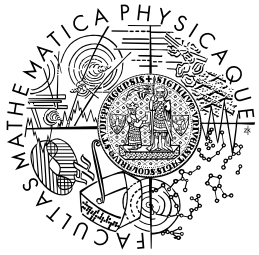


FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

DOCTORAL THESIS

Mgr. Marie Turčičová

Covariance estimation for filtering in high dimension

Department of probability and mathematical statistics

Supervisor of the doctoral thesis:	RNDr. Jan Mandel, CSc.
Consultant of the doctoral thesis:	RNDr. Kryštof Eben, CSc.
Study programme:	Probability and statistics, econometrics and financial mathematics

Prague 2020

I declare that I carried out this doctoral thesis independently, and only with the cited sources, literature and other professional sources. It has not been used to obtain another or the same degree.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Sb., the Copyright Act, as amended, in particular the fact that the Charles University has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 subsection 1 of the Copyright Act.

In ..... date ....

Author's signature

Foremost, I would like to express my sincere gratitude to my supervisor, RNDr. Jan Mandel, CSc., and to my consultant RNDr. Kryštof Eben, CSc., who patiently guided me throughout my PhD studies. I am grateful for their constant support, kindness and all the time they spent with me during our regular Tuesday online meetings. Their guidance proved invaluable for both my research and writing of this thesis. I am very thankful for their pertinent remarks, that have raised the quality of this thesis, but also for their useful advice concerning the scientific research principles in general. I very much appreciate their influence and I always tried to learn as much as possible from both of them.

Speaking of teachers, I attended many interesting lectures during my studies and met many great teachers who attracted my attention to their field at some point. Here, I would like to express my deepest gratitude to all of them.

My special thanks goes to my colleagues from the Institute of Computer Science, CAS, especially to those from the former Department of Nonlinear Modelling for their constant support and friendly encouragement.

Most of all, however, I am deeply grateful to my parents, Hana and Branko, for their tireless support throughout my whole studies - from my first school day up to the completion of this thesis. Their love and care will never be forgotten. I am very much thankful that they gave me the chance to devote myself to things that I wanted and enjoyed. I also want to extend my thanks to my whole family and other people important in my life, who contributed to the completion of my studies directly or indirectly.

Last but not least, my sincere thanks goes to all my friends, who also contributed to the submission of this thesis, either with a piece of advice or just with a word of encouragement both of which were so much needed at many times.

The research was partially supported by NSF grants DMS-1216481 and ICER-1664175, by Grant Agency of the Czech Republic grant 13-34856S, and also by grants SVV No. 305-09/267315 (year 2013), 260105 (year 2014), 260225 (year 2015), 260334 (year 2016), 260454 (year 2017) and 260454 (year 2018).

Title: Covariance estimation for filtering in high dimension

Author: Mgr. Marie Turčičová

Department: Department of probability and mathematical statistics

Supervisor: RNDr. Jan Mandel, CSc., Department of Mathematical and Statistical Sciences, University of Colorado Denver

Consultant: RNDr. Kryštof Eben, CSc., Department of Complex Systems, Institute of Computer Science, CAS

Abstract: Estimating large covariance matrices from small samples is an important problem in many fields. Among others, this includes spatial statistics and data assimilation. In this thesis, we deal with several methods of covariance estimation with emphasis on regularization and covariance models useful in filtering problems. We prove several properties of estimators and propose a new filtering method. After a brief summary of basic estimating methods used in data assimilation, the attention is shifted to covariance models. We show a distinct type of hierarchy in nested models applied to the spectral diagonal covariance matrix: explicit estimators of parameters are computed by the maximum likelihood method and asymptotic variance of these estimators is shown to decrease when the maximization is restricted to a subspace that contains the true parameter value. A similar result is obtained for general M-estimators. For more complex covariance models, maximum likelihood method cannot provide explicit parameter estimates. In the case of a linear model for a precision matrix, however, consistent estimator in a closed form can be computed by the score matching method. Modelling of the precision matrix is particularly beneficial in Gaussian Markov random fields (GMRF), which possess a sparse precision matrix. The score matching estimator is a key component of the ensemble filtering algorithms proposed in the second part of the thesis, that is devoted to data assimilation. In every time step, the proposed *Score matching filter with Gaussian resampling (SMF-GR)* provides a consistent (in the large ensemble limit) estimator of the mean and covariance matrix of the true forecast distribution, under the condition that the original process can be assumed to be a GMRF. Further, we propose a filtering method called *Score matching ensemble filter (SMEF)*, which is based on regularization of the well-known Ensemble Kalman filter (EnKF). The filter performs very well even for some particular examples of non-Gaussian systems with nonlinear dynamic.

Keywords: score matching, Gaussian Markov random field, nested parameter spaces, ensemble Kalman filter, high dimension

Contents

Introduction

In many fields of applied science, e.g. finance, medicine, image processing or climate studies, it is common to encounter situations where we are confronted with a very high dimension of vectors of interest, be it a data vector corresponding to an observation or a state vector, say, of a dynamical system. Often this dimension highly exceeds the size of the sample we have at our disposal. If we are interested in relations between variables, e.g. in filtering tasks or in data assimilation, using sample covariance as a legitimate estimate of the true covariance matrix is problematic. Obviously the main shortcoming of sample covariance is its low rank, which complicates using many of the standard methods. Moreover, if we need to estimate the precision matrix (inverse of the covariance matrix), the naive estimator in the form of the inverse of sample covariance is unavailable. Further, a sample covariance matrix of low rank usually contains spurious covariances that distort the covariance structure among individual variables. Therefore, it is desirable to find alternative covariance estimators that are more accurate and better-conditioned than the sample covariance matrix. Any technique leading to a covariance estimate that is regular and positive definite will be called covariance regularization in this thesis.

Covariance estimation and the quality of the estimate form a key component of data assimilation algorithms in meteorological sciences. In this context, the dimension of the state vector describing the atmosphere or ocean is in the order of millions or larger, however, due to the computational cost, the size of available sample (usually called ensemble) is in the order of tens. For example, the Canadian Meteorological Centre (CMC) uses an ensemble with 20 members, and the European Centre for Medium-Range Weather Forecasts (ECMWF) uses 51 members.

In data assimilation, where the state vector is composed from several 2D or 3D spatial fields, a common approach to regularization of the sample covariance matrix is localization, which is usually achieved by imposing sparsity of the covariance estimate. A simple localization method consists in multiplying the sample covariance matrix term by term by a gradual cut-off matrix (Buehner and Charron [12], Furrer and Bengtsson [24]) in order to suppress off-diagonal entries corresponding to long-range spurious covariances. If the random field is weakly stationary, we may keep the diagonal of the sample covariance matrix only (Section 1.2) after transformation to spectral domain. Such diagonal approximation in the spectral domain is also beneficial in filtering algorithms (Parrish and Derber [59], Kasanický et al. [37]). Beside localization, current filtering methods use non-parametric estimating methods as shrinkage and ad hoc techniques for dimension reduction. A summary of estimation methods used in data assimilation is provided in Section 2.1.

In this thesis, we focus on introducing sparsity into covariance matrices or their inverses by means of suitable parametric covariance models. We deal with two specific aspects of covariance modelling, namely, hierarchy of nested parametric models and modelling of precision matrices with applications in filtering. In Chapter 3, we study nested parametric models estimated by the maximum likelihood method and show a hierarchical structure of their asymptotic covari-

ance matrices. In particular, the asymptotic variance of the maximum likelihood estimate (MLE) is proved to decrease when the maximization is restricted to a subspace that contains the true parameter value. We apply this result to nested models for a diagonal covariance matrix arising, e.g., in the context of the diagonal approximation mentioned above. In the case of a covariance matrix of a weakly stationary random field after the spectral transform, sample covariance matrix represents the MLE of the most general model and MLE of an unconstrained diagonal matrix is its submodel. We also compute the MLE for parameters of two specific models describing the decay of diagonal elements. In accordance with the theory, such models, if realistic, outperform the simple estimate in form of a diagonal of sample covariance matrix. The hierarchical property is illustrated by means of a simulation and the Fisher information matrices representing the inverse of asymptotic covariances of the computed estimators are provided as well.

The second part of the thesis shares the parametric approach with the first part and deals with modelling of the precision matrix. A very general linear model for the precision matrix is investigated and applied to Gaussian Markov Random Fields (GMRFs). Since conditional independence of variables implies zero corresponding elements in the precision matrix, sparsity is taken into account as well. In Section 4.6 we compute explicit formulas for estimators of the parameters of this linear model by an estimation method which arose originally in the area of graphical models and which is called *score matching*. Beside these formulas, we show continuity of the score matching estimators to random perturbations for the exponential family of distributions. The score matching estimators belong to the class of M-estimators. This motivated an extension of the results on hierarchical structure of asymptotic variances from maximum likelihood estimators to M-estimators. The closed form estimator for a precision matrix of a GMRF becomes a key component for the new filtering algorithms proposed in Chapter 7. Both of these filters are intended for a dynamical system whose state vector can be represented by a GMRF in every time step. The first proposed filter is the *Score matching filter with Gaussian resampling* (SMF-GR) and it performs very well under this assumption. Moreover, we prove that SMF-GR provides a consistent estimator for the mean and covariance matrix of the true forecast distribution in every time step. The second proposed filter is called *Score matching ensemble filter* (SMEF), since it consists of the Ensemble Kalman filter employing the score matching estimate of the precision matrix. This algorithm appears to be more robust than SMF-GR. Simulations suggest that it works very well for small samples and even for a non-Gaussian and non-Markov system like the Lorenz 96 model. It seems that the score matching covariance estimate improves the filtering process significantly even though it was derived under the assumption of normality.

Chapters 1, 2 and 6 contain summary of the known methods and provide background for other chapters. Chapters 3, 4, 5 and 7 offer short introductions to the problem areas followed by new results, most of which have been published in Turčičová et al. [67] and Turčičová et al. [66].

1. Covariance and its properties

Consider a random vector $\boldsymbol{X}$ defined on a probability space $(\Omega, \mathcal{A}, P)$ with values in $\mathbb{R}^n$. In the context of data assimilation, $\boldsymbol{X}$ represents the state of some dynamical system and it is usually understood as a discretization of a continuous random field defined on a one, two or three-dimensional bounded domain. The discretization is realized by evaluating the random field on a uniform mesh covering the domain and stacking these values vertically in a single column.

When $\boldsymbol{X} \in L^2(\Omega, \mathcal{A}, P) =: L_2(\Omega)$, we can define its *covariance matrix*

$$\operatorname{cov} \boldsymbol{X} = \mathsf{E}(\boldsymbol{X} - \mathsf{E}\,\boldsymbol{X})(\boldsymbol{X} - \mathsf{E}\,\boldsymbol{X})^\top = \mathsf{E}(\boldsymbol{X}\boldsymbol{X}^\top) - (\mathsf{E}\,\boldsymbol{X})(\mathsf{E}\,\boldsymbol{X})^\top.$$

In the whole thesis, we will assume that $\boldsymbol{X} \in L^2(\Omega, \mathcal{A}, P)$ and denote $\Sigma \equiv \operatorname{cov} \boldsymbol{X}$. From the definition, covariance matrix is a symmetric positive semidefinite matrix.

In the following two sections, we provide few specific properties of covariance matrices or covariance operators that are needed later in the thesis.

1.1 Covariance operator on a compact domain

When the dimension n is very large (in practice, n is often of order 10^6 or more), even in numeric processing of such fields, effects that are typical for continuous fields emerge. These limiting properties become important for high-dimensional covariance matrices, and it is useful to keep in mind the link to covariance operators.

Domain $\mathfrak{D} \subset \mathbb{R}^d$ is defined as an open and connected set. Assume that $\mathfrak{D}$ is bounded and define $\mathcal{C}(\overline{\mathfrak{D}}) = \{g \colon \overline{\mathfrak{D}} \to \mathbb{R},\ g \text{ continuous}\}$, where $\overline{\mathfrak{D}}$ denotes the closure of $\mathfrak{D}$. Consider a random field X on $\overline{\mathfrak{D}}$ as a collection of random variables $X(\boldsymbol{s}, \cdot) \colon (\Omega, \mathcal{A}, P) \to \mathbb{R}$ indexed by $\boldsymbol{s} \in \overline{\mathfrak{D}}$. It is assumed that $X(\cdot, \omega) \in \mathcal{C}(\overline{\mathfrak{D}})$ for a fixed $\omega \in \Omega$, and $X(\boldsymbol{s}, \cdot) \in L_2(\Omega)$ for a fixed $\boldsymbol{s} \in \overline{\mathfrak{D}}$. In order to make the notation shorter, we will omit the variable ω and use $X(\boldsymbol{s})$ instead of $X(\boldsymbol{s}, \omega)$. The expected value of X is defined pointwise, i.e. $(\mathsf{E}\,X)(\boldsymbol{s}) = \mathsf{E}(X(\boldsymbol{s}))\ \forall \boldsymbol{s} \in \overline{\mathfrak{D}}$.

The *covariance function* $c\,(\boldsymbol{t}, \boldsymbol{s})$ of X is defined as $c\,(\boldsymbol{t}, \boldsymbol{s}) = \operatorname{cov}\,(X(\boldsymbol{t}), X(\boldsymbol{s}))\,,$ $\forall \boldsymbol{t}, \boldsymbol{s} \in \overline{\mathfrak{D}}$ and it is assumed to be continuous on $\overline{\mathfrak{D}} \times \overline{\mathfrak{D}}$. The covariance function forms the kernel of a *covariance operator* $\mathrm{T} : \mathcal{C}(\overline{\mathfrak{D}}) \to \mathcal{C}(\overline{\mathfrak{D}})$, which is defined, for $u \in \mathcal{C}(\overline{\mathfrak{D}})$, by

$$\mathrm{T} : u \mapsto v, \quad v\,(\boldsymbol{t}) = \int_{\mathfrak{D}} c\,(\boldsymbol{t}, \boldsymbol{s})\, u\,(\boldsymbol{s})\, d\boldsymbol{s}, \quad \boldsymbol{t} \in \overline{\mathfrak{D}}. \tag{1.1}$$

Equation (1.1) also defines T as a bounded operator $\mathrm{T} \colon L_2(\mathfrak{D}) \to L_2(\mathfrak{D})$ and from the definition of covariance, it follows that T is positive, i.e.,

$$\langle \mathrm{T}\, u, u \rangle_{L_2(\mathfrak{D})} \geq 0,\ \forall u \in L_2(\mathfrak{D}). \tag{1.2}$$

Since the domain $\mathfrak{D}$ is bounded, the covariance operator T is compact as an operator on $\mathcal{C}(\overline{\mathfrak{D}})$ as well as an operator on $L_2(\mathfrak{D})$. Thus, it has countably many eigenvalues λ_k and eigenvectors $u_k \in L_2(\mathfrak{D}),\ k \in \mathbb{N}$, defined by $\lambda_k u_k = \mathrm{T}\, u_k,\ u_k \neq 0$. Also, all $u_k \in \mathcal{C}(\overline{\mathfrak{D}})$. The set $\{\lambda_k\}_{k \in \mathbb{N}}$ is bounded and has only

zero as an accumulation point. It follows from (1.2) that all eigenvalues λ_k are non-negative.

An important condition that is automatically fulfilled for finite-dimensional random vectors but becomes nontrivial in infinite dimension is the *trace-class property* specified in the following lemma, which follows immediately from the classical Mercer's theorem (e.g., König [39, Theorem 3.a.1]).

Lemma 1 (trace-class property)**.** *Under the assumptions made above, the covariance operator* T *is of the trace-class, i.e., it holds that*

$$\mathrm{tr}(\mathrm{T}) \equiv \sum_{k=1}^{\infty} \lambda_k < \infty.$$

1.1.1 Spectral convergence of covariance operators

As indicated at the beginning of this section, the random vector $\boldsymbol{X}$ often arises by a discretization of a random field $X = (X(\boldsymbol{s}),\ \boldsymbol{s} \in \overline{\mathfrak{D}})$. In what follows, we will briefly describe the spectral convergence of the covariance matrices of $\boldsymbol{X}$ to the covariance operator of X. This result should be kept in mind together with the trace-class property (Lemma 1) when looking for a proper model for high-dimensional covariance matrices. Eigenvalues of the model covariance matrix (sorted in descending order) should rapidly decay to zero for all n, in order to fulfil the trace-class property in the limit case.

Consider a uniform mesh of points $\boldsymbol{s}_i \in \overline{\mathfrak{D}}$, $i = 1, \ldots, n$, with spacing h_n. The discretization of $u \in \mathcal{C}(\overline{\mathfrak{D}})$ is a vector $\boldsymbol{u}_n = (u(\boldsymbol{s}_1), \ldots, u(\boldsymbol{s}_n))^\top$. Analogously, the discretization of a continuous random field X is the random vector $\boldsymbol{X}_n = (X(\boldsymbol{s}_1), \ldots, X(\boldsymbol{s}_n))^\top$ with covariance matrix $\Sigma_n = [\sigma_{ij}]_{i,j=1}^n$ consisting of elements

$$\sigma_{ij} = c(\boldsymbol{s}_i, \boldsymbol{s}_j) = \mathrm{cov}\left(X(\boldsymbol{s}_i), X(\boldsymbol{s}_j)\right).$$

Replacing the integral in (1.1) by the numerical quadrature scheme

$$\int_{\mathfrak{D}} c(\boldsymbol{t}, \boldsymbol{s})\, u(\boldsymbol{s})\, d\boldsymbol{s} \approx \sum_{j=1}^{n} w_{n,j} c(\boldsymbol{t}, \boldsymbol{s}_j) u(\boldsymbol{s}_j)$$

with weights $\{w_{n,j}\}_{j=1}^n$, we can interpret matrix-vector multiplication by the covariance matrix as a numerical approximation of the covariance operator. Define the discrete operator $\mathrm{T}_n : \mathcal{C}(\overline{\mathfrak{D}}) \to \mathcal{C}(\overline{\mathfrak{D}})$ by

$$\mathrm{T}_n : u \mapsto v, \quad v(\boldsymbol{t}) = \sum\nolimits_{j=1}^{n} h_n^d c(\boldsymbol{t}, \boldsymbol{s}_j)\, u(\boldsymbol{s}_j). \tag{1.3}$$

Since $v = \mathrm{T}_n u$ depends on the values of $u(\boldsymbol{s}_j)$, $j = 1, \ldots, n$, only, operator T_n has rank at most n and $v = \mathrm{T}_n u$ is determined uniquely by the values $v(\boldsymbol{s}_j)$, $j = 1, \ldots, n$. Values $v(\boldsymbol{t})$ elsewhere are interpolated naturally by the kernel $c(\cdot, \cdot)$ itself. In numerical analysis, (1.3) is known as *Nyström interpolation formula* (Atkinson and Han [3, Section 12.4]). The eigenvalue equation for T_n,

$$\lambda_n \tilde{u} = \mathrm{T}_n \tilde{u}, \qquad \tilde{u} \in \mathcal{C}(\overline{\mathfrak{D}}),$$

is equivalent to

$$\lambda_n \tilde{u}(\boldsymbol{t}_i) = \sum\nolimits_{j=1}^{n} h_n^d c(\boldsymbol{t}_i, \boldsymbol{s}_j)\, \tilde{u}(\boldsymbol{s}_j), \quad i = 1, \ldots, n,$$

which is in turn equivalent to

$$\lambda_n \tilde{\boldsymbol{u}}_n = h_n^d \Sigma_n \tilde{\boldsymbol{u}}_n, \qquad \tilde{\boldsymbol{u}}_n \in \mathbb{R}^n.$$

Thus, the eigenvalues of the discretized operator T_n are the same as the eigenvalues of the scaled covariance matrix $h_n^d \Sigma_n$.

It is known from the theory of collectively compact operators that on a sequence of meshes with $h_n \to 0$ for $n \to \infty$, the eigenvalues of T_n converge to the eigenvalues of T (Atkinson [4, 2]). A brief contemporary review on this topic can be found in the introduction of Huang et al. [31].

1.2 Spectral representation of n-periodic stationary random sequence

In this section, assume for simplicity that $\overline{\mathfrak{D}}$ is a line segment, which is covered by a uniform mesh of n nodes with spacing h. Further assume that $n = 2m + 1$ and that the nodes $\{s_k\}$ are indexed by $k \in \{-m, \ldots, -1, 0, 1, \ldots, m\} \equiv \mathbb{M}$, i.e., $\overline{\mathfrak{D}} = [-mh, mh]$ and $s_k = kh,\ k \in \mathbb{M}$. Denote $X_k = X(s_k),\ k \in \mathbb{M}$. Now, extend $\overline{\mathfrak{D}}$ periodically to the whole $\mathbb{R}$, so that the resulting infinite mesh consists of nodes $s_k = kh$ indexed by $k \in \mathbb{Z}$. The associated random element $\boldsymbol{X}_\infty = (X_k)_{k \in \mathbb{Z}}$ is then an n-periodic random sequence satisfying

$$X_{n+j} = X_j,\ \forall j \in \mathbb{Z}.$$

In geophysical sciences, the random sequence $\boldsymbol{X}_\infty$ represents the state of some dynamical system, which can be often assumed to be *weakly stationary*, i.e., it holds that

$$\begin{aligned} \mathsf{E}\, X_j &= \text{const.}, \quad \forall j \in \mathbb{Z}, \\ \operatorname{cov}(X_j, X_k) &= c(s_j, s_k) = \tilde{c}(|s_j - s_k|), \quad \forall j, k \in \mathbb{Z}, \end{aligned}$$

where the covariance function $\tilde{c}(\cdot)$ depends only on the distance between s_j and s_k. Without loss of generality, consider $\mathsf{E}\, X_j = 0,\ \forall j \in \mathbb{Z}$. In order to simplify the calculations, assume for the moment that $\boldsymbol{X}_\infty$ is complex. Each component of $\boldsymbol{X}_\infty$ has the spectral representation of the form (Brockwell and Davis [11])

$$X_k = \int_{-\pi}^{\pi} e^{i\nu s_k} dZ(\nu) \equiv \lim_{\max |\nu_j - \nu_{j-1}| \to 0} \sum_{j=1}^{r} e^{i\nu'_j s_k} (Z(\nu_j) - Z(\nu_{j-1})), \tag{1.4}$$

where $-\pi = \nu_0 < \nu_1 < \ldots < v_{r-1} < \nu_r = \pi$ is a partition of $[-\pi, \pi]$, ν'_j is an arbitrary point from the subinterval $[\nu_{j-1}, \nu_j]$, and $Z(\nu)$ is a centered random process with uncorrelated increments.

The periodicity condition $X_{-m} = X_{m+1}$ restricts the values of $\nu \in [-\pi, \pi]$, so

$$\nu_\ell = \frac{2\pi\ell}{(2m+1)h} = \frac{2\pi\ell}{nh}, \quad \ell \in \mathbb{M}.$$

Therefore, $\boldsymbol{X}_\infty$ has discrete spectrum and the right-hand side of (1.4) turns into a simple sum of uncorrelated harmonic oscillations (Yaglom [73])

$$X_k = \sum_{\ell=-m}^{m} Z_\ell e^{i\frac{2\pi\ell s_k}{nh}} = \sum_{\ell=-m}^{m} Z_\ell e^{i\frac{2\pi\ell k}{n}}, \tag{1.5}$$

where $\{Z_\ell\}_{\ell\in\mathbb{M}}$ are random variables such that $\mathsf{E}\, Z_\ell = 0,\ \ell \in \mathbb{M}$, and $\mathsf{E}(Z_k\overline{Z}_\ell) = 0,\ \ell, k \in \mathbb{M},\ \ell \neq k$.

Denote by Σ the $n \times n$ covariance matrix of $\boldsymbol{X}_n = (X_{-m}, \ldots, X_m)^\top$. Then its (j,k)-th entry is of the form

$$\begin{aligned}\sigma_{jk} = \mathsf{E}(X_j\overline{X}_k) &= \mathsf{E}\sum_{\ell_1=-m}^{m}\sum_{\ell_2=-m}^{m} Z_{\ell_1}\overline{Z}_{\ell_2}e^{i\frac{2\pi}{n}(\ell_1 j-\ell_2 k)}\\ &= \sum_{\ell=-m}^{m}\mathsf{E}\,|Z_\ell|^2 e^{i\frac{2\pi\ell}{n}(j-k)} = \sum_{\ell=-m}^{m} e^{i\frac{2\pi\ell}{n}j}\,\mathsf{E}\,|Z_\ell|^2 e^{-i\frac{2\pi\ell}{n}k},\end{aligned}$$

where we used that the random variables Z_ℓ are uncorrelated. Denote by D the $n\times n$ matrix with $\mathsf{E}\,|Z_{-m}|^2, \ldots, \mathsf{E}\,|Z_m|^2$ on its diagonal and by F_c the matrix with rows consisting of vectors $\boldsymbol{u}^{(\ell)} = [u_j^{(\ell)}]_{j\in\mathbb{M}}$, $\ell \in \mathbb{M}$, such that $u_j^{(\ell)} = e^{i\frac{2\pi\ell}{n}j}$. Then $\Sigma = F_c D\overline{F}_c^{\top}$, where $\overline{F}_c^{\top}$ is the adjoint matrix to F_c, is the spectral decomposition of Σ. Hence, $\boldsymbol{u}^{(\ell)}$ are eigenvectors of Σ and $\lambda_\ell = \mathsf{E}\,|Z_\ell|^2 \in \mathbb{R}$ the associated eigenvalues.

When $\boldsymbol{X}_n$ is real, then $X_k = \overline{X}_k$, and (1.5) implies that $Z_\ell = \overline{Z}_{-\ell}$. Therefore,

$$\lambda_{-\ell} = \mathsf{E}\,|Z_{-\ell}|^2 = \mathsf{E}\,|\overline{Z}_\ell|^2 = \mathsf{E}\,|Z_\ell|^2 = \lambda_\ell. \tag{1.6}$$

In order to obtain a real basis, define real orthogonal vectors $\boldsymbol{v}^{(\ell)} = [v_j^{(\ell)}]_{j\in\mathbb{M}}$, $\boldsymbol{w}^{(\ell)} = [w_j^{(\ell)}]_{j\in\mathbb{M}}$ by linear combinations of vectors $\boldsymbol{u}^{(\ell)}$ of the complex basis:

$$v_j^{(\ell)} = \frac{u_j^{(\ell)} + u_j^{(-\ell)}}{2} = \cos\left(\frac{2\pi\ell j}{n}\right), \quad \ell = 0, \ldots, m, \tag{1.7}$$

$$w_j^{(\ell)} = \frac{u_j^{(\ell)} - u_j^{(-\ell)}}{2i} = \sin\left(\frac{2\pi\ell j}{n}\right), \quad \ell = 1, \ldots, m. \tag{1.8}$$

Indeed, for $k \neq \ell$,

$$\left\langle \boldsymbol{v}^{(k)}, \boldsymbol{v}^{(\ell)}\right\rangle_n = \left\langle \boldsymbol{w}^{(k)}, \boldsymbol{w}^{(\ell)}\right\rangle_n = \left\langle \boldsymbol{v}^{(k)}, \boldsymbol{w}^{(\ell)}\right\rangle_n = 0$$

and

$$\left\langle \boldsymbol{v}^{(\ell)}, \boldsymbol{w}^{(\ell)}\right\rangle_n = \left\langle \frac{\boldsymbol{u}^{(\ell)} + \boldsymbol{u}^{(-\ell)}}{2}, \frac{\boldsymbol{u}^{(\ell)} - \boldsymbol{u}^{(-\ell)}}{2i}\right\rangle_n = \frac{i}{4}\left(\left\|\boldsymbol{u}^{(\ell)}\right\|_n^2 - \left\|\boldsymbol{u}^{(-\ell)}\right\|_n^2\right) = 0,$$

where $\langle\cdot,\cdot\rangle_n$ is the standard inner product in $\mathbb{R}^n$.

This new basis $\mathcal{V} = \{\boldsymbol{v}^{(0)}, \boldsymbol{v}^{(\ell)}, \boldsymbol{w}^{(\ell)}, \ell = 1, \ldots, m\}$ has again $n = 2m+1$ elements and matrix Σ can again be shown to be diagonal in this basis. The covariance between the coefficients of $\boldsymbol{X}_n$ in the basis $\mathcal{V}$ equals

$$\begin{aligned}\mathsf{E}\left(\left\langle \boldsymbol{X}_n, \boldsymbol{v}^{(k)}\right\rangle_n \left\langle \boldsymbol{X}_n, \boldsymbol{v}^{(\ell)}\right\rangle_n\right) &= \mathsf{E}\left(\left(\boldsymbol{v}^{(k)}\right)^\top \boldsymbol{X}_n\boldsymbol{X}_n^\top \boldsymbol{v}^{(\ell)}\right) = \left(\boldsymbol{v}^{(k)}\right)^\top \Sigma\boldsymbol{v}^{(\ell)}\\ &= \left(\boldsymbol{v}^{(k)}\right)^\top \Sigma\left(\frac{\boldsymbol{u}^{(\ell)} + \boldsymbol{u}^{(-\ell)}}{2}\right) = \left(\boldsymbol{v}^{(k)}\right)^\top \lambda_\ell\left(\frac{\boldsymbol{u}^{(\ell)} + \boldsymbol{u}^{(-\ell)}}{2}\right)\\ &= \left(\boldsymbol{v}^{(k)}\right)^\top \lambda_\ell\boldsymbol{v}^{(\ell)} = \begin{cases}\lambda_\ell, & \text{if } k = \ell,\\ 0, & \text{otherwise,}\end{cases}\\ \mathsf{E}\left(\left\langle \boldsymbol{X}_n, \boldsymbol{w}^{(i)}\right\rangle_n \left\langle \boldsymbol{X}_n, \boldsymbol{w}^{(j)}\right\rangle_n\right) &= \begin{cases}\lambda_j, & \text{if } i = j,\\ 0, & \text{otherwise,}\end{cases}\\ \mathsf{E}\left(\left\langle \boldsymbol{X}_n, \boldsymbol{v}^{(k)}\right\rangle_n \left\langle \boldsymbol{X}_n, \boldsymbol{w}^{(j)}\right\rangle_n\right) &= 0\end{aligned}$$

for $k, \ell = 0, \ldots, m$ and $i, j = 1, \ldots, m$, where we used the eigenvalue symmetry (1.6) and the orthogonality of the basis vectors.

Denote by $F = [\boldsymbol{w}^{(m)}, \ldots, \boldsymbol{w}^{(1)}, \boldsymbol{v}^{(0)}, \boldsymbol{v}^{(1)}, \ldots, \boldsymbol{v}^{(m)}]$ the matrix with basis vectors in its columns. Then

$$\Sigma = FDF^\top$$

is the spectral decomposition of Σ.

This result can be generalized for a domain $\mathfrak{D} \subset \mathbb{R}^d$, $d \geq 1$. Since the matrix F represents the discrete Fourier transform, we can conclude that the covariance matrix of a periodic stationary random sequence can be diagonalized by the discrete Fourier transform. We will take advantage of this result later in Section 2.2.1.

For other than periodic boundary condition, the discrete Fourier transform leads only to approximate diagonality of Σ, since the spectral coefficients Z_ℓ are uncorrelated only in the limit $n \to \infty$ (Dwivedi and Rao [21]).

2. Covariance regularization in high dimension

As mentioned in the introduction, estimation of a large covariance matrix or its inverse from a small sample is an important task in many applied fields. In the situation of low sample size, the sample covariance matrix

$$S = \frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{X}_i - \bar{\boldsymbol{X}}\right)\left(\boldsymbol{X}_i - \bar{\boldsymbol{X}}\right)^{\top}, \tag{2.1}$$

which is the most common covariance estimator, is known to perform poorly.

The main problems appearing here are rank deficiency and spurious covariances. When the matrix dimension n is larger than the number N of available observations, S has a low rank and hence is not even invertible. Although, estimation of the precision matrix Σ^{-1} is crucial in many situations. The latter undesirable phenomenon is the occurrence of high covariances between variables with small true dependency. When $\boldsymbol{X}$ is a meteorological field, these *spurious covariances* typically appear between meteorological variables at distant locations and arise only as a result of small sample size. In the context of spatial statistics, suppression of long-term covariances is called *localization*.

In order to avoid the drawbacks listed above, the estimating process in case of $n >> N$ usually requires an extra contribution. In this thesis, by a *regularization* method, we understand any estimating technique leading to positive definite covariance estimator without spurious covariances. Some of the methods consist of consecutive transform of sample covariance S, and some of them impose a specific covariance structure based on additional assumptions.

This chapter contains an overview of regularization methods suitable for high-dimensional covariance matrices, with special attention to methods used in data assimilation. We proceed from non-parametric methods, which usually work element-wise, to parametric models, which take an advantage of some specific property of the random vector. All estimators are based on a random sample $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_N$.

2.1 Non-parametric methods

In this section, we present three estimating methods that does not assume any particular distribution of $\boldsymbol{X}$, neither a particular structure of Σ. They start with the sample covariance matrix (2.1) and by means of element-wise operations transform it into a regular matrix and suppress the spurious covariances. The estimated covariance matrix is sparse and positive definite (at least asymptotically). Consistency result is usually achieved under additional condition on the relation between n and N.

2.1.1 Shrinkage

In case of sample size deficiency, the eigenstructure of the sample covariance matrix S tends to be systematically distorted (Muirhead [56]) in the sense that

the largest (smallest) eigenvalues are overestimated (underestimated). Below are two methods from a wide range of attempts that deal with correction of this phenomenon.

Linear shrinkage estimator

Ledoit and Wolf [42] assume that $\boldsymbol{X}$ has zero mean and they proposed a covariance estimator that is regular and better conditioned than the sample covariance matrix, without assuming any particular structure. The needed assumptions relate to boundedness of the ratio n/N and finite moments of $\boldsymbol{X}$. The proposed shrinkage estimator has the form

$$S_{\text{shr}} = \rho\nu I + (1-\rho)S_0, \tag{2.2}$$

where $0 < \rho < 1$, $\nu > 0$, and $S_0 = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i^\top$ is a sample covariance matrix estimating covariance of a zero-mean variable. Therefore, the resulting estimator is given by shrinking the sample covariance S_0 towards a diagonal matrix. Note that νI can be interpreted as a shrinkage target and the weight ρ as a shrinkage intensity. In Ledoit and Wolf [42], the optimal estimates of the parameters ρ, ν are sought by minimization of expected quadratic loss

$$\min_{\rho,\nu} \mathsf{E}\, \|\rho\nu I + (1-\rho)S_0 - \Sigma\|_F^2 .$$

The calculation takes into account that $\mathsf{E}\, S_0 = \Sigma$. Ledoit and Wolf [42, Lemma 2.1 and Theorem 2.1] found that optimal coefficients of the linear combination (2.2) are

$$\nu = \frac{1}{n}\operatorname{tr}(\Sigma), \qquad\qquad \rho = \frac{\beta^2}{\alpha^2+\beta^2} = \frac{\beta^2}{\delta^2}, \tag{2.3}$$

where

$$\beta^2 = \frac{1}{n}\mathsf{E}\, \|S_0 - \Sigma\|_F^2 , \qquad \alpha^2 = \frac{1}{n}\|\Sigma - \nu I\|_F^2 , \qquad \delta^2 = \frac{1}{n}\mathsf{E}\, \|S_0 - \nu I\|_F^2 ,$$

which, unfortunately, depend on the unknown covariance matrix Σ. However, consistent estimators of the parameters $\nu, \alpha, \beta, \delta$ are provided in Ledoit and Wolf [42]. Using these estimators in the formula (2.3) gives estimators $\hat{\nu}, \hat{\rho}$, which can be plugged into (2.2) in order to get an optimal covariance estimator. For $\hat{\rho} \neq 0$, positive definiteness of the identity matrix ensures the resulting matrix to be positive definite as well and therefore invertible. Moreover, in simulations carried out by Ledoit and Wolf [42], S_{shr} performed incomparably better than the sample covariance matrix S_0.

The interpretation of (2.2) in Ledoit and Wolf [42] is based on the dispersion of covariance matrix eigenvalues. The sample eigenvalues (i.e. eigenvalues of the sample covariance) are more dispersed around their mean than the true ones (Ledoit and Wolf [42], Muirhead [56]). By using the convex combination (2.2), the sample eigenvalues are shrunk towards their mean, which results in an improved estimator.

When $\boldsymbol{X}$ represents a discretization of a continuous random process, then, by making the discretization finer and finer, its covariance matrix tends to the

covariance operator of the original process. From that point of view, using of the identity matrix as a shrinkage target is problematic because in the limiting case ($n \to \infty$), it does not satisfy the trace-class property (Lemma 1).

The estimator (2.2) can be generalized to

$$S^*_{\text{shr}} = \rho T + (1 - \rho) S_0,$$

where T is a target matrix endowed with standard covariance properties like full rank and positive definiteness. Usually, T is chosen to be diagonal.

Shrinkage effect can be achieved also implicitly by computing sample covariance matrix from a random sample $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_N$ augmented by additional independent random vectors sampled from a distribution with covariance matrix T. Construction of these new random vectors depends on the assumed statistical properties of $\boldsymbol{X}$. For example in spatial modelling of meteorological variables at distant locations, it turned out to be useful to generate random vectors with spatial Markov property with covariance matrix from the Matérn family. A fruitful source of such vectors is a stochastic diffusion equation, whose stationary solution is just like that (Simpson et al. [64]). This equation can make generating of new vectors easier, especially in high-dimension.

Condition-number-regularized covariance estimation

An estimator proposed in Won and Kim [71] and Won et al. [72] falls into a broad family of shrinkage estimators, however it additionally assumes a specific distribution of $\boldsymbol{X}$. The regularization effect is achieved by bounding the condition number of the estimate by a regularization parameter κ_{max}. This ensures the resulting matrix to be invertible and well-conditioned. Since the condition number is defined as a ratio of the largest and smallest eigenvalue, this method corrects for overestimation of the largest eigenvalues and underestimation of the small eigenvalues simultaneously. The resulting estimator is called a *condition-number-regularized covariance estimator* and it is formulated as the MLE restricted on the subspace of matrices with condition number bounded by κ_{max}, i.e.,

$$\max_{\Sigma} \ell(\Sigma) \quad \text{subject to } \operatorname{cond}(\Sigma) \equiv \frac{\lambda_{max}(\Sigma)}{\lambda_{min}(\Sigma)} \leq \kappa_{max}, \tag{2.4}$$

where $\lambda_{max}(\Sigma)$, resp. $\lambda_{min}(\Sigma)$, is the largest, resp. the smallest, eigenvalue of the covariance matrix Σ. An implicit condition is that Σ be symmetric and positive definite. Therefore, the idea of this method is to search a MLE in a subspace defined by covariance matrices with condition number smaller or equal to the true condition number.

Let $l_1 \geq \ldots \geq l_n \geq 0$ be the ordered eigenvalues of the sample covariance S, so that $QLQ^\top$ with $L = \operatorname{diag}(l_1, \ldots, l_n)$ and $QQ^\top = Q^\top Q = I_n$. For a given $\kappa_{max} < \operatorname{cond}(S)$, the unique solution of the problem (2.4) is a matrix $S_{\text{con}} = Q\Lambda^* Q^\top$ (Won and Kim [71], Won et al. [72]), where the diagonal matrix Λ^* is formed by

$$\lambda_i^* = \begin{cases} \tau & \text{if } l_i \leq \tau, \\ l_i & \text{if } \tau < l_i < \kappa_{max}\tau, \\ \kappa_{max}\tau & \text{if } l_i \geq \kappa_{max}\tau. \end{cases}$$

Therefore, the sample eigenvalues l_i are truncated when they are smaller than τ or larger than $\kappa_{max}\tau$. The optimal lower cut-off level τ equals

$$\tau = \frac{\sum_{i=1}^{k_1} l_i/\kappa_{max} + \sum_{i=k_2}^{n} l_i}{k_1 + n - k_2 + 1},$$

where $k_1 \in \{1, \ldots, n\}$ is the largest index such that $l_{k_1} > \kappa_{max}\tau$ and k_2 is the smallest index such that $l_{k_2} < \tau$. Hence, τ is an average of the (scaled and) truncated eigenvalues. Note that when $\kappa_{max} \geq \mathrm{cond}(S)$, then $S_{\mathrm{con}} = S$.

An optimal κ_{max} is selected by maximization of the expected likelihood, which is approximated by using K-fold cross-validation. Details of the computational process are provided in Won et al. [72]. The authors also proved that κ_{max} selected in this way is a consistent estimator for the true condition number.

2.1.2 Tapering

An effective and simple way of localization of sample covariance matrix is multiplying S by a real sparse positive definite matrix M. The estimator is of the form

$$S_{\mathrm{tap}} = S \circ M, \tag{2.5}$$

where $\circ$ denotes the Schur product. This method is called *tapering* or *banding* (in Pourahmadi [60]) and matrix S_{tap} is called *tapered matrix*. Due to the sparsity of M, the matrix S_{tap} is sparse as well, which brings many computational advantages.

When M is real and positive definite, than, due to the Lemma 2, the matrix S_{tap} is real and positive definite, too.

Lemma 2 (Horn and Johnson [29, Theorem 7.5.3])**.** *Let A, B be real matrices of type $n \times n$. If A is positive definite and B is positive semidefinite with positive entries on the main diagonal, then $A \circ B$ is positive definite.*

In the context of data assimilation, regularization of a covariance matrix by means of Schur product has been proposed in Houtekamer and Mitchell [30] and Hamill et al. [28], based on covariance modelling by means of a covariance function. Let $X = (X(\boldsymbol{s}), \boldsymbol{s} \in \overline{\mathfrak{D}})$ be a continuous random field defined on a bounded domain $\mathfrak{D} \subset \mathbb{R}^3$ (e.g. covering part of the Earth's atmosphere) and $\mathbf{X} = (X_1, \ldots, X_n)^\top$ represents some discretization of X, i.e. $X_i = X(\boldsymbol{s}_i)$, $\boldsymbol{s}_i \in \overline{\mathfrak{D}}$. Denote by $c(\boldsymbol{s}, \boldsymbol{t}) = \mathrm{cov}(X(\boldsymbol{s}), X(\boldsymbol{t}))$ the covariance function of X. It holds that for every discretization $\mathbf{X}$, a matrix with entries $c(\boldsymbol{s}_i, \boldsymbol{s}_j)$ is positive semidefinite and, on the contrary, when a covariance matrix with entries $c(\boldsymbol{s}_i, \boldsymbol{s}_j)$ is positive definite for arbitrary set of points $\boldsymbol{s}_i$, then c define a covariance function. In Houtekamer and Mitchell [30], the matrix M is constructed so that it is positive definite and its (i, j)-th entry M_{ij} equals $\varrho(\|\boldsymbol{s}_i - \boldsymbol{s}_j\|_3)$, where $\varrho\colon [0, \infty) \to [0, 1]$ is a function with compact support and $\|\cdot\|_3$ is a norm in $\mathbb{R}^3$. In other words, ϱ can be identified with a well-defined correlation function and so (2.5) models a covariance matrix of a random field as the Schur product of sample covariance matrix with a correlation matrix defined by means of ϱ. Construction of such a function (in particular compactly supported) is a non-trivial task, which is dealt with in Gaspari and Cohn [25]. Hamill et al. [28] use a function ϱ that

is smooth and monotonically decreasing to zero, specifically, it is a polynomial of the 5th degree (Gaspari and Cohn [25, expression (4.10)]). The shape of the curve resembles the right half of a Gaussian curve that takes zero values at a finite distance, which ensures the sparsity of M.

