2230, 1990.

P. Rosenau. Tempered diffusion: A transport process with propagating fronts and inertial delay. *Phys. Rev. A*, 46(12):7371–7374, 1992.

P. Rosenau, P. S. Hagan, R. L. Northcutt, and D. S. Cohen. Delayed diffusion due to flux limitation. *Phys. Lett. A*, 142(1):26–30, 1989.

J. L. Roux. English translation of 'L. Boltzmann, über die..., Sitzb. d. Kaiserlichen Akademie der Wissenschaften, Mathematisch-Naturwissen Cl. LXXVI, Abt. II, 1877, pp. 373-435.'. Tech. rep., Université de Nice Sophia Antipolis, 2002.

J. J. K. Ó. Ruanaidh and W. J. Fitzgerald. *Numerical Bayesian Methods Applied to Signal Processing*. Springer, 1996.

D. Salomé. *Statistical inference via fiducial methods*. Ph. d. thesis, Rijksuniversiteit Groningen, 1998.

C. E. Shannon. A mathematical theory of communication. *Bell Sys. Tech. J.*, 27: 379–423, 1948.

J. Skilling. Bayesian numerical analysis. In *Physics and Probability*, pages 207–221. Cambridge Univ. Press, 1993.

J. Sporring and J. Weickert. Information measures in scale spaces. *IEEE Trans. Inf. Theory*, 45(3):1051–1058, 1999.

S. Teboul, L. Blanc-Féraud, G. Aubert, and M. Barlaud. Variational approach for edge-preserving regularization using coupled PDE's. *IEEE Trans. Im. Proc.*, 7 (3):387–397, 1998.

R. Vein and P. Dale. *Determinants and Their Applications in Mathematical Physics*. Springer, 1998.

J. Von Neumann. *Mathematical Foundations of Quantum Mechanics*. Princeton Univ. Press, 1955.

J. Weickert. *Anisotropic diffusion in image processing*. Ph. d. thesis, Kaiserslautern Univ., 1996.

J. Weickert and B. Benhamouda. A semidiscrete nonlinear scale-space theory and its relation to the Perona–Malik paradox. In *Adv. Comp. Vis.*, pages 1–10. Springer, 1997.

J. Weickert, B. M. ter Haar Romeny, and M. A.Viergever. Efficient and reliable schemes for nonlinear diffusion filtering. *IEEE Trans. on Image Processing*, 7 (3):398–410, March 1998.

J. Weickert, S. Ishikawa, and A. Imiya. Linear scale-space has first been proposed in Japan. *J. Math. Imag. Vision*, 10(3):237–252, 1999.

Appendix A

More Technical and Philosophical Findings

A.1 Novel Method of Evaluating Determinants

It can be seen from Chapter 2 that determining singularity and positivity of a quadratic form is a fundamental problem that has a lot of implications to physics and engineering. It is important to work with the concept of matrix minor (determinant), which provides well-known conditions of necessity and sufficiency.

This section contributes with a novel way of evaluating matrix determinants. An inspiration comes from Section 2.5 which states an open problem for characterizing discrete multivariate diffusions which are sign variation diminishing. This calls for methods of evaluating minors with nontrivial sparsity patterns. The essential discoveries of this chapter do not solve this problem, but it is believed that this could be a helpful path.

The idea is to first reduce the matrix to the one whose structure is well-known, i.e. its determinant and inverse can be expressed explicitly. The principal result is the expansion given in Proposition A.2, which should be seen as an improvement to the method of Bourgin [1936]. Everything before Proposition A.2 should be considered as an introductory text, and the material that follows it is novel.

The method should be useful when the application of Taylor series makes sense. Thus, the focus is not on the efficiency of a numerical evaluation of determinants of large matrices. Instead, the aim is to provide simple approximate expansions for nontrivial quantities which are of relevance to the diffusion scale-spaces.

Bourgin [1936] noticed an original way to reveal explicit dependencies of the matrix determinant on its specific elements. If arbitrary matrix elements are given multipliers ν, thereby resulting in the matrix $\mathbf{A}(\nu)$ with an obvious relationship

$\mathbf{A} \equiv \mathbf{A}(1)$, the determinant can be approached via Taylor series:

$$|\mathbf{A}(\nu)| = |\mathbf{A}(0)| + \sum_{m=1}^{n} \frac{\nu^m}{m!} \frac{\partial^m |\mathbf{A}(0)|}{\partial \nu^m}\,, \tag{A.1}$$

where it should be obvious that $\partial^m |\mathbf{A}(0)|/\partial\nu^m$ is $\partial^m |\mathbf{A}(\nu)|/\partial\nu^m$ at $\nu = 0$.

When applying the product-derivative rule to the determinant, written as a sum of signed permutations, the derivative of a determinant can be expressed as:

$$\frac{\partial |\mathbf{A}(\nu)|}{\partial \nu} = \sum_{i=1}^{n} \begin{vmatrix} a_{1,1} & a_{1,2} & \dots & a_{i-1,n} \\ \vdots & \vdots & & \vdots \\ a_{i-1,1} & a_{i-1,2} & \dots & a_{i-1,n} \\ a'_{i,1} & a'_{i,2} & \dots & a'_{i,n} \\ a_{i+1,1} & a_{i+1,1} & \dots & a_{i+1,1} \\ \vdots & \vdots & & \vdots \\ a_{n,1} & a_{n,2} & \dots & a_{n-1,n} \end{vmatrix}, \quad a'_{i,1} \equiv \frac{\partial a_{i,1}}{\partial \nu}\,. \tag{A.2}$$

A specific multiplication pattern in $\mathbf{A}(\nu)$ was called a diagram, and, in words of Bourgin [1936]: "*The diagram provides a classification and suggests a direction for further systematic investigation of determinants and their expansions*". Apparently, this idea has not been much investigated. A direct way of taking derivatives produces $\sum_{k=0}^{n} n^k$ of such terms, which already equals to 40 when $n = 3$.

Whenever the group of elements in a matrix can be identified as a single parameter, Eq. (A.1) can produce a simple approximation, an exact analytical expression, an insight to symmetry, a novel inequality, or a less usual presentation of an already existing identity.

It will prove to be useful to apply the Faà di Bruno formula for the derivative of the composite function. The formula allows to express the Taylor series of a given function via the Taylor series of a simpler one.

Proposition A.1 (The Faà di Bruno formula)**.** *Assuming that all functions are differentiable a sufficient number of times, the m-th derivative of the composite function $h(t) = g(f(t))$ at $t = a$ equals to, clf. [Charalambides, 2002, p. 420],*

$$h_m = \sum_{k=0}^{m} g_k B_{m,k}(f_1, f_2, \dots, f_{m-k+1}), \tag{A.3}$$

where

$$h_m \equiv \frac{\partial^m h(t)}{\partial t^m}\Big|_{t=a}, \quad g_m \equiv \frac{\partial^m g(w)}{\partial w^m}\Big|_{w=f(a)}, \quad f_m \equiv \frac{\partial^m f(t)}{\partial t^m}\Big|_{t=a}, \tag{A.4}$$

and $B_{m,k}$ is the exponential partial Bell polynomial.

When g_k does not depend on k, the sum of exponential partial Bell polynomials yields the exponential Bell polynomial $B_m(\mathbf{x})$, which can be expressed as the

determinant of the Hessenberg matrix, see e.g. [Johnson, 2002]:

$$\begin{aligned} B_m(\mathbf{x}) &\equiv B_m(x_1, x_2, \ldots, x_m) \\ &\equiv \sum_{k=0}^{m} B_{m,k}(x_1, x_2, \ldots, x_{m-k+1}) \\ &= \begin{vmatrix} \binom{m-1}{0}x_1 & -1 & 0 & \ldots & 0 & 0 \\ \binom{m-1}{1}x_2 & \binom{m-2}{0}x_1 & -1 & \ldots & 0 & 0 \\ \binom{m-1}{2}x_3 & \binom{m-2}{1}x_2 & \binom{m-3}{0}x_1 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \binom{m-1}{m-2}x_{m-1} & \binom{m-2}{m-3}x_{m-2} & \binom{m-3}{m-4}x_{m-3} & \ldots & \binom{1}{0}x_1 & -1 \\ \binom{m-1}{m-1}x_m & \binom{m-2}{m-2}x_{m-1} & \binom{m-3}{m-3}x_{m-2} & \ldots & \binom{1}{1}x_2 & \binom{0}{0}x_1 \end{vmatrix}. \end{aligned} \tag{A.5}$$

The recurrence between the successive principal minors of a general Hessenberg matrix is known. If $\mathbf{A} \in \mathbb{R}^{n\times n}$ is such that $a_{ij} = 0$ for $j - i > 1$, then the identities in [Vein and Dale, 1998, p. 91] along with $|\mathbf{A}| = |\mathbf{A}^T|$ yield:

$$A(\gamma_{1:n}|\gamma_{1:n}) = (-1)^{n-1}\sum_{k=0}^{n-1}(-1)^k a_{n,k+1}\prod_{s=k+1}^{n-1} a_{s,s+1}\; A(\gamma_{1:k}|\gamma_{1:k}), \tag{A.6}$$

where $A(\gamma_{1:0}|\gamma_{1:0}) \equiv 0$, and $\prod_{s=k+1}^{n-1} a_{s+1,s} \equiv 1$ for $k = n-1$.

It deserves to be mentioned that Eq. (A.6) is a special case of the Jacobi equation whose general form has been stated by Kawasaki [2001]. Beautiful connections of this discrete equation to Jacobi's conjugate point theory, the Riccati equation and 'global analysis' were first discussed by Gelfand and Yaglom [1960], and Gelfand and Fomin [1963].

The essential mathematical result of this section can now be established. A remaining material will provide some of its applications.

Proposition A.2 (Main identity)**.** *Given* $\mathbf{A} \in \mathbb{R}^{n\times n}$*, multiply any of its elements by* ν*. The following is true:*

$$|\mathbf{A}(\nu)| = |\mathbf{A}(0)|\sum_{m=0}^{n}\frac{\nu^m}{m!}B_m(\boldsymbol{x}), \tag{A.7}$$

where $B_0(\boldsymbol{x}) \equiv 1$ *and the components of* $\boldsymbol{x}$ *are:*

$$x_i = (-1)^{i-1}(i-1)!\,\mathrm{tr}\left(\left(\mathbf{A}^{-1}(0)\frac{\partial\mathbf{A}(0)}{\partial\nu}\right)^i\right), \quad i = 1, 2, \ldots, n. \tag{A.8}$$

Proof. The cofactor expansion leads to a well-known formula, see e.g. [Magnus and Neudecker, 1999]:

$$\frac{\partial|\mathbf{A}(\nu)|}{\partial\nu} = |\mathbf{A}(\nu)|\,\mathrm{tr}\left(\mathbf{A}^{-1}(\nu)\frac{\partial\mathbf{A}(\nu)}{\partial\nu}\right). \tag{A.9}$$

Here one may suspect that the logarithm of the determinant is more amenable to successive differentiation. The derivatives of the higher order require evaluating

$\partial \mathbf{A}^{-1}(\nu)/\partial\nu$. Applying the product derivative rule to $\mathbf{A}(\nu)\mathbf{A}^{-1}(\nu) = \mathbf{I}$ results in:

$$\frac{\partial \mathbf{A}^{-1}(\nu)}{\partial \nu} = -\mathbf{A}^{-1}(\nu)\frac{\partial \mathbf{A}(\nu)}{\partial \nu}\mathbf{A}^{-1}(\nu). \tag{A.10}$$

A successive differentiation of the log-determinant with the use of Eq. (A.10) yields:

$$\frac{\partial^i \ln|\mathbf{A}(\nu)|}{\partial \nu^i} = (-1)^{i-1}(i-1)!\,\mathrm{tr}\left(\left(\mathbf{A}^{-1}(\nu)\frac{\partial \mathbf{A}(\nu)}{\partial \nu}\right)^i\right), \quad i \in \mathbb{Z}_+. \tag{A.11}$$

At this stage, it is useful to conclude the following:

1. There is a finite number of nonzero derivatives $\partial^i|\mathbf{A}(\nu)|/\partial\nu^i$ in the Taylor series of $\mathbf{A}(\nu)$, but they are not easily computable.
2. There is an infinite number of nonzero terms $\partial^i|\ln \mathbf{A}(\nu)|/\partial\nu^i$ in the Taylor series of $\ln \mathbf{A}(\nu)$, but each of them is easy to compute.

The suggestion is to expand each derivative $\partial^i|\mathbf{A}(\nu)|/\partial\nu^i$ in $\partial^k|\ln \mathbf{A}(\nu)|/\partial\nu^i$ with $k = 0, 1, \ldots i$ by means of the Faà di Bruno formula. Applying $h(t) = e^{f(t)}$ with Eq. (A.3) expresses the derivatives of $h(t)$ as the functions whose arguments are the derivatives of $f(t) = \ln h(t)$. In this case, $g(w) = e^w$ and $g_k(w) = g(w)$ for all k. One may set $h(\nu) = |\mathbf{A}(\nu)|$, and $g_k = g(f(0)) = |\mathbf{A}(0)|$ is constant. Therefore,

$$\frac{\partial^m|\mathbf{A}(\nu)|}{\partial\nu^m} = |\mathbf{A}(0)|\sum_{k=0}^{m} B_{m,k}(x_1, x_2, \ldots, x_{m-k+1}), \tag{A.12}$$

where the quantities x_i $(i = 1, 2, \ldots, m)$ are provided by Eq. (A.11). A substitution to Eq. (A.12) establishes the result. □

Computation of $|\mathbf{A}|$ reduces to the computation of $|\mathbf{A}(0)|$ and the trace in Eq. (A.8), which largely depends on the chosen zero pattern. There are 2^{n^2-1} such patterns when $\mathbf{A} \in \mathbb{R}^{n\times n}$, which gives a freedom in explorations.