In Furrer and Bengtsson [24], M is chosen to minimize the mean square error

$$\mathrm{MSE}(S_{\mathrm{tap}}) = \mathsf{E}\,\|\Sigma - S_{\mathrm{tap}}\|_F^2 = \mathsf{E}\left(\mathrm{tr}\left((\Sigma - S \circ M)^2\right)\right).$$

To ensure that the estimator $S \circ M$ is positive definite, the above minimization should be carried over the set of positive definite matrices M, which is a non-trivial problem. Therefore, this constraint is usually ignored, which allows a term by term minimization leading to an explicit formula for entries m_{ij} of M,

$$m_{ij} = \frac{\sigma_{ij}^2}{\sigma_{ij}^2 + (\sigma_{ij}^2 + \sigma_{ii}\sigma_{jj})/N},$$

where $[\sigma_{ij}]_{i,j=1}^n$ are entries of Σ. Given sample data, a plug-in estimator based on the entries s_{ij} of S is straightforwardly obtained. Note that m_{ij} tends to one at rate $1/N$. Finally, the resulting M is made to be positive definite by some heuristic approaches (e.g. keep only the positive eigenvalues of M and set the remaining ones equal to any small number $\varepsilon > 0$). Also, sparseness may be introduced by setting $m_{ij} = 0$ whenever $s_{ij} \approx 0$. Second attempt (inspired by Houtekamer and Mitchell [30], Hamill et al. [28]) is to parametrize the matrix M by a valid correlation function describing the correlation range and estimate its parameters by minimizing of MSE. This method ensures the resulting matrix to be positive definite, however, it does not introduce sparseness, which makes it less computationally attractive.

2.1.3 Thresholding

High-dimensional covariance matrices of random vectors representing meteorological fields are usually supposed to be sparse and to contain many zero entries. The idea of thresholding (Bickel and Levina [9]) is to neglect the small covariances in order to get an improved estimate

$$T_t(S) \equiv \left(s_{ij}\mathbf{1}_{[|s_{ij}|\geq t]} : i,j = 1,\ldots,p\right),$$

where $T_t(S)$ denotes the thresholding operator applied to the sample covariance and $t > 0$ is the chosen threshold. Bickel and Levina [9] recommend to choose t according to the following procedure. The available sample is split randomly into two parts of size $N_1 = N\left(1 - \frac{1}{\log N}\right)$ and $N_2 = \frac{N}{\log N}$ and the associated sample covariance matrices S_{N_1} and S_{N_2} are computed. This step is repeated K times and t is chosen so as to minimize

$$R(t) = \frac{1}{K}\sum_{k=1}^{K} \|T_t(S_{N_1,k}) - S_{N_2,k}\|_F^2.$$

A big advantage of this method is its simple implementation. A potential disadvantage is the loss of positive definiteness. However, it has been shown in Bickel and Levina [9] that the thresholded estimator is consistent in the operator norm as long as the true covariance matrix is sparse (in a suitable sense), the variable $\boldsymbol{X}$ is Gaussian (or sub-Gaussian) and $(\log n)/N \to 0$.

2.2 Parametric methods

When $\boldsymbol{X}$ can be assumed to have additional statistical properties like particular distribution and specific covariance structure, its covariance matrix can be estimated by using a proper covariance model.

For meteorological random fields, it is often possible to assume normal distribution, covariance stationarity and spatial Markov property (see Definition 1). Each of these properties offer a potential improvement of the covariance estimator by imposing a special parametric structure, whose parameters are estimated by standard statistical methods. Accuracy of the resulting estimate and its performance in further application (e.g., data assimilation) depend on how realistic those additional assumptions are.

In a bid to avoid complex models with a large number of parameters, many estimating methods are based on transforming $\boldsymbol{X}$ to a space where its covariance matrix is approximately diagonal. This leads to a large reduction in the number of parameters.

Possibilities of using parametric models in covariance modelling are very wide. A particular model can be used for Σ itself, its inverse, or Σ after some decomposition or transformation. From the large number of options, we have chosen two remarkable methods that are closely connected with the contribution of this thesis. These methods are briefly summarized in the following subsections and will be revisited in later chapters of the thesis.

2.2.1 Regularization in spectral domain

This approach is based on the Karhunen-Loève expansion

$$\boldsymbol{X} = \mathsf{E}\,\boldsymbol{X} + \sum_{j=1}^{n} d_j^{1/2} \xi_j \boldsymbol{v}_j, \tag{2.6}$$

where $\{d_j\}_{k=1}^n$ are coefficients, $\{\xi_j\}_{j=1}^n$ are pairwise uncorrelated random variables with zero mean and unit variance and $\{\boldsymbol{v_j}\}_{j=1}^n$ are orthonormal vectors in $\mathbb{R}^n$. Then, the covariance matrix of $\boldsymbol{X}$ can be written as

$$\Sigma = FDF^\top, \tag{2.7}$$

where $D = \operatorname{diag}(d_1, \ldots, d_n)$ and columns of the matrix F are formed by vectors $\boldsymbol{v}_j$. Since (2.7) represents the spectral decomposition of Σ, (2.6) represents $\boldsymbol{X}$ in the basis of its principal components. However, for large n, most of the sample eigenvalues are zero and estimation of the theoretical decomposition (2.6) is difficult. Thus, it is better to base a regularization method on an appropriate deterministic basis (or, more generally, on frames).

For a given F, modelling of Σ through (2.7) can be based on estimating the diagonal matrix

$$D = F^\top \Sigma F \equiv \mathcal{F}(\Sigma).$$

If we do not accept any other assumptions on D, it is possible, as in the previous section, to apply some non-parametric method on the sample covariance matrix S transformed to the spectral space, i.e., on $F^\top S F = \mathcal{F}(S)$. Of course, this matrix is never perfectly diagonal, but there remain (in practice usually small)

non-zero entries. The easiest way of the regularization of $\mathcal{F}(S)$ is using only its diagonal part (Kasanický et al. [37]). Essentially, it is equivalent to tapering of $\mathcal{F}(S)$ according to (2.5) for $M = I$.

When a specific distribution of $\boldsymbol{X}$ is assumed, the diagonal entries $\{d_{jj}\}_{j=1}^n$ of the spectral covariance matrix D can be estimated, for instance, by the method of maximum likelihood. In the case of normal distribution, the MLE based on a random sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N$ is (Turčičová et al. [67])

$$\hat{d}_{jj} = \frac{1}{N} \sum_{i=1}^{N} X_{ij}^2, \quad j = 1, \ldots, n,$$

where X_{ij} denotes the j-th entry of $\boldsymbol{X}_i$, $i = 1, \ldots, N$.

Until now, no assumption on the specific choice of the orthogonal matrix F has been made. In many practical applications, the random vector $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)^\top$ can be assumed to be weakly stationary, i.e., for $k = 1, \ldots, n$,

$$\begin{aligned} \mathsf{E}\, X_k &= \text{const.}, \\ \operatorname{cov}(X_k, X_{k+h}) &= \tilde{c}(h), \quad \forall h \in \mathbb{R} \end{aligned}$$

for some function $\tilde{c}$. As shown in Section 1.2, the covariance matrix Σ of a weakly stationary random vector $\boldsymbol{X}$ can be diagonalized by the discrete Fourier transform. Therefore, $\Sigma = \operatorname{cov} \boldsymbol{X}$ has the spectral decomposition (2.7) with F representing the discrete Fourier transform.

In meteorological and geophysical applications, the random vector $\boldsymbol{X}$ often represents a discretization of a continuous random field $X = (X(\boldsymbol{s}), \boldsymbol{s} \in \overline{\mathfrak{D}})$, where $\mathfrak{D}$ is a spatial domain covering part of the Earth. Usually, $\mathfrak{D}$ is a subset of $\mathbb{R}^d$ for $d = 1, 2, 3$. Following Section 1.2, assume for simplicity that $\overline{\mathfrak{D}}$ is a line segment $[-mh, mh]$ with $n = 2m + 1$ nodes $s_k = kh$, where $h > 0$ and $k \in \mathbb{M} = \{-m, \ldots, -1, 0, 1, \ldots, m\}$. The matrix F representing the discrete Fourier transform consists of orthonormal vectors $\boldsymbol{v}^{(\ell)} = [v_j^{(\ell)}]_{j \in \mathbb{M}}$ and $\boldsymbol{w}^{(\ell)} = [w_j^{(\ell)}]_{j \in \mathbb{M}}$ with entries

$$\begin{aligned} v_j^{(0)} &= \frac{1}{\sqrt{n}}, \\ v_j^{(\ell)} &= \sqrt{\frac{2}{n}} \cos\left(\frac{2\pi \ell j}{n}\right), \quad \ell = 1, \ldots, m, \\ w_j^{(\ell)} &= \sqrt{\frac{2}{n}} \sin\left(\frac{2\pi \ell j}{n}\right), \quad \ell = 1, \ldots, m, \end{aligned}$$

obtained by normalizing eigenvectors (1.7) and (1.8). Now, we can take the advantage of the fact that the vectors $\boldsymbol{v}^{(\ell)}$ and $\boldsymbol{w}^{(\ell)}$ are also eigenvectors of the discrete Laplace operator $\mathrm{L} : \mathbb{R}^n \to \mathbb{R}^n$, which is in one-dimensional case identical to the operator of second derivative and can be represented by the matrix

$$\mathrm{L} = \frac{1}{h^2} \begin{pmatrix} -2 & 1 & 0 & \ldots & 1 \\ 1 & -2 & 1 & \ldots & 0 \\ & & \ddots & & \\ 0 & \ldots & 1 & -2 & 1 \\ 1 & \ldots & 0 & 1 & -2 \end{pmatrix}. \tag{2.8}$$

Then for $\boldsymbol{x} = [x_j]_{j\in\mathbb{M}} \in \mathbb{R}^n$, the j-th element of the vector $\mathrm{L}(\boldsymbol{x}) \equiv \mathrm{L}\,\boldsymbol{x}$ equals

$$\mathrm{L}(\boldsymbol{x})_j = \frac{x_{j-1} - 2x_j + x_{j+1}}{h^2},$$

where $j \in \mathbb{M}$. For values x_{-m-1} and x_{m+1} located at s_{-m-1} and s_{m+1} (beyond the boundary of $\overline{\mathfrak{D}}$), we can considered various boundary conditions. The definition (2.8) corresponds to the periodic boundary condition for which $x_{-m-1} = x_m$ and $x_{m+1} = x_1$.

Eigenvalues of L associated to the definition (2.8) are

$$\lambda_\ell = -\frac{4}{h^2}\sin^2\left(\frac{\pi\ell}{n}\right), \quad \ell = 0, 1, \ldots, m, \tag{2.9}$$

and it holds

$$\begin{aligned}\mathrm{L}(\boldsymbol{v}^{(\ell)}) &= \lambda_\ell \boldsymbol{v}^{(\ell)}, \quad \ell = 0, 1, \ldots, m,\\ \mathrm{L}(\boldsymbol{w}^{(\ell)}) &= \lambda_\ell \boldsymbol{w}^{(\ell)}, \quad \ell = 1, \ldots, m.\end{aligned}$$

By denoting

$$\begin{aligned}F &= [\boldsymbol{w}^{(m)}, \ldots, \boldsymbol{w}^{(1)}, \boldsymbol{v}^{(0)}, \boldsymbol{v}^{(1)}, \ldots, \boldsymbol{v}^{(m)}],\\ \Lambda &= \mathrm{diag}\,(\lambda_m, \ldots, \lambda_1, \lambda_0, \lambda_1, \ldots, \lambda_m),\end{aligned}$$

we get the spectral decomposition $\mathrm{L} = F\Lambda F^\top$.

Having $\mathrm{L} = F\Lambda F^\top$ and $\Sigma = FDF^\top$, it is natural to model Σ as a function of the discrete Laplace operator. When $\mathrm{L} = F\Lambda F^\top$, then for a continuous matrix f, it is possible to define a matrix $f(\mathrm{L})$ by the spectral decomposition $f(\mathrm{L}) = Ff(\Lambda)F^\top$, where $f(\Lambda) = \mathrm{diag}\,(f(\lambda_m), \ldots, f(\lambda_1), f(\lambda_0), f(\lambda_1), \ldots, f(\lambda_m))$.

The eigenvalues d_{jj} (forming the diagonal of D) can thus be modelled by a suitable function of the eigenvalues (2.9) of L, i.e., $d_{jj} = f(\lambda_{|j-m-1|}),\ j = 1, \ldots, n$. Recall that $\boldsymbol{X}$ is assumed to represent a discretization of a continuous random field X. According to Lemma 1, the covariance operator T associated to X needs to have finite trace, i.e., the sum of its eigenvalues needs to be finite. Under the assumption that the kernel $c(\boldsymbol{s}, \boldsymbol{t}) = \mathrm{cov}(X(\boldsymbol{s}), X(\boldsymbol{t}))$, $\boldsymbol{s}, \boldsymbol{t} \in \overline{\mathfrak{D}}$, of T is continuous, it was shown in Section 1.1.1 that the eigenvalues $\{d_{jj}\}_{j=1}^n$ of Σ converge to the eigenvalues of the covariance operator T as $n \to \infty$, i.e., as the discretization is getting finer and finer. In order to fulfil the trace-class property in the limit case, the eigenvalues $\{f(\lambda_k)\}_{k=0}^m$ should also decrease rapidly to zero[1] with increasing k even for every m finite. Since $\{\lambda_k\}_{k=0}^m$ is a decreasing sequence of negative numbers, the function f needs to have a sufficiently fast decay for $\lambda \to -\infty$. The exponential decay is used, e.g., by Mirouze and Weaver [55]. One specific exponential model is treated in Section 3.3 for normally distributed $\boldsymbol{X}$. Another possible choice of a covariance model is a power model, where the eigenvalues of the covariance are assumed to be a negative power of $\{-\lambda_k\}_{k=0}^m$, e.g., Berner et al. [8], Gaspari et al. [26], Simpson et al. [64]. When using those models for modelling D, the number of parameters is reduced from n to the number of parameters of the particular model. The model adjusts the eigenstructure of the

[1]For results on the use of random fields, whose covariance operator does not have finite trace, in data assimilation, we refer to Kasanický [36].

covariance matrix in a similar way as shrinkage, smooth down the shape of the estimated $\hat{d}_{kk}$ and so contributes to the noise reduction.

The spectral diagonal approach can be particularly beneficial for the Ensemble Kalman filter (Section 6.2, Algorithm 1), when $R = H = I_n$ (Kasanický et al. [37]). More general and practical methods that use the Laplace operator can be found in Lindgren et al. [46], Mirouze and Weaver [55], etc.

The idea of covariance diagonalization in a transformed space appears also in Courtier et al. [16] for a continuous field defined on a sphere and in Pannekoucke et al. [58] for a discrete field defined on a 1D and 2D cyclic domain.

2.2.2 Regularization in inverse space

In many applications, the need for a precision matrix Σ^{-1} is stronger than that for Σ itself. Moreover, modelling of the precision matrix can be more convenient. One of the ideas that provides a sparse estimate of the precision matrix is based on the following result. For a subset of indices $A \subset \{1, \dots, n\}$ denote by $\boldsymbol{X}_{-A}$ the subfield $(X_i : i \in \{1, 2, \dots, n\} \setminus A)$.

Lemma 3 (Rue and Held [63, Theorem 2.2])**.** *Let $\boldsymbol{X}$ be normally distributed with mean $\boldsymbol{\mu}$ and precision matrix $\Sigma^{-1} > 0$. Then for $i \neq j$,*

$$X_i \perp X_j | \boldsymbol{X}_{-\{i,j\}} \quad \Leftrightarrow \quad (\Sigma^{-1})_{ij} = 0,$$

where $(\Sigma^{-1})_{ij}$ denotes the (i, j)-th entry of Σ^{-1}.

This result can be particularly beneficial for normal random vectors with the spatial Markov property. Below, we adopt its definition from Rue and Held [63] and consider $\boldsymbol{X}$ equipped with an adjacency structure of an undirected graph. For each X_k, denote by $\mathfrak{N}_{X_k} \subset \boldsymbol{X}_{-k}$ the set of neighbours of X_k.

Definition 1 (spatial Markov property)**.** *The random vector $\boldsymbol{X}$ is said to have the spatial Markov property if for every $k \in \{1, \dots, n\}$, the conditional distribution of X_k depends only on the neighbourhood $\mathfrak{N}_{X_k}$, i.e., for every k and every Borel set B,*

$$\mathsf{P}(X_k \in B | \boldsymbol{X}_{-k}) = \mathsf{P}(X_k \in B | \mathfrak{N}_{X_k}).$$

That is, each variable X_k of a Markov field $\boldsymbol{X}$ is conditionally independent on variables outside $\mathfrak{N}_{X_k}$.

This observation provides a powerful tool for a sparse representation of the covariance matrix of a GMRF, since its inverse (the precision matrix) is a sparse, band matrix. For illustration, in Figure 2.1, we provide three examples of two-dimensional Gaussian Markov fields (with dimension 10×10) and the corresponding precision matrices. The random field corresponding to the Figure 2.1a, where only the four closest neighbours are considered, will be called the *first order GMRF* in this thesis.

GMRFs naturally arise in the area of graphical models, where a common task is to estimate the precision matrix and its associated graph from data. For further details, we refer to e.g., Dempster [19], Whittaker [70], Giudici and Green [27]. More recent results on estimation of graphical models in high-dimension can be found in Lin et al. [45], where the score matching method, which will be studied in Chapter 4, is modified by using ℓ_1 penalty in order to accommodate sparsity of the graph.

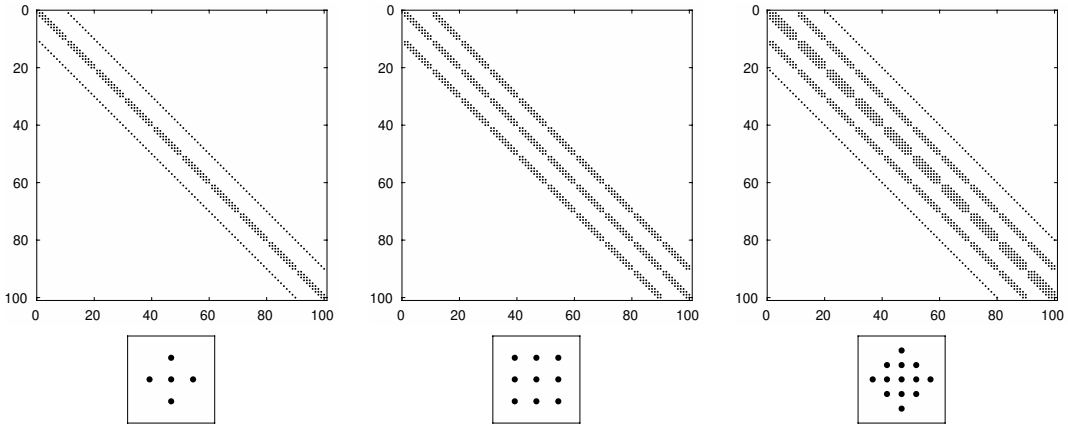


Figure 2.1: Block band-diagonal structure of inverse covariance matrix of a 10×10 GMRF with columns stacked vertically. In the bottom row: diagrams of a gridpoint and its 4, 8, 12 nearest neighbours.

Linear model for the precision matrix

Consider the model

$$\Sigma^{-1} = \beta_1 A_1 + \ldots + \beta_r A_r, \tag{2.10}$$

where $A_1, \ldots, A_r$ are known, linearly independent (in the space of matrices) and sparse matrices of type $(n \times n)$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_r)^\top$ are unknown parameters. Suitable choice of $A_1, \ldots, A_r$ are matrices A_{ij} with one on the (i, j)-th position and zero elsewhere, or symmetric matrices containing non-zero elements only on selected subdiagonals or their parts.

Parameters $\boldsymbol{\beta}$ can be estimated by the maximum likelihood method (Ueno and Tsuchiya [68]), where the numerical maximization is needed. Closed formula does not exist. The score matching estimation method (Hyvärinen [33]) makes it possible to compute a closed form estimate (Turčičová et al. [66]), which is described in Chapter 4 in greater detail. Under further assumptions, both these methods provide consistent estimators. However, positive definiteness of the resulting matrix estimate is guaranteed only asymptotically (as follows from consistency).

3. Nested maximum likelihood estimators

The principal result of this chapter is the observation that if parameters of a distribution are fitted as the MLE, then, under some assumptions on the true parameters, the estimate using fewer parameters is asymptotically more (or equally) accurate. Although, our result is asymptotic, the difference in accuracy is often significant even for small samples. This observation points out the importance of searching for a covariance model that is as accurate as possible for the given problem because overparametrization can be very harmful. A simulation study comparing the accuracy of covariance submodels in spectral and inverse space is provided at the end of the chapter. The results contained in this chapter were published in Turčičová et al. [67].

3.1 Asymptotic variance of the maximum likelihood estimator

First, we briefly review some standard results of maximum likelihood method following Lehmann and Casella [43]. Suppose $\mathbb{X}_N = [\boldsymbol{X}_1, \dots, \boldsymbol{X}_N]$ is a random sample from a distribution on $\mathbb{R}^n$ with density $f(\boldsymbol{x}|\boldsymbol{\theta})$ with unknown parameter vector $\boldsymbol{\theta}$ in a parameter space $\boldsymbol{\Theta} \subset \mathbb{R}^s$. The maximum likelihood estimate $\hat{\boldsymbol{\theta}}_N$ of the true parameter $\boldsymbol{\theta}_0$ is defined by maximizing the likelihood

$$\hat{\boldsymbol{\theta}}_N = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}|\mathbb{X}_N), \quad \mathcal{L}(\boldsymbol{\theta}|\mathbb{X}_N) = \prod_{i=1}^{N} \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{X}_i), \quad \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{x}) = f(\boldsymbol{x}|\boldsymbol{\theta}),$$

or, equivalently, maximizing the log-likelihood

$$\hat{\boldsymbol{\theta}}_N = \arg\max_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}|\mathbb{X}_N), \quad \ell(\boldsymbol{\theta}|\mathbb{X}_N) = \sum_{i=1}^{N} \ell(\boldsymbol{\theta}|\boldsymbol{X}_i), \quad \ell(\boldsymbol{\theta}|\boldsymbol{x}) = \log f(\boldsymbol{x}|\boldsymbol{\theta}).$$

We adopt the usual assumptions (Lehmann and Casella [43, Section 6.3 and 6.5]) that

(A1) the true parameter $\boldsymbol{\theta}_0$ lies in an open subset $\tilde{\boldsymbol{\Theta}}$ of $\boldsymbol{\Theta}$,

(A2) the density f determines the parameter $\boldsymbol{\theta}$ uniquely in the sense that $f(\boldsymbol{x}|\boldsymbol{\theta}_1) = f(\boldsymbol{x}|\boldsymbol{\theta}_2)$ a.e. if and only if $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$,

(A3) $\mathcal{X} = \{\boldsymbol{x} : f(\boldsymbol{x}|\boldsymbol{\theta}) > 0\}$ does not depend on $\boldsymbol{\theta}$,

(A4) the derivative $\frac{\partial^3}{\partial\theta_i\partial\theta_j\partial\theta_k} f(\boldsymbol{x}|\boldsymbol{\theta})$ exists for all $\boldsymbol{\theta} \in \tilde{\boldsymbol{\Theta}}$, for almost all $\boldsymbol{x} \in \mathcal{X}$ and for every $i, j, k = 1, \dots, s$,

(A5) $\int_{\mathcal{X}} \frac{\partial^2}{\partial\theta_i\partial\theta_j} f(\boldsymbol{x}|\boldsymbol{\theta}) d\boldsymbol{x} = 0$ for all $\boldsymbol{\theta} \in \tilde{\boldsymbol{\Theta}}$ and every $i, j = 1, \dots, s$,

(A6) for all $i, j, k = 1, \dots, s$, there exists a function $M_{ijk}(\boldsymbol{x}) \geq 0$ such that $\mathsf{E}\, M_{ijk}(\boldsymbol{x}) < \infty$ and $\left|\frac{\partial^3}{\partial\theta_i\partial\theta_j\partial\theta_k} \log f(\boldsymbol{x}|\boldsymbol{\theta})\right| \leq M_{ijk}(\boldsymbol{x})$ for all $\boldsymbol{\theta} \in \tilde{\boldsymbol{\Theta}}$ and almost all $\boldsymbol{x} \in \mathcal{X}$.

Then the error of the estimate is asymptotically normal (Lehmann and Casella [43, Theorem 5.1, p. 463]),

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}_n(\mathbf{0}, C_{\boldsymbol{\theta}_0}), \text{ as } N \to \infty, \tag{3.1}$$

where

$$C_{\boldsymbol{\theta}_0} = \mathcal{I}_{\boldsymbol{\theta}_0}^{-1}, \quad \mathcal{I}_{\boldsymbol{\theta}_0} = \mathsf{E}\left(\nabla_{\boldsymbol{\theta}}^{\top}\ell(\boldsymbol{\theta}_0|\boldsymbol{X})\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}_0|\boldsymbol{X})\right), \quad \boldsymbol{X} \sim f\left(\boldsymbol{x}|\boldsymbol{\theta}_0\right). \tag{3.2}$$

The matrix $\mathcal{I}_{\boldsymbol{\theta}_0}$ is called the Fisher information matrix for the parametrization $\boldsymbol{\theta}$. Here, $\boldsymbol{X}$, $\boldsymbol{x}$, and $\boldsymbol{\theta}$ are columns, while the gradient $\nabla_{\boldsymbol{\theta}}\ell$ of ℓ with respect to the parameter $\boldsymbol{\theta}$ is a row vector, which is compatible with the dimensioning of Jacobian matrices below. The column vector $(\nabla_{\boldsymbol{\theta}}\ell)^{\top}$ is denoted by $\nabla_{\boldsymbol{\theta}}^{\top}\ell$.

3.2 Asymptotic variance of nested estimators

Now, suppose we have an additional information that the true parameter $\boldsymbol{\theta}_0$ lies in a subspace of $\boldsymbol{\Theta}$, which is parametrized by $r \leq s$ parameters $(\varphi_1, \ldots, \varphi_r)^{\top} = \boldsymbol{\varphi}$. Denote by $J_{\boldsymbol{\varphi}}(\boldsymbol{\theta}(\boldsymbol{\varphi}))$ the $s \times r$ Jacobian matrix with entries $\frac{\partial\theta_i}{\partial\varphi_j}$. In the next theorem, the asymptotic covariance of the maximum likelihood estimator for $\boldsymbol{\varphi}$

$$\hat{\boldsymbol{\varphi}}_N = \arg\max_{\boldsymbol{\varphi}} \ell\left(\boldsymbol{\varphi}|\mathbb{X}_N\right), \quad \ell\left(\boldsymbol{\varphi}|\mathbb{X}_N\right) = \sum_{i=1}^{N} \ell\left(\boldsymbol{\varphi}|\boldsymbol{X}_i\right), \quad \ell\left(\boldsymbol{\varphi}|\boldsymbol{x}\right) = \log f\left(\boldsymbol{x}|\boldsymbol{\theta}\left(\boldsymbol{\varphi}\right)\right),$$

is derived based on the asymptotic covariance of $\boldsymbol{\theta}$ in (3.1).

Theorem 4. *Assume that the map $\boldsymbol{\varphi} \mapsto \boldsymbol{\theta}(\boldsymbol{\varphi})$ is one-to-one from $\boldsymbol{\Phi} \subset \mathbb{R}^r$ to $\boldsymbol{\Theta}$, the map $\boldsymbol{\varphi} \mapsto \boldsymbol{\theta}(\boldsymbol{\varphi})$ is continuously differentiable, $J_{\boldsymbol{\varphi}}(\boldsymbol{\theta}(\boldsymbol{\varphi}))$ is full rank for all $\boldsymbol{\varphi} \in \boldsymbol{\Phi}$, and $\boldsymbol{\theta}_0 = \boldsymbol{\theta}(\boldsymbol{\varphi}_0)$ with $\boldsymbol{\varphi}_0$ in the interior of $\boldsymbol{\Phi}$. Then,*

$$\sqrt{N}(\hat{\boldsymbol{\varphi}}_N - \boldsymbol{\varphi}_0) \xrightarrow{d} \mathcal{N}_r\left(\mathbf{0}, C_{\boldsymbol{\varphi}_0}\right) \text{ as } N \to \infty, \tag{3.3}$$

where $C_{\boldsymbol{\varphi}_0} = \mathcal{I}_{\boldsymbol{\varphi}_0}^{-1}$, with $\mathcal{I}_{\boldsymbol{\varphi}_0}$ the Fisher information matrix of the parametrization $\boldsymbol{\varphi}$ given by

$$\mathcal{I}_{\boldsymbol{\varphi}_0} = J_{\boldsymbol{\varphi}}(\boldsymbol{\theta}(\boldsymbol{\varphi}_0))^{\top}\, \mathcal{I}_{\boldsymbol{\theta}_0} J_{\boldsymbol{\varphi}}(\boldsymbol{\theta}(\boldsymbol{\varphi}_0)).$$

Proof. From (3.2) and the chain rule

$$\nabla_{\boldsymbol{\varphi}}\ell(\boldsymbol{\varphi}|\boldsymbol{X}) = \nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}|\boldsymbol{X})J_{\boldsymbol{\varphi}}(\boldsymbol{\theta}(\boldsymbol{\varphi})),$$

it follows

$$\begin{aligned}
\mathcal{I}_{\boldsymbol{\varphi}_0} &= \mathsf{E}\left(\nabla_{\boldsymbol{\varphi}}^{\top}\ell(\boldsymbol{\varphi}_0|\boldsymbol{X})\nabla_{\boldsymbol{\varphi}}\ell(\boldsymbol{\varphi}_0|\boldsymbol{X})\right) \\
&= J_{\boldsymbol{\varphi}}(\boldsymbol{\theta}(\boldsymbol{\varphi}_0))^{\top}\, \mathsf{E}\left(\nabla_{\boldsymbol{\theta}}^{\top}\ell(\boldsymbol{\theta}_0|\boldsymbol{X})\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}_0|\boldsymbol{X})\right) J_{\boldsymbol{\varphi}}(\boldsymbol{\theta}(\boldsymbol{\varphi}_0)) \\
&= J_{\boldsymbol{\varphi}}(\boldsymbol{\theta}(\boldsymbol{\varphi}_0))^{\top}\, \mathcal{I}_{\boldsymbol{\theta}_0} J_{\boldsymbol{\varphi}}(\boldsymbol{\theta}(\boldsymbol{\varphi}_0)).
\end{aligned}$$

The asymptotic distribution (3.3) is now (3.1) applied to $\boldsymbol{\varphi}$. □

When the parameter $\boldsymbol{\theta}$ is the quantity of interest in an application, it is useful to express the estimate and its variance in terms of the original parameter $\boldsymbol{\theta}$ rather than the subspace parameter $\boldsymbol{\varphi}$.

Corollary. Under the assumptions of Theorem 4,

$$\sqrt{N}\left(\boldsymbol{\theta}\left(\hat{\boldsymbol{\varphi}}_N\right)-\boldsymbol{\theta}_0\right) \xrightarrow{d} \mathcal{N}_n\left(\mathbf{0}, C_{\boldsymbol{\theta}(\boldsymbol{\varphi}_0)}\right) \text{ as } N\to\infty,$$

where

$$\begin{aligned} C_{\boldsymbol{\theta}(\boldsymbol{\varphi}_0)} &= J_{\boldsymbol{\varphi}}(\boldsymbol{\theta}(\boldsymbol{\varphi}_0))\,\mathcal{I}_{\boldsymbol{\varphi}_0}^{-1} J_{\boldsymbol{\varphi}}(\boldsymbol{\theta}(\boldsymbol{\varphi}_0))^{\top} \\ &= J_{\boldsymbol{\varphi}}(\boldsymbol{\theta}(\boldsymbol{\varphi}_0))\left(J_{\boldsymbol{\varphi}}(\boldsymbol{\theta}(\boldsymbol{\varphi}_0))^{\top}\,\mathcal{I}_{\boldsymbol{\theta}_0} J_{\boldsymbol{\varphi}}(\boldsymbol{\theta}(\boldsymbol{\varphi}_0))\right)^{-1} J_{\boldsymbol{\varphi}}(\boldsymbol{\theta}(\boldsymbol{\varphi}_0))^{\top}. \end{aligned} \tag{3.4}$$

Proof. The lemma follows from (3.3) by the delta method (Rao [62, p. 387]), since the map $\boldsymbol{\varphi}\mapsto\boldsymbol{\theta}(\boldsymbol{\varphi})$ is continuously differentiable. □

Remark 1. The matrix $C_{\boldsymbol{\theta}(\boldsymbol{\varphi}_0)}$ is singular, so it cannot be written as the inverse of another matrix, but it can be understood as the inverse $\mathcal{I}^{-1}_{\boldsymbol{\theta}(\boldsymbol{\varphi}_0)}$ of the Fisher information matrix for $\boldsymbol{\varphi}$, embedded in the larger parameter space $\boldsymbol{\Theta}$.

The next theorem shows that in case of two parametrizations $\boldsymbol{\varphi}$ and $\boldsymbol{\theta}$ which are nested, the smaller parametrization has smaller or equal asymptotic covariance than the larger one. For symmetric matrices A and B, $A\le B$ means that $A-B$ is positive semidefinite.

Theorem 5. *Suppose that $\boldsymbol{\varphi}$ satisfies the assumptions in Theorem 4. Then,*

$$C_{\boldsymbol{\theta}(\boldsymbol{\varphi}_0)} \le C_{\boldsymbol{\theta}_0}. \tag{3.5}$$

In addition, if $\boldsymbol{U}\sim\mathcal{N}_n(\mathbf{0}, C_{\boldsymbol{\theta}(\boldsymbol{\varphi}_0)})$ and $\boldsymbol{V}\sim\mathcal{N}_n\left(\mathbf{0}, C_{\boldsymbol{\theta}_0}\right)$ are random vectors with the asymptotic distributions of the estimates $\boldsymbol{\theta}\left(\hat{\boldsymbol{\varphi}}_N\right)$ and $\hat{\boldsymbol{\theta}}_N$, then

$$\mathsf{E}\,\|\boldsymbol{U}\|_n^2 = \operatorname{tr}\left(C_{\boldsymbol{\theta}(\boldsymbol{\varphi}_0)}\right) \le \operatorname{tr}\left(C_{\boldsymbol{\theta}^0}\right) = \mathsf{E}\,\|\boldsymbol{V}\|_n^2, \tag{3.6}$$

where $\|\boldsymbol{V}\|_n$ is the standard Euclidean norm in $\mathbb{R}^n$.

Proof. Denote $A=\mathcal{I}_{\boldsymbol{\theta}_0}$, $B=J_{\boldsymbol{\varphi}}(\boldsymbol{\theta}(\boldsymbol{\varphi}_0))$ and define

$$P_B = A^{1/2}B(B^{\top}AB)^{-1}B^{\top}A^{1/2}.$$

The matrix P_B is symmetric and idempotent, hence it is an orthogonal projection. In addition,

$$\operatorname{Range} P_B = \operatorname{Range} A^{1/2}B \subset \operatorname{Range} I_n.$$

Consequently, $P_B\le I$ holds from standard properties of orthogonal projections, and (3.5) follows.

To prove (3.6), note that for random vector $\boldsymbol{X}$ with $\mathsf{E}\,\boldsymbol{X}=\mathbf{0}$ and finite second moment, $\mathsf{E}\,\|\boldsymbol{X}\|_n^2 = \operatorname{tr}(\operatorname{cov}\boldsymbol{X})$. The proof is concluded by using the fact that for symmetric matrices, $A\le B$ implies $\operatorname{tr}A\le\operatorname{tr}B$, cf. e.g., Carlen [15]. □

Remark 2. In the practically interesting cases when there is a large difference in the dimensions of the parameters $\boldsymbol{\varphi}$ and $\boldsymbol{\theta}$, many eigenvalues in the covariance of the estimation error become zero. The computational tests in Section 3.4 show that the resulting decrease of the estimation error can be significant.

3.3 Application: nested covariance models

A frequent assumption in data assimilation is the weak stationarity, which leads to diagonality in spectral space (Courtier et al. [16], Pannekoucke et al. [58]), as described in Section 1.2. Besides, part of the assimilation methods that dominate today's practice of meteorological services (the so called variational methods) usually employ a covariance model based on a series of transformations leading to independence of variables (Bannister [5], Michel and Auligné [54]). One way or other, it results in an estimation problem for a diagonal covariance matrix. The distribution is not normal but formulas derived from the normal distribution are used anyway.

In what follows, we introduce the particular covariance structures, state some known facts on full and diagonal covariance, propose parametric models for the diagonal and compute corresponding MLE.

3.3.1 Sample covariance

Assume that the top-level parameter space $\boldsymbol{\Theta}$ consists of all symmetric positive definite matrices, resulting in the parametrization Σ with $\frac{n(n+1)}{2}$ independent parameters. Recall that $\mathbb{X}_N = [\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N]$ denotes a matrix with columns formed by vectors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N$. The likelihood of Σ given a sample $\mathbb{X}_N$ from $\mathcal{N}_n(\mathbf{0}, \Sigma_0)$ is

$$\mathcal{L}\left(\Sigma|\mathbb{X}_N\right) = \frac{1}{\left(\det \Sigma\right)^{N/2} \left(2\pi\right)^{nN/2}} e^{-\frac{1}{2}\operatorname{tr}\left(\Sigma^{-1}\mathbb{X}_N\mathbb{X}_N^\top\right)}.$$

If $N \geq n$, it is well known (e.g. Muirhead [57, p. 83]) that the likelihood is maximized at the sample covariance matrix

$$S_{(\boldsymbol{\mu}=0)} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i^\top. \tag{3.7}$$

The Fisher information matrix for the maximum likelihood covariance estimator (Magnus and Neudecker [49, p. 356]) is

$$\mathcal{I}_{\operatorname{vec}(\Sigma_0)} = \frac{1}{2}\Sigma_0^{-1} \otimes \Sigma_0^{-1},$$

where $\otimes$ stands for the Kronecker product and vec is an operator that transforms a matrix into a vector by stacking the columns of the matrix one underneath the other. This matrix has dimension $n^2 \times n^2$.

Remark 3. If $S_{(\boldsymbol{\mu}=0)}$ is singular, $\mathcal{L}\left(S_{(\boldsymbol{\mu}=0)}|\mathbb{X}_N\right)$ cannot be evaluated because that requires the inverse of $S_{(\boldsymbol{\mu}=0)}$. Also, in that case, the likelihood $\mathcal{L}\left(\Sigma|\mathbb{X}_N\right)$ is not bounded above on the set of all $\Sigma > 0$, thus the maximum of $\mathcal{L}\left(\Sigma|\mathbb{X}_N\right)$ does not exist. To show that, consider an orthonormal change of basis so that the vectors in $\operatorname{span}\left(\mathbb{X}_N\right)$ come first, write vectors and matrices in the corresponding 2×2 block form, and let

$$S'_{(\boldsymbol{\mu}=0)} = \begin{bmatrix} S'_{11} & 0 \\ 0 & 0 \end{bmatrix}, \quad S'_{11} > 0.$$

Then $\lim_{a\to 0^+} \mathbb{X}_N^\top \left(S'_{(\boldsymbol{\mu}=0)} + aI\right)^{-1} \mathbb{X}_N$ exists, but $\lim_{a\to 0^+} \det\left(S'_{(\boldsymbol{\mu}=0)} + aI\right) = 0$, thus

$$\lim_{a\to 0^+} \mathcal{L}\left(S'_{(\boldsymbol{\mu}=0)} + aI \middle| \mathbb{X}_N\right) = \infty.$$

Note that when the likelihood is redefined in terms of the subspace span $(\mathbb{X}_N)$ only, the sample covariance can be obtained by maximization on the subspace (Rao [62, p. 527]).

Suppose

$$\boldsymbol{X} \sim \mathcal{N}_n(\mathbf{0}, D_0), \tag{3.8}$$

where $\boldsymbol{X}$ denotes the random vector after an appropriate transform and D_0 is a diagonal matrix.

When the true covariance is diagonal (i.e., $\Sigma_0 \equiv D_0$, cf. (3.8)), a significant improvement can be achieved by setting the off-diagonal terms of sample covariance (3.7) to zero, which is equivalent to tapering of $S_{(\boldsymbol{\mu}=0)}$ according to (2.5) with $M = I_n$, i.e.,

$$S_{\text{tap}} = \operatorname{diag}\left(S_{(\boldsymbol{\mu}=0)}\right). \tag{3.9}$$

It is known that using only the diagonal of the unbiased sample covariance

$$\tilde{S}_{(\boldsymbol{\mu}=0)} = \frac{1}{N-1}\sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i^\top$$

results in smaller (or equal) Frobenius norm of the error pointwise,

$$\mathsf{E}\left\|\operatorname{diag}\left(\tilde{S}_{(\boldsymbol{\mu}=0)}\right) - D_0\right\|_F \le \mathsf{E}\left\|\tilde{S}_{(\boldsymbol{\mu}=0)} - D_0\right\|_F, \tag{3.10}$$

cf. Furrer and Bengtsson [24] for the case when the mean is assumed to be known like here, and Kasanický et al. [37] for the unbiased sample covariance and unknown mean.

3.3.2 Diagonal covariance

The parameter space $\boldsymbol{\Theta}_1 \subset \boldsymbol{\Theta}$ consisting of all diagonal matrices with positive diagonal, with n parameters $\boldsymbol{d} = (d_1, \ldots, d_n)^\top$, can be viewed as a simple class of models for either covariance or its inverse. The log-likelihood function for $D = \operatorname{diag}(d_1, \ldots, d_n)$ as covariance with a given random sample $\mathbb{X}_N = [\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N]$ from $\mathcal{N}_n(\mathbf{0}, D_0)$ is

$$\ell(D|\mathbb{X}_N) = -\frac{N}{2}\log\left((2\pi)^n \det D\right) - \frac{1}{2}\sum_{i=1}^{N} \boldsymbol{X}_i^\top D^{-1} \boldsymbol{X}_i$$

and has its maximum at

$$\hat{d}_k = \frac{1}{N}\sum_{i=1}^{N} X_{i,k}^2, \quad k = 1, \ldots, n,$$

where $X_{i,k}$ denotes the k-th entry of $\boldsymbol{X}_i$. The sum of squares $S_k^2 = \sum_{i=1}^{N} X_{i,k}^2$ is a sufficient statistic for the variance d_k. Thus, the maximum likelihood estimator of covariance in the class of diagonal matrices is

$$\hat{D}_N^{(1)} = \frac{1}{N}\operatorname{diag}\left(S_1^2, \ldots, S_n^2\right). \tag{3.11}$$

Denote $D_0^{(1)} = \operatorname{diag}(d_{01}, \dots, d_{0n})$. It is easy to compute the Fisher information matrix explicitly,

$$\mathcal{I}_{D_0^{(1)}} = \operatorname{diag}\left(\frac{1}{2d_{01}^2}, \dots, \frac{1}{2d_{0n}^2}\right),$$

which is an $n \times n$ matrix and gives the asymptotic covariance of the estimation error

$$\frac{1}{N} C_{D_0^{(1)}} = \frac{1}{N} \mathcal{I}^{-1}_{D_0^{(1)}} = \frac{1}{N} \operatorname{diag}\left(2d_{01}^2, \dots, 2d_{0n}^2\right)$$

from (3.1).