Another invariance property emerges from Eq. (A.10):

$$\mathbf{A}(\nu)\frac{\partial \mathbf{A}^{-1}(\nu)}{\partial \nu} = -\frac{\partial \mathbf{A}(\nu)}{\partial \nu}\mathbf{A}^{-1}(\nu). \tag{A.13}$$

This identity allows to obtain a logical statement. Consider the matrix $\mathbf{B} \in \{0,1\}^{n\times n}$ which consists of only binary elements, and let $\mathbf{B}_c \in \{0,1\}^{n\times n}$ denote its complement in the sense that the zero-unity pattern of $\mathbf{B}$ corresponds to the unity-zero pattern in $\mathbf{B}_c$. Setting $\nu = 0$ in Eq. (A.13) then indicates that

$$(\mathbf{A} \circ \mathbf{B})(\mathbf{A}^{-1} \circ \mathbf{B}_c) = -(\mathbf{A} \circ \mathbf{B}_c)(\mathbf{A}^{-1} \circ \mathbf{B}). \tag{A.14}$$

Thus, Proposition A.2 states that any matrix and its inverse complement each other when contributing to the matrix determinant. Given that there is at least one multiplier, there can be no elements of the matrix and its inverse present in the expansion such that they would share the same set of indices.

As an example, consider one of the simplest diagrams (ν-patterns):

$$\mathbf{A}(\nu) = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & \nu a_{nn} \end{pmatrix}. \tag{A.15}$$

When writing the first equations for $s = 1, 2, 3, \ldots$, one may notice that

$$\left(\mathbf{A}(0)^{-1}\frac{\partial \mathbf{A}(0)}{\partial \nu}\right)^s = \begin{pmatrix} \mathbf{0}_{n\times(n-1)} & \boldsymbol{\Delta}^s_{n\times 1} \end{pmatrix}, \tag{A.16}$$

where the elements of the vector $\boldsymbol{\Delta}^s$ are:

$$\Delta_i^s \equiv [\mathbf{A}^{-1}(0)]_{i,n} a_{nn} \left([\mathbf{A}^{-1}(0)]_{nn} a_{nn}\right)^{s-1} \tag{A.17}$$

The application of Proposition A.2 allows to expand $|\mathbf{A}|$ as a multivariate polynomial of the variables $x_s = (-1)^{s-1}(s-1)!\Delta_n^s$:

$$B_1 = x_1 = \Delta_n, \tag{A.18}$$

$$B_2 = \begin{vmatrix} x_1 & -1 \\ x_2 & x_1 \end{vmatrix} = x_1^2 + x_2 = \Delta_n^2 - \Delta_n^2 = 0, \tag{A.19}$$

$$B_3 = \begin{vmatrix} x_1 & -1 & 0 \\ 2x_2 & x_1 & -1 \\ x_3 & x_2 & x_1 \end{vmatrix} = x_1^3 + 3x_1x_2 + x_3 = \Delta_n^3 - 3\Delta_n^2\Delta_n + 2!\Delta_n^3 = 0\,. \tag{A.20}$$

One may see that all $B_m(\mathbf{x}) = 0$ for $m \geq 2$.

Thus, Eq. (A.7) simplifies to:

$$|\mathbf{A}(\nu)| = |\mathbf{A}(0)|(1 + \nu\Delta_n), \tag{A.21}$$

where $\Delta_n = [\mathbf{A}^{-1}(0)]_{nn} a_{nn}$.

Furthermore, the Laplace expansion of the matrix determinant into the cofactors of the last column (row) indicates that

$$|\mathbf{A}(\nu)| = |\mathbf{A}(0)| + \nu a_{nn} A(\gamma_{1:n-1}|\gamma_{1:n-1}). \tag{A.22}$$

Upon eliminating the determinant $|\mathbf{A}(0)|$ from Eqs. (A.21) and (A.22):

$$|\mathbf{A}| = \left(\nu[\mathbf{A}^{-1}(0)]_{nn}^{-1} + a_{nn}\right) A(\gamma_{1:n-1}|\gamma_{1:n-1}), \tag{A.23}$$

where $[\cdot]_{ij}^{-1}$ is equivalent to $1/[\cdot]_{ij}$.

On the other hand, let us divide the matrix into four blocks:

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{n-1} & \mathbf{a}_1 \\ \mathbf{a}_2^T & a_{nn} \end{pmatrix}. \tag{A.24}$$

If $\mathbf{A}_{n-1} \equiv \mathbf{A}(\gamma_{1:n-1}|\gamma_{1:n-1})$ is nonsingular, then:

$$[\mathbf{A}^{-1}]_{nn}^{-1} = a_{nn} - \mathbf{a}_2^T \mathbf{A}_{n-1}^{-1} \mathbf{a}_1. \tag{A.25}$$

A simple identity $[\mathbf{A}^{-1}(1)]^{-1}_{nn} = [\mathbf{A}^{-1}(0)]^{-1}_{nn} + a_{nn}$ now emerges, and one finally arrives at:

$$|\mathbf{A}| = [\mathbf{A}^{-1}]^{-1}_{nn} A(\gamma_{1:n-1}|\gamma_{1:n-1}). \tag{A.26}$$

This is the Schur complement formula for the determinant of a partitioned matrix. It was derived in this section by employing Proposition A.2 with a specific pattern in Eq. (A.15).

An example can be given when the second Bell polynomial is the only one which is nonzero. For this purpose, let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be tridiagonal such that

$$\mathbf{A}(\nu) = \begin{pmatrix} a_{11} & a_{12} & & & & & & & \\ a_{21} & a_{22} & a_{23} & & & & & & \\ & \ddots & \ddots & \ddots & & & & & \\ & & a_{i,i-1} & a_{ii} & \nu a_{i,i+1} & & & & \\ & & & \nu a_{i+1,i} & a_{i+1,i+1} & a_{i+1,i+2} & & & \\ & & & & \ddots & \ddots & \ddots & & \\ & & & & & a_{n-1,n} & a_{n-1,n-1} & a_{n-1,n} \\ & & & & & & a_{n,n-1} & a_{n,n} \end{pmatrix}. \tag{A.27}$$

The matrix $|\mathbf{A}(0)|$ becomes block-diagonal. Therefore,

$$|\mathbf{A}(0)| = A(\gamma_{1:i}|\gamma_{1:i}) A(\gamma_{i+1:n}|\gamma_{i+1:n}). \tag{A.28}$$

The matrix $\mathbf{C} \equiv \mathbf{A}^{-1}(0)\partial\mathbf{A}(0)/\partial\nu$ have only four nonzero elements:

$$c_{i-1,i+1} = [\mathbf{A}^{-1}(0)]_{i-1,i} a_{i,i+1}, \qquad c_{i+1,i} = [\mathbf{A}^{-1}(0)]_{i+1,i+1} a_{i+1,i}, \tag{A.29}$$

$$c_{i,i+1} = [\mathbf{A}^{-1}(0)]_{i,i} a_{i,i+1}, \qquad c_{i+2,i} = [\mathbf{A}^{-1}(0)]_{i+2,i+1} a_{i+1,i}. \tag{A.30}$$

Thus, $\operatorname{tr}\mathbf{C} = 0$. For the odd s, this nonzero pattern will be preserved, whereas when s is even, the only nonzero elements are $c_{i-1,i}$, $c_{i,i}$, $c_{i+1,i+1}$ and $c_{i+2,i+1}$.

A recursion between different matrix powers is very simple. For the even s,

$$[\mathbf{C}^s]_{i+1,i+1} = [\mathbf{C}^{s-1}]_{i+1,i} c_{i,i+1}, \qquad [\mathbf{C}^s]_{i-1,i} = [\mathbf{C}^{s-1}]_{i-1,i+1} c_{i+1,i}, \tag{A.31}$$

$$[\mathbf{C}^s]_{i+2,i+1} = [\mathbf{C}^{s-1}]_{i+2,i} c_{i,i+1}, \qquad [\mathbf{C}^s]_{i,i} = [\mathbf{C}^{s-1}]_{i,i+1} c_{i+1,i}. \tag{A.32}$$

When s is odd,

$$[\mathbf{C}^s]_{i-1,i+1} = [\mathbf{C}^{s-1}]_{i-1,i} c_{i,i+1}, \qquad [\mathbf{C}^s]_{i+1,i} = [\mathbf{C}^{s-1}]_{i+1,i+1} c_{i+1,i}, \tag{A.33}$$

$$[\mathbf{C}^s]_{i,i+1} = [\mathbf{C}^{s-1}]_{i,i} c_{i,i+1}, \qquad [\mathbf{C}^s]_{i+2,i} = [\mathbf{C}^{s-1}]_{i+2,i+1} c_{i+1,i}. \tag{A.34}$$

It follows from Eqs. (A.29)–(A.34) that

$$\operatorname{tr}\mathbf{C}^s = \begin{cases} 2(c_{i+1,i}c_{i,i+1})^{s/2} & \text{if } \textit{s is even,} \\ 0 & \text{if } \textit{s is odd.} \end{cases} \tag{A.35}$$

The reader may now check that all the exponential Bell polynomials in Eq. (A.7) are equal to zero, except

$$B_2(x_1, x_2) = x_2 + x_1^2 = x_2 = -2c_{i+1,i}c_{i,i+1}. \tag{A.36}$$

Therefore,

$$|\mathbf{A}(\nu)| = A(\gamma_{1:i}|\gamma_{1:i})A(\gamma_{i+1:n}|\gamma_{i+1:n}) \cdot \left(1 - \nu^2 a_{i,i+1}a_{i+1,i}[\mathbf{A}^{-1}(0)]_{i,i}[\mathbf{A}^{-1}(0)]_{i+1,i+1}\right). \tag{A.37}$$

A rather complex formula is now reduced to a simple equation which reveals how the elements $a_{i,i+1}a_{i+1,i}$ contribute to the matrix determinant. One could notice that in the case of a positive definite matrix, the term in the parentheses in Eq. (A.37) should be positive, which provides a novel criterion of positive definiteness.

Eq. (A.37) can also be rewritten by using a simplified case of a general Jacobi identity, clf. [Gantmacher and Krein, 2002]:

$$[\mathbf{A}^{-1}]_{ij} = \frac{(-1)^{i+j}A(\boldsymbol{\gamma}_n\backslash j|\boldsymbol{\gamma}_n\backslash i)}{A(\boldsymbol{\gamma}_n|\boldsymbol{\gamma}_n)}. \tag{A.38}$$

This is also known as 'Cramer's rule'. Substituting expressions for the elements $[\mathbf{A}^{-1}(0)]_{i,i}$ and $[\mathbf{A}^{-1}(0)]_{i+1,i+1}$ allows to obtain the statement with matrix minors:

$$|\mathbf{A}(\nu)| = A(\gamma_{1:i}|\gamma_{1:i})A(\gamma_{i+1:n}|\gamma_{i+1:n}) - \nu^2 a_{i,i+1}a_{i+1,i}A(\gamma_{1:i-1}|\gamma_{1:i-1})A(\gamma_{i+2:n}|\gamma_{i+2:n}). \tag{A.39}$$

This expression should also be positive when the matrix $\mathbf{A}(\nu)$ is positive definite. Notice the resemblance of Eq. (A.39) with the equation for the determinant of a 2×2 matrix.

Consider now a tridiagonal matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ and let us multiply all the elements lying on the sub- and super-diagonals adjacent to the main diagonal by ν:

$$\mathbf{A}(\nu) = \begin{pmatrix} a_{11} & \nu a_{12} & & & & \\ \nu a_{21} & a_{22} & \nu a_{23} & & & \\ & \ddots & \ddots & \ddots & & \\ & & \nu a_{n-1,n} & a_{n-1,n-1} & \nu a_{n-1,n} \\ & & & \nu a_{n,n-1} & a_{n,n} \end{pmatrix}. \tag{A.40}$$

Clearly, $|\mathbf{A}(0)| = \prod_{i=1}^{n} a_{ii}$ and $\mathbf{C} \equiv \mathbf{A}^{-1}(0)\partial\mathbf{A}(0)/\partial\nu$ is tridiagonal with the main diagonal of zero entries, the super-diagonal is given by:

$$a_{11}^{-1}a_{12}, \;\; a_{22}^{-1}a_{23}, \;\; \ldots \;\; a_{n-1,n-1}^{-1}a_{n-1,n}, \tag{A.41}$$

whereas the sub-diagonal is:

$$a_{22}^{-1}a_{21}, \;\; a_{33}^{-1}a_{32}, \;\; \ldots \;\; a_{n,n}^{-1}a_{n,n-1}. \tag{A.42}$$

The evaluation of $\operatorname{tr}\mathbf{C}^s$ can be performed by using a general algorithm of Chu [1985], but one expects simpler calculations when the main diagonal is zero.

An example of the diagram-based method to solve this problem is shown in Fig. A.1 when $\mathbf{C} \in \mathbb{R}^{6\times 6}$.

One first plots a rectangular grid on the axis whose vertical-side indicates the s-th power of the matrix and the horizontal axis marks the i-th index of the c_{ii}-th

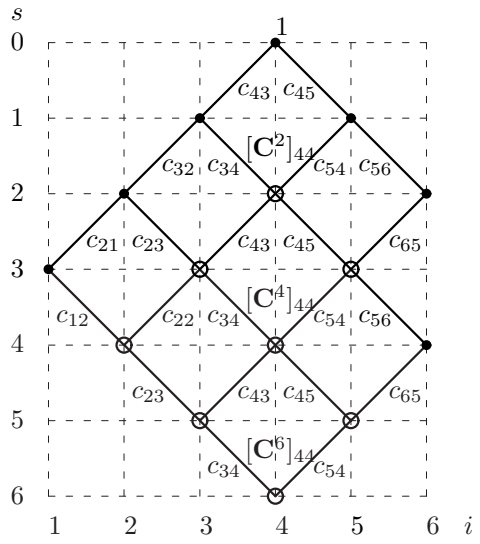


Figure A.1: A diagram for the calculation of $[\mathbf{C}^6]_{44}$.

element of the matrix $\mathbf{C}$. In order to obtain a symbolic expansion for the i-th diagonal element $[\mathbf{C}^s]_{ii}$, one marks all the possible paths with diagonal edges that enable the point $(0, i)$ to be connected to the point (s, i). The diagram is then read from top to bottom where each edge denotes a multiplication and an empty circle splits products into sums, e.g.:

$$[\mathbf{C}^2]_{44} = c_{43}c_{34} + c_{45}c_{54}, \tag{A.43}$$

$$[\mathbf{C}^4]_{44} = \left([\mathbf{C}^2]_{44}c_{43} + c_{43}c_{32}c_{23}\right) c_{34} + \left([\mathbf{C}^2]_{44}c_{45} + c_{45}c_{56}c_{65}\right) c_{54}. \tag{A.44}$$

In order to evaluate the trace, one must: (i) construct n diagrams, (ii) write down the expressions for the diagonal elements, and (iii) sum them.