3.3.3 Diagonal covariance with prescribed decay by 3 parameters

A more specific situation appears when we have an additional information that the matrix $D_0 = \operatorname{diag}(d_{01}, \dots, d_{0n})$ is not only diagonal, but its diagonal entries have a prescribed decay. For instance, this decay can be governed by a model of the form $d_k = ((c_1 + c_2 h_k) f_k(\alpha))^{-1}$, $k = 1, \dots, n$, where c_1, c_2 and α are unknown parameters, $h_1, \dots, h_n$ are known positive numbers, and $f_1, \dots, f_n$ are known differentiable functions. For easier computation it is useful to work with

$$\tau_k = \frac{1}{d_k} = (c_1 + c_2 h_k) f_k(\alpha).$$

Maximum likelihood estimators for the true parameters c_{01}, c_{02}, and α_0 can be computed efficiently from the likelihood

$$\ell(D|\mathbb{X}_N) = -\frac{N}{2} n \log(2\pi) + \frac{N}{2} \sum_{k=1}^{n} \log \tau_k - \frac{1}{2} \sum_{k=1}^{n} \tau_k S_k^2$$

by using the chain rule. It holds that

$$\begin{aligned} \frac{\partial \ell}{\partial c_1} &= \sum_{k=1}^{n} \frac{\partial \ell}{\partial \tau_k} \frac{\partial \tau_k}{\partial c_1} = \sum_{k=1}^{n} \left(\frac{N}{2\tau_k} - \frac{S_k^2}{2}\right) \frac{\partial \tau_k}{\partial c_1} \\ &= \frac{N}{2} \sum_{k=1}^{n} \left(\frac{1}{(c_1 + c_2 h_k) f_k(\alpha)} - \frac{1}{N} S_k^2\right) f_k(\alpha). \end{aligned}$$

Setting this derivative equal to zero we get

$$\sum_{k=1}^{n} \left(\frac{1}{c_1 + c_2 h_k} - \frac{1}{N} S_k^2 f_k(\alpha)\right) = 0. \tag{3.12}$$

Analogously,

$$\frac{\partial \ell}{\partial c_2} = \sum_{k=1}^{n} \frac{\partial \ell}{\partial \tau_k} \frac{\partial \tau_k}{\partial c_2} = \frac{N}{2} \sum_{k=1}^{n} \left(\frac{1}{(c_1 + c_2 h_k) f_k(\alpha)} - \frac{1}{N} S_k^2\right) h_k f_k(\alpha),$$

so the equation for estimating the parameter c_2 is

$$\sum_{k=1}^{n} \left(\frac{h_k}{c_1 + c_2 h_k} - \frac{1}{N} S_k^2 h_k f_k(\alpha)\right) = 0. \tag{3.13}$$

Similarly,

$$\frac{\partial \ell}{\partial \alpha} = \sum_{k=1}^{n} \frac{\partial \ell}{\partial \tau_k} \frac{\partial \tau_k}{\partial \alpha} = \frac{N}{2} \sum_{k=1}^{n} \left(\frac{1}{(c_1 + c_2 h_k) f_k(\alpha)} - \frac{1}{N} S_k^2 \right) (c_1 + c_2 h_k) \frac{\partial f_k(\alpha)}{\partial \alpha}$$
$$= \frac{N}{2} \sum_{k=1}^{n} \left(\frac{1}{f_k(\alpha)} - \frac{1}{N} S_k^2 (c_1 + c_2 h_k) \right) \frac{\partial f_k(\alpha)}{\partial \alpha}$$

and setting the derivative to zero, we get

$$\sum_{k=1}^{n} \left(\frac{1}{f_k(\alpha)} \frac{\partial f_k(\alpha)}{\partial \alpha} - \frac{1}{N} S_k^2 (c_1 + c_2 h_k) \frac{\partial f_k(\alpha)}{\partial \alpha} \right) = 0. \tag{3.14}$$

The maximum likelihood estimator for D_0 is then given by

$$\hat{D}^{(2)} = \operatorname{diag} \left\{ ((\hat{c}_1 + \hat{c}_2 h_k) f_k(\hat{\alpha}))^{-1}, k = 1, \ldots, n \right\}, \tag{3.15}$$

where $(\hat{c}_1, \hat{c}_2, \hat{\alpha})$ is the solution of the system (3.12, 3.13, 3.14). This expression corresponds to searching a maximum likelihood estimator of D_0 in the subspace $\boldsymbol{\Theta}_2 \subset \boldsymbol{\Theta}_1 \subset \boldsymbol{\Theta}$ formed by matrices of the form

$$\operatorname{diag} \left\{ ((c_1 + c_2 h_k) f_k(\alpha))^{-1}, k = 1, \ldots, n \right\}.$$

For completeness, the asymptotic covariance of the estimation error of

$$D_0^{(2)} = \operatorname{diag} \left\{ d_k(c_{01}, c_{02}, \alpha_0), k = 1, \ldots, n \right\},$$

based on a sample of size N is

$$\frac{1}{N} C_{D_0^{(2)}} = \frac{1}{N} \nabla \boldsymbol{d}(c_{01}, c_{02}, \alpha_0) \mathcal{I}_{c_{01}, c_{02}, \alpha_0}^{-1} \left(\nabla \boldsymbol{d}(c_{01}, c_{02}, \alpha_0) \right)^{\top}$$

from (3.4), where the Fisher information matrix $\mathcal{I}_{c_{01}, c_{02}, \alpha_0}$ is the 3×3 matrix

$$\mathcal{I}_{c_1, c_2, \alpha} = \begin{bmatrix} \frac{1}{2} \sum_{k=1}^{n} \frac{1}{(c_1 + c_2 h_k)^2} & \frac{1}{2} \sum_{k=1}^{n} \frac{h_k}{(c_1 + c_2 h_k)^2} & \frac{1}{2} \sum_{k=1}^{n} \frac{1}{(c_1 + c_2 h_k) f_k(\alpha)} \frac{\partial f_k(\alpha)}{\partial \alpha} \\ \frac{1}{2} \sum_{k=1}^{n} \frac{h_k}{(c_1 + c_2 h_k)^2} & \frac{1}{2} \sum_{k=1}^{n} \frac{h_k^2}{(c_1 + c_2 h_k)^2} & \frac{1}{2} \sum_{k=1}^{n} \frac{h_k}{(c_1 + c_2 h_k) f_k(\alpha)} \frac{\partial f_k(\alpha)}{\partial \alpha} \\ \frac{1}{2} \sum_{k=1}^{n} \frac{1}{(c_1 + c_2 h_k) f_k(\alpha)} \frac{\partial f_k(\alpha)}{\partial \alpha} & \frac{1}{2} \sum_{k=1}^{n} \frac{h_k}{(c_1 + c_2 h_k) f_k(\alpha)} \frac{\partial f_k(\alpha)}{\partial \alpha} & \frac{1}{2} \sum_{k=1}^{n} \frac{1}{f_k^2(\alpha)} \left(\frac{\partial f_k(\alpha)}{\partial \alpha} \right)^2 \end{bmatrix}$$

evaluated at $(c_{01}, c_{02}, \alpha_0)$ and

$$\boldsymbol{d}(c_{01}, c_{02}, \alpha_0) = (d_1(c_{01}, c_{02}, \alpha_0), \ldots, d_n(c_{01}, c_{02}, \alpha_0))^{\top}$$
$$= \left(((c_{01} + c_{02} h_1) f_1(\alpha_0))^{-1}, \ldots, ((c_{01} + c_{02} h_n) f_n(\alpha_0))^{-1} \right)^{\top}.$$

3.3.4 Diagonal covariance with prescribed decay by 2 parameters

Consider an even more specific model for diagonal elements with two parameters: $d_k = (c f_k(\alpha))^{-1}$, i.e. $\tau_k = c f_k(\alpha)$, $k = 1, \ldots, n$, where c and α are unknown

parameters. Maximum likelihood estimators for c_0 and α_0 can be computed similarly as in the previous case. The estimating equations have the form

$$\begin{aligned}\frac{1}{c} &= \frac{1}{n}\sum_{k=1}^{n}\frac{1}{N}S_k^2 f_k(\alpha),\\ \frac{1}{c}\sum_{k=1}^{n}\frac{1}{f_k(\alpha)}\frac{\partial f_k(\alpha)}{\partial\alpha} &= \sum_{k=1}^{n}\frac{1}{N}S_k^2\frac{\partial f_k(\alpha)}{\partial\alpha},\end{aligned}$$

which can be rearranged to

$$\frac{1}{c} = \frac{1}{n}\sum_{k=1}^{n}\frac{1}{N}S_k^2 f_k(\alpha), \tag{3.16}$$

$$0 = \sum_{k=1}^{n} S_k^2 f_k(\alpha)\left(\frac{1}{f_k(\alpha)}\frac{\partial f_k(\alpha)}{\partial\alpha} - \frac{1}{n}\sum_{j=1}^{n}\frac{1}{f_j(\alpha)}\frac{\partial f_j(\alpha)}{\partial\alpha}\right). \tag{3.17}$$

Equation (3.17) is an implicit formula for estimating α_0. Its result can be used for estimating c_0 through (3.16). The maximum likelihood estimator for D_0 is then given by

$$\hat{D}^{(3)} = \operatorname{diag}\left((\hat{c}f_1(\hat{\alpha}))^{-1},\ldots,(\hat{c}f_n(\hat{\alpha}))^{-1}\right), \tag{3.18}$$

where $\hat{c}$ and $\hat{\alpha}$ are MLEs of c_0 and α_0. It corresponds to searching a maximum likelihood estimator of D_0 in the subspace $\boldsymbol{\Theta}_3 \subset \boldsymbol{\Theta}_2 \subset \boldsymbol{\Theta}_1 \subset \boldsymbol{\Theta}$ formed by diagonal matrices of the form $\operatorname{diag}\{(cf_k(\alpha))^{-1}, k = 1,\ldots,n\}$.

The covariance of the asymptotic distribution of the estimator $\hat{D}^{(3)}$ is

$$\frac{1}{N}C_{D_0^{(3)}} = \frac{1}{N}\nabla\boldsymbol{d}(c_0,\alpha_0)\mathcal{I}_{c_0,\alpha_0}^{-1}\left(\nabla\boldsymbol{d}(c_0,\alpha_0)\right)^{\top},$$

from (3.4), where Fisher information matrix $\mathcal{I}_{c_0,\alpha_0}$ is the 2×2 matrix

$$\mathcal{I}_{c,\alpha} = \begin{bmatrix} \frac{n}{2c^2} & \frac{1}{2c}\sum_{k=1}^{n}\frac{1}{f_k(\alpha)}\frac{\partial f_k(\alpha)}{\partial\alpha} \\ \frac{1}{2c}\sum_{k=1}^{n}\frac{1}{f_k(\alpha)}\frac{\partial f_k(\alpha)}{\partial\alpha} & \frac{1}{2}\sum_{k=1}^{n}\frac{1}{f_k^2(\alpha)}\left(\frac{\partial f_k(\alpha)}{\partial\alpha}\right)^2 \end{bmatrix}$$

evaluated at c_0, α_0 and

$$\boldsymbol{d}(c_0,\alpha_0) = \left[d_1(c_0,\alpha_0),\ldots,d_n(c_0,\alpha_0)\right]^{\top} = \left[(c_0f_1(\alpha_0))^{-1},\ldots,(c_0f_n(\alpha_0))^{-1}\right]^{\top}.$$

3.4 Computational study

In Section 3.2, it has been shown that in the sense of asymptotic variance and second moment (mean-squared) error, the maximum likelihood estimator computed in a smaller space containing the true parameter is more (or equally) precise. For small samples, this behaviour is illustrated by means of simulations.

3.4.1 Simulation of fields with diagonal covariance

The simulation is carried out with a covariance stationary random field. As mentioned in Section 2.2.1, the spectral covariance of such a field is diagonal and

can be modelled by means of eigenvalues of Laplace operator. The simulation set-up was similar to the Section 3.3. First, a diagonal matrix D_0 was prepared, whose diagonal entries decay according to the model $d_k = \frac{1}{c}e^{\alpha\lambda_k}, k = 1, \ldots, n$, where c and α are parameters and λ_k are the eigenvalues of discrete Laplace operator in two dimensions on 10×10 nodes (so $n = 100$). Note that all λ_k, $k = 1, \ldots, n$, are negative. Such models are useful in modelling smooth random fields, e.g., in meteorology. Then, random samples were generated from $\mathcal{N}_n(\mathbf{0}, D_0)$ with sample sizes $N = 5, \ldots, 20$. For each sample, four covariance matrix estimators were computed:

- sample covariance matrix $S_{(\boldsymbol{\mu}=0)}$, cf. (3.7),
- MLE $\hat{D}^{(1)}$ in the space of diagonal matrices, cf. (3.11),
- MLE $\hat{D}^{(2)} = \mathrm{diag}\{(\hat{c}_1 - \hat{c}_2\lambda_k)^{-1}e^{\hat{\alpha}\lambda_k}, k = 1, \ldots, n\}$ with 3 parameters c_1, c_2 and α, cf. (3.15), and
- MLE $\hat{D}^{(3)} = \mathrm{diag}\{\hat{c}^{-1}e^{\hat{\alpha}\lambda_k}, k = 1, \ldots, n\}$ with 2 parameters c and α, cf. (3.18).

The difference of each estimator from the true matrix D_0 was measured in the Frobenius norm. To reduce the sampling noise, 50 replications have been done for each sample size and the mean of squared Frobenius norm can be found in Figure 3.1.

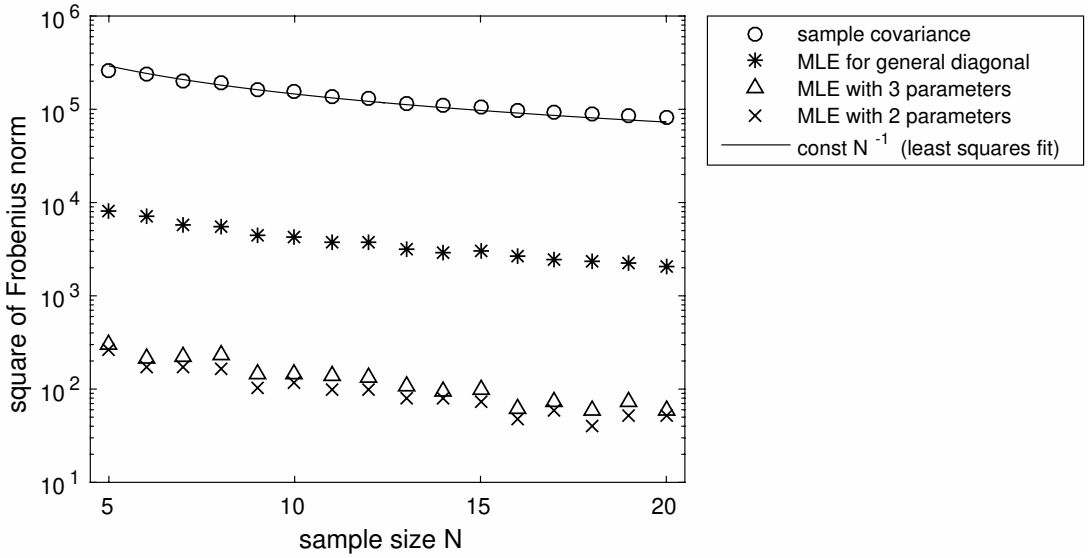


Figure 3.1: Nested covariance models (simulation): Errors of estimators $\hat{D}^{(j)} \in \boldsymbol{\Theta}_j$, $j = 1, 2, 3$, of a diagonal covariance D_0 measured by $\|\hat{D}^{(j)} - D_0\|_F^2$. The error of $S_{(\boldsymbol{\mu}=0)}$ is also added. The random field had dimension $n = 10 \times 10$. Eigenvalues of D_0 decay exponentially, i.e. $d_k = c_0^{-1}e^{\alpha_0\lambda_k}$, where $\lambda_k < 0$, $k = 1, \ldots, n$, with parameters $c_0 = 1/30$ and $\alpha_0 = 0.002$. The full line is the order of convergence $\mathrm{const}(N^{-1})$ fitted to the error of the sample covariance.

For the diagonal MLE, given by (3.11), (3.15), and (3.18), we can expect from (3.6) that these estimators should satisfy asymptotically

$$\mathsf{E}\left\|\hat{D}^{(j)} - D_0\right\|_F^2 \approx \frac{1}{N}\,\mathrm{tr}\left(\mathcal{I}^{-1}_{D_0^{(j)}}\right), \quad j = 1, 2, 3,$$

even if convergence in distribution does not imply convergence of moments without additional assumptions. This conjecture can be supported by a comparison of Figures 3.3 and 3.2, where the same decay is observed. From the nesting, it is known that (cf. (3.6))

$$\mathrm{tr}\left(\mathcal{I}^{-1}_{D_0^{(3)}}\right) \le \mathrm{tr}\left(\mathcal{I}^{-1}_{D_0^{(2)}}\right) \le \mathrm{tr}\left(\mathcal{I}^{-1}_{D_0^{(1)}}\right)$$

and it can be expected that the Frobenius norm should decrease for more restrictive models, that is,

$$\mathsf{E}\left\|\hat{D}^{(3)} - D_0\right\|_F^2 \le \mathsf{E}\left\|\hat{D}^{(2)} - D_0\right\|_F^2 \le \mathsf{E}\left\|\hat{D}^{(1)} - D_0\right\|_F^2, \tag{3.19}$$

which is confirmed by the simulations (see Figure 3.2, resp. 3.3).

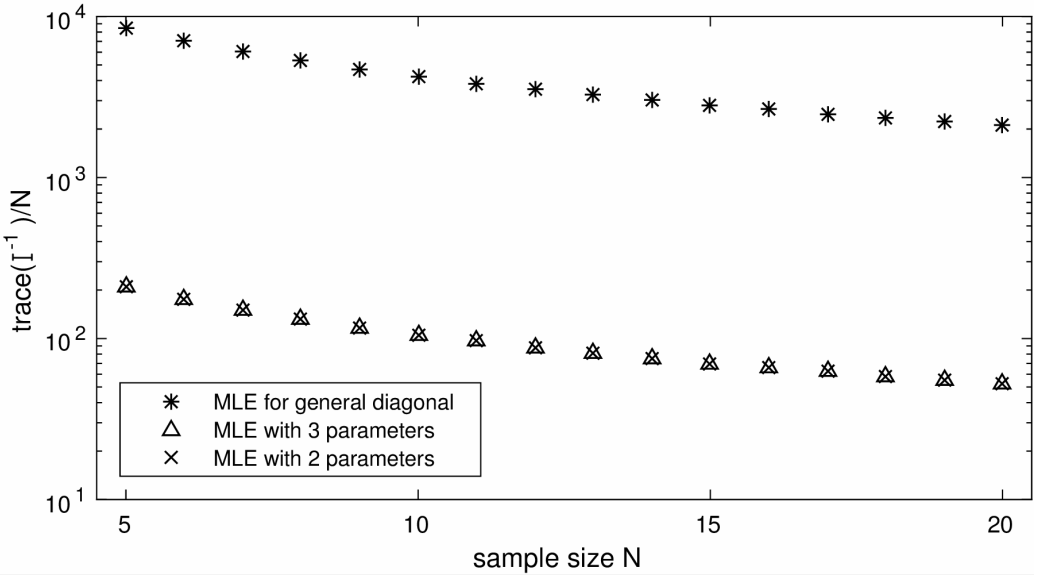


Figure 3.2: Nested covariance models (simulation): Comparison of sums of estimated asymptotic variances $\frac{1}{N}\,\mathrm{tr}(\mathcal{I}^{-1}_{\hat{D}^{(j)}})$ for three estimators $\hat{D}^{(j)} \in \mathbf{\Theta}_j$, $j = 1, 2, 3$, of a diagonal matrix $D_0 = \mathrm{diag}\{c_0^{-1} e^{\alpha_0 \lambda_k}, k = 1, \ldots, n\}$, where $c_0 = 1/30$ and $\alpha_0 = 0.002$.

The comparisons (3.19) of the Frobenius norm of the error in the mean squared complement the pointwise comparison (3.10) between the sample covariance and its diagonal. Relying on MLE for that comparison is not practical, because the sample size of interest here is $N < n$, and, consequently, $S_{(\boldsymbol{\mu}=0)}$ is singular and cannot be cast as MLE with an accompanying Fisher information matrix, cf. Remark 3 in Section 3.3.1. But it is evident that for small sample sizes, estimators computed in the proper subspace perform better. Hence, the hierarchical order seems to hold even when $N < n$.

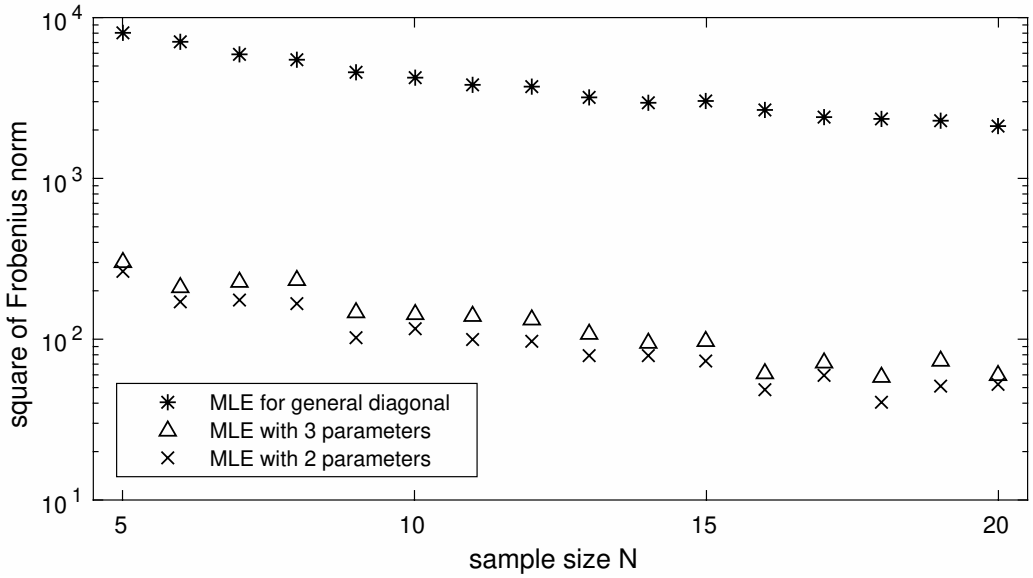


Figure 3.3: Nested covariance models (simulation): Averaged errors $\|\hat{D}^{(j)} - D_0\|_F^2$ (based on 50 replications) of estimators $\hat{D}^{(j)} \in \mathbf{\Theta}_j,\ j = 1, 2, 3$, of a diagonal matrix $D_0 = \mathrm{diag}\{c_0^{-1} e^{\alpha_0 \lambda_k}, k = 1, \ldots, n\}$, where $c_0 = 1/30$ and $\alpha_0 = 0.002$.

3.4.2 Simulation of sparse inverse covariance of GMRF

The second simulation is related to a simple GMRF (cf. Section 2.2.2) and it illustrates another way to bring in assumed covariance structure. In the GMRF on a rectangular mesh, a variable on a gridpoint is conditionally independent on the rest of the gridpoints, given values on neighbouring gridpoints. It follows from Lemma 3 that nonzero entries in the inverse of the covariance matrix can be only between neighbour gridpoints. Adding more details, we start with 4 neighbours (as in Figure 2.1a), and adding neighbours gives rise to a sequence of nested covariance models. If the columns of the mesh are stacked vertically, their inverse covariance matrix will have a band-diagonal structure as has been already seen at Figure 2.1.

The inverse covariance model for GMRF fitted by MLE was introduced by Ueno and Tsuchiya [68] and applied on data from oceanography. The corresponding Fisher information matrix may be found as the negative of the Hessian matrix (Ueno and Tsuchiya [68, eq. (C17)]). Later, in Section 4.6, we will estimate parameters of such inverse covariance model also by the score matching method.

The simulation has been carried out as follows. First, a sample of realizations of the GMRF has been generated with dimensions 10×10 (resulting in $n = 100$) and inverse covariance structure as in Figure 2.1. The values on the diagonals of the precision matrix have been set to constant, since we assume the correlation with left and right neighbour to be identical, as well as the correlation with upper and lower neighbour. In particular, the main diagonal of precision matrix was set to the value 5, the elements that correspond to the dependence between lower and upper neighbours were set to -0.2 and the elements describing the dependence between left and right neighbours were set to 0.5. This leads to a sequence of nested models with 3 parameters for 4 neighbours, 5 parameters for 8 neighbours

and 7 parameters for 12 neighbours,

The structure of Σ_0^{-1} with 4 neighbours (Figure 2.1a) was set as the "truth" and random samples were generated from $\mathcal{N}_n(\mathbf{0}, \Sigma_0)$ with sample sizes $N = 10, 15, 20, \ldots, 55$. As already said, the values on first, second and tenth diagonal have been set as 5, -0.2 and 0.5. For each sample, we computed successively the MLE with 3, 5 and 7 unknown parameters numerically by Newton's method, as described in Ueno and Tsuchiya [68].

The difference of each estimate from the true matrix Σ_0 was measured again in the Frobenius norm. In order to reduce the sampling error, 50 simulations of the same size were generated and the mean of squared Frobenius norm was computed. The results can be found in Figure 3.4.

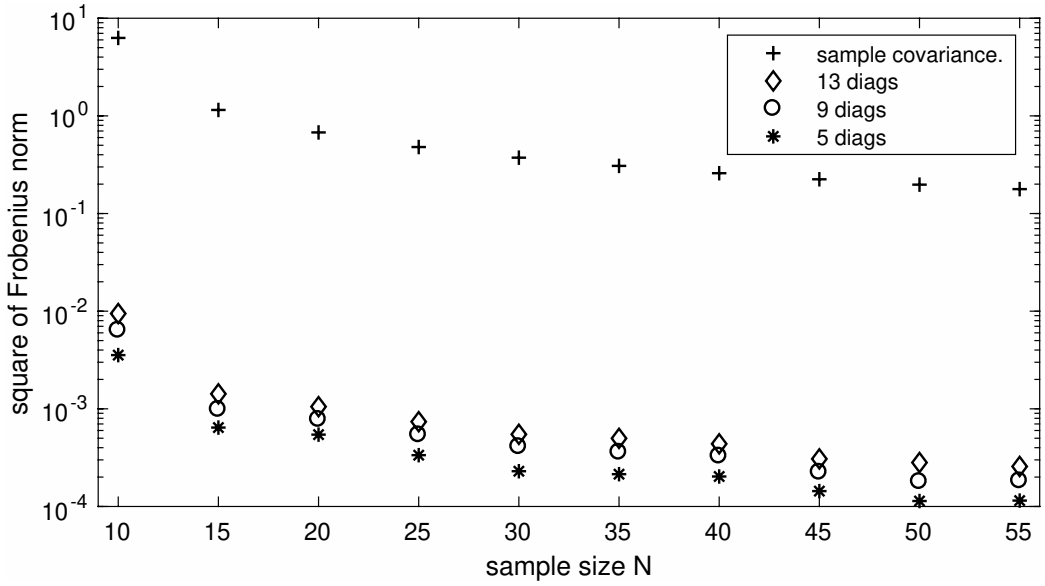


Figure 3.4: Nested covariance models for a GMRF (simulation): Errors of MLEs of four nested covariance models measured in the Frobenius norm. Compared estimators: sample covariance and models with 4, 8, 12 neighbours, i.e. 5, 9, 13 nonzero diagonals in the inverse covariance matrix.

As expected, the MLE with 3 parameters outperforms the estimates with 5 and 7 parameters and the Frobenius norm for sample covariance stays one order worse than all parametric estimates.

4. Score matching estimators

When estimating parameters of a distribution, the maximum likelihood method is usually the first choice. However, sometimes this method is not suitable because of some computational difficulties, e.g., when only numerical maximization is possible or when the normalization constant in the density function is unknown. In that situations, the score matching estimation method published by Hyvärinen [33] in 2005 provides a consistent estimator that may be easy to accomplish.

Because the score matching estimation method is less known than the maximum likelihood method, a brief summary of this method, following Forbes and Lauritzen [23] and Hyvärinen [33, 34], is provided in Section 4.2. The general result from Hyvärinen [33] is supplemented by later results from Forbes and Lauritzen [23] concerning the case when the model distribution belongs to the exponential family and the estimator is available in a closed form. This method is applied to estimating parameters of a precision matrix of GMRF that will be taken an advantage of later, in Chapter 7, where we propose three filtering algorithms based on this estimator. Score matching provides an explicit estimating formula that is easy to compute and the resulting matrix is a consistent estimator.

Asymptotic variances of score matching estimators corresponding to nested parametrizations follow a hierarchical structure similar to the MLE in the previous chapter. This is proved in Chapter 5 in a slightly more general case of M-estimators.

4.1 Notation

Let $\boldsymbol{X}$ be a random vector with values in a set $\mathcal{X} \subset \mathbb{R}^n$. In this chapter, the expected value of $\boldsymbol{X}$ with respect to a probability density $p(\boldsymbol{x})$ is denoted by

$$\mathsf{E}_{\boldsymbol{X}\sim p}\left(\boldsymbol{X}\right) = \int_{\mathcal{X}} \boldsymbol{x} p(\boldsymbol{x}) d\boldsymbol{x}.$$

The Euclidean norm of $\boldsymbol{x} \in \mathbb{R}^n$ is denoted by $\|\boldsymbol{x}\|_n$, which is a shortcut of $\|\boldsymbol{x}\|_{\mathbb{R}^n}$, and the inner product by $\langle \boldsymbol{x}, \boldsymbol{v} \rangle_n$, $\boldsymbol{x}, \boldsymbol{v} \in \mathbb{R}^n$. A column vector $(x_1, \ldots, x_n)^\top$ is sometimes written as $[x_j]_{j=1}^n$, in order to make the notation shorter. A matrix consisting of columns $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m \in \mathbb{R}^n$ is denoted by $[\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m]$ and a matrix of elements v_{jk}, $j = 1, \ldots, m$, $k = 1, \ldots, n$, by $[v_{jk}]_{j,k=1}^{m,n}$, or $[v_{jk}]_{j,k=1}^{n}$, when $m = n$.

For a scalar function g of vector argument $\boldsymbol{x} = (x_1, \ldots, x_n)^\top \in \mathbb{R}^n$, $\Delta_{\boldsymbol{x}}$ stands for the Laplacian and $\nabla_{\boldsymbol{x}}$ for the gradient,

$$\Delta_{\boldsymbol{x}} g(\boldsymbol{x}) = \sum_{i=1}^n \frac{\partial^2 g}{\partial x_i^2}(\boldsymbol{x}), \quad \nabla_{\boldsymbol{x}} g(\boldsymbol{x}) = \left(\frac{\partial g}{\partial x_1}(\boldsymbol{x}), \ldots, \frac{\partial g}{\partial x_n}(\boldsymbol{x}) \right).$$

If the gradient is needed as a column vector, we denote $\nabla_{\boldsymbol{x}}^\top g(\boldsymbol{x}) = \left(\nabla_{\boldsymbol{x}} g(\boldsymbol{x})\right)^\top$. The Jacobian matrix of a vector function $h = (h_1, \ldots, h_m)^\top \colon \mathbb{R}^n \to \mathbb{R}^m$ is denoted by

$$J_{\boldsymbol{x}}(h(\boldsymbol{x})) = \left[\frac{\partial h_i}{\partial x_j}(\boldsymbol{x}) \right]_{i,j=1}^{m,n}.$$

4.2 Score matching estimation method

For the unknown probability density function $p(\boldsymbol{x})$ of the random vector $\boldsymbol{X}$, consider a parametrized density model of the form

$$f(\boldsymbol{x}|\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} q(\boldsymbol{x}|\boldsymbol{\theta}), \quad Z(\boldsymbol{\theta}) = \int_{\mathcal{X}} q(\boldsymbol{x}|\boldsymbol{\theta}) d\boldsymbol{x}, \tag{4.1}$$

where the normalization constant $Z(\boldsymbol{\theta})$ may be difficult to compute, and $\boldsymbol{\theta}$ varies over $\boldsymbol{\Theta}$, which is an open set in a finite dimensional vector space L. The objective is to find an estimate $\hat{\boldsymbol{\theta}} \in \boldsymbol{\Theta}$ of $\boldsymbol{\theta}$ and to approximate $p(\boldsymbol{x})$ by $f(\boldsymbol{x}|\hat{\boldsymbol{\theta}})$ without the use of $Z(\boldsymbol{\theta})$.

The idea of score matching estimation (Hyvärinen [33]) is to make inference about $\boldsymbol{\theta}$ using the gradient with respect to $\boldsymbol{x}$ of the log-density

$$\nabla_{\boldsymbol{x}} \log f(\boldsymbol{x}|\boldsymbol{\theta}) \tag{4.2}$$

instead of the density itself. The function (4.2) is called the *score function* in Hyvärinen [33] because it is the Fisher score function (Barndorff-Nielsen and Cox [7, expr. (2.5)], Hyvärinen [34]), with respect to a hypothetical location parameter: assuming an additional location parameter vector $\boldsymbol{\xi}$, (4.2) can be obtained by taking the gradient of $\log f(\boldsymbol{x} - \boldsymbol{\xi}|\boldsymbol{\theta})$ with respect to $\boldsymbol{\xi}$ and evaluate it at $\boldsymbol{\xi} = \mathbf{0}$.

The principal observation is that

$$\nabla_{\boldsymbol{x}} \log f(\boldsymbol{x}|\boldsymbol{\theta}) = \nabla_{\boldsymbol{x}} \left(\log q(\boldsymbol{x}|\boldsymbol{\theta}) - \log Z(\boldsymbol{\theta}) \right) = \nabla_{\boldsymbol{x}} \log q(\boldsymbol{x}|\boldsymbol{\theta}), \tag{4.3}$$

thus the score function $\nabla_{\boldsymbol{x}} \log f(\boldsymbol{x}|\boldsymbol{\theta})$ does not depend on $Z(\boldsymbol{\theta})$. The parameter $\boldsymbol{\theta}$ in $f(\boldsymbol{x}|\boldsymbol{\theta})$ is then estimated by matching the score function of the model to the score function of the data by minimizing the expectation of the squared distance

$$\begin{aligned} \mathcal{S}(\boldsymbol{\theta}) &= \int_{\mathcal{X}} \left\| \nabla_{\boldsymbol{x}}^{\top} \log q(\boldsymbol{x}|\boldsymbol{\theta}) - \nabla_{\boldsymbol{x}}^{\top} \log p(\boldsymbol{x}) \right\|_n^2 p(\boldsymbol{x}) d\boldsymbol{x} \\ &= \mathsf{E}_{\boldsymbol{X} \sim p} \left\| \nabla_{\boldsymbol{x}}^{\top} \log q(\boldsymbol{X}|\boldsymbol{\theta}) - \nabla_{\boldsymbol{x}}^{\top} \log p(\boldsymbol{X}) \right\|_n^2 . \end{aligned} \tag{4.4}$$

Estimating parameters by matching the model and data scores gave the procedure its name. The Score Matching Estimator (SME) of $\boldsymbol{\theta}$ is

$$\hat{\boldsymbol{\theta}} = \operatorname*{argmin}_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathcal{S}(\boldsymbol{\theta}).$$

SME relies on the following assumptions:

(B1) $p(\boldsymbol{x})$ and $\nabla_{\boldsymbol{x}} \log q(\boldsymbol{x}|\boldsymbol{\theta})$ are differentiable in $\mathcal{X}$,

(B2) $\mathsf{E}_{\boldsymbol{X} \sim p} \left\| \nabla_{\boldsymbol{x}}^{\top} \log q(\boldsymbol{x}|\boldsymbol{\theta}) \right\|_n^2$ is finite for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$,

(B3) $\mathsf{E}_{\boldsymbol{X} \sim p} \left\| \nabla_{\boldsymbol{x}}^{\top} \log p(\boldsymbol{x}) \right\|_n^2$ is finite, and

(B4) function $g(\boldsymbol{x}|\boldsymbol{\theta}) = \log q(\boldsymbol{x}|\boldsymbol{\theta}) \colon \mathbb{R}^n \to \mathbb{R}$ satisfies that $p(\boldsymbol{x}) \nabla_{\boldsymbol{x}}^{\top} g(\boldsymbol{x}|\boldsymbol{\theta}) \to \mathbf{0}$ for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ when $\boldsymbol{x} \to \partial\mathcal{X}$ and the boundary $\partial\mathcal{X}$ of $\mathcal{X}$ is sufficiently regular for integration by parts, in particular

$$\int_{\mathcal{X}} \left\langle \nabla_{\boldsymbol{x}}^{\top} p(\boldsymbol{x}), \nabla_{\boldsymbol{x}}^{\top} g(\boldsymbol{x}|\boldsymbol{\theta}) \right\rangle_n d\boldsymbol{x} = - \int_{\mathcal{X}} p(\boldsymbol{x}) \Delta_{\boldsymbol{x}} g(\boldsymbol{x}|\boldsymbol{\theta}) d\boldsymbol{x}, \quad \forall \boldsymbol{\theta} \in \boldsymbol{\Theta}.$$

Remark 4. When $g(\boldsymbol{x}|\boldsymbol{\theta})$ is polynomial for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, assumption (B4) is satisfied for a large class of probability distributions. For example for the normal distribution, where $\mathcal{X} = \mathbb{R}^n$, the Fubini theorem implies

$$\begin{aligned}\int_{\mathbb{R}^n} \left\langle \nabla_{\boldsymbol{x}}^\top p(\boldsymbol{x}), \nabla_{\boldsymbol{x}}^\top g(\boldsymbol{x}|\boldsymbol{\theta}) \right\rangle_n d\boldsymbol{x} &= \sum_{j=1}^n \int_{\mathbb{R}} \frac{\partial p(\boldsymbol{x})}{\partial x_j} \frac{\partial g(\boldsymbol{x}|\boldsymbol{\theta})}{\partial x_j} dx_j = \\ &= \sum_{j=1}^n \left(\left[p(\boldsymbol{x}) \frac{\partial g(\boldsymbol{x}|\boldsymbol{\theta})}{\partial x_j} \right]_{x_j=-\infty}^{x_j=\infty} - \int_{\mathbb{R}} \frac{\partial^2 g(\boldsymbol{x}|\boldsymbol{\theta})}{\partial x_j^2} p(\boldsymbol{x}) dx_j \right) \\ &= -\int_{\mathbb{R}^n} p(\boldsymbol{x}) \Delta_{\boldsymbol{x}} g(\boldsymbol{x}|\boldsymbol{\theta}) d\boldsymbol{x},\end{aligned}$$

because $p(\boldsymbol{x}) \frac{\partial g(\boldsymbol{x}|\boldsymbol{\theta})}{\partial x_j} \to 0$ as $x_j \to \pm\infty$ due to the exponential decay of the Gaussian density.

When assumptions (B1)-(B4) hold, it can be shown (Hyvärinen [33, Theorem 1]) by integration by parts that the objective function (4.4) equals to

$$\mathcal{S}(\boldsymbol{\theta}) = \mathsf{E}_{\boldsymbol{X}\sim p} \left[\frac{1}{2} \left\| \nabla_{\boldsymbol{x}}^\top \log q(\boldsymbol{X}|\boldsymbol{\theta}) \right\|_n^2 + \Delta_{\boldsymbol{x}} \log q(\boldsymbol{X}|\boldsymbol{\theta}) \right] + c, \tag{4.5}$$

where $c = \mathsf{E}_{\boldsymbol{X}\sim p} \left\| \nabla_{\boldsymbol{x}}^\top \log p(\boldsymbol{X}) \right\|_n^2$ does not depend on $\boldsymbol{\theta}$. Thus, the squared distance of the model score function from the data score function can be computed as an expectation of certain functions of the unnormalized model density $q(\boldsymbol{x}|\boldsymbol{\theta})$.