A complete list of diagrams for the computation of the 4-th order trace when $\mathbf{C} \in \mathbb{R}^{5\times5}$ is shown in Fig. A.2. Each of them summarizes a symbolic computation of the elements $[\mathbf{C}^s]_{ii}$ in an easier way than writing algebraic expressions such as Eqs. (A.43) and (A.44).

Moreover, a number of qualitative properties can be read directly from the diagrams:

1. $\operatorname{tr} \mathbf{C}^s = 0$ for any positive odd integer s.
2. If $\mathbf{A} = \mathbf{A}^T$, then $\operatorname{tr} \mathbf{C}^s \geq 0$.
3. $\operatorname{tr} \mathbf{C}^s$ requires at most $O(ns^2)$ multiplications.

By looking at the tabulation of the partial exponential Bell polynomials in [Charalambides, 2002, p. 417], one can notice that the first property greatly simplifies

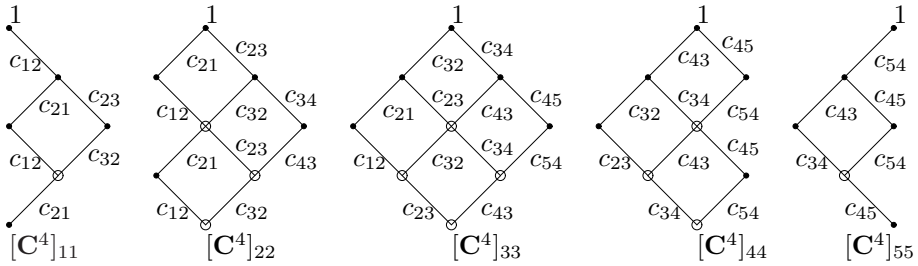


Figure A.2: Diagrams for the evaluation of the diagonal elements of $\mathbf{C}^4$.

Eq. (A.7): (i) each polynomial $B_{m,k}(\mathbf{x}) = 0$ for the odd m, and (ii) $B_{m,k}(\mathbf{x})$ is nonzero only for $k = 1, 2, \ldots m/2$.

For example, the determinant of a tridiagonal matrix $\mathbf{A} \in \mathbb{R}^{10\times 10}$ reads:

$$
\begin{aligned}
|\mathbf{A}(\nu)| = \prod_{i=1}^{n} a_{ii} \Big(& 1 - \frac{\nu^2}{2!}\,\mathrm{tr}\,\mathbf{C}^2 + \frac{\nu^4}{4!}\left(-3!\,\mathrm{tr}\,\mathbf{C}^4 + 3\,\mathrm{tr}^2\,\mathbf{C}^2\right) \\
& + \frac{\nu^6}{6!}\left(-5!\,\mathrm{tr}\,\mathbf{C}^6 + 15\cdot 3!\,\mathrm{tr}\,\mathbf{C}^2\,\mathrm{tr}\,\mathbf{C}^4 - 15\,\mathrm{tr}^3\,\mathbf{C}^2\right) \\
& + \frac{\nu^8}{8!}\big(- 7!\,\mathrm{tr}\,\mathbf{C}^8 + 28\cdot 5!\,\mathrm{tr}\,\mathbf{C}^2\,\mathrm{tr}\,\mathbf{C}^6 + 35(3!)^2\,\mathrm{tr}^2\,\mathbf{C}^4 \\
& \qquad - 210\cdot 3!\,\mathrm{tr}^2\,\mathbf{C}^2\,\mathrm{tr}\,\mathbf{C}^4 + 105\,\mathrm{tr}^4\,\mathbf{C}^2\big) \\
& + \frac{\nu^{10}}{10!}\big(- 9!\,\mathrm{tr}\,\mathbf{C}^{10} + 45\cdot 7!\,\mathrm{tr}\,\mathbf{C}^2\,\mathrm{tr}\,\mathbf{C}^8 + 210\cdot 3!\cdot 5!\,\mathrm{tr}\,\mathbf{C}^4\,\mathrm{tr}\,\mathbf{C}^6 \\
& \qquad - 630\cdot 5!\,\mathrm{tr}^2\,\mathbf{C}^2\,\mathrm{tr}\,\mathbf{C}^6 - 1575(3!)^2\,\mathrm{tr}\,\mathbf{C}^2\,\mathrm{tr}^2\,\mathbf{C}^4 \\
& \qquad + 3150(3!)\,\mathrm{tr}^3\,\mathbf{C}^2\,\mathrm{tr}\,\mathbf{C}^4 - 945\,\mathrm{tr}^5\,\mathbf{C}^2\big)\Big).
\end{aligned}
\tag{A.45}
$$

Here $\mathrm{tr}^s(\cdot) \equiv (\mathrm{tr}(\cdot))^s$. One could note that a general case of Eq. (A.7), written with the terms ranging up to the 10-th order, would barely fit into a single page.

Eq. (A.45) provides a combinatorial method of evaluating Taylor series. As an example, consider the determinant in [Itzykson and Zuber, 1980]:

$$
\begin{vmatrix}
1 & -\nu & & & \\
-\nu & 1 & -\nu & & \\
 & \ddots & \ddots & \ddots & \\
 & & -\nu & 1 & -\nu \\
 & & & -\nu & 1
\end{vmatrix}_{n\times n}
= \frac{\lambda_+^{n+1}(\nu) - \lambda_-^{n+1}(\nu)}{\lambda_+(\nu) - \lambda_-(\nu)}\,,
\tag{A.46}
$$

where

$$
\lambda_\pm = \frac{1}{2}\left(1 \pm \sqrt{1 - 4\nu^2}\right).
\tag{A.47}
$$

This determinant occurs in the analysis of the harmonic oscillator, Brownian motion, discrete cosine transform, etc. It is first essential to observe that since all the nonzero elements of the matrix $\mathbf{C}$ equal to a unity (in modulus), the evaluation of

traces reduces to the number of paths which join the starting and ending points in the diagrams such as those shown in Fig. A.2.

Consider first evaluating $\operatorname{tr}\mathbf{C}^2$. For an even size of the matrix n, there are

- 1 diagram of type $[\mathbf{C}^2]_{11}$.
- $n-2$ diagrams of type $[\mathbf{C}^2]_{22}$.
- 1 diagram of type $[\mathbf{C}^2]_{nn}$.

Counting the number of paths in these diagrams yields:

$$\operatorname{tr}\mathbf{C}^2 = 1 + 2\cdot(n-2) + 1 = 2n-2. \tag{A.48}$$

In the case of $\operatorname{tr}\mathbf{C}^4$ and even size n, Figs. A.1 and A.2 with their description allow to deduce that there are:

- 1 diagram of type $[\mathbf{C}^4]_{11}$.
- 1 diagram of type $[\mathbf{C}^4]_{22}$.
- $n-4$ diagrams of type $[\mathbf{C}^4]_{33}$.
- 1 diagram of type $[\mathbf{C}^4]_{n-1,n-1}$.
- 1 diagram of type $[\mathbf{C}^4]_{nn}$.