Given a sample $\mathbb{X}_N = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N\}$ from the density p, the expected value in (4.5) can be approximated by the sample mean,

$$\mathcal{S}_N(\boldsymbol{\theta}|\mathbb{X}_N) = \frac{1}{N} \sum_{i=1}^N \left(\frac{1}{2} \left\| \nabla_{\boldsymbol{x}}^\top \log q(\boldsymbol{X}_i|\boldsymbol{\theta}) \right\|_n^2 + \Delta_{\boldsymbol{x}} \log q(\boldsymbol{X}_i|\boldsymbol{\theta}) \right) + c_N(\mathbb{X}_N), \tag{4.6}$$

where $c_N(\mathbb{X}_N) = \frac{1}{N} \sum_{i=1}^N \left\| \nabla_{\boldsymbol{x}}^\top \log p(\boldsymbol{X}_i) \right\|_n^2$ does not depend on $\boldsymbol{\theta}$. The coefficient $1/N$ and the constant c_N, for a fixed sample $\mathbb{X}_N$, do not affect the point where the minimum in (4.6) is attained. Thus, we obtain the empirical estimate

$$\hat{\boldsymbol{\theta}}_N = \operatorname*{argmin}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left(\sum_{i=1}^N \frac{1}{2} \left\| \nabla_{\boldsymbol{x}}^\top \log q(\boldsymbol{X}_i|\boldsymbol{\theta}) \right\|_n^2 + \Delta_{\boldsymbol{x}} \log q(\boldsymbol{X}_i|\boldsymbol{\theta}) \right). \tag{4.7}$$

4.3 Exponential family

Following Forbes and Lauritzen [23], suppose in addition, that the density model (4.1) belongs to the exponential family , i.e.,

$$\log f(\boldsymbol{x}|\boldsymbol{\theta}) = \langle T(\boldsymbol{x}), \boldsymbol{\theta} \rangle_L - a(\boldsymbol{\theta}) + b(\boldsymbol{x}), \tag{4.8}$$

where $\langle \cdot, \cdot \rangle_L$ is the inner product of the vector space $L \supset \boldsymbol{\Theta}$. Further assume that $\boldsymbol{\Theta}$ is such that $Z(\boldsymbol{\theta}) < \infty$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Function $T(\boldsymbol{x})$ is the canonical sufficient statistics and $\boldsymbol{\theta}$ is the canonical parameter. For density from the exponential family, the expressions in the objective function (4.5),

$$\begin{aligned}\nabla_{\boldsymbol{x}} \log q(\boldsymbol{x}|\boldsymbol{\theta}) &= \nabla_{\boldsymbol{x}} \langle T(\boldsymbol{x}), \boldsymbol{\theta} \rangle_L + \nabla_{\boldsymbol{x}} b(\boldsymbol{x}), \\ \Delta_{\boldsymbol{x}} \log q(\boldsymbol{x}|\boldsymbol{\theta}) &= \Delta_{\boldsymbol{x}} \langle T(\boldsymbol{x}), \boldsymbol{\theta} \rangle_L + \Delta_{\boldsymbol{x}} b(\boldsymbol{x}),\end{aligned} \tag{4.9}$$

are linear functions of $\boldsymbol{\theta}$. Substituting into (4.6), we obtain an empirical scoring function quadratic in $\boldsymbol{\theta}$,

$$\begin{aligned}\mathcal{S}_N(\boldsymbol{\theta}|\mathbb{X}_N) &= \frac{1}{N}\sum_{i=1}^{N}\Big[\frac{1}{2}\left\|\nabla_{\boldsymbol{x}}^{\top}\langle T(\boldsymbol{X}_i),\boldsymbol{\theta}\rangle_L + \nabla_{\boldsymbol{x}}^{\top} b(\boldsymbol{X}_i)\right\|_n^2 + \Delta_{\boldsymbol{x}}\langle T(\boldsymbol{X}_i),\boldsymbol{\theta}\rangle_L \\ &\quad + \Delta_{\boldsymbol{x}} b(\boldsymbol{X}_i)\Big] + c_N\left(\mathbb{X}_N\right) \\ &= \frac{1}{N}\sum_{i=1}^{N}\Big[\frac{1}{2}\left\|\nabla_{\boldsymbol{x}}^{\top}\langle T(\boldsymbol{X}_i),\boldsymbol{\theta}\rangle_L\right\|^2 + \left\langle\nabla_{\boldsymbol{x}}^{\top} b(\boldsymbol{X}_i), \nabla_{\boldsymbol{x}}^{\top}\langle T(\boldsymbol{X}_i),\boldsymbol{\theta}\rangle_L\right\rangle_n \\ &\quad + \Delta_{\boldsymbol{x}}\langle T(\boldsymbol{X}_i),\boldsymbol{\theta}\rangle_L\Big] + c_N^*\left(\mathbb{X}_N\right), \end{aligned} \tag{4.10}$$

where

$$c_N^*\left(\mathbb{X}_N\right) = \frac{1}{N}\sum_{i=1}^{N}\left\|\nabla_{\boldsymbol{x}}^{\top}\log p(\boldsymbol{X}_i)\right\|_n^2 + \frac{1}{N}\sum_{i=1}^{N}\Big[\frac{1}{2}\left\|\nabla_{\boldsymbol{x}}^{\top} b(\boldsymbol{X}_i)\right\|_n^2 + \Delta_{\boldsymbol{x}} b(\boldsymbol{X}_i)\Big]$$

does not depend on $\boldsymbol{\theta}$. For a fixed $\boldsymbol{x}$, define linear operator $D(\boldsymbol{x})$ by

$$D(\boldsymbol{x})\colon L \to \mathbb{R}^n, \quad D(\boldsymbol{x})\boldsymbol{\theta} = \nabla_{\boldsymbol{x}}^{\top}\langle T(\boldsymbol{x}),\boldsymbol{\theta}\rangle_L, \tag{4.11}$$

its adjoint operator $D^*(\boldsymbol{x})$ by

$$D^*(\boldsymbol{x})\colon \mathbb{R}^n \to L, \quad \langle\boldsymbol{\theta}, D^*(\boldsymbol{x})\boldsymbol{v}\rangle_L = \langle D(\boldsymbol{x})\boldsymbol{\theta},\boldsymbol{v}\rangle_n, \quad \forall\boldsymbol{\theta}\in L,\ \forall\boldsymbol{v}\in\mathbb{R}^n, \tag{4.12}$$

and the Laplacian vector $\Delta_{\boldsymbol{x}} T(\boldsymbol{x})$ by

$$\Delta_{\boldsymbol{x}} T(\boldsymbol{x}) \in L: \quad \langle\Delta_{\boldsymbol{x}} T(\boldsymbol{x}),\boldsymbol{\theta}\rangle_L = \Delta_{\boldsymbol{x}}\langle T(\boldsymbol{x}),\boldsymbol{\theta}\rangle_L, \quad \forall\boldsymbol{\theta}\in L. \tag{4.13}$$

Then, (4.10) becomes

$$\begin{aligned}\mathcal{S}_N(\boldsymbol{\theta}|\mathbb{X}_N) = \frac{1}{N}\sum_{i=1}^{N}\Big[\frac{1}{2}\left\|D(\boldsymbol{X}_i)\boldsymbol{\theta}\right\|_n^2 + \langle D^*(\boldsymbol{X}_i)\nabla_{\boldsymbol{x}}^{\top} b(\boldsymbol{X}_i),\boldsymbol{\theta}\rangle_L + \\ + \langle\Delta_{\boldsymbol{x}} T(\boldsymbol{X}_i),\boldsymbol{\theta}\rangle_L\Big] + c_N^*\left(\mathbb{X}_N\right). \end{aligned} \tag{4.14}$$

Since the feasible set $\boldsymbol{\Theta}$ is open and the quadratic form $\frac{1}{N}\sum_{i=1}^{N}\left\|D(\boldsymbol{X}_i)\boldsymbol{\theta}\right\|_n^2$ is positive semidefinite, $\mathcal{S}_N(\boldsymbol{\theta}|\mathbb{X}_N)$ attains minimum on $\boldsymbol{\Theta}$ if and only if

$$\nabla_{\boldsymbol{\theta}}^{\top}\mathcal{S}_N(\boldsymbol{\theta}|\mathbb{X}_N) = \mathbf{0}. \tag{4.15}$$

In addition, if $\sum_{i=1}^{N}\left\|D(\boldsymbol{X}_i)\boldsymbol{\theta}\right\|_n^2$ is positive definite and the minimum of $\mathcal{S}_N(\boldsymbol{\theta}|\mathbb{X}_N)$ on $\boldsymbol{\Theta}$ exists, then the minimum is unique. Equation (4.15) provides the linear estimating equation for $\boldsymbol{\theta}$

$$\sum_{i=1}^{N}\left(D^*(\boldsymbol{X}_i)D(\boldsymbol{X}_i)\right)\boldsymbol{\theta} + \sum_{i=1}^{N}\Big(D^*(\boldsymbol{X}_i)\nabla_{\boldsymbol{x}}^{\top} b(\boldsymbol{X}_i) + \Delta_{\boldsymbol{x}} T(\boldsymbol{X}_i)\Big) = \mathbf{0}, \tag{4.16}$$

where $D^*(\boldsymbol{x})D(\boldsymbol{x})$ is a linear map on L and $D^*(\boldsymbol{x})\nabla_{\boldsymbol{x}}^{\top} b(\boldsymbol{x}) + \Delta_{\boldsymbol{x}} T(\boldsymbol{x}) \in L$. If $\sum_{i=1}^{N}\left(D^*(\boldsymbol{X}_i)D(\boldsymbol{X}_i)\right)$ is invertible, (4.16) has a unique solution in L,

$$\begin{aligned}\hat{\boldsymbol{\theta}}_N &= \operatorname*{argmin}_{\boldsymbol{\theta}\in L}\mathcal{S}_N(\boldsymbol{\theta}|\mathbb{X}_N) \\ &= -\left(\sum_{i=1}^{N}D^*(\boldsymbol{X}_i)D(\boldsymbol{X}_i)\right)^{-1}\sum_{i=1}^{N}\Big(D^*(\boldsymbol{X}_i)\nabla_{\boldsymbol{x}}^{\top} b(\boldsymbol{X}_i) + \Delta_{\boldsymbol{x}} T(\boldsymbol{X}_i)\Big). \end{aligned} \tag{4.17}$$

In the Section 4.6, we will take an advantage of (4.17), and derive SME for the mean and covariance matrix of a Gaussian Markov random vector.

Hereafter, we will assume that the true density $p(\boldsymbol{x}) = f(\boldsymbol{x}|\boldsymbol{\theta}_0)$ for a unique $\boldsymbol{\theta}_0$, where $f(\boldsymbol{x}|\boldsymbol{\theta})$ belongs to the exponential family. Under that condition, the SME (4.7) can be shown to be consistent. In Theorem 6 below, an exact statement based on Forbes and Lauritzen [23] with a detailed proof is provided for the SME (4.17), because it will be helpful later on, in Section 4.5.

In addition to assumptions (B1)-(B4), we need the exponential family distribution to satisfy that

(C1) $\mathsf{E}_{\boldsymbol{X}\sim p} \left\|\Delta_{\boldsymbol{x}} T(\boldsymbol{X})\right\|_L < \infty$,

(C2) $\mathsf{E}_{\boldsymbol{X}\sim p} \left\|\nabla_{\boldsymbol{x}}^\top b(\boldsymbol{X})\right\|_n^2 < \infty$,

(C3) $\mathsf{E}_{\boldsymbol{X}\sim p} \left\|D(\boldsymbol{X})\right\|_{op}^2 < \infty$,
where $\left\|D(\boldsymbol{X})\right\|_{op}^2 = \sup\left\{\left\|D(\boldsymbol{X})\boldsymbol{\theta}\right\|_n^2 :\ \boldsymbol{\theta} \in L,\ \left\|\boldsymbol{\theta}\right\|_L \le 1\right\}$, and

(C4) assumption (B4) holds with $g(\boldsymbol{x}|\boldsymbol{\theta}) = \langle T(\boldsymbol{x}), \boldsymbol{\theta}\rangle_L$ in place of g.

Remark 5. For distributions with $b(\boldsymbol{x}) = 0$, assumption (C2) is fulfilled automatically and assumption (C4) coincides with (B4).

Since $p(\boldsymbol{x}) = f(\boldsymbol{x}|\boldsymbol{\theta}_0)$, assumption (B3) follows from (B2). Moreover, assumption (B2) follows from (C2) and (C3), because

$$\begin{aligned}
\mathsf{E}_{\boldsymbol{X}\sim p} \left\|\nabla_{\boldsymbol{x}}^\top \log q(\boldsymbol{X}|\boldsymbol{\theta})\right\|_n^2 &= \mathsf{E}_{\boldsymbol{X}\sim p} \left\|D(\boldsymbol{X})\boldsymbol{\theta} + \nabla_{\boldsymbol{x}}^\top b(\boldsymbol{x})\right\|_n^2 \\
&\le \mathsf{E}_{\boldsymbol{X}\sim p} \left(\left\|D(\boldsymbol{X})\boldsymbol{\theta}\right\|_n + \left\|\nabla_{\boldsymbol{x}}^\top b(\boldsymbol{x})\right\|_n\right)^2 \\
&\le 2\,\mathsf{E}_{\boldsymbol{X}\sim p} \left\|D(\boldsymbol{X})\right\|_{op}^2 \left\|\boldsymbol{\theta}\right\|_L^2 + 2\,\mathsf{E}_{\boldsymbol{X}\sim p} \left\|\nabla_{\boldsymbol{x}}^\top b(\boldsymbol{x})\right\|_n^2.
\end{aligned}$$

Theorem 6 (Consistency of SME, Forbes and Lauritzen [23])**.** *Assume that $p(\boldsymbol{x}) = f(\boldsymbol{x}|\boldsymbol{\theta}_0)$ for a unique $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$ and that $q(\boldsymbol{x}|\boldsymbol{\theta}) > 0$ for all $\boldsymbol{x} \in \mathcal{X}$ and all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Further assume that $f(\boldsymbol{x}|\boldsymbol{\theta})$ satisfies conditions (B1), (B4), (C1)-(C4) and that $\mathsf{E}_{\boldsymbol{X}\sim p}(D^*(\boldsymbol{X})D(\boldsymbol{X}))$ is invertible. Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N$ be the random sample from $p(\boldsymbol{x})$. Then, the SME (4.17) exists with probability approaching one as $N \to \infty$, and it is a consistent estimator of $\boldsymbol{\theta}_0$, i.e.*

$$\hat{\boldsymbol{\theta}}_N \xrightarrow[N\to\infty]{P} \boldsymbol{\theta}_0.$$

Proof. Denote by $\{\boldsymbol{\ell}_k\}_{k=1}^s$ an orthonormal basis of the vector space L. Then the components $\{\phi_k\}_{k=1}^s$ of any $\boldsymbol{\phi} \in L$ in this basis are given by

$$\boldsymbol{\phi} = \sum_{k=1}^{s} \underbrace{\langle \boldsymbol{\phi}, \boldsymbol{\ell}_k\rangle_L}_{\phi_k} \boldsymbol{\ell}_k.$$

Take $\boldsymbol{\phi}$ equal to the last sum in (4.17), i.e.,

$$\boldsymbol{\phi}(\boldsymbol{X}) = D^*(\boldsymbol{X})\nabla_{\boldsymbol{x}}^\top b(\boldsymbol{X}) + \Delta_{\boldsymbol{x}} T(\boldsymbol{X}) \in L \tag{4.18}$$

and evaluate the coefficients $\mathsf{E}_{\boldsymbol{X}\sim p}\left(\phi_k(\boldsymbol{X})\right)$ for $k = 1, \ldots, s$,

$$\mathsf{E}_{\boldsymbol{X}\sim p}\left(\phi_k(\boldsymbol{X})\right) = \mathsf{E}_{\boldsymbol{X}\sim p}\langle \boldsymbol{\phi}(\boldsymbol{X}), \boldsymbol{\ell}_k\rangle_L = \int_{\mathcal{X}} p(\boldsymbol{x})\langle D^*(\boldsymbol{x})\nabla_{\boldsymbol{x}}^\top b(\boldsymbol{x}), \boldsymbol{\ell}_k\rangle_L d\boldsymbol{x} + \\ + \int_{\mathcal{X}} p(\boldsymbol{x})\langle \Delta_{\boldsymbol{x}} T(\boldsymbol{x}), \boldsymbol{\ell}_k\rangle_L d\boldsymbol{x}, \quad (4.19)$$

where the Laplacian is defined by (4.13),

$$\langle \Delta_{\boldsymbol{x}} T(\boldsymbol{x}), \boldsymbol{\ell}_k\rangle_L = \Delta_{\boldsymbol{x}}\langle T(\boldsymbol{x}), \boldsymbol{\ell}_k\rangle_L.$$

The first integral in (4.19) is finite as a consequence of the Cauchy-Schwartz inequality in $L_2(\Omega)$

$$\begin{aligned}\int_{\mathcal{X}} p(\boldsymbol{x}) \left\| D^*(\boldsymbol{x})\nabla_{\boldsymbol{x}}^\top b(\boldsymbol{x})\right\|_L d\boldsymbol{x} &\le \int_{\mathcal{X}} p(\boldsymbol{x}) \left\|D^*(\boldsymbol{x})\right\|_{op} \cdot \left\|\nabla_{\boldsymbol{x}}^\top b(\boldsymbol{x})\right\|_L d\boldsymbol{x} \\ &\le \left(\int_{\mathcal{X}} p(\boldsymbol{x}) \left\|D^*(\boldsymbol{x})\right\|_{op}^2 d\boldsymbol{x}\right)^{\frac{1}{2}} \left(\int_{\mathcal{X}} p(\boldsymbol{x}) \left\|\nabla_{\boldsymbol{x}}^\top b(\boldsymbol{x})\right\|_n^2 d\boldsymbol{x}\right)^{\frac{1}{2}} \\ &= \left(\mathsf{E}_{\boldsymbol{X}\sim p} \left\|D^*(\boldsymbol{X})\right\|_{op}^2\right)^{\frac{1}{2}} \left(\mathsf{E}_{\boldsymbol{X}\sim p} \left\|\nabla_{\boldsymbol{x}}^\top b(\boldsymbol{X})\right\|_n^2\right)^{\frac{1}{2}}\end{aligned}$$

combined with assumptions (C2) and (C3) and by the linearity of the inner product

$$\int_{\mathcal{X}} p(\boldsymbol{x})\langle D^*(\boldsymbol{x})\nabla_{\boldsymbol{x}}^\top b(\boldsymbol{x}), \boldsymbol{\ell}_k\rangle_L d\boldsymbol{x} = \left\langle \int_{\mathcal{X}} p(\boldsymbol{x}) D^*(\boldsymbol{x})\nabla_{\boldsymbol{x}}^\top b(\boldsymbol{x}) d\boldsymbol{x}, \boldsymbol{\ell}_k\right\rangle_L.$$

The second integral in (4.19) is finite due to assumption (C1) and again by the linearity of the inner product. Moreover, the integration by parts imply

$$\int_{\mathcal{X}} p(\boldsymbol{x})\Delta_{\boldsymbol{x}}\langle T(\boldsymbol{x}), \boldsymbol{\ell}_k\rangle_L d\boldsymbol{x} = -\int_{\mathcal{X}} \left\langle \nabla_{\boldsymbol{x}}^\top p(\boldsymbol{x}), \nabla_{\boldsymbol{x}}^\top \langle T(\boldsymbol{x}), \boldsymbol{\ell}_k\rangle_L\right\rangle_n d\boldsymbol{x}, \quad (4.20)$$

due to assumption (C4). From (4.11) and (4.12), it follows that

$$\begin{aligned}\left\langle \nabla_{\boldsymbol{x}}^\top p(\boldsymbol{x}), \nabla_{\boldsymbol{x}}^\top \langle T(\boldsymbol{x}), \boldsymbol{\ell}_k\rangle_L\right\rangle_n &= \left\langle \nabla_{\boldsymbol{x}}^\top p(\boldsymbol{x}), D(\boldsymbol{x})\boldsymbol{\ell}_k\right\rangle_n \\ &= \left\langle D^*(\boldsymbol{x})\nabla_{\boldsymbol{x}}^\top p(\boldsymbol{x}), \boldsymbol{\ell}_k\right\rangle_L,\end{aligned}$$

and hence (4.20) results in

$$\int_{\mathcal{X}} p(\boldsymbol{x})\Delta_{\boldsymbol{x}}\langle T(\boldsymbol{x}), \boldsymbol{\ell}_k\rangle_L d\boldsymbol{x} = -\int_{\mathcal{X}} \left\langle D^*(\boldsymbol{x})\nabla_{\boldsymbol{x}}^\top p(\boldsymbol{x}), \boldsymbol{\ell}_k\right\rangle_L d\boldsymbol{x}. \quad (4.21)$$

Due to (4.3), $p(\boldsymbol{x}) = f(\boldsymbol{x}|\boldsymbol{\theta}_0)$ implies $\nabla_{\boldsymbol{x}} \log p(\boldsymbol{x}) = \nabla_{\boldsymbol{x}} \log q(\boldsymbol{x}|\boldsymbol{\theta}_0)$. By substituting

$$\nabla_{\boldsymbol{x}}^\top p(\boldsymbol{x}) = p(\boldsymbol{x})\nabla_{\boldsymbol{x}}^\top \log p(\boldsymbol{x}) = p(\boldsymbol{x})\left(D(\boldsymbol{x})\boldsymbol{\theta}_0 + \nabla_{\boldsymbol{x}}^\top b(\boldsymbol{x})\right),$$

from (4.9) and (4.11),

$$\nabla_{\boldsymbol{x}}^\top \log p(\boldsymbol{x}) = D(\boldsymbol{x})\boldsymbol{\theta}_0 + \nabla_{\boldsymbol{x}}^\top b(\boldsymbol{x}) \in \mathbb{R}^n$$

into (4.21), we get

$$\begin{aligned}\int_{\mathcal{X}} p(\boldsymbol{x})\Delta_{\boldsymbol{x}}\langle T(\boldsymbol{x}), \boldsymbol{\ell}_k\rangle_L d\boldsymbol{x} &= -\int_{\mathcal{X}} p(\boldsymbol{x}) \left\langle D^*(\boldsymbol{x})\nabla_{\boldsymbol{x}}^\top \log p(\boldsymbol{x}), \boldsymbol{\ell}_k\right\rangle_L d\boldsymbol{x} \\ &= -\int_{\mathcal{X}} p(\boldsymbol{x}) \left\langle D^*(\boldsymbol{x})\left(D(\boldsymbol{x})\boldsymbol{\theta}_0 + \nabla_{\boldsymbol{x}}^\top b(\boldsymbol{x})\right), \boldsymbol{\ell}_k\right\rangle_L d\boldsymbol{x}\end{aligned}$$

for every $k = 1, \ldots, s$. This is the same as

$$\mathsf{E}_{\boldsymbol{X}\sim p}\langle \Delta_{\boldsymbol{x}} T(\boldsymbol{X}), \boldsymbol{\ell}_k\rangle_L = -\,\mathsf{E}_{\boldsymbol{X}\sim p}\left\langle D^*(\boldsymbol{X})D(\boldsymbol{X})\boldsymbol{\theta}_0 + D^*(\boldsymbol{X})\nabla_{\boldsymbol{x}}^\top b(\boldsymbol{X}), \boldsymbol{\ell}_k\right\rangle_L. \tag{4.22}$$

Substituting (4.22) into (4.19) results in

$$\mathsf{E}_{\boldsymbol{X}\sim p}\left(\phi_k(\boldsymbol{X})\right) = -\,\mathsf{E}_{\boldsymbol{X}\sim p}\langle D^*(\boldsymbol{X})D(\boldsymbol{X})\boldsymbol{\theta}_0, \boldsymbol{\ell}_k\rangle_L. \tag{4.23}$$

Multiplication of both sides of (4.23) by $\boldsymbol{\ell}_k$ and summation over k gives

$$\mathsf{E}_{\boldsymbol{X}\sim p}(\boldsymbol{\phi}(\boldsymbol{X})) = -\,\mathsf{E}_{\boldsymbol{X}\sim p}\left(D^*(\boldsymbol{X})D(\boldsymbol{X})\right)\boldsymbol{\theta}_0. \tag{4.24}$$

Finally, by comparing (4.18) and (4.24), we get

$$\mathsf{E}_{\boldsymbol{X}\sim p}\left(D^*(\boldsymbol{X})\nabla_{\boldsymbol{x}}^\top b(\boldsymbol{X}) + \Delta_{\boldsymbol{x}} T(\boldsymbol{X})\right) = -\,\mathsf{E}_{\boldsymbol{X}\sim p}\left(D^*(\boldsymbol{X})D(\boldsymbol{X})\right)\boldsymbol{\theta}_0. \tag{4.25}$$

By the Khinchin's weak law of large numbers

$$\frac{1}{N}\sum_{i=1}^N \left(D^*(\boldsymbol{X}_i)\nabla_{\boldsymbol{x}}^\top b(\boldsymbol{X}_i) + \Delta_{\boldsymbol{x}} T(\boldsymbol{X}_i)\right) \xrightarrow[N\to\infty]{P} -\,\mathsf{E}_{\boldsymbol{X}\sim p}\left(D^*(\boldsymbol{X})D(\boldsymbol{X})\right)\boldsymbol{\theta}_0$$

and similarly

$$\frac{1}{N}\sum_{i=1}^N D^*(\boldsymbol{X}_i)D(\boldsymbol{X}_i) \xrightarrow[N\to\infty]{P} \mathsf{E}_{\boldsymbol{X}\sim p}\left(D^*(\boldsymbol{X})D(\boldsymbol{X})\right).$$

The weak law of large numbers assumes that $\mathsf{E}_{\boldsymbol{X}\sim p}\left[D^*(\boldsymbol{X})D(\boldsymbol{X})\right]$ is finite, which is ensured by (C3) together with the Cauchy-Schwarz inequality in $L_2(\Omega)$,

$$\begin{aligned}\mathsf{E}_{\boldsymbol{X}\sim p}\left\|D^*(\boldsymbol{X})D(\boldsymbol{X})\right\|_{op} &\le \mathsf{E}_{\boldsymbol{X}\sim p}\left(\left\|D(\boldsymbol{X})\right\|_{op}\cdot\left\|D^*(\boldsymbol{X})\right\|_{op}\right)\\ &\le \left(\mathsf{E}_{\boldsymbol{X}\sim p}\left\|D(\boldsymbol{X})\right\|_{op}^2\right)^{\frac{1}{2}}\left(\mathsf{E}_{\boldsymbol{X}\sim p}\left\|D^*(\boldsymbol{X})\right\|_{op}^2\right)^{\frac{1}{2}}.\end{aligned}$$

Since $\mathsf{E}_{\boldsymbol{X}\sim p}\left(D^*(\boldsymbol{X})D(\boldsymbol{X})\right)$ is invertible by assumption, we get

$$\hat{\boldsymbol{\theta}}_N = -\left(\sum_{i=1}^N D^*(\boldsymbol{X}_i)D(\boldsymbol{X}_i)\right)^{-1}\sum_{i=1}^N \left(D^*(\boldsymbol{X}_i)\nabla_{\boldsymbol{x}}^\top b(\boldsymbol{X}_i) + \Delta_{\boldsymbol{x}} T(\boldsymbol{X}_i)\right) \xrightarrow[N\to\infty]{P} \boldsymbol{\theta}_0,$$

where the inverse exists with probability approaching one (cf., Lemma 7 below for $B_N = \frac{1}{N}\sum_{i=1}^N D^*(\boldsymbol{X}_i)D(\boldsymbol{X}_i)$ and $A = \mathsf{E}_{\boldsymbol{X}\sim p}\left(D^*(\boldsymbol{X})D(\boldsymbol{X})\right)$). □

Lemma 7. *Suppose that $\{B_N\}_{N\in\mathbb{N}}$ are random operators with values in a finite dimensional normed space $\mathcal{V}$. Assume that $B_N \xrightarrow{P} A$ as $N\to\infty$ and that A^{-1} exists. Then,* $\Pr\left(B_N^{-1}\ \textit{exists}\right)\to 1$ *and $B_N^{-1}\xrightarrow{P} A^{-1}$ as $N\to\infty$.*

Proof. Note that if $\left\|A - B_N\right\|_{op} \le 1/\left\|A^{-1}\right\|_{op}$ for a given $N\in\mathbb{N}$, then B_N^{-1} exists, and

$$\left\|A^{-1} - B_N^{-1}\right\|_{op} \le \frac{\left\|A^{-1}\right\|_{op}^2\left\|A - B_N\right\|_{op}}{1 - \left\|A^{-1}\right\|_{op}\left\|A - B_N\right\|_{op}}.$$

The last inequality follows from

$$A^{-1} - B_N^{-1} = A^{-1}(B_N - A)B_N^{-1},$$
$$B_N^{-1} = A^{-1}(I - (A - B_N)A^{-1})^{-1},$$

which together yields

$$A^{-1} - B_N^{-1} = A^{-1}(B_N - A)A^{-1}(I - (A - B_N)A^{-1})^{-1}.$$

Let $\varepsilon > 0$ and $\delta > 0$. Without loss of generality, suppose $\delta < 1/\left(2\left\|A^{-1}\right\|_{op}\right)$. Since $B_N \xrightarrow[N\to\infty]{P} A$ by assumption, there exists N_1 such that for every $N \geq N_1$, $\Pr\left[\left\|A - B_N\right\|_{op} < \delta\right] \geq 1 - \varepsilon$. Thus, for every $N \geq N_1$, with probability at least $1 - \varepsilon$, B_N^{-1} exists, and

$$\begin{aligned}\left\|A^{-1} - B_N^{-1}\right\|_{op} &\leq \frac{\left\|A^{-1}\right\|_{op}^2 \left\|B_N - A\right\|_{op}}{1 - \left\|A - B_N\right\|_{op}\left\|A^{-1}\right\|_{op}} \\ &\leq \left\|A^{-1}\right\|_{op}^2 \frac{\delta}{1 - \delta\left\|A^{-1}\right\|_{op}} < 2\left\|A^{-1}\right\|_{op}^2 \delta.\end{aligned} \tag{4.26}$$

Since $\left\|A - B_N\right\|_{op} < \delta$ implies $\left\|A^{-1} - B_N^{-1}\right\|_{op} < 2\left\|A^{-1}\right\|_{op}^2 \delta$, then

$$1 - \varepsilon \leq \Pr\left[\left\|A - B_N\right\|_{op} < \delta\right] \leq \Pr\left[\left\|A^{-1} - B_N^{-1}\right\|_{op} < 2\left\|A^{-1}\right\|_{op}^2 \delta\right],$$

which means that $B_N^{-1} \xrightarrow[N\to\infty]{P} A^{-1}$. □

4.4 SME in matrices and vectors

When L is the entire space $\mathbb{R}^s$ and $\langle\cdot,\cdot\rangle_L$ is the usual Euclidean inner product $\langle\cdot,\cdot\rangle_s$, the SME simplifies further (Hyvärinen [34]). In the notation here, which is from Forbes and Lauritzen [23], the canonical density (4.8) becomes

$$\log f(\boldsymbol{x}|\boldsymbol{\theta}) = \sum\nolimits_{k=1}^{s} T_k(\boldsymbol{x})\theta_k - a(\boldsymbol{\theta}) + b(\boldsymbol{x}),$$

the linear operator D in (4.11) becomes

$$D(\boldsymbol{x})\boldsymbol{\theta} = \left[\frac{\partial}{\partial x_j}\sum\nolimits_{k=1}^{s} T_k(\boldsymbol{x})\theta_k\right]_{j=1}^{n} = \left[\sum\nolimits_{k=1}^{s}\frac{\partial T_k(\boldsymbol{x})}{\partial x_j}\theta_k\right]_{j=1}^{n} = J_{\boldsymbol{x}}(T(\boldsymbol{x}))^{\top}\boldsymbol{\theta},$$

where

$$J_{\boldsymbol{x}}(T(\boldsymbol{x})) = \begin{bmatrix} \frac{\partial T_1}{\partial x_1} & \cdots & \frac{\partial T_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial T_s}{\partial x_1} & \cdots & \frac{\partial T_s}{\partial x_n} \end{bmatrix}$$

is the Jacobian matrix of T, so $D(\boldsymbol{x})$ is the Jacobian transposed,

$$D(\boldsymbol{x}) = J_{\boldsymbol{x}}(T(\boldsymbol{x}))^{\top},$$

the adjoint operator in (4.12) becomes simply the Jacobian itself,

$$D^*(\boldsymbol{x}) = J_{\boldsymbol{x}}(T(\boldsymbol{x}))\,. \tag{4.27}$$

and (4.13) becomes the Laplacian applied to T entry by entry,

$$\Delta_{\boldsymbol{x}} T(\boldsymbol{x}) = \begin{pmatrix} \Delta_{\boldsymbol{x}} T_1(\boldsymbol{x}) \\ \vdots \\ \Delta_{\boldsymbol{x}} T_s(\boldsymbol{x}) \end{pmatrix} \tag{4.28}$$

and the estimate (4.17) becomes

$$\hat{\boldsymbol{\theta}}_N = -\left(\sum_{i=1}^{N} J_{\boldsymbol{x}}(T(\boldsymbol{X}_i))\, J_{\boldsymbol{x}}(T(\boldsymbol{X}_i))^\top\right)^{-1} \sum_{i=1}^{N} \left(J_{\boldsymbol{x}}(T(\boldsymbol{X}_i))\, \nabla_{\boldsymbol{x}}^\top b(\boldsymbol{X}_i) + \Delta_{\boldsymbol{x}} T(\boldsymbol{X}_i)\right).$$

where $\nabla_{\boldsymbol{x}} b\,(\boldsymbol{x}) = \left(\frac{\partial b(\boldsymbol{x})}{\partial x_1}, \ldots, \frac{\partial b(\boldsymbol{x})}{\partial x_n}\right)$.

4.5 Continuity of SME

In Chapter 7, where the filtering method using SME is proposed, we will need continuity of SME with respect to random perturbations. We start with two well-known statements that will be used several times in this thesis. Then, we continue with a weak law of large numbers for triangular arrays. In addition to the notation from Section 4.1, we denote by $\operatorname{cov}_{\boldsymbol{X}\sim f(\cdot|\boldsymbol{\theta})}(\boldsymbol{X})$ the covariance of $\boldsymbol{X}$ with respect to its probability density $f\,(\boldsymbol{x}|\boldsymbol{\theta})$.

Lemma 8. *Let $\boldsymbol{X}$ be a random variable with values in $\mathbb{R}^n$, which has the density $f(\boldsymbol{x})$. Assume that $h : \mathbb{R}^n \to \mathbb{R}^m$ is a Borel measurable function such that $\mathsf{E}_{\boldsymbol{X}\sim f}\left(h(\boldsymbol{X})h(\boldsymbol{X})^\top\right)$ exists. Then, $\mathsf{E}_{\boldsymbol{X}\sim f}\, h(\boldsymbol{X})$ exists.*

Proof. From the Cauchy inequality,

$$\begin{aligned}
\left(\mathsf{E}_{\boldsymbol{X}\sim f} \left\|h(\boldsymbol{X})\right\|_m\right)^2 &= \left(\mathsf{E}_{\boldsymbol{X}\sim f} \left(\left\|h(\boldsymbol{X})\right\|_m \cdot 1\right)\right)^2 \leq \mathsf{E}_{\boldsymbol{X}\sim f}\left(\left\|h(\boldsymbol{X})\right\|_m^2\right) \mathsf{E}_{\boldsymbol{X}\sim f}\,(1) \\
&= \mathsf{E}_{\boldsymbol{X}\sim f} \left\|h(\boldsymbol{X})\right\|_m^2 = \mathsf{E}_{\boldsymbol{X}\sim f} \operatorname{tr}\left(h(\boldsymbol{X})h(\boldsymbol{X})^\top\right) \\
&= \operatorname{tr}\left(\mathsf{E}_{\boldsymbol{X}\sim f}\left(h(\boldsymbol{X})h(\boldsymbol{X})^\top\right)\right),
\end{aligned}$$

which is finite by assumption. □

Theorem 9 (Continuous mapping, Van der Vaart [69, Theorem 2.3 (ii)])**.** *Let $g : \mathbb{R}^k \to \mathbb{R}^m$ be continuous at every point of a set C such that $\Pr(\boldsymbol{X} \in C) = 1$. If $\boldsymbol{X}_N \xrightarrow[N\to\infty]{P} \boldsymbol{X}$, then $g(\boldsymbol{X}_N) \xrightarrow[N\to\infty]{P} g(\boldsymbol{X})$.*

Lemma 10. *Suppose that $f\,(\boldsymbol{x}|\boldsymbol{\theta})$ is a parametric probability density with respect to Lebesgue measure on a measurable set $\mathcal{X} \subset \mathbb{R}^n$ with parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset L$, with $\boldsymbol{\Theta}$ open, such that $\operatorname{cov}_{\boldsymbol{X}\sim f(\cdot|\boldsymbol{\theta})}(\boldsymbol{X})$ exists and is continuous from $\boldsymbol{\Theta}$ to $\mathbb{R}^{n\times n}$. Suppose that $\boldsymbol{\theta}_N$ are random parameters with values in $\boldsymbol{\Theta}$ such that $\boldsymbol{\theta}_N \xrightarrow{P} \boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$ as $N \to \infty$, and, for each N, $\left\{\boldsymbol{X}_i^N : i = 1, \ldots, k_N\right\}$ is a sample from $f\,(\boldsymbol{x}|\boldsymbol{\theta}_N)$, with $k_N \to \infty$ as $N \to \infty$. Then,*

$$\frac{1}{k_N} \sum_{i=1}^{k_N} \boldsymbol{X}_i^N - \mathsf{E}_{\boldsymbol{X}_1^N\sim f(\cdot|\boldsymbol{\theta}_N)}\left(\boldsymbol{X}_1^N\right) \xrightarrow[N\to\infty]{P} \mathbf{0}.$$

Proof. Since $\operatorname{cov}_{\boldsymbol{X}_1^N \sim f(\cdot|\boldsymbol{\theta}_N)}\left(\boldsymbol{X}_1^N\right)$ exists by assumption, $\mathsf{E}_{\boldsymbol{X}_1^N \sim f(\cdot|\boldsymbol{\theta}_N)}\left(\boldsymbol{X}_1^N\right)$ exists from Lemma 8. In order to simplify the notation, we will denote

$$\begin{aligned}\mathsf{E}_{\boldsymbol{X}_1^N \sim f(\cdot|\boldsymbol{\theta}_N)}\left(\boldsymbol{X}_1^N\right) &= \mathsf{E}\left(\boldsymbol{X}_1^N|\boldsymbol{\theta}_N\right), \quad \text{and}\\ \operatorname{cov}_{\boldsymbol{X}_1^N \sim f(\cdot|\boldsymbol{\theta}_N)}\left(\boldsymbol{X}_1^N\right) &= \operatorname{cov}\left(\boldsymbol{X}_1^N|\boldsymbol{\theta}_N\right) = C(\boldsymbol{\theta}_N).\end{aligned}$$

Further, denote

$$\boldsymbol{W}_i^N = \boldsymbol{X}_i^N - \mathsf{E}\left(\boldsymbol{X}_1^N|\boldsymbol{\theta}_N\right), \quad \bar{\boldsymbol{W}}^N = \frac{1}{k_N}\sum\nolimits_{i=1}^{k_N} \boldsymbol{W}_i^N.$$

We need to show that $\bar{\boldsymbol{W}}^N \xrightarrow{P} \mathbf{0}$ as $N \to \infty$. Fix N and $\boldsymbol{\theta}_N$. Then,

$$\begin{aligned}\mathsf{E}\left(\boldsymbol{W}_1^N|\boldsymbol{\theta}_N\right) &= \mathbf{0},\\ \operatorname{cov}\left(\boldsymbol{W}_1^N|\boldsymbol{\theta}_N\right) &= \operatorname{cov}\left(\boldsymbol{X}_1^N|\boldsymbol{\theta}_N\right) = C\left(\boldsymbol{\theta}_N\right),\end{aligned}$$

and, since $\boldsymbol{W}_i^N$ are uncorrelated, by the standard L^2 law of large numbers argument,

$$\begin{aligned}\mathsf{E}\left(\left\|\bar{\boldsymbol{W}}^N\right\|_n^2|\boldsymbol{\theta}_N\right) &= \frac{1}{k_N^2}\sum\nolimits_{i=1}^{k_N}\mathsf{E}\left(\left\|\boldsymbol{W}_i^N\right\|_n^2|\boldsymbol{\theta}_N\right) = \frac{1}{k_N}\mathsf{E}\left(\left\|\boldsymbol{W}_1^N\right\|_n^2|\boldsymbol{\theta}_N\right)\\ &= \frac{1}{k_N}\operatorname{tr}\left(\operatorname{cov}\left(\boldsymbol{W}_1^N|\boldsymbol{\theta}_N\right)\right) = \frac{1}{k_N}\operatorname{tr}\left(C\left(\boldsymbol{\theta}_N\right)\right).\end{aligned}$$

Let $\varepsilon > 0$ and $\delta > 0$. Using the Markov inequality, we have

$$\Pr\left(\left\|\bar{\boldsymbol{W}}^N\right\|_n^2 \geq \varepsilon^2|\boldsymbol{\theta}_N\right) \leq \frac{\mathsf{E}\left(\left\|\bar{\boldsymbol{W}}^N\right\|_n^2|\boldsymbol{\theta}_N\right)}{\varepsilon^2} = \frac{\operatorname{tr}\left(C\left(\boldsymbol{\theta}_N\right)\right)}{\varepsilon^2 k_N}.$$

Since $C\left(\boldsymbol{\theta}\right)$ is continuous function of $\boldsymbol{\theta}$, there exists $\eta > 0$ such that

$$\operatorname{tr}\left(C\left(\boldsymbol{\theta}\right)\right) < \operatorname{tr}\left(C\left(\boldsymbol{\theta}_0\right)\right) + 1, \quad \text{if } \left\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\right\|_L < \eta.$$

Since $\boldsymbol{\theta}_N \xrightarrow{P} \boldsymbol{\theta}_0$, there exists N_1 such that

$$\Pr\left(\left\|\boldsymbol{\theta}_N - \boldsymbol{\theta}_0\right\|_L \geq \eta\right) < \frac{\delta}{2}, \quad \text{if } N \geq N_1.$$

Then, by the law of total probability,

$$\begin{aligned}\Pr\left(\left\|\bar{\boldsymbol{W}}^N\right\|_n \geq \varepsilon\right) &= \underbrace{\Pr\left(\left\|\bar{\boldsymbol{W}}^N\right\|_n \geq \varepsilon|\left\|\boldsymbol{\theta}_N - \boldsymbol{\theta}_0\right\|_L \geq \eta\right)}_{\leq 1}\underbrace{\Pr\left(\left\|\boldsymbol{\theta}_N - \boldsymbol{\theta}_0\right\|_L \geq \eta\right)}_{<\frac{\delta}{2}}\\ &+ \underbrace{\Pr\left(\left\|\bar{\boldsymbol{W}}^N\right\|_n \geq \varepsilon|\left\|\boldsymbol{\theta}_N - \boldsymbol{\theta}_0\right\|_L < \eta\right)}_{\leq \frac{\operatorname{tr} C(\boldsymbol{\theta}_0)+1}{\varepsilon^2 k_N}}\underbrace{\Pr\left(\left\|\boldsymbol{\theta}_N - \boldsymbol{\theta}_0\right\|_L < \eta\right)}_{\leq 1}\\ &< \frac{\delta}{2} + \frac{\operatorname{tr}\left(C\left(\boldsymbol{\theta}_0\right)\right)+1}{\varepsilon^2 k_N}, \quad \text{if } N \geq N_1.\end{aligned}$$

Since $k_N \to \infty$, there exists N_2 such that

$$\frac{\operatorname{tr}\left(C\left(\boldsymbol{\theta}_0\right)\right)+1}{\varepsilon^2 k_N} < \frac{\delta}{2}, \quad \text{if } N \geq N_2.$$

Then, finally,

$$\Pr\left(\left\|\bar{\boldsymbol{W}}^N\right\|_n \geq \varepsilon\right) < \frac{\delta}{2} + \frac{\delta}{2} = \delta, \quad \text{if } N > \max\left\{N_1, N_2\right\},$$

which proves that $\bar{\boldsymbol{W}}^N \xrightarrow[N\to\infty]{P} \mathbf{0}$. □

We now apply Lemma 10 to functions of samples from exponential family distribution.

Lemma 11. *Suppose that $f(\boldsymbol{x}|\boldsymbol{\theta})$ is a parametric probability density on $\mathcal{X} \subset \mathbb{R}^n$ of the exponential family (4.8) with parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset L$, $\boldsymbol{\Theta}$ open. Further, suppose that $h : \mathbb{R}^n \to \mathbb{R}^m$ is Borel measurable and such that*

$$\int_{\mathcal{X}} \|h(\boldsymbol{x})\|_m^2 f(\boldsymbol{x}|\boldsymbol{\theta})\, d\boldsymbol{x} < \infty \quad \text{for all } \boldsymbol{\theta} \in \boldsymbol{\Theta}. \tag{4.29}$$

Finally, suppose that $\boldsymbol{\theta}_N$ are random parameters with values in $\boldsymbol{\Theta}$ such that $\boldsymbol{\theta}_N \xrightarrow[N\to\infty]{P} \boldsymbol{\theta}_0$, and, for each N, $\left\{\boldsymbol{X}_i^N : i = 1, \ldots, k_N\right\}$ is a sample from $f(\boldsymbol{x}|\boldsymbol{\theta}_N)$, and $k_N \xrightarrow[N\to\infty]{} \infty$. Then,

$$\frac{1}{k_N}\sum_{i=1}^{k_N} h\left(\boldsymbol{X}_i^N\right) \xrightarrow[N\to\infty]{P} \mathsf{E}_{\boldsymbol{X}\sim f(\cdot|\boldsymbol{\theta}_0)}\, h(\boldsymbol{X}).$$

Proof. We use Lemma 10 with $h\left(\boldsymbol{X}_i^N\right)$ in place of $\boldsymbol{X}_i^N$. From (4.29) and Lemma 8, the integrals

$$\begin{aligned} M(\boldsymbol{\theta}) &= \int_{\mathcal{X}} h(\boldsymbol{x})\, h(\boldsymbol{x})^\top f(\boldsymbol{x}|\boldsymbol{\theta})\, d\boldsymbol{x} = e^{-a(\boldsymbol{\theta})} \int_{\mathcal{X}} e^{\langle T(\boldsymbol{x}),\boldsymbol{\theta}\rangle_L} h(\boldsymbol{x})\, h(\boldsymbol{x})^\top e^{b(\boldsymbol{x})} d\boldsymbol{x} \\ m(\boldsymbol{\theta}) &= \int_{\mathcal{X}} h(\boldsymbol{x})\, f(\boldsymbol{x}|\boldsymbol{\theta})\, d\boldsymbol{x} = e^{-a(\boldsymbol{\theta})} \int_{\mathcal{X}} e^{\langle T(\boldsymbol{x}),\boldsymbol{\theta}\rangle_L} h(\boldsymbol{x})\, e^{b(\boldsymbol{x})} d\boldsymbol{x} \end{aligned} \tag{4.30}$$

exist for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Since $M(\boldsymbol{\theta})$ and $m(\boldsymbol{\theta})$ are the Fourier-Laplace transform of the sufficient statistics $T(\boldsymbol{x})$, they are analytic in $\boldsymbol{\Theta}$ (Lehmann and Romano [44, Theorem 2.7.1], Barndorff-Nielsen [6, Theorem 7.2]) and, in particular continuous. Thus,

$$C(\boldsymbol{\theta}) = \operatorname{cov}_{\boldsymbol{X}\sim f(\cdot|\boldsymbol{\theta})}\left(h(\boldsymbol{X})\right) = M(\boldsymbol{\theta}) - m(\boldsymbol{\theta})\, m(\boldsymbol{\theta})^\top$$

exists and is continuous from $\boldsymbol{\Theta}$ to $\mathbb{R}^{n\times n}$. Now, from Lemma 10,

$$\frac{1}{k_N}\sum_{i=1}^{k_N} h\left(\boldsymbol{X}_i^N\right) - \mathsf{E}_{\boldsymbol{X}_1^N\sim f(\cdot|\boldsymbol{\theta}_N)}\left(h\left(\boldsymbol{X}_1^N\right)\right) \xrightarrow[N\to\infty]{P} \mathbf{0}.$$

Next, writing the expectation as the integral (4.30),

$$\mathsf{E}_{\boldsymbol{X}_1^N\sim f(\cdot|\boldsymbol{\theta}_N)}\left(h\left(\boldsymbol{X}_1^N\right)\right) = m(\boldsymbol{\theta}_N),$$

which is a continuous function of $\boldsymbol{\theta}$, we have

$$\mathsf{E}_{\boldsymbol{X}_1^N\sim f(\cdot|\boldsymbol{\theta}_N)}\left(h\left(\boldsymbol{X}_1^N\right)\right) = m(\boldsymbol{\theta}_N) \xrightarrow[N\to\infty]{P} m(\boldsymbol{\theta}_0) = \mathsf{E}_{\boldsymbol{X}\sim f(\cdot|\boldsymbol{\theta}_0)}\left(h(\boldsymbol{X})\right),$$

by the continuous mapping theorem (cf., Theorem 9). □

For the following theorem, we need stronger versions of assumptions (C1)-(C3). Assume that

(C1b) $\mathsf{E}_{\boldsymbol{X}\sim p}\left\|\Delta_{\boldsymbol{x}}T(\boldsymbol{X})\right\|_L^2<\infty$,

(C2b) $\mathsf{E}_{\boldsymbol{X}\sim p}\left\|\nabla_{\boldsymbol{x}}^\top b\left(\boldsymbol{X}\right)\right\|_n^4<\infty$, and

(C3b) $\mathsf{E}_{\boldsymbol{X}\sim p}\left\|D(\boldsymbol{X})\right\|_{op}^4<\infty$.

We are now ready to prove continuity of SME to random perturbations of the parameter of the exponential family.

Theorem 12. *Suppose that $p(\boldsymbol{x})=f(\boldsymbol{x}|\boldsymbol{\theta}_0)$ for a unique $\boldsymbol{\theta}_0\in\boldsymbol{\Theta}$, where $f\left(\boldsymbol{x}|\boldsymbol{\theta}\right)=q(\boldsymbol{x}|\boldsymbol{\theta})/Z(\boldsymbol{\theta})$ is a parametric density on $\mathcal{X}$ from an exponential family (4.8) that satisfies assumptions:*

(B1) $f(\boldsymbol{x}|\boldsymbol{\theta}_0)$ and $\nabla_{\boldsymbol{x}}\log q(\boldsymbol{x}|\boldsymbol{\theta})$ are differentiable in $\mathcal{X}$,

(B4) $f(\boldsymbol{x}|\boldsymbol{\theta}_0)\nabla_{\boldsymbol{x}}^\top\log q(\boldsymbol{x}|\boldsymbol{\theta})\to\mathbf{0}$ when $\boldsymbol{x}\to\partial\mathcal{X}$ and the boundary $\partial\mathcal{X}$ of $\mathcal{X}$ is sufficiently regular for integration by parts,

(C4) $f(\boldsymbol{x}|\boldsymbol{\theta}_0)\nabla_{\boldsymbol{x}}^\top\langle T(\boldsymbol{x}),\boldsymbol{\theta}\rangle_L\to\mathbf{0}$ when $\boldsymbol{x}\to\partial\mathcal{X}$,

(C1b) $\mathsf{E}_{\boldsymbol{X}\sim f(\cdot|\boldsymbol{\theta}_0)}\left\|\Delta_{\boldsymbol{x}}T(\boldsymbol{X})\right\|_L^2<\infty$,

(C2b) $\mathsf{E}_{\boldsymbol{X}\sim f(\cdot|\boldsymbol{\theta}_0)}\left\|\nabla_{\boldsymbol{x}}^\top b\left(\boldsymbol{X}\right)\right\|_n^4<\infty$, and

(C3b) $\mathsf{E}_{\boldsymbol{X}\sim f(\cdot|\boldsymbol{\theta}_0)}\left\|D(\boldsymbol{X})\right\|_{op}^4<\infty$.

Let $\boldsymbol{\theta}_N$, $N=1,2,\ldots$, be random parameters with values in an open set $\boldsymbol{\Theta}\subset L$ such that $\boldsymbol{\theta}_N\xrightarrow[N\to\infty]{P}\boldsymbol{\theta}_0\in\boldsymbol{\Theta}$. Assume that the inverse of $\mathsf{E}_{\boldsymbol{X}\sim f(\cdot|\boldsymbol{\theta}_0)}\left(D^*(\boldsymbol{X})D(\boldsymbol{X})\right)$ exists. For each $N\in\mathbb{N}$, denote by $\hat{\boldsymbol{\theta}}_N$ the SME computed by (4.17) using a sample $\boldsymbol{X}_1^N,\ldots,\boldsymbol{X}_N^N$ from $f(\boldsymbol{x}|\boldsymbol{\theta}_N)$. Then $\hat{\boldsymbol{\theta}}_N\xrightarrow[N\to\infty]{P}\boldsymbol{\theta}_0$.

Proof. From Lemma 11 with $h\left(\boldsymbol{x}\right)=D^*(\boldsymbol{x})D(\boldsymbol{x})$ and assumption (C3b), it follows that

$$\frac{1}{N}\sum_{i=1}^N D^*(\boldsymbol{X}_i^N)D(\boldsymbol{X}_i^N)\xrightarrow[N\to\infty]{P}\mathsf{E}_{\boldsymbol{X}\sim f(\cdot|\boldsymbol{\theta}_0)}\left(D^*(\boldsymbol{X})D(\boldsymbol{X})\right). \tag{4.31}$$

Next, we apply Lemma 11 with $h\left(\boldsymbol{x}\right)=D^*(\boldsymbol{x})\nabla_{\boldsymbol{x}}^\top b(\boldsymbol{x})+\Delta_{\boldsymbol{x}}T(\boldsymbol{x})$. By Cauchy inequality and (C2b) together with (C3b),

$$\left(\int_{\mathcal{X}}\left\|D^*(\boldsymbol{x})\nabla_{\boldsymbol{x}}^\top b\left(\boldsymbol{x}\right)\right\|_L^2 f\left(\boldsymbol{x}|\boldsymbol{\theta}\right)d\boldsymbol{x}\right)^2\le\int_{\mathcal{X}}\left\|D^*\left(\boldsymbol{x}\right)\right\|_{op}^4 f\left(\boldsymbol{x}|\boldsymbol{\theta}\right)d\boldsymbol{x}\cdot$$
$$\cdot\int_{\mathcal{X}}\left\|\nabla_{\boldsymbol{x}}^\top b\left(\boldsymbol{x}\right)\right\|_n^4 f\left(\boldsymbol{x}|\boldsymbol{\theta}\right)d\boldsymbol{x}<\infty,$$

which yields (4.29). Thus, by Lemma 11,

$$\frac{1}{N}\sum_{i=1}^N D^*(\boldsymbol{X}_i^N)\nabla_{\boldsymbol{x}}^\top b(\boldsymbol{X}_i^N)+\Delta_{\boldsymbol{x}}T(\boldsymbol{X}_i^N)\xrightarrow[N\to\infty]{P}$$
$$\mathsf{E}_{\boldsymbol{X}\sim f(\cdot|\boldsymbol{\theta}_0)}\left(D^*(\boldsymbol{X})\nabla_{\boldsymbol{x}}^\top b(\boldsymbol{X})+\Delta_{\boldsymbol{x}}T(\boldsymbol{X})\right).$$

It was shown in (4.25) (inside the proof of Theorem 6) that

$$\mathsf{E}_{\boldsymbol{X}\sim f(\cdot|\boldsymbol{\theta}_0)}\left(D^*\left(\boldsymbol{X}\right)D\left(\boldsymbol{X}\right)\right)\boldsymbol{\theta}_0+\mathsf{E}_{\boldsymbol{X}\sim f(\cdot|\boldsymbol{\theta}_0)}\left(D^*\left(\boldsymbol{X}\right)\nabla_{\boldsymbol{x}}^{\top}b\left(\boldsymbol{X}\right)+\Delta_{\boldsymbol{x}}T\left(\boldsymbol{X}\right)\right)=\mathbf{0}. \tag{4.32}$$

Thus, from (4.31), (4.32), and the continuous mapping theorem (cf., Theorem 9),

$$\begin{aligned}&\left(\sum_{i=1}^{N}D^*(\boldsymbol{X}_i^N)D(\boldsymbol{X}_i^N)\right)^{-1}\sum_{i=1}^{N}\left(D^*(\boldsymbol{X}_i^N)\nabla_{\boldsymbol{x}}^{\top}b(\boldsymbol{X}_i^N)+\Delta_{\boldsymbol{x}}T(\boldsymbol{X}_i^N)\right)\\&\xrightarrow[N\to\infty]{P}\left(\mathsf{E}_{\boldsymbol{X}\sim f(\cdot|\boldsymbol{\theta}_0)}\left(D^*(\boldsymbol{X})D(\boldsymbol{X})\right)\right)^{-1}\mathsf{E}_{\boldsymbol{X}\sim f(\cdot|\boldsymbol{\theta}_0)}\left(D^*(\boldsymbol{X})\nabla_{\boldsymbol{x}}^{\top}b(\boldsymbol{X})+T(\boldsymbol{X})\right)=\boldsymbol{\theta}_0.\end{aligned}$$

Existence of the inverse on the left-hand side follows from Lemma 7 with $B_N=\frac{1}{N}\sum_{i=1}^{N}D^*(\boldsymbol{X}_i^N)D(\boldsymbol{X}_i^N)$ and $A=\mathsf{E}_{\boldsymbol{X}\sim f(\cdot|\boldsymbol{\theta}_0)}\left(D^*(\boldsymbol{X})D(\boldsymbol{X})\right)$. □

4.6 SME in Gaussian Markov random vector

In this section, the covariance matrix of a GMRF (defined in Section 2.2.2) will be estimated by the score matching method together with its expected value. First, an unconstrained covariance matrix, which corresponds to a degenerate case with no conditional independence between the entries of the random vector, is considered for completeness. In that case, it is shown that the score matching estimation method leads to the same estimator as the maximum likelihood method. Covariance regularization can be achieved through a linear model (2.10) for the precision matrix after replacing the parameter $\boldsymbol{\beta}=(\beta_1,\beta_2,\ldots,\beta_r)^{\top}$ by its SME, which is addressed in Section 4.6.2.

4.6.1 Unconstrained covariance matrix

Random vector following normal distribution $\mathcal{N}_n(\boldsymbol{\mu}_0,\Sigma_0)$ with unconstrained Σ_0 can be considered as a special case of GMRF. The computational procedure is based on Hyvärinen [33] with missing arguments added.

The SME of $(\boldsymbol{\mu}_0,\Sigma_0)$ in this case can be most easily computed by minimizing the score matching objective function (4.6). In order to compute the minimum, we take derivative of the objective function with respect to a symmetric matrix, which is addressed in the following lemma.

Lemma 13. *Define* $\mathbb{S}_n=\{B\in\mathbb{R}^{n\times n}:B=B^{\top}\}$ *with an inner product* $\langle A,B\rangle_{\mathbb{S}_n}=\operatorname{tr}(AB)\ \forall A,B\in\mathbb{S}_n$. *For* $A\in\mathbb{S}_n$ *define*

$$\begin{aligned}&g_1\colon\mathbb{S}_n\to\mathbb{R},\ g_1(Z)=\operatorname{tr}(ZAZ),\quad \textit{and}\\&g_2\colon\mathbb{S}_n\to\mathbb{R},\ g_2(Z)=\operatorname{tr}(Z).\end{aligned}$$

Then, the derivative $g_1'(Z)$ *is represented by* $ZA+AZ$ *and* $g_2'(Z)$ *by* I_n.

Proof. The derivative of $g_1(Z)$, resp. $g_2(Z)$, in direction $H \in \mathbb{S}_n$ is

$$\begin{aligned} g_1'(Z)H &= \lim_{t\to 0} \frac{g_1(Z+tH) - g_1(Z)}{t} \\ &= \lim_{t\to 0} \frac{\operatorname{tr}\left(ZAZ + tZAH + tHAZ + t^2HAH\right) - \operatorname{tr}(ZAZ)}{t} \\ &= \lim_{t\to 0} \frac{1}{t} \operatorname{tr}\left(tZAH + tHAZ + t^2HAH\right) \\ &= \operatorname{tr}(ZAH) + \operatorname{tr}(HAZ) = 2\operatorname{tr}(ZAH), \\ g_2'(Z)H &= \lim_{t\to 0} \frac{g_2(Z+tH) - g_2(Z)}{t} = \lim_{t\to 0} \frac{\operatorname{tr}(Z+tH) - \operatorname{tr}(Z)}{t} = \operatorname{tr}(H), \end{aligned}$$

since the trace is linear and invariant under transposition and since A, Z and H are symmetric.

Linear functionals $g_1'(Z)$ and $g_2'(Z)$ has the Riesz representation

$$\begin{aligned} g_1'(Z)H &= \langle B_1, H\rangle_{\mathbb{S}_n}, \quad \forall H \in \mathbb{S}_n, \\ g_2'(Z)H &= \langle B_2, H\rangle_{\mathbb{S}_n}, \quad \forall H \in \mathbb{S}_n, \end{aligned}$$

for some $B_1, B_2 \in \mathbb{S}_n$. The right choice is $B_1 = ZA + AZ$ and $B_2 = I_n$, which is easy to verify,

$$\begin{aligned} g_1'(Z)H &= \langle B_1, H\rangle_{\mathbb{S}_n} = \langle ZA + AZ, H\rangle_{\mathbb{S}_n} = \operatorname{tr}((ZA + AZ)H) \\ &= \operatorname{tr}(ZAH) + \operatorname{tr}(AZH) = 2\operatorname{tr}(ZAH) \\ g_2'(Z)H &= \langle B_2, H\rangle_{\mathbb{S}_n} = \langle I_n, H\rangle_{\mathbb{S}_n} = \operatorname{tr}(H), \end{aligned}$$

where we again used the linearity of trace and its invariance under transposition and cyclic permutation, which provide $\operatorname{tr}(AZH) = \operatorname{tr}(HAZ) = \operatorname{tr}\left((HAZ)^\top\right) = \operatorname{tr}(ZAH)$, due to the symmetry of Z, A, H. Hence, $g_1'(Z)$ is represented by $ZA + AZ \in \mathbb{S}_n$ and $g_2'(Z)$ by $I_n \in \mathbb{S}_n$. □

In order to show that the minimum of the objective function is unique, we will need Lemma 14.

Lemma 14. *Suppose $S \in \mathbb{R}^{n\times n}$ is symmetric positive definite and define operator $\mathcal{T} : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ by $\mathcal{T}(P) = PS + SP$. Then, $\mathcal{T}$ is injective.*

Proof. A linear mapping is injective if and only if its kernel is $\{0\}$. Therefore, the objective is to show that the only solution of

$$PS + SP = 0 \tag{4.33}$$

is $P = 0$.

Since $S > 0$, all its eigenvalues are positive and there exists a basis of $\mathbb{R}^n$ consisting of eigenvalues of S. By multiplying (4.33) by an eigenvector $\boldsymbol{v} \in \mathbb{R}^n$ of S, we get

$$\begin{aligned} \mathbf{0} &= P(S\boldsymbol{v}) + S(P\boldsymbol{v}) \\ \mathbf{0} &= P(\lambda\boldsymbol{v}) + S(P\boldsymbol{v}), \end{aligned}$$

which implies $S(P\boldsymbol{v}) = -\lambda(P\boldsymbol{v})$. Since all the eigenvalues of S are positive, i.e. $\lambda > 0$, $P\boldsymbol{v}$ cannot be an eigenvector of S and it follows that $P\boldsymbol{v} = \mathbf{0}$. This holds for every eigenvector of S and so $P = 0$.

□

We are now ready to prove that the score matching estimate of the covariance matrix equals to the sample covariance.

Theorem 15 (Hyvärinen [33, Section 3.1])**.** *Let* $\mathbb{X}_N = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_N\}$ *be a random sample from* $\mathcal{N}_n(\boldsymbol{\mu}_0, \Sigma_0)$ *with regular* Σ_0*. Then, the score matching estimator of* $(\boldsymbol{\mu}_0, \Sigma_0)$ *based on* $\mathbb{X}_N$ *is*

$$(\hat{\boldsymbol{\mu}}, \hat{\Sigma}) = \left(\bar{\boldsymbol{X}}, S\right),$$

if the sample covariance matrix S *is regular.*

Proof. It will be more convenient to work with a parameter $\boldsymbol{\theta} = (\boldsymbol{\mu}, P)$, where P stands for the precision matrix, instead of $(\boldsymbol{\mu}, \Sigma)$. By assumption, P is symmetric and positive definite. The logarithm of the density of $\mathcal{N}_n(\boldsymbol{\mu}, \Sigma)$ is

$$\log f(\boldsymbol{x}|\boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi) + \frac{1}{2}\log(\det P) - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top P(\boldsymbol{x}-\boldsymbol{\mu})$$

and its part that depends on the data is

$$\log q(\boldsymbol{x}|\boldsymbol{\theta}) = -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top P(\boldsymbol{x}-\boldsymbol{\mu}).$$

Since the gradient of a scalar function $\boldsymbol{y}^\top A\boldsymbol{y},\ A \in \mathbb{S}_n$, with respect to $\boldsymbol{y} \in \mathbb{R}^n$ is

$$\nabla_{\boldsymbol{y}}^\top(\boldsymbol{y}^\top A\boldsymbol{y}) = 2A\boldsymbol{y}, \tag{4.34}$$

we get

$$\begin{aligned}
\nabla_{\boldsymbol{x}}^\top \log q(\boldsymbol{x}|\boldsymbol{\theta}) &= -P(\boldsymbol{x}-\boldsymbol{\mu}) = P(\boldsymbol{\mu}-\boldsymbol{x}) \\
\Delta_{\boldsymbol{x}} \log q(\boldsymbol{x}|\boldsymbol{\theta}) &= \nabla_{\boldsymbol{x}} \cdot (\nabla_{\boldsymbol{x}} \log q(\boldsymbol{x}|\boldsymbol{\theta})) = \nabla_{\boldsymbol{x}} \cdot \left((\boldsymbol{\mu}-\boldsymbol{x})^\top P\right) = -\operatorname{tr}(P),
\end{aligned}$$

and the objective function (4.6) turns into

$$\begin{aligned}
\mathcal{S}_N(\boldsymbol{\theta}|\mathbb{X}_N) &= \frac{1}{N}\sum_{i=1}^N \left(\frac{1}{2}\left\|P(\boldsymbol{\mu}-\boldsymbol{X}_i)\right\|_n^2 - \operatorname{tr}(P)\right) + c_N(\mathbb{X}_N) \\
&= \frac{1}{N}\sum_{i=1}^N \left(\frac{1}{2}(\boldsymbol{\mu}-\boldsymbol{X}_i)^\top P^2(\boldsymbol{\mu}-\boldsymbol{X}_i) - \operatorname{tr}(P)\right) + c_N(\mathbb{X}_N) \qquad (4.35) \\
&= \operatorname{tr}\left(P\left(\frac{1}{2N}\sum_{i=1}^N (\boldsymbol{\mu}-\boldsymbol{X}_i)(\boldsymbol{\mu}-\boldsymbol{X}_i)^\top\right)P - P\right) + c_N(\mathbb{X}_N) \\
&= \operatorname{tr}\left(\frac{1}{2}PS_{\boldsymbol{\mu}}P - P\right) + c_N(\mathbb{X}_N), \qquad (4.36)
\end{aligned}$$

where $S_{\boldsymbol{\mu}} = \frac{1}{N}\sum_{i=1}^N (\boldsymbol{X}_i-\boldsymbol{\mu})(\boldsymbol{X}_i-\boldsymbol{\mu})^\top$ and $c_N(\mathbb{X}_N)$ is a constant that does not depend on parameters.

To obtain the SME, we minimize $\mathcal{S}_N(\boldsymbol{\theta}|\mathbb{X}_N)$ over $(\boldsymbol{\mu}, P)$ with symmetric positive definite P. For this purpose, compute the derivatives of $\mathcal{S}_N(\boldsymbol{\theta}|\mathbb{X}_N)$ with respect to $\boldsymbol{\mu}$ and P, which give the conditions for the minimizer $(\hat{\boldsymbol{\mu}}, \hat{P})$. The derivative with respect to $\boldsymbol{\mu}$ can be most easily computed from (4.35) by using (4.34):

$$\frac{\partial \mathcal{S}_N}{\partial \boldsymbol{\mu}}(\boldsymbol{\theta}|\mathbb{X}_N) = P^2(\boldsymbol{\mu}-\bar{\boldsymbol{X}}) = \boldsymbol{0}, \tag{4.37}$$

where $\bar{\boldsymbol{X}} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{X}_i$. By using Lemma 13 with $Z = P$ and $A = S_{\boldsymbol{\mu}}$, we can compute the derivative with respect to P from (4.36),

$$\frac{\partial \mathcal{S}_N}{\partial P}(\boldsymbol{\theta}|\mathbb{X}_N) = \frac{1}{2}\left(PS_{\boldsymbol{\mu}} + S_{\boldsymbol{\mu}}P\right) - I_n = 0. \tag{4.38}$$

It is easy to see that $\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{X}}$ and $\hat{P} = S_{\hat{\boldsymbol{\mu}}}^{-1}$ is a solution of (4.37, 4.38). By Lemma 14, $\hat{P}$ is determined by (4.38) uniquely. Since $\hat{P}$ is positive definite, from (4.37), $\hat{\boldsymbol{\mu}}$ is also unique. Hence, the score matching estimator of $\boldsymbol{\theta}_0 = (\boldsymbol{\mu}_0, P_0)$ is

$$\hat{\boldsymbol{\theta}} = (\bar{\boldsymbol{X}}, S_{\hat{\boldsymbol{\mu}}}^{-1}),$$

where $S_{\hat{\boldsymbol{\mu}}} = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})^\top = S$. □

4.6.2 Linear model for the precision matrix

Suppose that the precision matrix of a GMRF follows the linear model (2.10),

$$\Sigma^{-1} = \beta_1 A_1 + \beta_2 A_2 + \ldots + \beta_r A_r. \tag{4.39}$$

This model is formed by a linear combination of linearly independent known matrices $\{A_k\}_{k=1}^r$, which are called *design matrices* by Ueno and Tsuchiya [68]. The unconstrained covariance matrix in Section 4.6.1 can be considered to follow the linear model (4.39) with design matrices $A_{ij}, i = 1, \ldots, n,\ j = 1, \ldots, i$, such that A_{ij} has value 1 at positions (i, j) and (j, i) and zeros elsewhere. By choosing a more restrictive set of design matrices, the linear model (4.39) can provide a regularization of the precision matrix. In Ueno and Tsuchiya [68], the parameters $\beta_1, \ldots, \beta_r$ of this model are estimated by the maximum likelihood method, which, however, is not a linear problem in this case, and the maximization has to be done numerically. In this section, it will be shown that the score matching method provides estimators in a closed form. Regularization of the covariance matrix by means of the Markov property has been also considered in Spantini et al. [65]. Linear estimation of the precision matrix by score matching was studied in Forbes and Lauritzen [23].

The following theorem provides the explicit formulas of SME of the mean value and a linear model for the precision matrix of normal distribution $\mathcal{N}_n(\boldsymbol{\mu}_0, \Sigma_0)$. This result is essential for filtering algorithms presented in Chapter 7. Since the computation is for one fixed N, we can drop the subscript N here.

Theorem 16 (Turčičová et al. [66])**.** *Assume that $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N$ is a sample from $\mathcal{N}_n(\boldsymbol{\mu}_0, \Sigma_0)$, suppose that Σ_0 is regular, and consider the model $\sum_{k=1}^r \beta_k A_k$ for the precision matrix Σ^{-1}, where $\{A_k\}_{k=1}^r$ are given symmetric and linearly independent matrices. Denote $\boldsymbol{\beta} = (\beta_1, ..., \beta_r)^\top$. Then, the score matching estimator of $(\boldsymbol{\mu}_0, \boldsymbol{\beta}_0)$ is*

$$(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\beta}}) = \left(\bar{\boldsymbol{X}}, \left([\mathrm{tr}\,(SA_kA_l)]_{k,l=1}^r\right)^{-1}\left(\mathrm{tr}(A_1), \ldots, \mathrm{tr}(A_r)\right)^\top\right), \tag{4.40}$$

where $\bar{\boldsymbol{X}}$ and S are the sample mean and sample covariance (2.1), assuming that the inverse exists.

Proof. Normal distribution belongs to the exponential family of distributions, so we can compute SME based on (4.17). Note that the model density $f(\boldsymbol{x}|\boldsymbol{\mu},\Sigma)$ is assumed to coincide with the true density $p(\boldsymbol{x})$ for some $(\boldsymbol{\mu}_0,\Sigma_0)$. Dimension of the parameter space is $s=n+r$. The density of $\mathcal{N}_n(\boldsymbol{\mu},\Sigma)$ with $\Sigma^{-1}=\sum_{k=1}^r \beta_k A_k$ is

$$f(\boldsymbol{x}|\boldsymbol{\mu},\Sigma)=\frac{q(\boldsymbol{x}|\boldsymbol{\mu},\Sigma)}{\int_{\mathbb{R}^n} q(\boldsymbol{x}|\boldsymbol{\mu},\Sigma)d\boldsymbol{x}}, \tag{4.41}$$

where

$$\begin{aligned}\log q(\boldsymbol{x}|\boldsymbol{\mu},\Sigma) &= -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \sum_{k=1}^r \beta_k A_k(\boldsymbol{x}-\boldsymbol{\mu}) \\ &= \left\langle \begin{pmatrix} \boldsymbol{x} \\ -\frac{1}{2}\left(\boldsymbol{x}^\top A_1\boldsymbol{x},\dots,\boldsymbol{x}^\top A_r\boldsymbol{x}\right)^\top \end{pmatrix}, \begin{pmatrix} \sum_{k=1}^r \beta_k A_k\boldsymbol{\mu} \\ \boldsymbol{\beta} \end{pmatrix} \right\rangle_{n+r} \\ &\quad -\frac{1}{2}\left\langle \boldsymbol{\mu}, \sum_{k=1}^r \beta_k A_k\boldsymbol{\mu} \right\rangle_n \\ &= \langle T(\boldsymbol{x}),\boldsymbol{\eta}\rangle_{n+r} - a(\boldsymbol{\eta})\end{aligned} \tag{4.42}$$

is of the exponential family with a new parametrization

$$\boldsymbol{\eta}=\begin{pmatrix}\boldsymbol{\eta}_1\\ \boldsymbol{\eta}_2\end{pmatrix}=\begin{pmatrix}\sum_{k=1}^r \beta_k A_k\boldsymbol{\mu}\\ \boldsymbol{\beta}\end{pmatrix} \tag{4.43}$$

with sufficient statistics

$$T(\boldsymbol{x})=\begin{pmatrix}\boldsymbol{x}\\ -\frac{1}{2}\left(\boldsymbol{x}^\top A_1\boldsymbol{x},\dots,\boldsymbol{x}^\top A_r\boldsymbol{x}\right)^\top\end{pmatrix} \tag{4.44}$$

and

$$a(\boldsymbol{\eta})=\frac{1}{2}\left\langle \boldsymbol{\mu}, \sum_{k=1}^r \beta_k A_k\boldsymbol{\mu}\right\rangle_n .$$

The original parameters $(\boldsymbol{\mu},\boldsymbol{\beta})$ define probability density by (4.41) if and only if

$$(\boldsymbol{\mu},\boldsymbol{\beta})\in\tilde{\boldsymbol{\Theta}}=\left\{(\boldsymbol{\mu},\boldsymbol{\beta})\,\middle|\,\boldsymbol{\mu}\in\mathbb{R}^n, \sum_{k=1}^r \beta_k A_k \text{ is positive definite}\right\}.$$

Define

$$\boldsymbol{\Theta}=\left\{\boldsymbol{\eta}\in\mathbb{R}^{n+r}\,\middle|\,\sum_{k=1}^r \eta_{2k}A_k \text{ is positive definite}\right\},$$

where $\boldsymbol{\eta}_2=[\eta_{2k}]_{k=1}^r$. Lemma 17 (below) shows that (4.43) defines a one-to-one correspondence between $\boldsymbol{\eta}\in\boldsymbol{\Theta}$ and $(\boldsymbol{\mu},\boldsymbol{\beta})\in\tilde{\boldsymbol{\Theta}}$.

Now, the estimate (4.17) can be evaluated. From (4.27),

$$D^*(\boldsymbol{x})=J_{\boldsymbol{x}}(T(\boldsymbol{x}))=J_{\boldsymbol{x}}\begin{pmatrix}\boldsymbol{x}\\ -\frac{1}{2}\left(\boldsymbol{x}^\top A_1\boldsymbol{x},\dots,\boldsymbol{x}^\top A_r\boldsymbol{x}\right)^\top\end{pmatrix}=\begin{bmatrix} I_n \\ -\left[A_1\boldsymbol{x},\dots,A_r\boldsymbol{x}\right]^\top\end{bmatrix}$$

and then D is the transpose,

$$D(\boldsymbol{x})=\left[I_n,-\left[A_1\boldsymbol{x},\dots,A_r\boldsymbol{x}\right]\right].$$

So,

$$D^*(\boldsymbol{x})D(\boldsymbol{x}) = \begin{bmatrix} I_n \\ -\left[A_1\boldsymbol{x},\ldots,A_r\boldsymbol{x}\right]^\top \end{bmatrix}\left[I_n, -\left[A_1\boldsymbol{x},\ldots,A_r\boldsymbol{x}\right]\right]$$
$$= \begin{bmatrix} I_n & -\left[A_1\boldsymbol{x},\ldots,A_r\boldsymbol{x}\right] \\ -\left[A_1\boldsymbol{x},\ldots,A_r\boldsymbol{x}\right]^\top & \left[\boldsymbol{x}^\top A_k A_l \boldsymbol{x}\right]_{k,l=1}^r \end{bmatrix}. \tag{4.45}$$

For every $k = 1,\ldots,r$, we have

$$\Delta_{\boldsymbol{x}}\left(-\frac{1}{2}\boldsymbol{x}^\top A_k\boldsymbol{x}\right) = \nabla_{\boldsymbol{x}}\cdot\left(\nabla_{\boldsymbol{x}}\left(-\frac{1}{2}\boldsymbol{x}^\top A_k\boldsymbol{x}\right)\right) = -\nabla_{\boldsymbol{x}}\cdot\left(\boldsymbol{x}^\top A_k\right) = -\operatorname{tr}(A_k)$$

and due to (4.28), the Laplacian of the sufficient statistics from (4.44) is

$$\Delta_{\boldsymbol{x}}T(\boldsymbol{x}) = \begin{pmatrix} \mathbf{0} \\ -\left(\operatorname{tr}(A_1),\ldots,\operatorname{tr}(A_r)\right)^\top \end{pmatrix}. \tag{4.46}$$

With a random sample $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_N$, we have the SME (4.17),

$$\hat{\boldsymbol{\eta}} = -\left(\frac{1}{N}\sum_{i=1}^N D^*(\boldsymbol{X}_i)D(\boldsymbol{X}_i)\right)^{-1}\frac{1}{N}\sum_{i=1}^N \Delta_{\boldsymbol{x}}T(\boldsymbol{X}_i) \tag{4.47}$$
$$= -\begin{bmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{bmatrix}^{-1}\begin{pmatrix} \mathbf{0} \\ \left(\operatorname{tr}(A_1),\ldots,\operatorname{tr}(A_r)\right)^\top \end{pmatrix},$$

where

$$\begin{bmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{bmatrix} = \begin{bmatrix} I_n & -\frac{1}{N}\sum_{i=1}^N\left[A_1\boldsymbol{X}_i,\ldots,A_r\boldsymbol{X}_i\right] \\ -\frac{1}{N}\sum_{i=1}^N\left[A_1\boldsymbol{X}_i,\ldots,A_r\boldsymbol{X}_i\right]^\top & \frac{1}{N}\sum_{i=1}^N[\boldsymbol{X}_i^\top A_k A_l\boldsymbol{X}_i]_{k,l=1}^r \end{bmatrix}.$$

Using the formula for the inverse of 2×2 block matrix,

$$\begin{bmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{bmatrix}^{-1} = \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix} = \begin{bmatrix} {E_{11}}^{-1}\left(I_n + E_{12}M_{22}E_{21}{E_{11}}^{-1}\right) & -{E_{11}}^{-1}E_{12}M_{22} \\ -M_{22}E_{21}{E_{11}}^{-1} & S_{22}^{-1} \end{bmatrix},$$

where

$$\begin{aligned} S_{22} &= E_{22} - E_{21}{E_{11}}^{-1}E_{12} \\ &= \left[\frac{1}{N}\sum_{i=1}^N \boldsymbol{X}_i^\top A_k A_\ell \boldsymbol{X}_i\right]_{k,\ell=1}^r - \\ &\qquad -\frac{1}{N}\left[\sum_{i=1}^N A_1\boldsymbol{X}_i,\ldots,\sum_{i=1}^N A_s\boldsymbol{X}_i\right]^\top\frac{1}{N}\left[\sum_{i=1}^N A_1\boldsymbol{X}_i,\ldots,\frac{1}{N}\sum_{i=1}^N A_s\boldsymbol{X}_i\right] \\ &= \left[\frac{1}{N}\sum_{i=1}^N \boldsymbol{X}_i^\top A_k A_\ell\boldsymbol{X}_i - \frac{1}{N}\sum_{i=1}^N \boldsymbol{X}_i^\top A_k\frac{1}{N}\sum_{j=1}^N A_\ell\boldsymbol{X}_j\right]_{k,\ell=1}^r \qquad (4.48) \\ &= \left[\operatorname{tr}\left(\left(\frac{1}{N}\sum_{i=1}^N \boldsymbol{X}_i\boldsymbol{X}_i^\top - \frac{1}{N}\sum_{j=1}^N\boldsymbol{X}_j\frac{1}{N}\sum_{i=1}^N\boldsymbol{X}_i^\top\right)A_k A_\ell\right)\right]_{k,\ell=1}^r \\ &= \left[\operatorname{tr}\left(SA_kA_\ell\right)\right]_{k,\ell=1}^r. \qquad (4.49) \end{aligned}$$

Since S_{22}^{-1} exists by assumption, the inverse in (4.47) exists, and

$$\begin{aligned}\hat{\boldsymbol{\eta}} &= -\begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix} \begin{pmatrix} \mathbf{0} \\ (\operatorname{tr}(A_1), \ldots, \operatorname{tr}(A_r))^\top \end{pmatrix} \\ &= \begin{bmatrix} -{E_{11}}^{-1} E_{12} S_{22}^{-1} \\ S_{22}^{-1} \end{bmatrix} (\operatorname{tr}(A_1), \ldots, \operatorname{tr}(A_r))^\top \\ &= \begin{bmatrix} -{E_{11}}^{-1} E_{12} \\ I_n \end{bmatrix} S_{22}^{-1} (\operatorname{tr}(A_1), \ldots, \operatorname{tr}(A_r))^\top,\end{aligned}$$

which gives

$$\hat{\boldsymbol{\eta}}_2 = \hat{\boldsymbol{\beta}} = \left([\operatorname{tr}(S A_k A_\ell)]_{k,\ell=1}^r \right)^{-1} (\operatorname{tr}(A_1), \ldots, \operatorname{tr}(A_r))^\top$$

and

$$\begin{aligned}\hat{\boldsymbol{\eta}}_1 &= -{E_{11}}^{-1} E_{12} \hat{\boldsymbol{\eta}}_2 = \frac{1}{N} \sum_{i=1}^N [A_1 \boldsymbol{X}_i, \ldots, A_r \boldsymbol{X}_i] \, \hat{\boldsymbol{\beta}} \\ &= \frac{1}{N} \sum_{i=1}^N \sum_{k=1}^r \hat{\beta}_k A_k \boldsymbol{X}_i = \sum_{k=1}^r \hat{\beta}_k A_k \bar{\boldsymbol{X}}.\end{aligned}$$

By (4.43), $\boldsymbol{\eta}_1 = \sum_{k=1}^r \beta_k A_k \boldsymbol{\mu}$, and since the mapping of the parameters $(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)$ and the original parameters $(\boldsymbol{\mu}, \boldsymbol{\beta})$ is one-to-one by Lemma 17 below, it follows that $\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{X}}$. □

Remark 6. Note that the form (4.48) will be cheaper to compute than the elegant form (4.49).

Lemma 17. *Let all the assumptions of Theorem 16 hold and denote*

$$\begin{aligned}\tilde{\boldsymbol{\Theta}} &= \left\{ (\boldsymbol{\mu}, \boldsymbol{\beta}) \,\middle|\, \boldsymbol{\mu} \in \mathbb{R}^n, \sum_{k=1}^r \beta_k A_k \text{ is positive definite} \right\} \\ \boldsymbol{\Theta} &= \left\{ \boldsymbol{\eta} = \begin{pmatrix} \boldsymbol{\eta}_1 \\ \boldsymbol{\eta}_2 \end{pmatrix} \,\middle|\, \boldsymbol{\eta}_1 \in \mathbb{R}^n, \sum_{k=1}^r \eta_{2k} A_k \text{ is positive definite} \right\}.\end{aligned}$$

Then

$$\boldsymbol{\eta} = \begin{pmatrix} \boldsymbol{\eta}_1 \\ \boldsymbol{\eta}_2 \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^r \beta_k A_k \boldsymbol{\mu} \\ \boldsymbol{\beta} \end{pmatrix} \tag{4.50}$$

defines a homeomorphism between $\boldsymbol{\eta} \in \boldsymbol{\Theta}$ *and* $(\boldsymbol{\mu}, \boldsymbol{\beta}) \in \tilde{\boldsymbol{\Theta}}$.

Proof. Evaluating (4.50), $(\boldsymbol{\mu}, \boldsymbol{\beta}) \in \tilde{\boldsymbol{\Theta}}$ gives a unique $\boldsymbol{\eta} \in \boldsymbol{\Theta}$. In the opposite direction, if $\boldsymbol{\eta} \in \boldsymbol{\Theta}$, then $\boldsymbol{\eta}_2 = \boldsymbol{\beta}$ and, since $\sum_{k=1}^r \beta_k A_k$ is nonsingular, $\boldsymbol{\mu} = \left(\sum_{k=1}^r \beta_k A_k\right)^{-1} \boldsymbol{\eta}_2$.

The mapping $(\boldsymbol{\mu}, \boldsymbol{\beta}) \mapsto \boldsymbol{\eta}$ is continuous from the continuity of vector space operation, while the continuity of the inverse mapping follows using also the continuity of the mapping $A \mapsto A^{-1}$, cf. (4.26). □

Following through the steps of the proof of Theorem 16 omitting $\boldsymbol{\mu}$ as a parameter, we get the following result.

Theorem 18. *When $\boldsymbol{\mu}_0$ is known, then the score matching estimator of $\boldsymbol{\beta}_0$ is*

$$\hat{\boldsymbol{\beta}} = \left(\left[\operatorname{tr}\left(S_{\boldsymbol{\mu}_0} A_k A_l\right) \right]_{k,l=1}^{r} \right)^{-1} \left(\operatorname{tr}(A_1), \ldots, \operatorname{tr}(A_r) \right)^\top \tag{4.51}$$

with $S_{\boldsymbol{\mu}_0} = \frac{1}{N} \sum_{i=1}^{N} (\boldsymbol{X}_i - \boldsymbol{\mu}_0)(\boldsymbol{X}_i - \boldsymbol{\mu}_0)^\top$ (if the inverse exists).