Counting paths produces the following result:

$$\operatorname{tr}\mathbf{C}^4 = 2 + 5 + 6\cdot(n-4) + 5 + 2 = 6n-10. \tag{A.49}$$

Substituting Eqs. (A.48) and (A.49) into Eq. (A.45), truncated up to the fourth order, yields the perturbative expansion of the determinant:

$$|\mathbf{A}(\nu)| = 1 - (n-1)\nu^2 + \frac{1}{4}(2n^2 - 7n + 7)\nu^4 + \dots \tag{A.50}$$

If $\nu = c/n$ with a constant c chosen so that the determinant is real, then, asymptotically, $|\mathbf{A}(\nu)| \sim 1$ with $O(n^{-1})$ error as $n \to \infty$.

Another example is Hill's determinant method in quantum mechanics. By substituting an 'ansatz' into the Schrödinger equation, one first obtains an infinite, but discrete, set of linear equations. The resulting eigenvalue problem of a discrete system can then be studied numerically by truncating its order up to a finite number. There exist a rich literature which discusses how Hill's determinant allows to estimate energy levels of a quantum-mechanical system.

In the case of the anharmonic oscillator with the coupling constant ν, Hill's determinant reads, see e.g. [Biswas et al., 1971]:

$$\mathbf{A}(\nu) = \begin{vmatrix} a_1 & b_1 & & & & \\ 0 & a_2 & b_2 & & & \\ -\nu & 0 & a_3 & b_3 & & \\ & \ddots & \ddots & \ddots & \ddots & \\ & & -\nu & 0 & a_{n-1} & b_{n-1} \\ & & & -\nu & 0 & a_n \end{vmatrix}, \tag{A.51}$$

where $a_k = \epsilon - 1 - 4k$, and $b_k = 2k(2k-1)$.

Biswas et al. [1971] apply the recurrence relationship to the determinant of the Hessenberg matrix and further perform numerical studies. Let us provide an answer to the following question: What is the coefficient in front of the term ν^m when expanding $|\mathbf{A}(\nu)|$ in Taylor series at $\nu = 0$?

The determinant an upper triangular matrix is the product of its diagonal elements:

$$|\mathbf{A}(0)| = \prod_{i=1}^{n} a_i \,. \tag{A.52}$$

The matrix $\mathbf{A}(0)$ is bidiagonal, and its inverse is upper-triangular, whose elements have been derived by Chatterjee [1974]:

$$[\mathbf{A}^{-1}(0)]_{ij} = \begin{cases} a_i^{-1} \sum_{k=i}^{j-1} (-a_{k+1}^{-1} b_k) & \text{if } i < j, \\ a_i^{-1} & \text{if } i = j, \\ 0 & \text{if } i > j. \end{cases} \tag{A.53}$$

The matrix $\mathbf{C} \equiv \mathbf{A}^{-1}(0)\partial\mathbf{A}(0)/\partial\nu$ becomes such that its elements below the second subdiagonal are zeroes. More explicitly,

$$[\mathbf{C}]_{ij} = \begin{cases} -[\mathbf{A}^{-1}(0)]_{i,j+2} & \text{if } i \le j+2, \\ 0 & \text{if } i > j+2. \end{cases} \tag{A.54}$$

Therefore, the m-th coefficient in the Taylor series expansion of Hill's determinant equals to $B_m(\mathbf{x})/m!$, where the i-th component of $\mathbf{x} \in \mathbb{R}^m$ is:

$$x_i = (-1)^{i-1}(i-1)! \operatorname{tr} \mathbf{C}^i \tag{A.55}$$

$$= (-1)^{i-1}(i-1)! \sum_{s_1, s_2, \ldots, s_i = 1}^{n} c_{s_i, s_1} c_{s_1, s_2} \cdots c_{s_{i-1}, s_i}. \tag{A.56}$$

This 'reduces' the determinant of a sparse Hessenberg matrix in Eq. (A.51) to the evaluation of the polynomials B_m which now include many determinants of non-sparse Hessenberg matrices. However, these determinants are no longer dependant on the parameter ν. They can be precomputed up to a high accuracy for moderate values of m and Taylor series can then be applied in establishing the positivity of Hill's determinant.

A.2 Some Difficulties With Statistical Inference

This section presents simple examples of difficult topics. This includes: (i) a probabilistic result on the asymptotic disappearance of the randomness of p-norms, which is supported by a novel unique feature of the Euclidean norm stated in Eqs. (A.58) and A.59, (ii) discussion of nonidentifiable parameters, (iii) integrating property of expectations, and (iv) counterexample which shows that the diagonal elements of the conditional covariance do not serve as a measure of confidence.

The first example deals with a conceptual problem of drawing the line between what is 'random' and 'deterministic'. Let $X_i \sim \mathrm{N}(0, \sigma^2)$ be independent identically distributed (i.i.d.) normal random variables. Define the p-norm variable according to:

$$D_p \equiv \left(\sum_{i=1}^{n} X_i^p \right)^{1/p}. \tag{A.57}$$

This seems to be a random quantity whose mean and variance are:

$$\langle D_p \rangle = c_1(p) n^{\frac{1}{p}} \sigma + O(n^{\frac{1}{p}-1}\sigma), \tag{A.58}$$

$$\mathrm{Var}(D_p) = c_2(p) n^{\frac{2}{p}-1} \sigma^2 + O(n^{\frac{2}{p}-2}\sigma^2). \tag{A.59}$$

The functions $c_i(p)$ depend only on p and their exact expressions are irrelevant here. A complete derivation of this result takes about three pages and is left as an exercise to the reader. It suffices to state that one should apply the marginalization property of probability densities and Taylor series of the function $f(\mathbf{x}) \equiv \|\mathbf{x}\|_p$.

The case $p = 2$ is a 'miracle': asymptotically, the variance becomes independent of the space dimension. To be more precise, $c_2(2) = 1/2$, and $\mathrm{Var}(D_2) = \sigma^2/2 + O(n^{-1}\sigma^2)$, whereas $c_1(2) = 2$ and $\langle D_2 \rangle = n^{1/2}\sigma$. This is a unique feature of the Euclidean norm.

It is common to discuss the ratio $\sqrt{\mathrm{Var}(D_2^2)}/\langle D_2^2 \rangle \sim O(n^{-1/2})$ as a measure of randomness, see e.g. [D. Landau and M. Lifshitz, 1958]. The mean $\langle D_2^2 \rangle \sim O(n)$ grows faster than the deviation $\sqrt{\mathrm{Var}\, D_2^2} \sim O(n^{1/2})$, and, in a way, the variable becomes less random with a growing n. The ratio $\sqrt{\mathrm{Var}(D_2)}/\langle D_2 \rangle \sim O(n^{-1})$ is even more striking, as the deviation becomes constant when n is large.

The maximum norm D_∞ cannot be studied similarly. Cramér [1945] ingeniously used the binomial variables to state the following asymptotic result for the vector of zero-mean Gaussian components with the unity variance:

$$\lim_{n\to\infty} \mathrm{Prob}\left(\frac{D_\infty - b(n)}{a(n)} < x \right) = e^{-e^{-x}}, \quad x \in \mathbb{R}, \tag{A.60}$$

where the parameters $b(n) \sim O(\ln n)$ and $a(n) \sim O((\ln n)^{-1})$ can be interpreted as location and scale. The probability distribution in Eq. (A.60) produces a finite mean and variance values for the random variable $(D_\infty - b(n))/a(n)$. Thus, one can draw the conclusion that $\mathrm{Var}\, D_\infty \sim O((\ln n)^{-1})$ which becomes negligible, and $\langle D_\infty \rangle \sim O(\ln n)$, which grows slower than $\langle D_2 \rangle \sim O(n^{1/2})$. This is not as slow as one might expect when fixing a large value of n and using $\langle D_p \rangle \sim O(n^{1/p})$ as $p \to \infty$. In summary, the randomness of the norm of the vector with random components diminishes as $n \to \infty$.