Proof. From (4.42), we have

$$\begin{aligned}
\log q(\boldsymbol{x}|\Sigma) = \log q(\boldsymbol{x}|\boldsymbol{\beta}) &= -\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu}_0)^\top \sum_{k=1}^{r} \beta_k A_k (\boldsymbol{x} - \boldsymbol{\mu}_0) \\
&= \left\langle -\frac{1}{2} \left((\boldsymbol{x} - \boldsymbol{\mu}_0)^\top A_1 (\boldsymbol{x} - \boldsymbol{\mu}_0), \ldots, (\boldsymbol{x} - \boldsymbol{\mu}_0)^\top A_r (\boldsymbol{x} - \boldsymbol{\mu}_0) \right)^\top, \boldsymbol{\beta} \right\rangle_r \\
&= \langle T(\boldsymbol{x}), \boldsymbol{\beta} \rangle_r
\end{aligned}$$

with the sufficient statistics

$$T(\boldsymbol{x}) = -\frac{1}{2} \left((\boldsymbol{x} - \boldsymbol{\mu}_0)^\top A_1 (\boldsymbol{x} - \boldsymbol{\mu}_0), \ldots, (\boldsymbol{x} - \boldsymbol{\mu}_0)^\top A_r (\boldsymbol{x} - \boldsymbol{\mu}_0) \right)^\top .$$

From (4.27),

$$\begin{aligned}
D^*(\boldsymbol{x}) &= J_{\boldsymbol{x}}(T(\boldsymbol{x})) \\
&= J_{\boldsymbol{x}} \left(-\frac{1}{2} \left((\boldsymbol{x} - \boldsymbol{\mu}_0)^\top A_1 (\boldsymbol{x} - \boldsymbol{\mu}_0), \ldots, (\boldsymbol{x} - \boldsymbol{\mu}_0)^\top A_r (\boldsymbol{x} - \boldsymbol{\mu}_0) \right)^\top \right) \\
&= -\left[A_1 (\boldsymbol{x} - \boldsymbol{\mu}_0), \ldots, A_r (\boldsymbol{x} - \boldsymbol{\mu}_0) \right]^\top
\end{aligned}$$

and then D is the transpose,

$$D(\boldsymbol{x}) = -\left[A_1 (\boldsymbol{x} - \boldsymbol{\mu}_0), \ldots, A_r (\boldsymbol{x} - \boldsymbol{\mu}_0) \right].$$

So,

$$D^*(\boldsymbol{x}) D(\boldsymbol{x}) = \left[(\boldsymbol{x} - \boldsymbol{\mu}_0)^\top A_k A_l (\boldsymbol{x} - \boldsymbol{\mu}_0) \right]_{k,l=1}^{r} .$$

Similarly as in (4.46),

$$\Delta_{\boldsymbol{x}} T(\boldsymbol{x}) = -\left(\operatorname{tr}(A_1), \ldots, \operatorname{tr}(A_r) \right)^\top .$$

With a random sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N$, we have

$$\begin{aligned}
\frac{1}{N} \sum_{i=1}^{N} D^*(\boldsymbol{X}_i) D(\boldsymbol{X}_i) &= \frac{1}{N} \sum_{i=1}^{N} (\boldsymbol{X}_i - \boldsymbol{\mu}_0)^\top A_k A_l (\boldsymbol{X}_i - \boldsymbol{\mu}_0) \\
&= \operatorname{tr} \left(\frac{1}{N} \sum_{i=1}^{N} (\boldsymbol{X}_i - \boldsymbol{\mu}_0)(\boldsymbol{X}_i - \boldsymbol{\mu}_0)^\top A_k A_l \right) = \operatorname{tr}\left(S_{\boldsymbol{\mu}_0} A_k A_l\right).
\end{aligned}$$

Then, the SME (4.17) is

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= -\left(\frac{1}{N} \sum_{i=1}^{N} D^*(\boldsymbol{X}_i) D(\boldsymbol{X}_i) \right)^{-1} \frac{1}{N} \sum_{i=1}^{N} \Delta_{\boldsymbol{x}} T(\boldsymbol{X}_i) \\
&= \left(\left[\operatorname{tr}\left(S_{\boldsymbol{\mu}_0} A_k A_l\right) \right]_{k,l=1}^{r} \right)^{-1} \left(\operatorname{tr}(A_1), \ldots, \operatorname{tr}(A_r) \right)^\top .
\end{aligned}$$

□

We now show that the SME of the distribution of GMRF is consistent. In order to emphasize that the sample covariance matrix depends on N, we will use the notation S_N instead of S in the following theorem. Similarly, we will use $\bar{\boldsymbol{X}}_N$ instead of $\bar{\boldsymbol{X}}$.

Theorem 19. *Assume that all the assumptions of Theorem 16 are satisfied and denote by $\hat{\boldsymbol{\beta}}_N$ the SME of $\boldsymbol{\beta}_0$ from (4.40) based on a sample of size N, if the inverse of $[\mathrm{tr}(S_N A_k A_l)]_{k,l=1}^r$ exists. Further assume that $\mathsf{E}_{\boldsymbol{X}\sim f(\cdot|\boldsymbol{\mu}_0,\boldsymbol{\beta}_0)}\left(D^*(\boldsymbol{X})D(\boldsymbol{X})\right)$ defined by (4.45) is invertible. Then, $(\bar{\boldsymbol{X}}_N, \hat{\boldsymbol{\beta}}_N)$ is a consistent estimator of $(\boldsymbol{\mu}_0, \boldsymbol{\beta}_0)$, and, in particular, $\sum_{k=1}^r \hat{\beta}_{Nk} A_k$ is a consistent estimator of Σ_0^{-1}.*

Proof. Denote $\boldsymbol{\beta}_0 = [\beta_{0k}]_{k=1}^r$. We will apply Theorem 6 to the parameter

$$\boldsymbol{\eta} = \begin{pmatrix} \sum_{k=1}^r \beta_k A_k \boldsymbol{\mu} \\ \boldsymbol{\beta} \end{pmatrix},$$

which provides

$$\hat{\boldsymbol{\eta}}_N \xrightarrow[N\to\infty]{P} \boldsymbol{\eta}_0 = \begin{pmatrix} \sum_{k=1}^r \beta_{0k} A_k \boldsymbol{\mu}_0 \\ \boldsymbol{\beta}_0 \end{pmatrix},$$

where $\Pr\left(\hat{\boldsymbol{\eta}}_N \text{ exists}\right) \to 1$ as $N \to \infty$. Then, the convergence

$$(\bar{\boldsymbol{X}}_N, \hat{\boldsymbol{\beta}}_N) \xrightarrow[N\to\infty]{P} (\boldsymbol{\mu}_0, \boldsymbol{\beta}_0)$$

will result from the continuous mapping theorem (cf., Theorem 9) applied to the mapping of $\boldsymbol{\eta}$ to $(\boldsymbol{\mu}, \boldsymbol{\beta})$, which is continuous by Lemma 17.

We need to verify the assumptions of Theorem 6. Recall that

$$\log f(\boldsymbol{x}|\boldsymbol{\eta}) = \langle T(\boldsymbol{x}), \boldsymbol{\eta}\rangle_{n+r} - a(\boldsymbol{\eta}) + b(\boldsymbol{x}),$$

where for normal distribution $b(\boldsymbol{x}) = 0$ and the sufficient statistics (4.44) for $\boldsymbol{\eta}$ is

$$T(\boldsymbol{x}) = \begin{bmatrix} \boldsymbol{x} \\ -\frac{1}{2}\left(\boldsymbol{x}^\top A_1 \boldsymbol{x}, \ldots, \boldsymbol{x}^\top A_r \boldsymbol{x}\right)^\top \end{bmatrix}.$$

The density of normal distribution and the gradient of its logarithm are differentiable, as required in assumption (B1). Since all moments of normal distribution are finite, assumptions (C1) and (C3) are fulfilled. Since $b(\boldsymbol{x}) = 0$ for normal distribution, (C2) is fulfilled automatically, and further, $\nabla_{\boldsymbol{x}} \log q(\boldsymbol{x}|\boldsymbol{\theta}) = \nabla_{\boldsymbol{x}} \langle T(\boldsymbol{x}), \boldsymbol{\theta}\rangle_{n+r}$, which implies that assumption (C4) coincides with (B4). It is evident that T is continuous and polynomial, and, therefore,

$$\lim_{\|\boldsymbol{x}\|_n \to \infty} f(\boldsymbol{x}|\boldsymbol{\eta}_0) \frac{\partial \log q(\boldsymbol{x}|\boldsymbol{\eta})}{\partial x_j} = \lim_{\|\boldsymbol{x}\|_n \to \infty} f(\boldsymbol{x}|\boldsymbol{\eta}_0) \frac{\partial}{\partial x_j} \sum_{k=1}^{n+r} T_k(\boldsymbol{x}) \eta_k = 0$$

for all $j = 1, \ldots, n$ and for any $\boldsymbol{\eta}$, because of the exponential decay of $f(\boldsymbol{x}|\boldsymbol{\eta}_0)$. Thus, assumption (B4) is satisfied (cf., Remark 4). The inverse of

$$\mathsf{E}_{\boldsymbol{X}\sim f(\cdot|\boldsymbol{\eta}_0)}\left(D^*(\boldsymbol{X})D(\boldsymbol{X})\right) = \mathsf{E}_{\boldsymbol{X}\sim f(\cdot|\boldsymbol{\eta}_0)} \begin{bmatrix} I_n & -[A_1\boldsymbol{X}, \ldots, A_r\boldsymbol{X}] \\ -[A_1\boldsymbol{X}, \ldots, A_r\boldsymbol{X}]^\top & \left[\mathrm{tr}\left(A_k A_l \boldsymbol{X}\boldsymbol{X}^\top\right)\right]_{k,l=1}^r \end{bmatrix}$$

exists by assumption. □

Corollary. Under the assumptions of Theorem 19, $\hat{\Sigma}_N = \left(\sum_{k=1}^r \hat{\beta}_{Nk} A_k\right)^{-1}$ (when the inverse exists) is a consistent estimator of Σ_0.

Proof. From Theorem 19,

$$\sum_{k=1}^r \hat{\beta}_{Nk} A_k \xrightarrow[N\to\infty]{P} \Sigma_0^{-1}.$$

Since Σ_0^{-1} exists by assumption, the consistency of $\hat{\Sigma}_N$ follows from Lemma 7. □

The matrices $A_k, k = 1, \ldots, r$, are usually chosen as sparse matrices, whose diagonals and subdiagonals can effectively model the appropriate precision matrix (e.g. as in Figure 2.1). The design matrices A_k also need to be selected in a way that the inverse in (4.40) (and hence the whole SME) exists.

Theorem 20. *The matrix* $[\operatorname{tr}(SA_kA_l)]_{k,l=1}^r$*, where* S *is the sample covariance of* $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N$*, is invertible if and only if the matrices* $A_k\left[\boldsymbol{X}_1 - \bar{\boldsymbol{X}}, \ldots, \boldsymbol{X}_N - \bar{\boldsymbol{X}}\right]$*,* $k = 1, \ldots, r$*, are linearly independent as elements of* $\mathbb{R}^{n\times N}$*.*

Proof. From (4.48),

$$\begin{aligned}
[\operatorname{tr}(SA_kA_\ell)]_{k,\ell=1}^r &= \left[\frac{1}{N}\sum_{i=1}^N \left(\boldsymbol{X}_i - \frac{1}{N}\sum_{j=1}^N \boldsymbol{X}_j\right)^\top A_k A_\ell \left(\boldsymbol{X}_i - \frac{1}{N}\sum_{j=1}^N \boldsymbol{X}_j\right)\right]_{k,\ell=1}^r \\
&= \frac{1}{N}\left[\sum_{i=1}^N \left(\boldsymbol{X}_i - \bar{\boldsymbol{X}}\right)^\top A_k^\top A_\ell \left(\boldsymbol{X}_i - \bar{\boldsymbol{X}}\right)\right]_{k,\ell=1}^r
\end{aligned}$$

which is a nonzero multiple of the Gram matrix of the vectors

$$A_k\left[\boldsymbol{X}_1 - \bar{\boldsymbol{X}}, \ldots, \boldsymbol{X}_N - \bar{\boldsymbol{X}}\right] = \left[A_k\left(\boldsymbol{X}_1 - \bar{\boldsymbol{X}}\right), \ldots, A_k\left(\boldsymbol{X}_N - \bar{\boldsymbol{X}}\right)\right] \in \mathbb{R}^{n\times N},$$

for $k = 1, \ldots, r$, with $\mathbb{R}^{n\times N}$ equipped with the Frobenius inner product, which can be written as

$$\langle [\boldsymbol{U}_1, \ldots, \boldsymbol{U}_N], [\boldsymbol{V}_1, \ldots, \boldsymbol{V}_N]\rangle_{n\times N} = \sum_{i=1}^N \boldsymbol{U}_i^\top \boldsymbol{V}_i,$$

where $\boldsymbol{U}_i, \boldsymbol{V}_i \in \mathbb{R}^n$ for $i = 1, \ldots, N$. □

Corollary. A necessary condition for $[\operatorname{tr}(SA_kA_l)]_{k,l=1}^r$ to be invertible is that the set of design matrices $\{A_k, k = 1, \ldots, r\}$ is linearly independent.

Corollary. If $[\operatorname{tr}(SA_kA_l)]_{k,l=1}^r$ is invertible for a set of design matrices $\mathcal{A} = \{A_k, k = 1, \ldots, r\}$, then it is invertible for any nonempty subset of $\mathcal{A}$.

Remark 7. The assumed model $\sum_{k=1}^r \beta_k A_k$ for Σ^{-1} does not contain any restriction that would ensure the positive definiteness of the resulting estimator, which is equivalent with positive definiteness of the associated covariance matrix estimator. Positive definiteness of $\sum_{k=1}^r \hat{\beta}_{Nk} A_k$ cannot be guaranteed, because the set of positive definite matrices is not closed and so the objective function may not achieve its minimum in this set. The matrix $\sum_{k=1}^r \hat{\beta}_{Nk} A_k$ only converge to

the positive definite matrix with probability tending to 1, due to the consistency stated in Theorem 19.

When the estimate of precision matrix is not required to be invertible, we can get a positive semidefinite estimate by taking a set of positive semidefinite design matrices $\{A_k\}_{k=1}^r$ and working only with nonnegative coefficients $\hat{\beta}_1, \dots, \hat{\beta}_r$. The existence and consistency of SME of a parameter from $\mathbb{R}_+^r$ was studied by Yu et al. [74]. However, matrices $\sum_{k=1}^r \beta_k A_k$ with $(\beta_1, \dots, \beta_r)^\top \in \mathbb{R}_+^r$ form only a subset of the space of all positive semidefinite matrices, so this method would reduce the parameter space dramatically.

Due to the fact that the positive definiteness of the resulting estimate cannot be guaranteed by any prior restrictions on the model, this problem needs to be addressed in a different way. In Section 7.4.2, positive definiteness of estimates based on small samples is ensured by the process of model selection.

4.6.3 SME of GMRF from a triangular array of samples

In the following theorem, we will apply the continuity result from Section 4.5 on the parameters $\boldsymbol{\mu}, \boldsymbol{\beta} = [\beta_k]_{k=1}^r$ of normal distribution $\mathcal{N}_n\left(\boldsymbol{\mu}, \left(\sum_{k=1}^r \beta_k A_k\right)^{-1}\right)$. It will be a key component in the proof of consistency of the filtering algorithm proposed in Section 7.1 and it may be also of independent interest.

Theorem 21. *Let $f(\cdot|\boldsymbol{\mu}_0, \boldsymbol{\beta}_0)$ be the density of $\mathcal{N}_n\left(\boldsymbol{\mu}_0, \left(\sum_{k=1}^r \beta_{0k} A_k\right)^{-1}\right)$, and assume that $\mathsf{E}_{\boldsymbol{X}\sim f(\cdot|\boldsymbol{\mu}_0,\boldsymbol{\beta}_0)}\left(D^*(\boldsymbol{X})D(\boldsymbol{X})\right)$ defined by (4.45) is invertible. Further, assume that $(\boldsymbol{\mu}_N, \boldsymbol{\beta}_N) \xrightarrow{P} (\boldsymbol{\mu}_0, \boldsymbol{\beta}_0)$ as $N \to \infty$ and denote $\Sigma_N^{-1} = \sum_{k=1}^r \beta_{N,k} A_k$. Let $\bar{\mathbf{X}}_N^N, \hat{\boldsymbol{\beta}}_N$ be the SMEs computed from formula (4.40) based on a random sample $\boldsymbol{X}_1^N, \dots, \boldsymbol{X}_N^N$ from $\mathcal{N}_n(\boldsymbol{\mu}_N, \Sigma_N)$. Then*

$$(\bar{\mathbf{X}}_N^N, \hat{\boldsymbol{\beta}}_N) \xrightarrow[N\to\infty]{P} (\boldsymbol{\mu}_0, \boldsymbol{\beta}_0). \tag{4.52}$$

Proof. The proof follows the same scheme as the proof of Theorem 19. The convergence (4.52) results from the Theorem 12 applied to the parameter

$$\boldsymbol{\eta} = \begin{pmatrix} \sum_{k=1}^r \beta_k A_k \boldsymbol{\mu} \\ \boldsymbol{\beta} \end{pmatrix},$$

followed by the continuous mapping theorem applied to the mapping of $\boldsymbol{\eta}$ to $(\boldsymbol{\mu}, \boldsymbol{\beta})$.

As in the proof of Theorem 19, we only need to verify the assumptions of Theorem 12 for normal distribution. Assumptions (B1), (B4) and (C4) have been already discussed within the proof of Theorem 19. Assumptions (C1b)-(C3b) represent stronger version of assumptions (C1)-(C3). However, since the normal distribution has finite moments of all orders, assumptions (C1b)-(C3b) are satisfied as well. The inverse of $\mathsf{E}_{\boldsymbol{X}\sim f(\cdot|\boldsymbol{\eta}_0)}\left(D^*(\boldsymbol{X})D(\boldsymbol{X})\right)$ exists by assumption. □

4.7 Computational study

The following simulations illustrate that, unlike the unconstrained SME in Section 4.6.1, the SME (4.40) gives only very similar (but not the same) results as

the maximum likelihood method for the same linear model (4.39), which was described in Ueno and Tsuchiya [68]. However, computing SME is incomparably faster since it is given as a solution of a system of linear equations and avoids numerical maximization of an intricate likelihood function.

4.7.1 Comparison of SME and MLE on simulated GMRF

In order to compare estimates computed by the score matching method with those obtained from maximum likelihood, we created a precision matrix $\Sigma_0^{-1} = \sum_{k=1}^{r} \beta_{0k} A_k$ displayed in Figure 4.2a corresponding to a first-order GMRF $\boldsymbol{X}$ with dimension 5×5 with 4 neighbours of every gridpoint (see the scheme in Figure 2.1a). The values on the main diagonal, resp. two subdiagonals, were generated from the uniform distribution on interval $[4, 8]$, resp. [-2,-0.5]. We compose our set of design matrices, $\{A_k\}_{k=1}^{r}$, from linearly independent matrices such that for each point (i, j) on the main diagonal and the subdiagonals, the corresponding A_k has value 1 at positions (i, j) and (j, i) and zeros elsewhere. Thus, the number of parameters is $r = 65$. A random sample $\mathbb{X}_N = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N\}$ was generated from $\mathcal{N}_n(\mathbf{0}, \Sigma_0)$. The SME of $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0r})^\top$ was computed from (4.51) and the MLE was obtained by numerical maximization of the log-likelihood

$$\ell(\boldsymbol{\beta}|\mathbb{X}_N) = -\frac{Nn}{2}\log(2\pi) + \frac{N}{2}\log\left(\det\left(\sum_{k=1}^{r}\beta_k A_k\right)\right) - \frac{1}{2}\sum_{i=1}^{N}\boldsymbol{X}_i^\top \sum_{k=1}^{r}\beta_k A_k \boldsymbol{X}_i,$$

both using the same sample. The numerical comparison of resulting estimates is depicted in Figure 4.1 and the corresponding estimates of the precision matrix in Figures 4.2b and 4.2c.

The score matching and maximum likelihood estimates do not exhibit substantial differences in their values. As expected for the sample size of 20, the error of both estimates was large for many of β_k's.

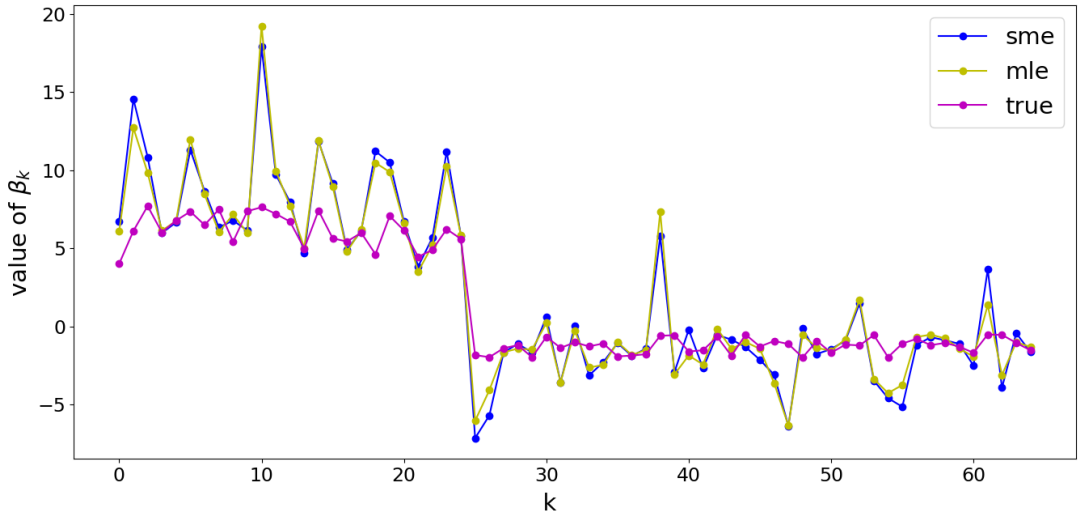
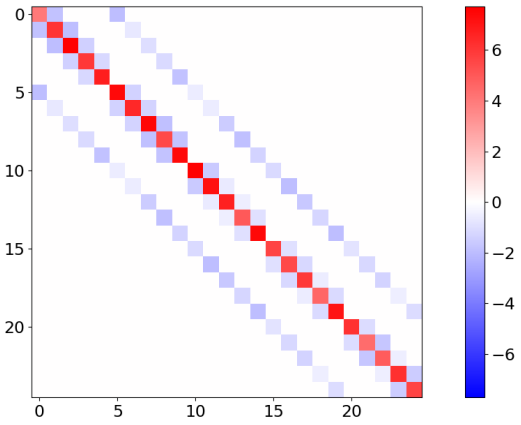
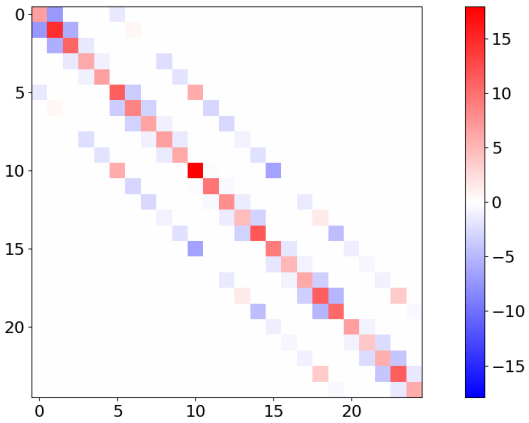


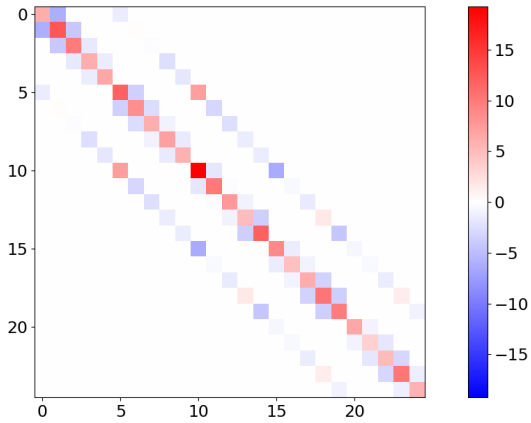
Figure 4.1: Simulated first-order GMRF(dimensions 5×5, columns stacked vertically): Comparison of the score matching (sme) and the maximum likelihood (mle) estimates with the true parameters $[\beta_{0k}]_{k=1}^{65}$. Sample size was $N = 20$. First 25 entries correspond to the diagonal of Σ_0^{-1}.



(a) True precision matrix.



(b) Estimate based on the score matching method.



(c) Estimate based on the maximum likelihood method.

Figure 4.2: Simulated first-order GMRF (dimensions 5×5, columns stacked vertically): The precision matrix and its estimates based on a sample of size $N = 20$.

4.7.2 An illustration of modelling covariance of real weather fields in wavelet domain

One of the standard approaches to modelling covariance matrix in atmospheric data assimilation lies in transforming the variables which enter the assimilation process (such as temperature, air pressure, wind velocity etc.) to another space, often spectral, such as Fourier or wavelet space. For spatial data, the wavelet transform is often preferred because it enables us to model not only wave (spectral) characteristics but also local properties tied to location in space. Similarly as in the Fourier transform, wavelet transform is based on a decomposition of $\boldsymbol{X}$,

$$\boldsymbol{X} = \mathrm{E}\,\boldsymbol{X} + \sum_{j=1}^{n} d_j^{1/2} \xi_j \boldsymbol{v}_j, \tag{4.53}$$

where $\boldsymbol{v}_j$ are now the vectors of the wavelet basis, ξ_j are random variables that have unit variance but that are not necessarily independent and d_j are deterministic real coefficients. We then have a decomposition of the covariance matrix

$$\Sigma = FDF^\top, \tag{4.54}$$

where matrix D may not be diagonal in general but there are important situations where it is approximately diagonal and/or sparse. For example, local stationarity of the random field may often be an appropriate assumption and Pannekoucke et al. [58] prove that "a wavelet diagonal approach amounts to locally averaging the correlations", i.e. modelling Σ with a diagonal matrix D in (4.54) represents a locally stationary approximation of the field $\boldsymbol{X}$. Another theoretical justification for employing models with sparse covariance matrix in wavelet space is in Matsuo et al. [53] and in the references therein. Roughly speaking, the decay of the off-diagonal elements of D in (4.54) is quantified under fairly general assumptions and it depends on the distance of locations.

We do not need a deeper insight into wavelet theory nor technical details here. For our purpose it is sufficient to keep in mind that the wavelet transform performs a multiscale decomposition of the field. A comprehensive treatment of wavelets is found in, e.g., Burrus et al. [14]. The theory is covered in Daubechies [18].

The transformation matrix F in (4.54) is composed of basis functions generated from *scaling functions*, which keep the lowest frequencies of the transform and ensure that the whole spectrum is covered, and *wavelets*, which keep higher frequencies and provide more detailed information. It is possible and sometimes advantageous to require the scaling functions and wavelets to be orthogonal. The scaling functions and wavelets are organized into *bands*. Each band involves a number of wavelets of the same frequency but shifted in space differently. In practice, the wavelet transform is computed by successive application of a filter bank consisting of a low-pass and a high-pass filter. In this manner, coefficients, which correspond to several *levels*, arise. The wavelet compression technique, common in engineering literature, lies in discarding some levels which are not essential for the task at hand. In this study, we will use *Daubechies 2* wavelets (often denoted as DB2) defined in Daubechies [17].

In the original physical space, the variables of interest can often be well approximated by a spatial GMRF. The scaling coefficients in the lowest frequency

sub-bands have a spatial covariance structure similar to the original physical space, and thus, they may be modelled by a GMRF as well. Hence, we propose to model the inverse of their covariance matrix by a linear model of type (2.10),

$$\Sigma^{-1} = \sum_{k=1}^{r} \beta_k A_k. \tag{4.55}$$

After estimating the parameters $\beta_1, \ldots, \beta_r$, we obtain a sparse estimate of the precision matrix of the low frequency wavelets. By inverting and transforming back to the physical space, we obtain a covariance matrix estimate adjusted for noise. The coefficients of scaling functions represent the coarse information in the data and so their covariance structure is strong. By contrast, the coefficients of high frequency wavelets representing the finer "details" have a much smaller variance and negligible correlation structure; we propose to neglect them. This wavelet compression is similar to denoising in image analysis, where the neglected coefficients usually correspond to the bands of higher frequency wavelets.

The particular data used in this study contains model fields of a control variable called *unbalanced temperature*[1]. To simplify the computations, we selected a domain with dimension 128×128. The data consists of 480 temperature fields - model fields in 12 consequential days and two different day-times (00 and 12 UTC) and we have an ensemble of 20 members for each time. For covariance modelling, we use departures from ensemble mean in the corresponding time. As it is a common practice, they are considered to be nearly independent. The field of sample variances in the chosen domain is given in Figure 4.3.

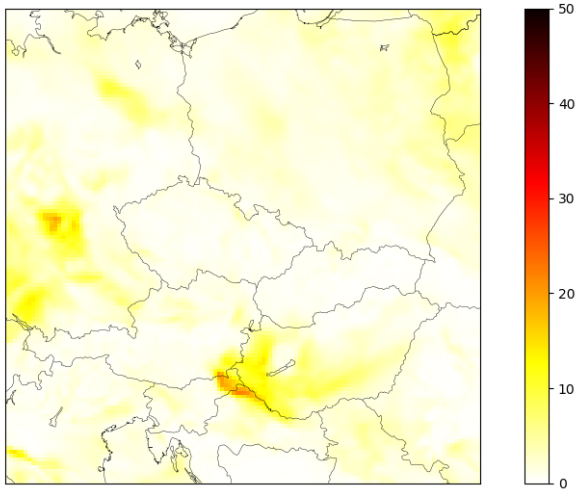


Figure 4.3: Real temperature data: Sample variances of unbalanced temperature computed from 480 temperature fields.

Since the domain is two-dimensional, the wavelet decomposition was made in each dimension separately, which results in four sub-bands of basis functions: A (scaling functions in both direction), H (scaling functions horizontally, wavelets vertically), V (scaling functions vertically, wavelets horizontally) and D (wavelets

[1]More precisely, potential temperature used in algorithms of data assimilation in numeric weather prediction. The data come from the Global Ensemble Forecast System of NCEP, USA, downscaled by the WRF model of NCAR, USA. We selected the lowest vertical level, 15 m above ground.

in both directions). We use Daubechies wavelets DB2 up to level 4. In Figure 4.4, it can be seen that only the level 4 coefficients in sub-bands A4, V4, H4 (three blocks of size 64×64) have significant variances. Therefore, the level 1, 2 and 3 covariance coefficients are discarded as well as the D sub-band in level 4. The sample covariance matrix of the remaining 192 coefficients is displayed in Figure 4.5. As seen from this figure, the major part of variance is concentrated in the "compressed image" A4. A much smaller part of the variance is in the sub-bands H4 and V4 (compressed in one directions and wavelet transformed in another). Since the variance in the remaining sub-bands (D4, complete levels 1,2,3) is insignificant, our model for the rest of the coefficients is zero.

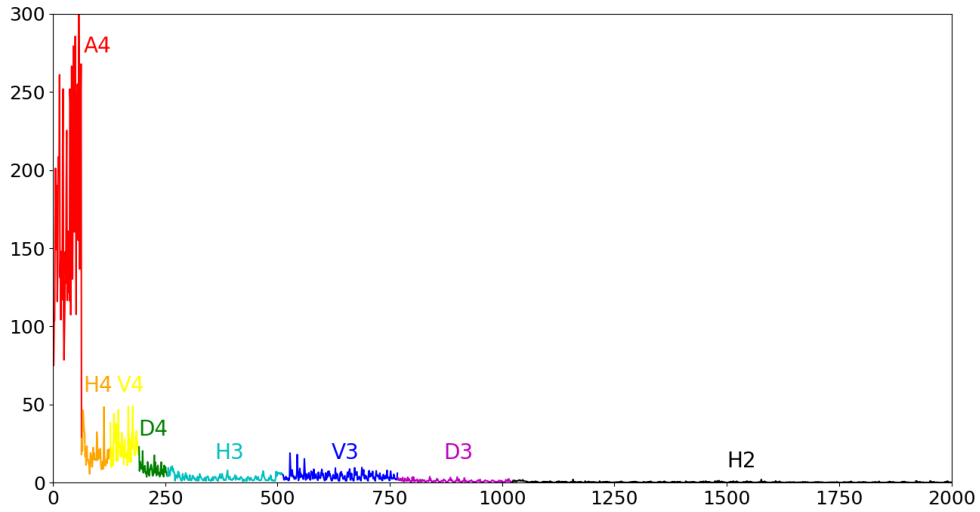


Figure 4.4: Real temperature data: Sample variances of wavelet coefficients corresponding to basis functions in different sub-bands. Based on 480 temperature fields. A4 represents the coarse informations and corresponds to scaling functions in both directions. We keep only coefficients from A4, H4 and V4. Coefficients from the remaining sub-bands were neglected.

The covariance structure in the wavelet space is different for particular sub-bands and our model for the precision matrix has to reflect that. For the A4 sub-band, we assume each gridpoint to have 12 neighbours, which results in a structure of 13 diagonals displayed in Figure 2.1c.

In Section 4.7, we used "elementary" design matrices with one or two ones only and zero otherwise. Here we use more restrictive choice of design matrices, which results in further reduction of the number of parameters and spatial smoothing of covariances. We divide each subdiagonal into sections corresponding to rows of the field of wavelet coefficients. Each section is modelled as a linear combination of B-spline basis along each section (the basis is displayed in Figure 4.6). Hence, every design matrix A_k in (4.55) consists of one B-spline on the subdiagonal section. For wavelet detail sub-bands H4 and V4, we use design matrices formed by constant diagonal segments corresponding to simple stencils of two neighbours. For the H4 sub-band, these neighbours are in horizontal direction, and for the V4 sub-band in the vertical direction. In total, we have $r = 201$ design matrices. The vector of parameters $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{200})^\top$ consists of coefficients associated with the B-splines for the precision matrix in the A4 sub-band and coefficients

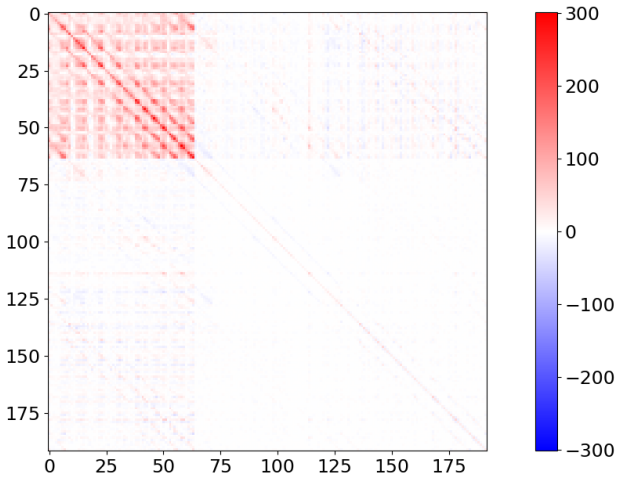


Figure 4.5: Real temperature data: Sample covariance matrix of coefficients corresponding to basis functions from A4, H4 and V4 sub-band. Based on 480 temperature fields.

associated with constant subdiagonals for the rest of level 4. Parameters were estimated by the maximum likelihood method (Ueno and Tsuchiya [68]) and the score matching method. Comparison of the coefficients estimates can be found in Figure 4.7. The resulting regularized precision matrix and regularized correlation matrix are depicted in Figure 4.8. Similarly as in the previous section, the estimates by SME and MLE do not show substantial differences.

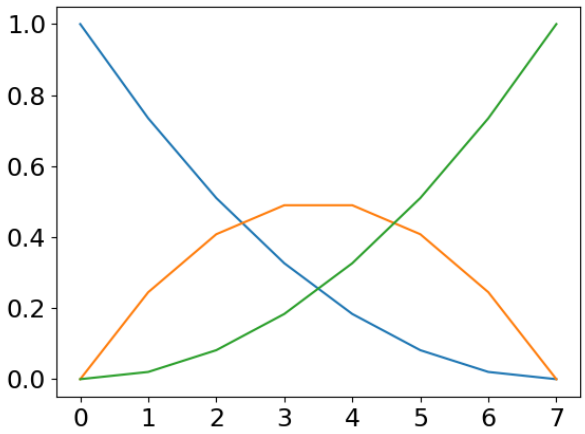


Figure 4.6: The B-spline basis for setting the design matrices intended for modelling the part of precision matrix corresponding to the A4 sub-band.

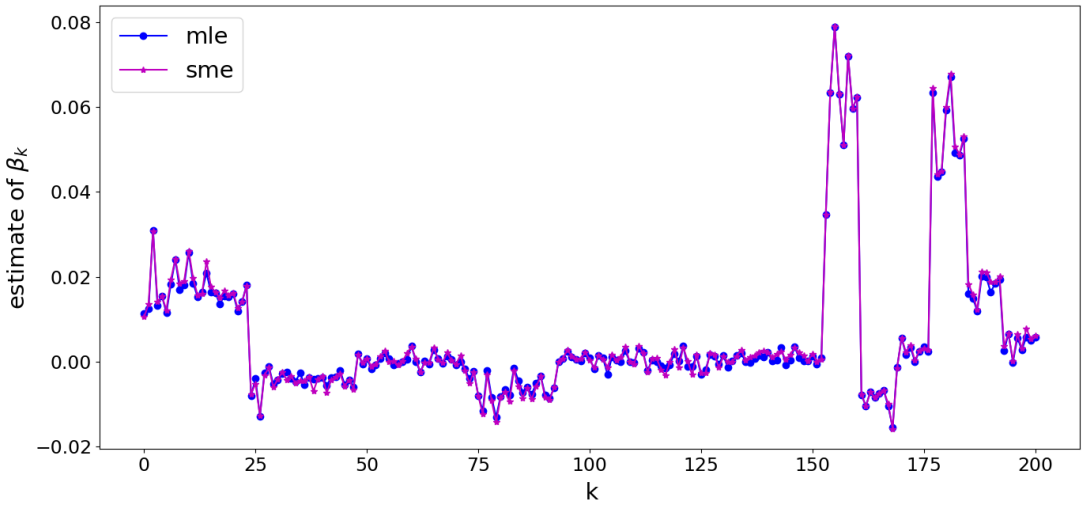
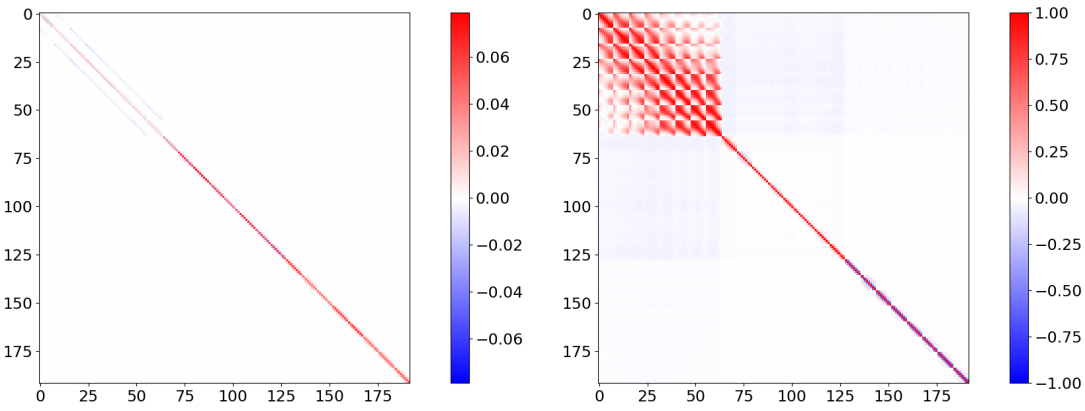
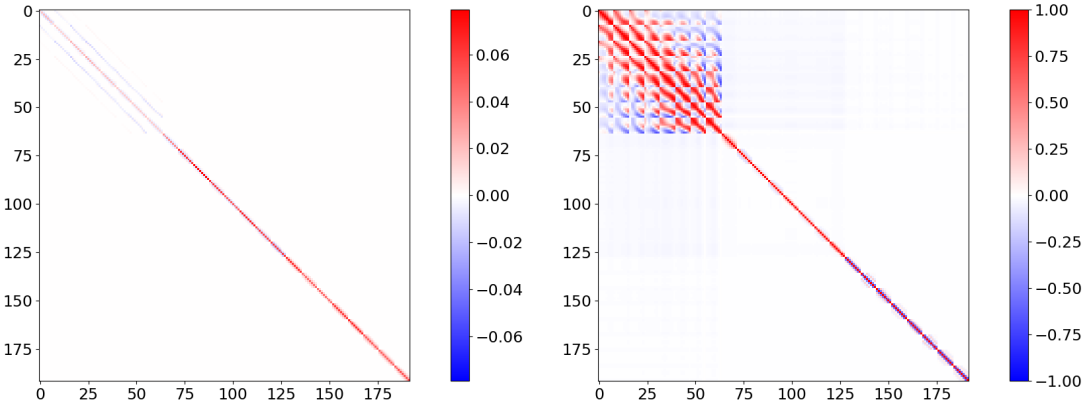


Figure 4.7: Real temperature data: Comparison of the maximum likelihood and score matching estimates of parameters $\beta_0, \ldots, \beta_{200}$ of the model $\Sigma^{-1} = \sum_{k=0}^{200} \beta_k A_k$ for the precision matrix of wavelet coefficients. Based on 480 temperature fields.



(a) MLE of the precision matrix.

(b) Correlation matrix associated to (a).



(c) SME of the precision matrix.

(d) Correlation matrix associated to (c).

Figure 4.8: Real temperature data: Estimates of the correlation and precision matrix of wavelet transform coefficients. Based on 480 temperature fields. Parameters estimated by the maximum likelihood and the score matching method.

5. Hierarchical structure of asymptotic variance of nested M-estimators

This chapter is concerned with asymptotic normality of M-estimators. Namely, it is proved that the asymptotic variance of nested M-estimators follow similar hierarchical structure as the MLE in Section 3.2. In particular, this result is applied to MLE and SME, which are special cases of M-estimators.

5.1 A brief introduction to M-estimators

We start with an introduction to M-estimators, based on Van der Vaart [69].

Consider a random sample $\mathbb{X}_N = \{\boldsymbol{X}_1, \dots, \boldsymbol{X}_N\}$ from a distribution depending on a parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^s$. *M-estimator* $\hat{\boldsymbol{\theta}}_N$ is defined as the maximizing value of the criterion function

$$\mathcal{M}_N(\boldsymbol{\theta}|\mathbb{X}_N) = \frac{1}{N}\sum_{i=1}^{N} m(\boldsymbol{X}_i, \boldsymbol{\theta}), \tag{5.1}$$

over $\boldsymbol{\Theta}$, where $m(\cdot, \boldsymbol{\theta}) : \mathbb{R}^n \to \mathbb{R}$ are known functions. The maximum is often sought by setting the derivative $\nabla_{\boldsymbol{\theta}}\mathcal{M}_N(\boldsymbol{\theta}|\mathbb{X}_N)$ equal to zero, i.e. the estimator $\hat{\boldsymbol{\theta}}_N$ satisfies

$$\nabla_{\boldsymbol{\theta}}\mathcal{M}_N(\hat{\boldsymbol{\theta}}_N|\mathbb{X}_N) = \mathbf{0}^\top, \tag{5.2}$$

provided that the derivatives exist. In order to simplify the notation, denote

$$\boldsymbol{\psi}(\boldsymbol{X}, \boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}^\top m(\boldsymbol{X}, \boldsymbol{\theta}), \tag{5.3}$$

which is a vector-valued function. Estimators satisfying systems of estimating equations of the type (5.2) are sometimes also called *Z-estimators*. Under mild conditions on $\mathcal{M}_N$, resp. $\boldsymbol{\psi}_N$, which are specified in Van der Vaart [69], M-estimator $\hat{\boldsymbol{\theta}}_N$ is consistent for $\boldsymbol{\theta}$. In the cases of our interest (MLE and SME), we have consistency from other arguments. Moreover, it can be proved that the M-estimator asymptotically follows the normal distribution:

Theorem 22 (Van der Vaart [69], Theorem 5.41)**.** *Let $\boldsymbol{X}$ be a random vector from the distribution $P_{\boldsymbol{\theta}_0}$ depending on a parameter $\boldsymbol{\theta}_0$ from an open subset $\boldsymbol{\Theta}$ of $\mathbb{R}^s$. For $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, let $\boldsymbol{\theta} \mapsto \boldsymbol{\psi}(\boldsymbol{x}, \boldsymbol{\theta})$ be twice continuously differentiable for every $\boldsymbol{x}$. Suppose that $\mathsf{E}\boldsymbol{\psi}(\boldsymbol{X}, \boldsymbol{\theta}_0) = \mathbf{0}$, $\mathsf{E}\|\boldsymbol{\psi}(\boldsymbol{X}, \boldsymbol{\theta}_0)\|_s^2 < \infty$ and that the expectation of the Jacobian matrix $\mathsf{E}\,J_{\boldsymbol{\theta}}(\boldsymbol{\psi}(\boldsymbol{X}, \boldsymbol{\theta}))$ exists and is non-singular at $\boldsymbol{\theta}_0$. Further assume that the second-order partial derivatives of $\boldsymbol{\psi}(\boldsymbol{x}, \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ are dominated by a fixed integrable function $d(\boldsymbol{x})$ for every $\boldsymbol{\theta}$ in a neighbourhood of $\boldsymbol{\theta}_0$. For each $N \in \mathbb{N}$ and a given sample $\boldsymbol{X}_1, \dots, \boldsymbol{X}_N$ from $P_{\boldsymbol{\theta}_0}$, suppose that $\hat{\boldsymbol{\theta}}_N$ satisfying $\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\psi}(\boldsymbol{X}_i, \hat{\boldsymbol{\theta}}_N) = \mathbf{0}$ is a consistent estimator of $\boldsymbol{\theta}_0$. Then,*

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0) \xrightarrow[N\to\infty]{d} \mathcal{N}_n(\mathbf{0}, C_{\boldsymbol{\theta}_0}),$$

where

$$C_{\boldsymbol{\theta}_0} = (\mathsf{E}\,J_{\boldsymbol{\theta}}(\boldsymbol{\psi}(\boldsymbol{X}, \boldsymbol{\theta}_0)))^{-1}\,\mathsf{E}\left[\boldsymbol{\psi}(\boldsymbol{X}, \boldsymbol{\theta}_0)(\boldsymbol{\psi}(\boldsymbol{X}, \boldsymbol{\theta}_0))^\top\right](\mathsf{E}\,J_{\boldsymbol{\theta}}(\boldsymbol{\psi}(\boldsymbol{X}, \boldsymbol{\theta}_0)))^{-1}. \tag{5.4}$$

The inverse of matrix (5.4) is sometimes called the Godambe information matrix because it plays the role of the Fisher information matrix for more general estimators and V. P. Godambe initiated the theory of unbiased estimating equations. It can be said that (5.4) is the inverse Godambe information.

5.2 Comparison of asymptotic variances of nested estimators

First, we will express the covariance matrix (5.4) in terms of function m. The column vector $\boldsymbol{\psi}(\boldsymbol{X},\boldsymbol{\theta}) = (\psi_1, \psi_2, \ldots, \psi_s)^\top$ in (5.3) has entries $\psi_i = \frac{\partial m(\boldsymbol{X},\boldsymbol{\theta})}{\partial \theta_i}$, $i = 1, 2, \ldots, s$, therefore

$$\boldsymbol{\psi}(\boldsymbol{X},\boldsymbol{\theta})(\boldsymbol{\psi}(\boldsymbol{X},\boldsymbol{\theta}))^\top = \nabla_{\boldsymbol{\theta}}^\top m(\boldsymbol{X},\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} m(\boldsymbol{X},\boldsymbol{\theta}) = \left[\frac{\partial m}{\partial \theta_i}\frac{\partial m}{\partial \theta_j}\right]_{i,j=1}^{s}.$$

The Jacobian matrix $J_{\boldsymbol{\theta}}(\boldsymbol{\psi}(\boldsymbol{X},\boldsymbol{\theta}))$ in (5.4) is a $s \times s$ matrix of the form

$$J_{\boldsymbol{\theta}}(\boldsymbol{\psi}(\boldsymbol{X},\boldsymbol{\theta})) = J_{\boldsymbol{\theta}}\left(\nabla_{\boldsymbol{\theta}}^\top m(\boldsymbol{X},\boldsymbol{\theta})\right) = \left[\frac{\partial^2 m(\boldsymbol{X},\boldsymbol{\theta})}{\partial \theta_j \partial \theta_i}\right]_{i,j=1}^{s} = H_{\boldsymbol{\theta}}(m(\boldsymbol{X},\boldsymbol{\theta})),$$

where $H_{\boldsymbol{\theta}}$ stands for the Hessian with respect to $\boldsymbol{\theta}$. Hence, the asymptotic covariance matrix (5.4) of an M-estimator is equal to

$$C_{\boldsymbol{\theta}_0} = (\mathsf{E}\, H_{\boldsymbol{\theta}}(m(\boldsymbol{X},\boldsymbol{\theta}_0)))^{-1}\, \mathsf{E}\left[\nabla_{\boldsymbol{\theta}}^\top m(\boldsymbol{X},\boldsymbol{\theta}_0) \nabla_{\boldsymbol{\theta}} m(\boldsymbol{X},\boldsymbol{\theta}_0)\right] (\mathsf{E}\, H_{\boldsymbol{\theta}}(m(\boldsymbol{X},\boldsymbol{\theta}_0)))^{-1}. \tag{5.5}$$

Denote by A the negative expectation of the Hessian matrix, i.e.,

$$A = -\,\mathsf{E}\, H_{\boldsymbol{\theta}}(m(\boldsymbol{X},\boldsymbol{\theta})). \tag{5.6}$$

If m is concave, then A is positive semidefinite, which will be assumed. Because A is symmetric,

$$\sqrt{N} A^{1/2}(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0) \xrightarrow[N\to\infty]{d} \mathcal{N}_n\left(\mathbf{0}, A^{1/2} C_{\boldsymbol{\theta}_0} A^{1/2}\right). \tag{5.7}$$

In order to compare precision of estimators based on nested parametrizations, we are interested in comparing their asymptotic variances as in Section 3.2. Here, the comparison is in terms of the sum of variances $\mathrm{tr}(A^{1/2} C_{\boldsymbol{\theta}_0} A^{1/2})$ of the asymptotic distribution (5.7). Similarly as in Section 3.2, suppose that the true parameter $\boldsymbol{\theta}_0$ lies in a subspace $\boldsymbol{\Phi}$ of $\boldsymbol{\Theta}$, which is parametrized by $r \le s$ parameters $(\varphi_1, \ldots, \varphi_r)^\top = \boldsymbol{\varphi}$, however the quantity of interest is the original parameter $\boldsymbol{\theta}$. Assume that $\boldsymbol{\theta}_0 = \boldsymbol{\theta}(\boldsymbol{\varphi}_0)$. The estimator $\boldsymbol{\theta}(\hat{\boldsymbol{\varphi}}_N)$, resulting from substituting the M-estimate $\hat{\boldsymbol{\varphi}}_N$ into the function $\boldsymbol{\theta}(\boldsymbol{\varphi})$, has the asymptotic covariance matrix $C_{\boldsymbol{\theta}(\boldsymbol{\varphi}_0)}$, which can be compared with $C_{\boldsymbol{\theta}_0}$. The next theorem shows that the asymptotic distribution of estimator $\boldsymbol{\theta}(\hat{\boldsymbol{\varphi}}_N)$ based on the smaller parametrization $\boldsymbol{\varphi}$ has total variance $\mathrm{tr}(A^{1/2} C_{\boldsymbol{\theta}(\boldsymbol{\varphi}_0)} A^{1/2})$ that is not larger than that of $\hat{\boldsymbol{\theta}}_N$. This corresponds to the comparison of asymptotic covariance matrices in the basis of eigenvectors of the matrix A scaled by the square roots of its eigenvectors.

Theorem 23. *Let $\boldsymbol{X}$ be a random vector with distribution $P_{\boldsymbol{\theta}_0}$, where $\boldsymbol{\theta}_0$ is a parameter belonging to an open parameter set $\boldsymbol{\Theta}$. Assume $\boldsymbol{\theta}_0$ being estimated by maximizing the criterion function (5.1) with $\boldsymbol{\psi}(\boldsymbol{X}, \boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}^{\top} m(\boldsymbol{X}, \boldsymbol{\theta})$ satisfying all the assumptions of Theorem 22. Denote $A = -\mathsf{E}\, H_{\boldsymbol{\theta}}(m(\boldsymbol{X}, \boldsymbol{\theta}))$. Suppose $\boldsymbol{\varphi} \mapsto \boldsymbol{\theta}(\boldsymbol{\varphi})$ is a one-to-one map from $\boldsymbol{\Phi} \subset \mathbb{R}^r$ to $\boldsymbol{\Theta}$ and continuously differentiable with $J_{\boldsymbol{\varphi}}(\boldsymbol{\theta}(\boldsymbol{\varphi}))$ that is non-singular for all $\boldsymbol{\varphi} \in \boldsymbol{\Phi}$. Assume $\boldsymbol{\theta}_0 = \boldsymbol{\theta}(\boldsymbol{\varphi}_0)$ with $\boldsymbol{\varphi}_0$ in the interior of $\boldsymbol{\Phi}$. Then*

$$\operatorname{tr}\left(A^{1/2} C_{\boldsymbol{\theta}(\boldsymbol{\varphi}_0)} A^{1/2}\right) \leq \operatorname{tr}\left(A^{1/2} C_{\boldsymbol{\theta}_0} A^{1/2}\right). \tag{5.8}$$

Proof. The estimator based on the parametrization $\boldsymbol{\theta}$ has the asymptotic covariance matrix (5.5). Analogically, the estimator of the submodel $\boldsymbol{\varphi}$ has the asymptotic covariance matrix

$$C_{\boldsymbol{\varphi}_0} = (\mathsf{E}\, H_{\boldsymbol{\varphi}}(m(\boldsymbol{X}, \boldsymbol{\varphi}_0)))^{-1} \mathsf{E}\left[\nabla_{\boldsymbol{\varphi}}^{\top} m(\boldsymbol{X}, \boldsymbol{\varphi}_0) \nabla_{\boldsymbol{\varphi}} m(\boldsymbol{X}, \boldsymbol{\varphi}_0)\right] (\mathsf{E}\, H_{\boldsymbol{\varphi}}(m(\boldsymbol{X}, \boldsymbol{\varphi}_0)))^{-1}. \tag{5.9}$$

From the chain rule, we obtain

$$\frac{\partial}{\partial \varphi_i} m(\boldsymbol{X}, \boldsymbol{\varphi}) = \frac{\partial}{\partial \varphi_i} m(\boldsymbol{X}, \boldsymbol{\theta}(\boldsymbol{\varphi})) = \sum_{k=1}^{s} \frac{\partial m(\boldsymbol{X}, \boldsymbol{\theta})}{\partial \theta_k} \frac{\partial \theta_k}{\partial \varphi_i},$$

which has the vector form

$$\nabla_{\boldsymbol{\varphi}} m(\boldsymbol{X}, \boldsymbol{\varphi}) = \nabla_{\boldsymbol{\theta}} m(\boldsymbol{X}, \boldsymbol{\theta}) J_{\boldsymbol{\varphi}}(\boldsymbol{\theta}),$$

where $J_{\boldsymbol{\varphi}}(\boldsymbol{\theta})$ is the $s \times r$ Jacobian matrix with entries $\frac{\partial \theta_k}{\partial \varphi_i}$. Now, we express the expected Hessian,

$$\mathsf{E}\, H_{\boldsymbol{\varphi}}(m(\boldsymbol{X}, \boldsymbol{\varphi})) = \mathsf{E}\left[\frac{\partial^2 m(\boldsymbol{X}, \boldsymbol{\varphi})}{\partial \varphi_i \partial \varphi_j}\right]_{i,j=1}^{r},$$

in (5.9) by using the chain rule as

$$\begin{aligned}
\mathsf{E}\, \frac{\partial^2 m(\boldsymbol{X}, \boldsymbol{\varphi})}{\partial \varphi_i \partial \varphi_j} &= \mathsf{E}\, \frac{\partial}{\partial \varphi_i}\left(\frac{\partial m(\boldsymbol{X}, \boldsymbol{\varphi})}{\partial \varphi_j}\right) = \mathsf{E}\, \frac{\partial}{\partial \varphi_i}\left(\sum_{k=1}^{s} \frac{\partial \theta_k}{\partial \varphi_j} \frac{\partial m(\boldsymbol{X}, \boldsymbol{\theta})}{\partial \theta_k}\right) \\
&= \sum_{k=1}^{s}\left(\frac{\partial}{\partial \varphi_i}\left(\frac{\partial \theta_k}{\partial \varphi_j}\right)\right) \mathsf{E}\left(\frac{\partial m(\boldsymbol{X}, \boldsymbol{\theta})}{\partial \theta_k}\right) + \sum_{k=1}^{s} \frac{\partial \theta_k}{\partial \varphi_j} \mathsf{E}\left(\frac{\partial}{\partial \varphi_i}\left(\frac{\partial m(\boldsymbol{X}, \boldsymbol{\theta})}{\partial \theta_k}\right)\right).
\end{aligned}$$

Evaluating the above expression at $\boldsymbol{\varphi}_0$, we obtain

$$\begin{aligned}
\mathsf{E}\, \frac{\partial^2 m(\boldsymbol{X}, \boldsymbol{\varphi})}{\partial \varphi_i \partial \varphi_j}\bigg|_{\boldsymbol{\varphi}=\boldsymbol{\varphi}_0} &= \sum_{k=1}^{s}\sum_{l=1}^{s} \frac{\partial \theta_k}{\partial \varphi_j} \mathsf{E}\left(\frac{\partial^2 m(\boldsymbol{X}, \boldsymbol{\theta})}{\partial \theta_k \partial \theta_l}\right) \frac{\partial \theta_l}{\partial \varphi_i}\bigg|_{\boldsymbol{\varphi}=\boldsymbol{\varphi}_0} \\
&= \left[J_{\boldsymbol{\varphi}}(\boldsymbol{\theta})^{\top}\, \mathsf{E}\, H_{\boldsymbol{\theta}}(m(\boldsymbol{X}, \boldsymbol{\theta})) J_{\boldsymbol{\varphi}}(\boldsymbol{\theta})\Big|_{\boldsymbol{\varphi}=\boldsymbol{\varphi}_0}\right]_{ij},
\end{aligned}$$

because $\mathsf{E}\, \nabla_{\boldsymbol{\theta}}^{\top} m(\boldsymbol{X}, \boldsymbol{\theta}) = \mathsf{E}\, \boldsymbol{\psi}(\boldsymbol{X}, \boldsymbol{\theta})$ is zero at $\boldsymbol{\theta}_0 = \boldsymbol{\theta}(\boldsymbol{\varphi}_0)$ by assumption. Thus, we have shown that there is the following equality:

$$\mathsf{E}\, H_{\boldsymbol{\varphi}}(m(\boldsymbol{X}, \boldsymbol{\varphi}_0)) = J_{\boldsymbol{\varphi}}(\boldsymbol{\theta}_0)^{\top}\, \mathsf{E}\, H_{\boldsymbol{\theta}}(m(\boldsymbol{X}, \boldsymbol{\theta}_0)) J_{\boldsymbol{\varphi}}(\boldsymbol{\theta}_0).$$

It follows that the asymptotic covariance matrix (5.9) of the submodel $\boldsymbol{\varphi}$ is

$$C_{\boldsymbol{\varphi}_0} = \left(J_{\boldsymbol{\varphi}}(\boldsymbol{\theta}_0)^\top \mathsf{E}\, H_{\boldsymbol{\theta}}(m(\boldsymbol{X},\boldsymbol{\theta}_0))J_{\boldsymbol{\varphi}}(\boldsymbol{\theta}_0)\right)^{-1} J_{\boldsymbol{\varphi}}(\boldsymbol{\theta}_0)^\top \mathsf{E}\left(\nabla_{\boldsymbol{\theta}}^\top m(\boldsymbol{X},\boldsymbol{\theta}_0)\nabla_{\boldsymbol{\theta}} m(\boldsymbol{X},\boldsymbol{\theta}_0)\right)\cdot \\ \cdot J_{\boldsymbol{\varphi}}(\boldsymbol{\theta}_0)\left(J_{\boldsymbol{\varphi}}(\boldsymbol{\theta}_0)^\top \mathsf{E}\, H_{\boldsymbol{\theta}}(m(\boldsymbol{X},\boldsymbol{\theta}_0))J_{\boldsymbol{\varphi}}(\boldsymbol{\theta}_0)\right)^{-1}.$$

The delta method (Lehmann and Romano [44, Theorem 11.2.14]) provides the following asymptotic covariance matrix of the estimator $\boldsymbol{\theta}(\hat{\boldsymbol{\varphi}}_N)$:

$$\begin{aligned} C_{\boldsymbol{\theta}(\boldsymbol{\varphi}_0)} &= J_{\boldsymbol{\varphi}}(\boldsymbol{\theta}_0)\, C_{\boldsymbol{\varphi}_0} J_{\boldsymbol{\varphi}}(\boldsymbol{\theta}_0)^\top \\ &= J_{\boldsymbol{\varphi}}(\boldsymbol{\theta}_0)\left(J_{\boldsymbol{\varphi}}(\boldsymbol{\theta}_0)^\top \mathsf{E}\, H_{\boldsymbol{\theta}}(m(\boldsymbol{X},\boldsymbol{\theta}_0))J_{\boldsymbol{\varphi}}(\boldsymbol{\theta}_0)\right)^{-1} J_{\boldsymbol{\varphi}}(\boldsymbol{\theta}_0)^\top \mathsf{E}\left(\nabla_{\boldsymbol{\theta}}^\top m(\boldsymbol{X},\boldsymbol{\theta}_0)\nabla_{\boldsymbol{\theta}} m(\boldsymbol{X},\boldsymbol{\theta}_0)\right) \\ &\quad \cdot J_{\boldsymbol{\varphi}}(\boldsymbol{\theta}_0)\left(J_{\boldsymbol{\varphi}}(\boldsymbol{\theta}_0)^\top \mathsf{E}\, H_{\boldsymbol{\theta}}(m(\boldsymbol{X},\boldsymbol{\theta}_0))J_{\boldsymbol{\varphi}}(\boldsymbol{\theta}_0)\right)^{-1} J_{\boldsymbol{\varphi}}(\boldsymbol{\theta}_0)^\top. \end{aligned}$$

In order to compare this with (5.5), denote

$$A = -\,\mathsf{E}\, H_{\boldsymbol{\theta}}(m(\boldsymbol{X},\boldsymbol{\theta}_0))$$

as in (5.6) and

$$\begin{aligned} B &= J_{\boldsymbol{\varphi}}(\boldsymbol{\theta}_0)\,, \\ C &= \mathsf{E}\left(\nabla_{\boldsymbol{\theta}}^\top m(\boldsymbol{X},\boldsymbol{\theta}_0)\nabla_{\boldsymbol{\theta}} m(\boldsymbol{X},\boldsymbol{\theta}_0)\right). \end{aligned}$$

Then, the covariance matrices can be written as

$$\begin{aligned} C_{\boldsymbol{\theta}_0} &= A^{-1}CA^{-1} \\ C_{\boldsymbol{\theta}(\boldsymbol{\varphi}_0)} &= B(B^\top AB)^{-1}B^\top CB(B^\top AB)^{-1}B^\top, \end{aligned}$$

and after multiplying both these matrices by $A^{1/2}$ from the right and left, we have

$$A^{1/2}C_{\boldsymbol{\theta}_0}A^{1/2} = D \tag{5.10}$$

$$A^{1/2}C_{\boldsymbol{\theta}(\boldsymbol{\varphi}_0)}A^{1/2} = PDP, \tag{5.11}$$

where $D = A^{-1/2}CA^{-1/2}$ is positive semidefinite and

$$P = A^{1/2}B(B^\top AB)^{-1}B^\top A^{1/2}$$

is symmetric and idempotent, and hence, an orthogonal projection.

If P and D have the same eigenvectors, i.e. if they commute, then $D \geq PDP$ and hence, $C_{\boldsymbol{\theta}_0} \geq C_{\boldsymbol{\theta}(\boldsymbol{\varphi}_0)}$ holds in the sense of comparison of symmetric positive semidefinite matrices, i.e., that $C_{\boldsymbol{\theta}_0} - C_{\boldsymbol{\theta}(\boldsymbol{\varphi}_0)}$ is positive semidefinite.

In general, we can choose an orthonormal basis, whose vectors form columns of an orthonormal matrix U, such that

$$U^\top PU = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}$$

and

$$U^\top DU = \begin{bmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{bmatrix}.$$

Then

$$U^\top PDPU = \begin{bmatrix} D_{11} & 0 \\ 0 & 0 \end{bmatrix}$$

and, consequently,

$$\operatorname{tr} D = \operatorname{tr}(U^\top DU) = \operatorname{tr} D_{11} + \operatorname{tr} D_{22} \geq \operatorname{tr} D_{11} = \operatorname{tr}(U^\top PDPU) = \operatorname{tr}(PDP)$$

because D is symmetric positive semidefinite so it has non-negative diagonal in any basis. Hence, by substituting from (5.10) and (5.11),

$$\operatorname{tr}\left(A^{1/2} C_{\boldsymbol{\theta}_0} A^{1/2}\right) \geq \operatorname{tr}\left(A^{1/2} C_{\boldsymbol{\theta}(\boldsymbol{\varphi}_0)} A^{1/2}\right).$$

□

5.3 Application to SME for normal distribution

Score matching estimators are a special case of M-estimators with

$$\mathcal{M}_N(\boldsymbol{\theta}|\mathbb{X}_N) = \frac{1}{N}\sum_{i=1}^{N} m(\boldsymbol{X}_i, \boldsymbol{\theta}) = -\mathcal{S}_N(\boldsymbol{\theta}|\mathbb{X}_N),$$

where

$$m(\boldsymbol{X}, \boldsymbol{\theta}) = -\left\|\nabla_{\boldsymbol{x}}^\top \log q(\boldsymbol{X}|\boldsymbol{\theta}) - \nabla_{\boldsymbol{x}}^\top \log p(\boldsymbol{X})\right\|_n^2. \tag{5.12}$$

When the true distribution $p(\boldsymbol{x})$ coincides with $f(\boldsymbol{x}|\boldsymbol{\theta}_0)$ for a unique $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$, it can be easily seen from (5.12) that $\boldsymbol{\theta}_0$ is the maximizing value of $\mathsf{E}\, m(\boldsymbol{X}, \boldsymbol{\theta})$. As we have that $\boldsymbol{\psi}(\boldsymbol{X}, \boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}^\top m(\boldsymbol{X}, \boldsymbol{\theta})$, it follows that $\mathsf{E}\, \boldsymbol{\psi}(\boldsymbol{X}, \boldsymbol{\theta}_0) = 0$.

When the model density $f(\boldsymbol{x}|\boldsymbol{\theta})$ belongs to the family of exponential distributions (4.8), $\mathcal{S}_N(\boldsymbol{\theta}|\mathbb{X}_N)$ is given in (4.14) and the function $\boldsymbol{\psi}(\boldsymbol{X}, \boldsymbol{\theta})$ is of the form

$$\boldsymbol{\psi}(\boldsymbol{X}, \boldsymbol{\theta}) = -D^*(\boldsymbol{X}) D(\boldsymbol{X}) \boldsymbol{\theta} - D^*(\boldsymbol{X}) \nabla_{\boldsymbol{x}}^\top b(\boldsymbol{X}) - \Delta_{\boldsymbol{x}} T(\boldsymbol{X}) \tag{5.13}$$

given in (4.16). Evidently, the function (5.13) is twice continuously differentiable in $\boldsymbol{\theta}$ for all $\boldsymbol{X}$. The second order partial derivatives of (5.13) with respect to $\boldsymbol{\theta}$ are zero, and therefore, they can be dominated by any constant function $d(\boldsymbol{X}) \equiv d \in (0, \infty)$. Since $\mathsf{E}\, d = d$, this dominating function is integrable.

In the case of the normal distribution $\mathcal{N}_n(\boldsymbol{\mu}_0, \Sigma_0)$ from Theorem 16,

$$\boldsymbol{\psi}(\boldsymbol{X}, \boldsymbol{\theta}) = \begin{bmatrix} -I_n & [A_1\boldsymbol{X}, \ldots, A_r\boldsymbol{X}] \\ [A_1\boldsymbol{X}, \ldots, A_r\boldsymbol{X}]^\top & -[\boldsymbol{X}^\top A_k A_l \boldsymbol{X}]_{k,l=1}^r \end{bmatrix} \boldsymbol{\theta} + \begin{bmatrix} \mathbf{0} \\ (\operatorname{tr}(A_1), \ldots, \operatorname{tr}(A_r))^\top \end{bmatrix}, \tag{5.14}$$

which results by substituting (4.45) and (4.46) in (5.13). The expectation of its Jacobian matrix,

$$\mathsf{E}\, J_{\boldsymbol{\theta}}(\boldsymbol{\psi}(\boldsymbol{X}, \boldsymbol{\theta})) = \mathsf{E} \begin{bmatrix} -I_n & [A_1\boldsymbol{X}, \ldots, A_r\boldsymbol{X}] \\ [A_1\boldsymbol{X}, \ldots, A_r\boldsymbol{X}]^\top & -[\boldsymbol{X}^\top A_k A_l \boldsymbol{X}]_{k,l=1}^r \end{bmatrix},$$

exists for all $\boldsymbol{\theta}$, and

$$\mathsf{E}\, J_{\boldsymbol{\theta}}(\boldsymbol{\psi}(\boldsymbol{X},\boldsymbol{\theta}_0)) = \begin{bmatrix} -I_n & [A_1\boldsymbol{\mu}_0,\dots,A_r\boldsymbol{\mu}_0] \\ [A_1\boldsymbol{\mu}_0,\dots,A_r\boldsymbol{\mu}_0]^\top & -\left[\mathrm{tr}\left(A_kA_l(\Sigma_0+\boldsymbol{\mu}_0\boldsymbol{\mu}_0^\top)\right)\right]_{k,l=1}^r \end{bmatrix}$$

is non-singular under the assumption that the matrix $[\mathrm{tr}(A_kA_l\Sigma_0)]_{k,l=1}^r$ is regular. This can be seen by using the formula for determinant of a block matrix,

$$\begin{aligned}\det(\mathsf{E}\, J_{\boldsymbol{\theta}}(\boldsymbol{\psi}(\boldsymbol{X},\boldsymbol{\theta}_0))) &= \det(I_n)\det\left(\left[\mathrm{tr}(A_kA_l(\Sigma_0+\boldsymbol{\mu}_0\boldsymbol{\mu}_0^\top))\right]_{k,l=1}^r - \left[\boldsymbol{\mu}_0^\top A_kA_l\boldsymbol{\mu}_0\right]_{k,l=1}^r\right)\\ &= 1\cdot\det\left([\mathrm{tr}(A_kA_l\Sigma_0)]_{k,l=1}^r\right).\end{aligned}$$

Since the normal distribution has finite moments of all orders, it follows from (5.14) that $\mathsf{E}\,\|\boldsymbol{\psi}(\boldsymbol{X},\boldsymbol{\theta}_0)\|_s^2 < \infty$.

Under the additional assumption that the matrix $[\mathrm{tr}(A_kA_l\Sigma_0)]_{k,l=1}^r$ is regular, all assumptions of Theorems 22 and 23 are satisfied and it follows that asymptotic covariance matrices of SMEs of the mean and the parameters of the linear model for precision matrix based on two nested parametrizations $\boldsymbol{\varphi}$ and $\boldsymbol{\theta}$ satisfy the hierarchical property (5.8).

5.4 Application to MLE

In this section, we show what Theorem 23 becomes for maximum likelihood estimators. Assume that the standard assumptions (A1)-(A5) listed at the beginning of Section 3.1 hold.

In the case of the maximum likelihood method, $m(\boldsymbol{X},\boldsymbol{\theta})$ is equal to the log-density and so

$$\boldsymbol{\psi}(\boldsymbol{X},\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}^\top \log f(\boldsymbol{x}|\boldsymbol{\theta}).$$

Hence, the middle term $\mathsf{E}\left[\boldsymbol{\psi}(\boldsymbol{X},\boldsymbol{\theta})\left(\boldsymbol{\psi}(\boldsymbol{X},\boldsymbol{\theta})\right)^\top\right]$ of (5.4) represents the common definition of the Fisher information matrix $\mathcal{I}_{\boldsymbol{\theta}}$. Its (i,j)-element is

$$[\mathcal{I}_{\boldsymbol{\theta}}]_{ij} = \mathsf{E}\left[\frac{\partial \log f(\boldsymbol{X}|\boldsymbol{\theta})}{\partial\theta_i}\frac{\partial \log f(\boldsymbol{X}|\boldsymbol{\theta})}{\partial\theta_j}\right] = \int_{\mathcal{X}} \frac{\partial_{\theta_i} f(\boldsymbol{x}|\boldsymbol{\theta})}{f(\boldsymbol{x}|\boldsymbol{\theta})}\frac{\partial_{\theta_j} f(\boldsymbol{x}|\boldsymbol{\theta})}{f(\boldsymbol{x}|\boldsymbol{\theta})} f(\boldsymbol{x}|\boldsymbol{\theta})d\boldsymbol{x}, \tag{5.15}$$

where $\mathcal{X} = \{\boldsymbol{x} : f(\boldsymbol{x}|\boldsymbol{\theta}) > 0\}$ and $\partial_{\theta_i} f(\boldsymbol{x}|\boldsymbol{\theta}) = \frac{\partial f(\boldsymbol{x}|\boldsymbol{\theta})}{\partial\theta_i}$. From

$$\frac{\partial^2 \log f(\boldsymbol{x}|\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j} = \partial_{\theta_i}\left(\frac{\partial_{\theta_j} f(\boldsymbol{x}|\boldsymbol{\theta})}{f(\boldsymbol{x}|\boldsymbol{\theta})}\right) = \frac{\partial^2_{\theta_i\theta_j} f(\boldsymbol{x}|\boldsymbol{\theta})}{f(\boldsymbol{x}|\boldsymbol{\theta})} - \frac{\partial_{\theta_j} f(\boldsymbol{x}|\boldsymbol{\theta})\partial_{\theta_i} f(\boldsymbol{x}|\boldsymbol{\theta})}{f(\boldsymbol{x}|\boldsymbol{\theta})^2},$$

it follows that

$$\frac{\partial_{\theta_j} f(\boldsymbol{x}|\boldsymbol{\theta})\partial_{\theta_i} f(\boldsymbol{x}|\boldsymbol{\theta})}{f(\boldsymbol{x}|\boldsymbol{\theta})^2} = \frac{\partial^2_{\theta_i\theta_j} f(\boldsymbol{x}|\boldsymbol{\theta})}{f(\boldsymbol{x}|\boldsymbol{\theta})} - \partial^2_{\theta_i\theta_j}\log f(\boldsymbol{x}|\boldsymbol{\theta}). \tag{5.16}$$

By substituting (5.16) into (5.15), we obtain

$$[\mathcal{I}_{\boldsymbol{\theta}}]_{ij} = \int_{\mathcal{X}} \partial^2_{\theta_i\theta_j} f(\boldsymbol{x}|\boldsymbol{\theta})d\boldsymbol{x} - \int_{\mathcal{X}} \left(\partial^2_{\theta_i\theta_j}\log f(\boldsymbol{x}|\boldsymbol{\theta})\right) f(\boldsymbol{x}|\boldsymbol{\theta})d\boldsymbol{x}. \tag{5.17}$$

By assumption (A5), it holds

$$\int_{\mathcal{X}} \partial^2_{\theta_i\theta_j} f(\boldsymbol{x}|\boldsymbol{\theta}) d\boldsymbol{x} = 0,$$

and hence, (5.17) turns into

$$[\mathcal{I}_{\boldsymbol{\theta}}]_{ij} = -\int_{\mathcal{X}} \left(\partial^2_{\theta_i\theta_j} \log f(\boldsymbol{x}|\boldsymbol{\theta})\right) f(\boldsymbol{x}|\boldsymbol{\theta}) d\boldsymbol{x}.$$

We thus obtain

$$\mathcal{I}_{\boldsymbol{\theta}} = -\,\mathsf{E}\, H_{\boldsymbol{\theta}}(\log f(\boldsymbol{X}|\boldsymbol{\theta})), \tag{5.18}$$

which we denoted by A in Theorem 23.

The asymptotic covariance matrix (5.5) of the MLE $\hat{\boldsymbol{\theta}}_N$ equals to

$$C_{\boldsymbol{\theta}_0} = (\mathcal{I}_{\boldsymbol{\theta}_0})^{-1} \mathcal{I}_{\boldsymbol{\theta}_0} (\mathcal{I}_{\boldsymbol{\theta}_0})^{-1} = \mathcal{I}^{-1}_{\boldsymbol{\theta}_0}, \tag{5.19}$$

which is a well-known result. Then, by using (5.18) and (5.19), we obtain

$$\operatorname{tr}\left(A^{1/2} C_{\boldsymbol{\theta}_0} A^{1/2}\right) = \operatorname{tr}\left(A C_{\boldsymbol{\theta}_0}\right) = \operatorname{tr}\left(\mathcal{I}_{\boldsymbol{\theta}_0} \mathcal{I}^{-1}_{\boldsymbol{\theta}_0}\right) = \operatorname{tr}(I_s) = s.$$

On the other hand, from (3.4) and the invariance of trace under cyclic permutation, we have that

$$\begin{aligned}\operatorname{tr}\left(A C_{\boldsymbol{\theta}(\boldsymbol{\varphi}_0)}\right) &= \operatorname{tr}\left(\mathcal{I}_{\boldsymbol{\theta}_0} J_{\boldsymbol{\varphi}}(\boldsymbol{\theta}(\boldsymbol{\varphi}_0)) \left(J_{\boldsymbol{\varphi}}(\boldsymbol{\theta}(\boldsymbol{\varphi}_0))^\top \mathcal{I}_{\boldsymbol{\theta}_0} J_{\boldsymbol{\varphi}}(\boldsymbol{\theta}(\boldsymbol{\varphi}_0))\right)^{-1} J_{\boldsymbol{\varphi}}(\boldsymbol{\theta}(\boldsymbol{\varphi}_0))^\top\right) \\ &= \operatorname{tr}(I_r) = r.\end{aligned}$$

Hence, (5.8) turns into

$$r \leq s.$$

Therefore, it simply compares dimensions of the parameter spaces. Note that in Theorem 5, we obtained a different result,

$$C_{\boldsymbol{\theta}_0} - C_{\boldsymbol{\theta}(\boldsymbol{\varphi}_0)} \geq 0,$$

and

$$\operatorname{tr}\left(C_{\boldsymbol{\theta}_0}\right) \geq \operatorname{tr}\left(C_{\boldsymbol{\theta}(\boldsymbol{\varphi}_0)}\right).$$

6. Data assimilation and ensemble Kalman filter

The purpose of this chapter is to provide a brief introduction to data assimilation problem and present one of the most famous algorithm called the Ensemble Kalman filter. The whole topic is discussed in the context of discrete-time dynamical systems, as it is the case in practical applications.

6.1 The linear data assimilation problem

The origin of data assimilation can be found in geographical sciences and the initial drivers for evolution of the field were atmospheric sciences, weather prediction and oceanography (Law et al. [40]). Nowadays, other applications are taking advantage of the methodology of data assimilation, e.g., in neuroscience, geophysical sciences, and oil industry.

In general, the data assimilation problem can be formulated for nonlinear dynamical systems with non-Gaussian perturbations. However, the theory presented in this and the following chapter will focus only on the linear and Gaussian case. This simpler formulation is also considered in practical algorithms in geophysical systems such as weather forecasting, since the general formulation would be beyond the current algorithmic and computational capability.

Consider a discrete-time stochastic linear dynamical system and observation model (Katzfuss et al. [38, eq. (6,7)])

$$\boldsymbol{X}_t = M\boldsymbol{X}_{t-1} + \boldsymbol{e}_t^{\boldsymbol{X}}, \quad t \in \mathbb{N}, \tag{6.1}$$

$$\boldsymbol{Y}_t = H\boldsymbol{X}_t + \boldsymbol{e}_t^{\boldsymbol{Y}}, \quad t \in \mathbb{N}, \tag{6.2}$$

with initial condition $\boldsymbol{X}_0 \sim \mathcal{N}_n(\boldsymbol{\mu}_0, \Sigma_0)$. Here, $\boldsymbol{X}_t$ is an unobservable system state and $\boldsymbol{Y}_t$ is its observation available up to an error $\boldsymbol{e}_t^{\boldsymbol{Y}}$. The model operator $M \in \mathbb{R}^{n\times n}$ represents the system dynamics, $H \in \mathbb{R}^{n\times m}$ is the observation operator, which selects m locations with available observations. The index $t \in \mathbb{N}$ denotes the time index. The additive random perturbations $\boldsymbol{e}_t^{\boldsymbol{X}} \sim \mathcal{N}_n(\boldsymbol{0}, Q)$ and the additive observation errors $\boldsymbol{e}_t^{\boldsymbol{Y}} \sim \mathcal{N}_m(\boldsymbol{0}, R)$ are independent mutually and also as a sequence of $t \in \mathbb{N}$.

The objective of data assimilation is to estimate the hidden system state $\boldsymbol{X}_t$ at particular time t based on the observed realizations $\boldsymbol{y}_t$ of $\boldsymbol{Y}_t$ and a prior knowledge. Mathematically, it is a problem of conditioning the random variable $\boldsymbol{X}_t$ on the observed data $\boldsymbol{y}_t$. In geosciences, the prior distribution of $\boldsymbol{X}_t$ conditioned on $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{t-1}$ is called the *forecast*, and the posterior, or filtering, distribution of $\boldsymbol{X}_t$ conditioned on $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_t$ is called the *analysis*.

For the system (6.1, 6.2), the forecast distribution is

$$\boldsymbol{X}_t^f = (\boldsymbol{X}_t | \boldsymbol{Y}_1 = \boldsymbol{y}_1, \ldots, \boldsymbol{Y}_{t-1} = \boldsymbol{y}_{t-1}) \sim \mathcal{N}_n\left(\boldsymbol{\mu}_t^f, \Sigma_t^f\right) \tag{6.3}$$

for some $\boldsymbol{\mu}_t^f$ and Σ_t^f, which results from the normality assumption on $e_t^{\boldsymbol{X}}$ and $e_t^{\boldsymbol{Y}}$ and from the invariance of normal distribution under linear transform.

Since the distribution of $\boldsymbol{Y}_t|\boldsymbol{X}_t$ is normal, then by application of the Bayes's formula (Law et al. [40, Section 1.1.4]), we obtain the filtering distribution at time t (Law et al. [40, Section 2.4 and 4.1]), as

$$\boldsymbol{X}_t^a = (\boldsymbol{X}_t|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_t) \sim \mathcal{N}_n(\boldsymbol{\mu}_t^a, \Sigma_t^a)$$

with parameters

$$\boldsymbol{\mu}_t^a = \boldsymbol{\mu}_t^f + \Sigma_t^f H^\top (H\Sigma_t^f H^\top + R)^{-1}(\boldsymbol{y}_t - H\boldsymbol{\mu}_t^f), \tag{6.4}$$

$$\Sigma_t^a = \Sigma_t^f - \Sigma_t^f H^\top (H\Sigma_t^f H^\top + R)^{-1} H\Sigma_t^f. \tag{6.5}$$

This can be rewritten using the Woodbury matrix formula (Law et al. [40, Lemma 4.4]) as

$$\boldsymbol{\mu}_t^a = \left((\Sigma_t^f)^{-1} + H^\top R^{-1} H\right)^{-1} \left((\Sigma_t^f)^{-1}\boldsymbol{\mu}_t^f + H^\top R^{-1}\boldsymbol{y}_t\right), \tag{6.6}$$

$$\Sigma_t^a = \left((\Sigma_t^f)^{-1} + H^\top R^{-1} H\right)^{-1}. \tag{6.7}$$

Applying model (6.1), we get the forecast distribution (6.3) at time $t+1$, with the parameters

$$\boldsymbol{\mu}_{t+1}^f = M\boldsymbol{\mu}_t^a + e_t^{\boldsymbol{X}}, \tag{6.8}$$

$$\Sigma_{t+1}^f = M\Sigma_t^a M^\top + Q. \tag{6.9}$$

The forecast covariance Σ_t^f is unknown, while the covariances R and Q of the observation and the model error, respectively, are assumed to be known.

The sequential algorithm given by equations (6.4, 6.5), or (6.6, 6.7), for assimilation of the data vector and (6.8, 6.9) for advancing the distribution parameters from time t to time $t+1$ is known as *Kalman filter* (Kalman [35]). The key difference between the update formulas in (6.4, 6.5) and those in (6.6, 6.7) is that in the former, matrix inversion takes place in the data space (with dimension m), while in the latter, matrix inversion takes place in the state space (with dimension n). Thus, in applications where $m << n$, the former formulation is more frequently employed. Alternatively, if the observations in the data vector are independent, it is possible to solve a system of size of ensemble Mandel et al. [51, p. 58] or assimilate them one by one (Anderson [1], Hunt et al. [32]).

6.2 Ensemble Kalman filter

In its original form, the Kalman filter is restricted to linear Gaussian problems, which makes it possible to represent the probability distributions only by their mean and covariance matrix. However, in weather prediction and similar applications, the state vector $\boldsymbol{X}_t, t \geq 0$, consists of the values of a simulation on a computational grid in a spatial domain and so its dimension is very high, often millions and more. In such a case, computing or even storing the exact covariance matrix of the system state is somewhat impractical. In addition, the model is nonlinear so the state is necessarily non-Gaussian and advancing the state covariance requires approximations.

One of the most successful data assimilation methods that addresses this problem is the Ensemble Kalman filter (EnKF) (Evensen [22]). EnKF is an approximation of the Kalman filter, in which the state probability distribution at time $t \geq 1$ is represented by a set of realizations $\boldsymbol{X}_{t1}, \ldots, \boldsymbol{X}_{tN}$. This set is called an *ensemble* instead of a sample because it does not fulfil the definition of a random sample - the vectors are usually not independent and may not be identically distributed.

Denote by i the ensemble member index and by N the ensemble size. The i-th ensemble member $\boldsymbol{X}_{ti}$ at time t is updated and advanced in time by the Kalman filter formulas (6.4) (or (6.6)) and (6.8), where the forecast covariance matrix Σ_t^f is estimated by the sample covariance matrix computed from the forecast ensemble. Specifically, the algorithm proceeds as follows:

Algorithm 1: Ensemble Kalman filter (with linear dynamics) (**EnKF**)

Initial condition: The initial ensemble $\boldsymbol{X}_{01}^a, \ldots, \boldsymbol{X}_{0N}^a$ is sampled from a given initial distribution $\mathcal{N}_n(\boldsymbol{\mu}_0, \Sigma_0)$.

for $t \geq 1$ **do**

Forecast: For all $i = 1, \ldots, N$:

$$\boldsymbol{X}_{ti}^f = M\boldsymbol{X}_{t-1,i}^a + \boldsymbol{e}_{ti}^{\boldsymbol{X}}, \text{ where } \boldsymbol{e}_{ti}^{\boldsymbol{X}} \sim \mathcal{N}_n(\boldsymbol{0}, Q)$$

Compute

$$\bar{\boldsymbol{X}}_t^f = \tfrac{1}{N}\sum_{i=1}^N \boldsymbol{X}_{ti}^f$$

$$\tilde{S}_t^f = \tfrac{1}{N-1}\sum_{i=1}^N \left(\boldsymbol{X}_{ti}^f - \bar{\boldsymbol{X}}_t^f\right)\left(\boldsymbol{X}_{ti}^f - \bar{\boldsymbol{X}}_t^f\right)^\top$$

$$\boldsymbol{Y}_{ti} \sim \mathcal{N}_m(\boldsymbol{y}_t, R),\ i = 1, \ldots, N \text{ (perturbed observations)}$$

Analysis: For all $i = 1, \ldots, N$:

$$\boldsymbol{X}_{ti}^a = \boldsymbol{X}_{ti}^f + \tilde{S}_t^f H^\top (H\tilde{S}_t^f H^\top + R)^{-1}(\boldsymbol{Y}_{ti} - H\boldsymbol{X}_{ti}^f)$$

end

The perturbed observations $\boldsymbol{Y}_{ti}, i = 1, \ldots, N$, are artificial observations found by perturbing the given observation $\boldsymbol{y}_t$ with additional noise. They are necessary in the calculation of the analysis ensemble, which otherwise has a too low variance (Burgers et al. [13]). Consequently, the analysis ensemble mean and sample covariance does not correspond to the parameters (6.4, 6.5) of the correct analysis distribution, which negatively affects the evolution of the filter.