As a second difficulty, suppose that one models the observations as a random process given by:

$$Y(t) = S(t) + N(t), \tag{A.61}$$

where the signal $S(t)$ is the Gaussian process of zero mean with

$$\mathrm{Cov}(S(t_i), S(t_j)) = \exp\left(-\epsilon |t_i - t_j|^\gamma\right), \tag{A.62}$$

and the noise $N(t)$ is the Gaussian process of zero mean with

$$\mathrm{Cov}(N(t_i), N(t_j)) = \sigma^2 \delta(t_i - t_j), \tag{A.63}$$

Given the observations of $Y(t)$, the problem of estimating the process $S(t)$ is known as the Wiener–Kolmogorov smoothing, see e.g. [Bode and Shannon, 1950].

Let us further neglect the effects of continuity and causality and further restrict the problem to obtaining the conditional expectation $\langle F|\mathbf{y}, \epsilon, \gamma, \sigma^2\rangle$ at the times t_1, t_2, …, t_n with the available observations $\hat{Y}(t_1)$, $\hat{Y}(t_2)$, …, and $\hat{Y}(t_n)$ which are gathered into a single vector $\mathbf{y} \in \mathbb{R}^n$. The solution becomes, clf.[Foster, 1961]:

$$\langle F|\mathbf{y}, \epsilon, \gamma, \sigma^2\rangle = \mathbf{\Sigma}_{ss}(\mathbf{\Sigma}_{ss} + \mathbf{\Sigma}_{nn})^{-1}\mathbf{y}. \tag{A.64}$$

Here $\mathbf{\Sigma}_{ss} \in \mathbb{R}^{n\times n}$ is the covariance matrix whose elements are $\mathrm{Cov}(S(t_i), S(t_j))$ and the noise covariance $\mathbf{\Sigma}_{nn} = \sigma^2\mathbf{I}$.

Experience shows that the parameters ϵ and σ^2 are identifiable if the correlation length-scale of $S(t)$ is larger than that of $N(t)$. Despite the warnings in [Dietrich, 1991], the maximum likelihood methodology is very useful.

The parameter γ affects the mean square continuity of the process sample paths:

$$\lim_{\Delta\to 0}\left\langle (N(t+\Delta) - N(t))^2 \right\rangle \sim \begin{cases} O(\Delta) & \text{if } \gamma = 1, \\ O(\Delta^2) & \text{if } \gamma = 2. \end{cases} \tag{A.65}$$

This difference has a very small impact on the smoothing quality in the sense of reducing the variance of $N(t)$ at the discrete set of the locations of observations. The identifiability of γ is in question.

An interesting clue to the choice of γ can be found by comparing the eigenvalues of the corresponding covariance matrices on the equidistantly mesh. For $\epsilon \to 0^+$ and $k = 2, 3, \ldots, n$, Cox and Chapman [1995] show that

$$\lambda_k(\epsilon) = \begin{cases} O(\epsilon) & \text{if } \gamma = 1, \\ O(\epsilon^{k-1}) & \text{if } \gamma = 2. \end{cases} \tag{A.66}$$

Thus, $\gamma = 1$ yields a better conditioned covariance, but there would be more to be gained from $\gamma = 2$ when using approximate rank reductions, see e.g. [Anderson, 1958]. It is also notable that the choice $\gamma = 1$ yields a tridiagonal inverse covariance matrix, which is computationally advantageous when n is large.

Another difficult point is a seemingly trivial extension of Eq. (A.64) via the 'Representer theorem', clf. [Maruyama et al., 1992]:

$$\langle F|\mathbf{y}, \epsilon, \gamma, \sigma^2\rangle = \sum_{i=1}^{n} c_i \mathrm{Cov}(S(t), S(t_i)), \tag{A.67}$$

where $c_i = [(\mathbf{\Sigma}_{ss} + \mathbf{\Sigma}_{nn})^{-1}\mathbf{y}]_i$. The equation simply states that the conditional expectation is spanned in a finite basis of the covariance functions.

Considering the definition of the covariance function as the 'dot-product between the sample paths', an interesting paradox emerges. The continuity classes of the

covariance function and its sample paths can be different. The two basic notions of the sample path continuity are: (i) the almost sure continuity and (ii) the mean square or L_2 continuity. Considering smoothing, one almost always assumes the L_2 continuity of the process paths which, at least in a stationary case, is equivalent to the continuity of a covariance function, clf. [Banerjee and Gelfand, 2003]. However, the continuity class may change as indicated in the following example.

Let the basis consist of sign functions $sgn(x - z)$ indexed with a continuous parameter $z \in [0, 1]$. A direct Hilbert product yields, see e.g. [Minka, 2000]:

$$\int_0^1 sgn(x_i - z)sgn(x_j - z)dz = 1 - 2|x_i - x_j|. \tag{A.68}$$

Any *piecewise-constant* function can be spanned by the basis of sign functions, but only a *piecewise-linear* function will be spanned by the corresponding Hilbert's dot product. Therefore, if a signal at the point x_0 has *a jump discontinuity*, postulating the process with the piecewise-constant paths will result in the loss of jump discontinuities when smoothing according to Eq. (A.67).

One example of a commonly misused Gaussian model deserves to be mentioned. Consider the process whose sample paths are spanned in a polynomially decaying Fourier basis:

$$S(t) = \sum_{d=1}^{m} w_d \phi_d(t), \quad w_d \sim \mathcal{N}(0, d^{-\gamma}). \tag{A.69}$$

The value of $\gamma \geq 1$ determines the continuity of the process path. The larger its value, the smoother it will be. It is a mistake to assume that this could be useful in recovering the step function from its sampled observations in white noise. One can apply the maximum likelihood estimation to find out the 'optimal' decay γ^* which may exist, but is hardly meaningful. The continuity class of the covariance functions is determined by the Hilbert product of weighted Fourier components, which does not include the functions with jump discontinuities. Even when $m \to \infty$, these covariance functions would suffer from the Wilbraham–Gibbs phenomenon. The model is not suitable for the recovery of jump discontinuities.

Use of confidence intervals often becomes another major source of confusion. If the conditional expectation produces the outcome of the smoothing, it seems like a sound idea to utilize the conditional covariance as a measure of confidence or accuracy of the result. However, this can be very misleading.

The conditional covariance of a general Gaussian model in Eq. (A.61) is, clf. Anderson [1958]:

$$\mathrm{Cov}\,(U|\mathbf{y}) = \mathbf{\Sigma}_{uu} - \mathbf{\Sigma}_{uu}(\mathbf{\Sigma}_{uu} + \mathbf{\Sigma}_{nn})^{-1}\mathbf{\Sigma}_{uu} \tag{A.70}$$

$$= (\mathbf{\Sigma}_{uu}^{-1} + \mathbf{\Sigma}_{nn}^{-1})^{-1}\,. \tag{A.71}$$

Let us further assume that the process $S(t)$ is the Brownian motion in the interval $t \in [0, 1]$ with the covariance $[\mathbf{\Sigma}_{ss}]_{ij} = \min(i, j)$. Let the noise be white with a variance σ^2. The smoothing is considered on a finite grid of samples in $t \in [0, 1]$, which makes all the matrices finite dimensional.

It is obvious that the conditioning reduces the process variance in the sense that $(\mathbf{\Sigma}_{ss}^{-1} + \mathbf{\Sigma}_{nn}^{-1})^{-1} - \mathbf{\Sigma}_{ss}$ is generally negative semidefinite. However, the increasing

pattern w.r.t. indices i and j carries to $[\mathrm{Cov}\,(U|\mathbf{y},\boldsymbol{\theta})]_{ij}$. The diagonal entries of this matrix are variances, which increase w.r.t. to $i = 1, 2, \ldots, n$. That, however, does not imply that we are less confident with the smoothing outcome located further away from the origin.

The problem can be 'fixed' by replacing the assumption of the Brownian motion with that of the Brownian bridge, whose covariance matrix $[\boldsymbol{\Sigma}_{ss}]_{ij} = \min(i,j) - ij$. The variance does no longer increase in time. However, does one increase 'confidence' of the smoothing outcome?

Firstly, it is essential to observe that both models of smoothing are discrete counterparts of a linear boundary value problem with the second order differential operator. The Gaussian model gives a stochastic (call it Monte Carlo or Feynman–Kac) interpretation. The assumption of the Brownian motion corresponds to mixed Dirichlet–von Neumann conditions. This is the constraint that the smoothing outcome will have a zero value at $t = 0$ and its derivative at $t = 1$ should be zero too. The assumption of the Brownian bridge corresponds to the Dirichlet boundary conditions of zero outcome at the both ends of the interval. It should now be clear that the conditional variance has little to do with 'confidence'.

Despite its apparent simplicity, this example leaves several open questions. Experiments would show that the maximum likelihood technique is useful in the automatic identification of noise variance σ^2. However, it is unclear whether it is applicable in the automatic choice of the boundary conditions. If the smoothing outcome is such that it has both, zero value at $t = 1$ and its derivative w.r.t. t is zero there too, then the difference between the two models is rather small. The assumption of Brownian bridge allows to remove white noise after the conditioning, with a slight side effect of pulling the solution towards zero values everywhere. So does the model with the assumption of the Brownian motion, and in the case of small variances of noise it is hard to discriminate between the two models.

Finally, it is also worth to emphasize certain ambiguities in probabilistic modeling. There is always a possibility of close alternatives which give similar conditioning on observations, but are not necessarily models with an additive noise.

Interestingly, Eqs. (3.11) and (3.12) define the smoothing outcome (on the level of conditional means), but they do not depend on the covariance matrix $\boldsymbol{\Sigma}_{uu}$. One is free to choose it. In addition, it is possible to define the models which would produce similar predictions in the sense that the conditional expectation Eq. (3.11) will remain the same and the criterion for the model parameters $\boldsymbol{\theta}$ in Eq. (3.11) might possess similar optima.

As an example, let us now take the covariance matrix of the signal from the model $\mathcal{H}_1$, i.e. $\boldsymbol{\Sigma}_{uu} \equiv \mathbf{K}_\theta$ and consider the joint covariance matrices with the structure which pertains to a non-additive noise:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \mathbf{P}_\theta^{-1} & \mathbf{I} \\ \mathbf{I} & \mathbf{P}_\theta \end{pmatrix}, \quad \mathbf{P}_\theta = \mathbf{I} + \theta_0 \mathbf{K}_\theta^{-1}. \tag{A.72}$$

One may check that Eq. (3.11) would yield the same conditional mean $\langle U|\mathbf{y},\boldsymbol{\theta}\rangle$ as the models $\mathcal{H}_1$ and $\mathcal{H}_2$. The marginal likelihoods in Eq. (3.12) are not in general identical, but they can produce qualitatively similar parameter estimates. Accord-

ing to the model $\mathcal{H}_1$, the covariance matrix of the osbservations is $\mathbf{\Sigma}_{yy} = \mathbf{K}_\theta + \theta_0 \mathbf{I}$, whereas Eq. (A.72) produces $\mathbf{\Sigma}_{yy} = \mathbf{I} + \theta_0 \mathbf{K}_\theta^{-1}$. They are seemingly different, but one should notice that the covariance is inverted twice in the Eq. (3.12).

The application of Schur complements to Eq. (A.72) reveals a certain curiosity. The inverse of the joint covariance matrix in Eq. (A.72) does not exist and the conditional covariance matrix in Eq. (3.13) has only zero entries, yet this probabilistic model supplies a well-defined marginal likelihood for the selection of the model parameters $\boldsymbol{\theta}$.

As the model in Eq. (A.72) can be ruled out due to the defect in its covariance matrix, consider a better proposition:

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{C}_\theta & \mathbf{A}_\theta \\ \mathbf{A}_\theta & \mathbf{C}_\theta^{-1} \end{pmatrix}, \quad \mathbf{C}_\theta = \mathbf{K}_\theta + \theta_0 \mathbf{I}, \quad \mathbf{A}_\theta = (\mathbf{C}_\theta \mathbf{P}_\theta)^{-1}, \tag{A.73}$$

Again, the conditional means remain the same. The marginal likelihood is now determined by the covariance $\mathbf{\Sigma}_{yy} = (\mathbf{K}_\theta + \theta_0 \mathbf{I})^{-1}$, which is an inverse of the one given by the model $\mathcal{H}_1$. Again, the behavior w.r.t. to the smoothing parameter is expected to be similar. This model is also a legitimate alternative in the sense that the diagonal elements of the conditional covariance matrix, given by Eq. (3.13), reduce under the conditioning.

One could also state the model which matches the conditional mean of the models $\mathcal{H}_1$ and $\mathcal{H}_2$, possesses a well-defined covariance matrix, but at the same time introduces bias in the parameter estimation:

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{P}_\theta^{-\alpha+2} & \mathbf{P}_\theta^{-\alpha-1} \\ \mathbf{P}_\theta^{-\alpha-1} & \mathbf{P}_\theta^{-\alpha} \end{pmatrix}, \quad \alpha = \mathbb{Z}^+. \tag{A.74}$$

For large $\alpha \gg 1$, the term $\mathbf{P}_\theta^{-\alpha}\mathbf{y}$ is a repetitive conditioning of the observations. It tends towards the steady state $\mathbf{u}_\infty^*$ which is typically a constant signal and can be independent of the model parameters $\boldsymbol{\theta}$. Thus, the first term in Eq. (3.12) can be constant independently of the scale $\theta = m$. On the other hand, the constant α would now multiply the logarithm of the determinant in Eq. (3.12). The preference would be given to a smaller value of the stopping time.

Formally, all of the above-presented alternatives, including the last one with a small parameter such as $\alpha = 1$, do not impose an additive noise assumption, yet the predictions will be similar if the argument based on the presence of 'double inverse' in Eq. (3.12) is valid.