When the state estimate is required, it can be obtained from the ensemble mean. The sample covariance matrix provides a quantification of uncertainty. Moreover, the mean and sample covariance of the analysis ensemble converge in the limit for large ensembles to the mean and covariance of the true filtering distribution in L_p for all $p \in [1, \infty)$ (Mandel et al. [52], Le Gland et al. [41]). Hence, in every time step, EnKF (with linear dynamic) provides consistent estimates of the true mean and covariance matrix of the analysis and forecast distribution.

Even though the algorithm is motivated as an approximation of the Kalman filter, which is restricted to Gaussian problems, distribution of the ensemble members is not prescribed to be Gaussian. Since the sample covariance matrix is computed from all ensemble members together, the first analysis step introduces dependence among the members and destroys their normality. Despite this fact, it was proved by Mandel et al. [52] and also by Le Gland et al. [41] that the

EnKF converges to the Kalman filter in the limit of infinite ensemble in the case of linear dynamics.

Since the EnKF does not need to maintain the state covariance matrix, it can be implemented efficiently for high-dimensional problems. Moreover, the algorithm can be used for nonlinear dynamical models and it can be modified also for nonlinear observation functions (e.g. Mandel et al. [50, p. 59]), which makes it very computationally appealing.

There exist also other algorithms that approximate the state probability distribution by means of an ensemble. Beside EnKF and its variants, which produce the analysis ensemble from the forecast ensemble and the data in a stochastic manner through the perturbed observations, an analysis ensemble with correct covariance can be formed also in a deterministic way by computing the square root of a matrix. The resulting unbiased square root filters (Livings et al. [47]) include, e.g., the ensemble transform Kalman filter (Bishop et al. [10], Hunt et al. [32]), the ensemble adjustment Kalman filter (Anderson [1]), etc.

6.3 Diagonal ensemble Kalman filter

Sometimes, the filter is performed in a space where the entries of the system state can be assumed to be independent, e.g., in the spectral or wavelet space (cf. Section 2.2.1). In that case, the covariance estimate can be improved by using only the diagonal of sample covariance (Parrish and Derber [59], Kasanický et al. [37]). The resulting algorithm is called the *diagonal EnKF* in this thesis and its algorithm is summarized below. In the next chapter, we will use this algorithm to demonstrate the importance of estimating some off-diagonal elements of the covariance matrix.

Algorithm 2: Diagonal ensemble Kalman filter (**diag EnKF**)

Initial condition: The initial ensemble $\boldsymbol{X}^a_{01}, \dots, \boldsymbol{X}^a_{0N}$ is sampled from a given initial distribution $\mathcal{N}_n(\boldsymbol{\mu}_0, \Sigma_0)$.

for $t \geq 1$ **do**

Forecast: For all $i = 1, \dots, N$:

$\boldsymbol{X}^f_{ti} = M\boldsymbol{X}^a_{t-1,i} + \boldsymbol{e}^{\boldsymbol{X}}_{ti}$, where $\boldsymbol{e}^{\boldsymbol{X}}_{ti} \sim \mathcal{N}_n(\boldsymbol{0}, Q)$

Compute

$\bar{\boldsymbol{X}}^f_t = \frac{1}{N}\sum_{i=1}^N \boldsymbol{X}^f_{ti}$

$D^f_t = \operatorname{diag}\left(\frac{1}{N-1}\sum_{i=1}^N \left(\boldsymbol{X}^f_{ti} - \bar{\boldsymbol{X}}^f_t\right)\left(\boldsymbol{X}^f_{ti} - \bar{\boldsymbol{X}}^f_t\right)^\top\right)$

$\boldsymbol{Y}_{ti} \sim \mathcal{N}_m(\boldsymbol{y}_t, R)$, $i = 1, \dots, N$ (perturbed observations)

Analysis: For all $i = 1, \dots, N$:

$\boldsymbol{X}^a_{ti} = \boldsymbol{X}^f_{ti} + D^f_t H^\top (H D^f_t H^\top + R)^{-1}(\boldsymbol{Y}_{ti} - H\boldsymbol{X}^f_{ti})$

end

7. Filtering algorithms for GMRF using score matching method

We briefly review the filter setting from Section 6.1 in order to recall the notation and terminology. Consider the state space model

$$\boldsymbol{X}_t = M\boldsymbol{X}_{t-1} + \boldsymbol{e}_t^{\boldsymbol{X}}, \quad t \in \mathbb{N}, \tag{7.1}$$

$$\boldsymbol{Y}_t = H\boldsymbol{X}_t + \boldsymbol{e}_t^{\boldsymbol{Y}}, \quad t \in \mathbb{N}, \tag{7.2}$$

where $\boldsymbol{X}_0 \sim \mathcal{N}_n(\boldsymbol{\mu}_0, \Sigma_0)$. Recall that $M \in \mathbb{R}^{n\times n}$, $H \in \mathbb{R}^{n\times m}$ and that the additive random errors $\boldsymbol{e}_t^{\boldsymbol{X}} \sim \mathcal{N}_n(\boldsymbol{0}, Q)$ and $\boldsymbol{e}_t^{\boldsymbol{Y}} \sim \mathcal{N}_m(\boldsymbol{0}, R)$ are independent mutually and also between as a sequence of $t \in \mathbb{N}$. The covariances Q and R are assumed to be known and diagonal.

The estimate of $\boldsymbol{X}_t$ given $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_t$ is specified by the *analysis distribution*,

$$\boldsymbol{X}_t^a = (\boldsymbol{X}_t | \boldsymbol{y}_1, \ldots, \boldsymbol{y}_t) \sim \mathcal{N}_n(\boldsymbol{\mu}_t^a, \Sigma_t^a) \tag{7.3}$$

with parameters

$$\boldsymbol{\mu}_t^a = \left((\Sigma_t^f)^{-1} + H^\top R^{-1} H\right)^{-1} \left((\Sigma_t^f)^{-1}\boldsymbol{\mu}_t^f + H^\top R^{-1}\boldsymbol{y}_t\right) \tag{7.4}$$

$$\Sigma_t^a = \left((\Sigma_t^f)^{-1} + H^\top R^{-1} H\right)^{-1}. \tag{7.5}$$

The prior estimate of $\boldsymbol{X}_{t+1}$ given $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_t$ is specified by the *forecast distribution*

$$\boldsymbol{X}_{t+1}^f = (\boldsymbol{X}_{t+1} | \boldsymbol{y}_1, \ldots, \boldsymbol{y}_t) \sim \mathcal{N}_n(\boldsymbol{\mu}_{t+1}^f, \Sigma_{t+1}^f) \tag{7.6}$$

with parameters

$$\boldsymbol{\mu}_{t+1}^f = M\boldsymbol{\mu}_t^a, \quad \Sigma_{t+1}^f = M\Sigma_t^a M^\top + Q.$$

Due to the high dimension of the problem, the estimation of Σ_t^f in practical applications is not a straightforward task and different regularization methods, many of them heuristic, are used to produce practically useful estimates.

Our objective is to build a filter resting on the assumption that $\boldsymbol{X}_t$ is a GMRF and using a linear model for its precision matrix. Parameters of the model are estimated by the score matching approach. This approach provides the explicit formula (4.17) for parameter estimators and avoids a heuristic regularization of the sample covariance.

7.1 Score matching filter with Gaussian resampling

Assume that $\boldsymbol{X}_0$ has the Markov property so that the inverse of Σ_0 is sparse. Further, assume that the matrix M, representing the dynamics, has sparse inverse so that $\boldsymbol{X}_1^f = \boldsymbol{X}_1 = M\boldsymbol{X}_0 + \boldsymbol{e}_1^{\boldsymbol{X}}$ is again a GMRF. For a short assimilation time step, this is a realistic assumption in meteorological sciences because values of

meteorological variables at one location are assumed to be influenced only by points from its immediate neighbourhood. By assuming also the sparsity of H, the distribution of the analysis $\boldsymbol{X}_1^a = \boldsymbol{X}_1|\boldsymbol{Y}_1$ resulting from Bayes's theorem has a sparse covariance matrix and therefore, $\boldsymbol{X}_1^a$ is also a GMRF. Sparsity of H means that observations are available only at a small number of locations, which is common in meteorological applications. By induction, we can consider the forecast $\boldsymbol{X}_t^f = (\boldsymbol{X}_t|\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{t-1})$ and the analysis $\boldsymbol{X}_t^a = (\boldsymbol{X}_t|\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_t)$ to be Gaussian Markov random fields for all $t \in \mathbb{N}$.

Some dynamical models may be assumed to be linear, however they are not accessible as a matrix. We may have use of the model only in the form of an algorithm that is able to forward a state vector in time. In this case, we are left with ensemble filtering algorithms. In Algorithm 3, we propose an ensemble filtering algorithm called the Score matching filter with Gaussian resampling (SMF-GR) that provides ensembles that approximate distributions (7.3) and (7.6). As opposed to the EnKF, SMF-GR does not need to perturb the observation and it preserves the normal distribution in every time step. Moreover, the algorithm estimates Σ_t^f through a linear model for its inverse (2.10) instead of using sample covariance and hence, it is better adapted for GMRFs. Parameters of the forecast precision matrix are estimated by the score matching method (estimator (4.40)), which gave the algorithm its name.

Algorithm 3: Score matching filter with Gaussian resampling (**SMF-GR**)

Initial condition: The initial ensemble $\boldsymbol{X}_{01}^a, \ldots, \boldsymbol{X}_{0N}^a$ is sampled from a given initial distribution $\mathcal{N}_n(\boldsymbol{\mu}_0, \Sigma_0)$.

for $t \geq 1$ **do**

Forecast: For all $i = 1, \ldots, N$:

$\boldsymbol{X}_{ti}^f = M\boldsymbol{X}_{t-1,i}^a + \boldsymbol{e}_{ti}^{\boldsymbol{X}}$, where $\boldsymbol{e}_{ti}^{\boldsymbol{X}} \sim \mathcal{N}_n(\boldsymbol{0}, Q)$

Compute

$\bar{\boldsymbol{X}}_t^f = \frac{1}{N}\sum_{i=1}^N \boldsymbol{X}_{ti}^f$

$S_t^f = \frac{1}{N}\sum_{i=1}^N \left(\boldsymbol{X}_{ti}^f - \bar{\boldsymbol{X}}_t^f\right)\left(\boldsymbol{X}_{ti}^f - \bar{\boldsymbol{X}}_t^f\right)^\top$

$\hat{\boldsymbol{\beta}}_t = (\hat{\beta}_{t1}, \ldots, \hat{\beta}_{tr})^\top = \left(\left[\mathrm{tr}\left(A_{tl}S_t^f A_{tk}^\top\right)\right]_{k,l=1}^{r_t}\right)^{-1}\left[\mathrm{tr}(A_{tk})\right]_{k=1}^{r_t}$

Analysis: For a given data vector $\boldsymbol{y}_t$, the analysis mean and covariance are found by substituting $\bar{\mathbf{X}}_t^f$ and $\hat{\boldsymbol{\beta}}_t$ into the formulas for the conditional mean (7.4) and covariance (7.5) :

$\hat{\boldsymbol{\mu}}_t^a = \left(\sum_{k=1}^{r_t}\hat{\beta}_{tk}A_{tk} + H^\top R^{-1}H\right)^{-1}\left(\sum_{k=1}^{r_t}\hat{\beta}_{tk}A_{tk}\bar{\mathbf{X}}_t^f + H^\top R^{-1}\boldsymbol{y}_t\right)$

$\hat{\Sigma}_t^a = \left(\sum_{k=1}^{r_t}\hat{\beta}_{tk}A_{tk} + H^\top R^{-1}H\right)^{-1}$

The ensemble $\boldsymbol{X}_{t1}^a, \ldots, \boldsymbol{X}_{tN}^a$ is then sampled from $\mathcal{N}_n(\hat{\boldsymbol{\mu}}_t^a, \hat{\Sigma}_t^a)$.

end

The set of design matrices $\{A_{tk} : k = 1, \ldots, r_t\}$ is selected in every time step t in order to capture the most important parts of Σ_t^f and to make $\sum_{k=1}^{r_t}\hat{\beta}_{tk}A_{tk}$ positive definite. This covariance selection process is addressed in Section 7.4.2.

The following theorem states that Algorithm 3 provides a consistent estimator of the mean and covariance of the forecast distribution in every time step.

A consistency result for the analysis distribution then follows immediately from the continuous mapping theorem (cf., Theorem 9).

Theorem 24 (SMF-GR)**.** *Assume the discrete-time stochastic linear dynamical system (7.1, 7.2) with $\boldsymbol{X}_0 \sim \mathcal{N}_n(\boldsymbol{\mu}_0, \Sigma_0)$ being a GMRF and with M and H sparse. Assume also that the forecast and analysis ensemble are generated by Algorithm 3, and, at every time t, the covariance matrix Σ_t^f of the forecast is regular and its inverse is in the span of the design matrices at time t, i.e.,*

$$(\Sigma_t^f)^{-1} = \sum_{k=1}^{r} \beta_{tk} A_{tk}, \tag{7.7}$$

for some $\beta_{tk} \in \mathbb{R}$. Suppose that $\mathsf{E}\left(D^*(\boldsymbol{X}_t^f)D(\boldsymbol{X}_t^f)\right)$ defined by (4.45) is invertible for all $t \geq 1$. For $\boldsymbol{X}_{t1}^f, \ldots, \boldsymbol{X}_{tN}^f$, let $\bar{\boldsymbol{X}}_t^f$ and $\hat{\boldsymbol{\beta}}_t^N$ be the SME computed from (4.40). Assume that $\sum_{k=1}^r \hat{\beta}_{tk}^N A_{tk}$ is invertible and denote $\hat{\Sigma}_t^N = \left(\sum_{k=1}^r \hat{\beta}_{tk}^N A_{tk}\right)^{-1}$. Then, for all $t \geq 1$,

$$(\bar{\boldsymbol{X}}_t^f, \hat{\Sigma}_t^N) \xrightarrow{P} (\boldsymbol{\mu}_t^f, \Sigma_t^f) \text{ as } N \to \infty. \tag{7.8}$$

Proof. At $t = 1$, Theorem 19 provides

$$(\bar{\boldsymbol{X}}_1^f, \hat{\boldsymbol{\beta}}_1^N) \xrightarrow[N\to\infty]{P} (\boldsymbol{\mu}_1^f, \boldsymbol{\beta}_1^f),$$

since $(\Sigma_t^f)^{-1}$ is assumed to be of form (7.7). Then, by the continuous mapping theorem (cf. Theorem 9), there is the convergence

$$\left(\bar{\boldsymbol{X}}_t^f, \sum_{k=1}^{r} \hat{\beta}_{tk}^N A_{tk}\right) \xrightarrow[N\to\infty]{P} \left(\boldsymbol{\mu}_1^f, (\Sigma_1^f)^{-1}\right)$$

and (7.8) for $t = 1$ follows from Lemma 7. Suppose now that (7.8) holds with $t-1$ in place of t for some $t \geq 1$. Then, by the continuous mapping theorem applied to the mapping $(\boldsymbol{\mu}_{t-1}^f, \Sigma_{t-1}^f) \mapsto (\boldsymbol{\mu}_{t-1}^a, \Sigma_{t-1}^a)$ defined by (7.4, 7.5), we obtain the convergence

$$(\hat{\boldsymbol{\mu}}_{t-1}^a, \hat{\Sigma}_{t-1}^a) \xrightarrow[N\to\infty]{P} (\boldsymbol{\mu}_{t-1}^a, \Sigma_{t-1}^a).$$

The forecast ensemble at time t is then a sample from $\mathcal{N}_n(M\hat{\boldsymbol{\mu}}_{t-1}^a, M\hat{\Sigma}_{t-1}^a M^\top + Q)$ and due to the sparsity of M and H, each its member is a GMRF. Consider $(\bar{\boldsymbol{X}}_t^f, \hat{\boldsymbol{\beta}}_t^N)$ computed from the forecast ensemble by using the expression (4.40). Then, it follows from Theorem 21 that $\bar{\boldsymbol{X}}_t^f$ and $\hat{\boldsymbol{\beta}}_t^N$ converge, i.e.,

$$(\bar{\boldsymbol{X}}_t^f, \hat{\boldsymbol{\beta}}_t^N) \xrightarrow[N\to\infty]{P} (\boldsymbol{\mu}_t^f, \boldsymbol{\beta}_t^f),$$

and therefore,

$$\left(\bar{\boldsymbol{X}}_t^f, \left(\sum_{k=1}^{r} \hat{\beta}_{tk}^N A_{tk}\right)^{-1}\right) \xrightarrow[N\to\infty]{P} (\boldsymbol{\mu}_t^f, \Sigma_t^f),$$

since the inverse was assumed to exist. □

7.2 Score matching ensemble filter

When there is a need for covariance regularization in filtering, the standard attempt is to insert the regularized covariance into the EnKF formula. This results in a filter that we call the Score matching ensemble filter (SMEF). The algorithm is summarized below.

Algorithm 4: Score matching ensemble filter (**SMEF**)

Initial condition: The initial ensemble $\boldsymbol{X}_{01}^a, \dots, \boldsymbol{X}_{0N}^a$ is sampled from a given initial distribution $\mathcal{N}_n(\boldsymbol{\mu}_0, \Sigma_0)$.

for $t \geq 1$ **do**

Forecast: For all $i = 1, \dots, N$:

$\boldsymbol{X}_{ti}^f = M\boldsymbol{X}_{t-1,i}^a + \boldsymbol{e}_{ti}^{\boldsymbol{X}}$, where $\boldsymbol{e}_{ti}^{\boldsymbol{X}} \sim \mathcal{N}_n(\boldsymbol{0}, Q)$

Compute

$\bar{\boldsymbol{X}}_t^f = \frac{1}{N}\sum_{i=1}^N \boldsymbol{X}_{ti}^f$

$S_t^f = \frac{1}{N}\sum_{i=1}^N \left(\boldsymbol{X}_{ti}^f - \bar{\boldsymbol{X}}_t^f\right)\left(\boldsymbol{X}_{ti}^f - \bar{\boldsymbol{X}}_t^f\right)^\top$

$\hat{\boldsymbol{\beta}}_t = (\hat{\beta}_{t1}, \dots, \hat{\beta}_{tr})^\top = \left(\left[\operatorname{tr}\left(A_{tl}S_t^f A_{tk}\right)\right]_{k,l=1}^{r_t}\right)^{-1}\left[\operatorname{tr} A_{tk}\right]_{k=1}^{r_t}$

$\boldsymbol{Y}_{ti} \sim \mathcal{N}_m(\boldsymbol{y}_t, R)$, $i = 1, \dots, N$ (perturbed observations)

Analysis: For all $i = 1, \dots, N$:

$\boldsymbol{X}_{ti}^a = \left(\sum_{k=1}^{r_t}\hat{\beta}_{tk}A_{tk} + H^\top R^{-1}H\right)^{-1}\left(\sum_{k=1}^{r_t}\hat{\beta}_{tk}A_{tk}\boldsymbol{X}_{ti}^f + H^\top R^{-1}\boldsymbol{Y}_{ti}\right).$

end

The set of design matrices $\{A_{tk}\}_{k=1}^{r_t}$ can again be chosen adaptively and therefore, it can change over time.

Unlike the SMF-GR, in this filter, the ensembles are not normally distributed even for $t = 1$ and therefore, the score matching estimator of $(\boldsymbol{\mu}_t^f, \boldsymbol{\beta}_t)$ cannot be shown to be consistent by the same method as in Theorem 24.

On the other hand, the filter from Algorithm 4 performs very well and in some situations can beat the standard EnKF (Algorithm 1) or the diagonal EnKF (Algorithm 2).

7.3 A non-ensemble score matching filter

In "small" models, we may be able to work with the matrix M, construct the model adjoint $M^\top$, or at least to evaluate the product of a sparse design matrix with the model matrix M. Then a non-ensemble filter that avoids generating analysis ensemble from the estimated posterior distribution may be set-up. The method proceeds directly in terms of the parameters $\boldsymbol{\mu}_t^f$, $\boldsymbol{\beta}_t$ and $\boldsymbol{\mu}_t^a$ of the forecast and analysis distribution.

At $t = 1$, let $\bar{\mathbf{X}}_1^f$ and $\hat{\boldsymbol{\beta}}_1$ be the score-matching estimators of $\boldsymbol{\mu}_1^f$ and $\boldsymbol{\beta}_1$ from Theorem 16 (i.e. constructed from an initial ensemble). For $t \geq 2$, we shall derive the forward model for $\boldsymbol{\mu}_t^f$ and $\boldsymbol{\beta}_t$ from the explicit form of the SME. Having the estimators $\bar{\mathbf{X}}_t^f$ and $\hat{\boldsymbol{\beta}}_t$, the analysis follows from conditioning. As in the previous section, the set of design matrices $\{A_k\}_{k=1}^r$ can change over time, which we again

point out by adding a subindex t.

The estimators of the forecast mean and covariance at time t are

$$\bar{\mathbf{X}}_t^f = M\bar{\mathbf{X}}_{t-1}^a,$$
$$\hat{\Sigma}_t^f = M\hat{\Sigma}_{t-1}^a M^\top + Q = M\left(\sum_{j=1}^{r_{t-1}} \hat{\beta}_{t-1,j} A_{t-1,j} + H^\top R^{-1} H\right)^{-1} M^\top + Q,$$

and therefore, the estimator of $\boldsymbol{\beta}_t$ is

$$\begin{aligned}\hat{\boldsymbol{\beta}}_t &= \left(\left[\operatorname{tr}\left(\left(M\left(\sum_{j=1}^{r_{t-1}} \hat{\beta}_{t-1,j} A_{t-1,j} + H^\top R^{-1} H\right)^{-1} M^\top + Q\right) A_{tk} A_{tl}\right)\right]_{k,l=1}^{r_t}\right)^{-1} \cdot \\ &\quad \cdot \left[\operatorname{tr}(A_{tk})\right]_{k=1}^{r_t} \\ &= \left(\left[\operatorname{tr}\left(\left(\sum_{j=1}^{r_{t-1}} \hat{\beta}_{t-1,j} A_{t-1,j} + H^\top R^{-1} H\right)^{-1} (A_{tk}M)^\top A_{tl} M + A_{tl} Q A_{tk}\right)\right]_{k,l=1}^{r_t}\right)^{-1} \cdot \\ &\quad \cdot \left[\operatorname{tr}(A_{tk})\right]_{k=1}^{r_t},\end{aligned}$$

where we used invariance of trace under cyclic permutation. Summarizing, we obtain Algorithm 5.

Algorithm 5: Non-ensemble score matching filter

Initial condition: From the initial ensemble $\boldsymbol{X}_{11}^f, \ldots, \boldsymbol{X}_{1N}^f$ compute the sample mean $\bar{\mathbf{X}}_1^f$ and sample covariance S_1. Then $\hat{\boldsymbol{\mu}}_1^f = \bar{\mathbf{X}}_1^f$ and $\hat{\boldsymbol{\beta}}_1 = \left(\left[\operatorname{tr}\left(S_1 A_{1k} A_{1l}\right)\right]_{k,l=1}^{r_1}\right)^{-1} \left[\operatorname{tr} A_{1k}\right]_{k=1}^{r_1}$, where $\{A_{1k}\}_{k=1}^{r_1}$ are selected so that $\left(\hat{\Sigma}_1^f\right)^{-1} := \sum_{k=1}^{r_1} \hat{\beta}_{1k} A_{1k}$ is positive definite. Finally, $\hat{\boldsymbol{\mu}}_1^a = \left(\sum_{k=1}^{r_1} \hat{\beta}_{1k} A_{1k} + H^\top R^{-1} H\right)^{-1} \left(\sum_{k=1}^{r_1} \hat{\beta}_{1k} A_{1k} \hat{\boldsymbol{\mu}}_1^f + H^\top R^{-1} \boldsymbol{y}_1\right)$.

for $t \geq 2$ **do**

Forecast:

$\hat{\boldsymbol{\mu}}_t^f = M\hat{\boldsymbol{\mu}}_{t-1}^a$

$\hat{\boldsymbol{\beta}}_t = \left(\left[\operatorname{tr}\left(\left(\sum_{j=1}^{r_{t-1}} \hat{\beta}_{t-1,j} A_{t-1,j} + H^\top R^{-1} H\right)^{-1} (A_{tk}M)^\top A_{tl} M + \right.\right.\right.$
$\left.\left.\left. + A_{tl} Q A_{tk}\right)\right]_{k,l=1}^{r_t}\right)^{-1} \cdot \left[\operatorname{tr}(A_{tk})\right]_{k=1}^{r_t}$

Analysis: For a given data vector $\boldsymbol{y}_t$:

$\hat{\boldsymbol{\mu}}_t^a = \left(\sum_{k=1}^{r_t} \hat{\beta}_{tk} A_{tk} + H^\top R^{-1} H\right)^{-1} \left(\sum_{k=1}^{r_t} \hat{\beta}_{tk} A_{tk} \hat{\boldsymbol{\mu}}_t^f + H^\top R^{-1} \boldsymbol{y}_t\right)$.

end

The consistency of the estimated analysis mean $\hat{\boldsymbol{\mu}}_t^a$ and covariance $\hat{\Sigma}_t^a := \left(\sum_{k=1}^{r_t} \hat{\beta}_{tk} A_{tk} + H^\top R^{-1} H\right)^{-1}$ follows from the consistency of $(\hat{\boldsymbol{\mu}}_1^f, \hat{\boldsymbol{\beta}}_1)$ and the continuous mapping theorem (cf., Theorem 9).

7.4 Computational study

In this section, we carry out a computational study comparing the performance of the proposed Score matching filter with Gaussian resampling and the Score

matching ensemble filter with the standard EnKF and the diagonal EnKF.

First, we consider a simple example of a Gaussian Markov system with linear dynamics, where the assumptions of Theorem 24 are nearly satisfied. In the second simulation, we test both these score matching filters on the Lorenz 96 model, which is neither Gaussian nor Markov but even in this case, the SMEF algorithm seems to be useful.

In both cases, the performance of the methods is measured by the root-mean-square-error of the analysis ensemble mean given by

$$\mathrm{RMSE}_t = \sqrt{\frac{1}{n}\left\|\mathbf{X}_t - \bar{\mathbf{X}}_t^a\right\|_n^2} \tag{7.9}$$

for every time step $t \geq 1$. Recall that n is the state vector dimension, $\mathbf{X}_t$ denotes the true system state and $\bar{\mathbf{X}}_t^a$ is the analysis ensemble mean produced by the given filtering algorithm.

7.4.1 Simple linear advection

Consider a dynamical system (7.1) with M being a simple linear advection model from Raanes et al. [61], which evolves according to a simple cyclic permutation with additive noise

$$X_{t+1,j} = X_{t,j-1} + e_{t+1,j}, \quad t \in \mathbb{N}, j = 1, \ldots, n, \tag{7.10}$$

where $X_{t,j}$ denotes the j-th component of the state vector $\boldsymbol{X}_t$ and $e_{t,j}$ the j-th component of the model error vector $\boldsymbol{e}_t \sim \mathcal{N}_n(\mathbf{0}, Q)$. We assume $X_{t,0} = X_{t,n}$, so the system domain is a circle. The initial precision matrix Σ_0^{-1} was set-up as a band matrix with two subdiagonals (involving the two corners), which corresponds to the first order Markov property on a circle. The structure of the initial precision and covariance matrix is depicted in Figure 7.1.

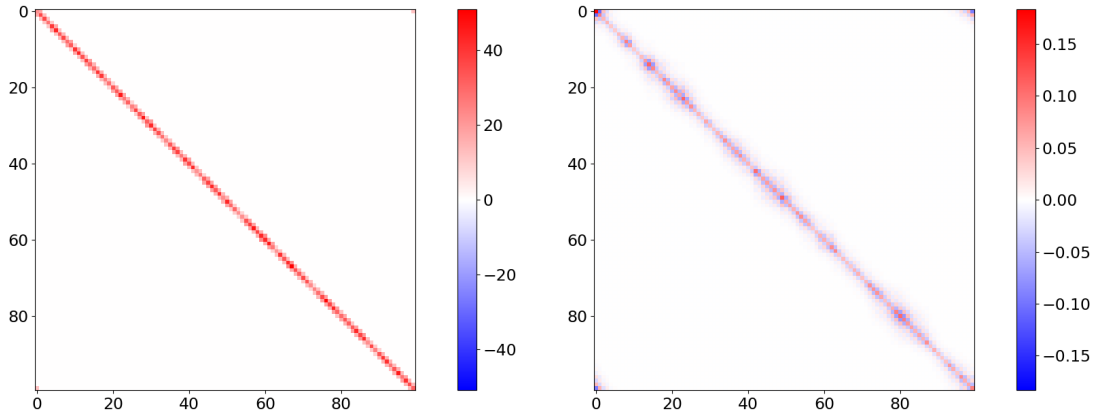
The matrix M which corresponds to (7.10) is sparse and orthogonal and without the presence of model error, the band structure of the precision matrix would have been preserved over time. The additive error contributing to the covariance matrix spoils the Markov property but for the values chosen below the departures are not large.

The initial state $\boldsymbol{X}_0$ with dimension $n = 100$ was drawn from $\mathcal{N}_n(\boldsymbol{\mu}_0, \Sigma_0)$, where $\boldsymbol{\mu}_0 = (\mu_{0,1}, \ldots, \mu_{0,n})^\top$ was generated as a sum of 25 sinusoids of random amplitude and phase (cf. Raanes et al. [61, expr. (62)])

$$\mu_{0,j} = \frac{1}{2}\sum_{k=1}^{25} a_k \sin\left(2\pi k\left[\frac{j}{n} + \varphi_k\right]\right). \tag{7.11}$$

The a_k and φ_k in (7.11) are drawn independently and uniformly from the interval $(0, 1)$ for each k. In the sequel, $\boldsymbol{\mu}_0$ is fixed. The model error covariance matrix is $Q = 0.01 \cdot \Sigma_0$ (Raanes et al. [61, expr. (63)]). Figure 7.2 illustrates the evolution of the state vector for the first three time steps.

We choose a simple linear observation operator H selecting every fifth component of $\boldsymbol{X}$, so the observation vector has dimension $m = 20$. The observation error covariance matrix is $R = 0.01 \cdot I_m$.



(a) Precision matrix Σ_0^{-1}.

(b) Covariance matrix Σ_0.

Figure 7.1: The initial precision and covariance matrix of a simulated random vector defined on a circle and possessing the first order Markov property.

The initial ensemble of N members arises from Gaussian perturbations of $\boldsymbol{X}_0$ with zero mean and covariance matrix Σ_0. Then the system evolves according to (7.1) and (7.2); at $t = 1$ the forecast ensemble has a multivariate normal distribution with mean $M\boldsymbol{\mu}_0$ and covariance matrix $\Sigma_1 = M\Sigma_0 M^\top + Q$, etc. The observations are assimilated by means of the standard EnKF (Algorithm 1), SMF-GR (Algorithm 3) and also by SMEF (Algorithm 4). The set of design matrices $\{A_{ij} \colon i = 1, \ldots, n-1,\ j = i, i+1\} \cup \{A_{n1}, A_{nn}\}$ consists of symmetric matrices A_{ij} that have value 1 at positions (i, j) and (j, i) and zeros elsewhere. In the reported simulation, the precision matrix estimated by the score matching method was positive definite in every time step.

Theorem 24 refers to the asymptotic behaviour of SMF-GR. However, for finite ensemble sizes, it is useful to centre the sampled analysis ensemble around the estimated mean $\hat{\boldsymbol{\mu}}_t^a$ (specified in Algorithm 3) in order to minimize the sampling error.

The performance of every filter in each time was measured by the RMSE (7.9) and plotted into Figure 7.3. It is evident that for smaller ensemble ($N = 50$), SMEF performs slightly better than SMF-GR, and that EnKF is the worst. When the ensemble size increases to the value of n (or more), we can observe that SMF-GR has the smallest RMSE. SMEF is slightly worse and EnKF has the worst performance. The mean RMSE computed as an average over all 500 time steps for each filter is in Table 7.1. For ensemble size smaller than 40, the precision matrix estimated by the score matching method sometimes happened to be negative semidefinite and the simulation was stopped. Since this section tends to illustrate the asymptotic behaviour of SMF-GR stated in Theorem 24, we decided to use only larger ensemble sizes. The problem of negative semidefiniteness for small ensemble size is addressed within Section 7.4.2.

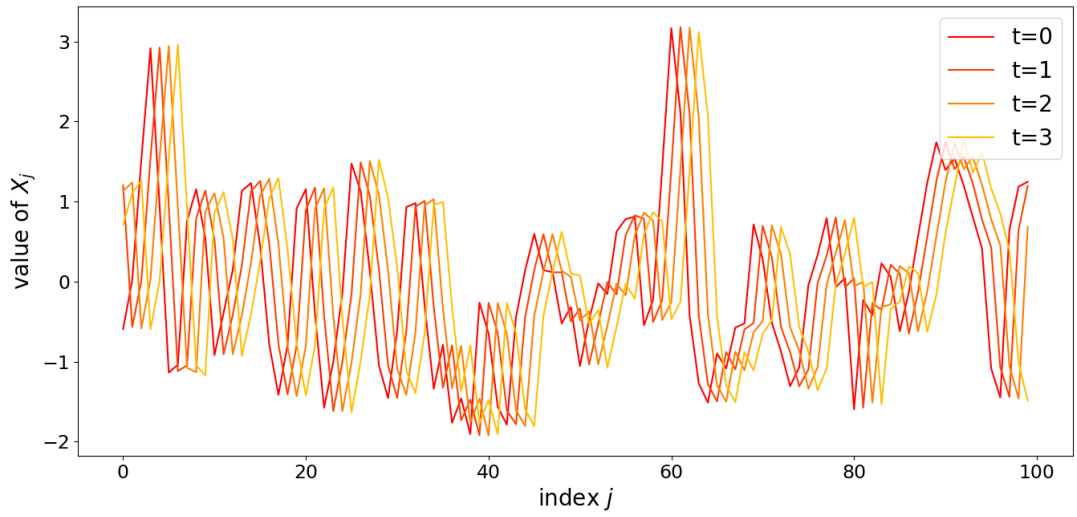


Figure 7.2: Linear advection: Random vector $\boldsymbol{X}_0 = (X_{0,j})_{j=1}^n \sim \mathcal{N}_n(\boldsymbol{\mu}_0, \Sigma_0)$ evolving in time by the model $X_{t+1,j} = X_{t,j-1} + e_{t+1,j}$, $\boldsymbol{e}_t = (e_{t,j})_{j=1}^n \sim \mathcal{N}_n(\boldsymbol{0}, Q)$. The mean $\boldsymbol{\mu}_0$ is specified in (7.11), Σ_0 is plotted in Figure 7.1b and $Q = 0.01 \cdot \Sigma_0$.

	ensemble size N		
	50	100	200
EnKF	0.0905	0.0720	0.0631
SMF-GR	0.0612	**0.0556**	**0.0518**
SMEF	**0.0573**	0.0571	0.0560

Table 7.1: Linear advection (simulation): RMSE of different filtering algorithms averaged from 500 time steps. Minimum in each column is displayed in bold font.

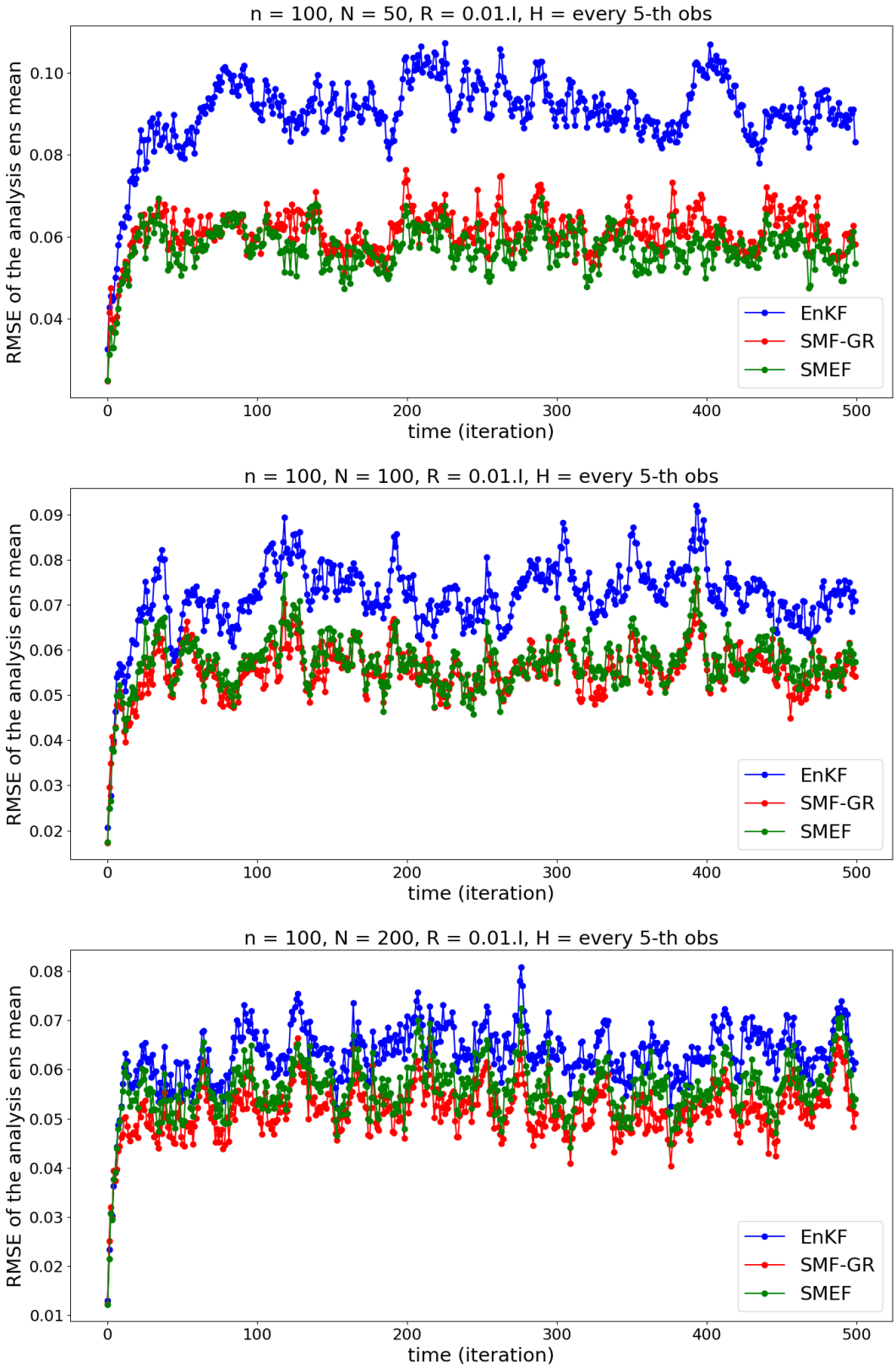


Figure 7.3: Linear advection (simulation): Comparison of SMF-GR (Algorithm 3), SMEF (Algorithm 4) and EnKF (Algorithm 1) for a Gauss Markov system with linear advection dynamics. The state vector dimension was $n = 100$ and the ensemble size was $N = 50, 100, 200$. Observations were available for every 5-th variable and the observation error has covariance $R = 0.01 \cdot I_m$.

7.4.2 Lorenz 96

This model was published by Lorenz [48] as a simplified one-dimensional equatorial atmospheric model. The system state at time t is represented by a random vector $\mathbf{X}_t = (X_{t,1}, \ldots, X_{t,n})^\top$ defined on a circle with $n = 40$ points. The time evolution of each component $X_{t,j}$, $j \in \{1, \ldots, n\}$, is defined by the equation

$$\begin{aligned} X_{t+1,j} &= X_{t,j} + \frac{dX_{t,j}}{dt}, \quad t \in \mathbb{N}, \text{ with} \\ \frac{dX_{t,j}}{dt} &= (X_{t,j+1} - X_{t,j-2})X_{t,j-1} - X_{t,j} + F, \end{aligned} \tag{7.12}$$

where $X_{t,-1} = X_{t,n-1}$, $X_{t,0} = X_{t,n}$, $X_{t,n+1} = X_{t,1}$. The components $X_{0,j}, j = 1, \ldots, n$, of the initial vector were sampled from uniform distribution on interval $\left[-\frac{1}{2}, \frac{1}{2}\right]$. The forcing term F was set to 8, which is a known value that causes chaotic behaviour. Due to this chaotic behaviour, the dynamic model does not need to contain any additive noise. The Lorenz system does not have the spatial Markov property in the traditional sense of Definition 1, however, equation (7.12) foreshadows some kind of relationship between each point and its three neighbours. This guess is further supported by the shape of inverse of the sample covariance matrix computed from a sample of 5000 random vectors resulting after 1000 steps of evolution by Lorenz 96, which is depicted in Figure 7.4. Even if the inverse of covariance matrix of forecast distribution could have slightly different structure than the matrix from Figure 7.4, we suggest to approximate it by a band matrix with one main diagonal and 3 subdiagonals on each side (including the two corners since the system is defined on a circle). The set of chosen design matrices

$$\begin{aligned} \mathcal{A} = \{A_{ij} \colon & i = 1, \ldots, n,\ j = i, i+1, i+2, i+3, \\ & \text{where } n + k \equiv k \text{ and } 1 - k \equiv n - k + 1 \text{ for } k = 1, 2, 3\} \end{aligned}$$

consists of symmetric matrices A_{ij} that have value 1 at positions (i, j) and (j, i) and zeros elsewhere.

Rather than using the covariance matrix from the free run (Figure 7.4), particle filters (Doucet et al. [20]) could be in principle used to approximate the exact filtering distribution, from which we could calculate the covariance and its inverse. However, the chosen set $\mathcal{A}$ seems to perform well in simulations. Since every design matrix corresponds only to one element (up to symmetry), the model is very flexible. We want to model the important parts of the precision matrix adaptively, and at the same time, we have to keep the estimates of covariance matrix positive definite. To this point we propose a selection method described in the following subsection.

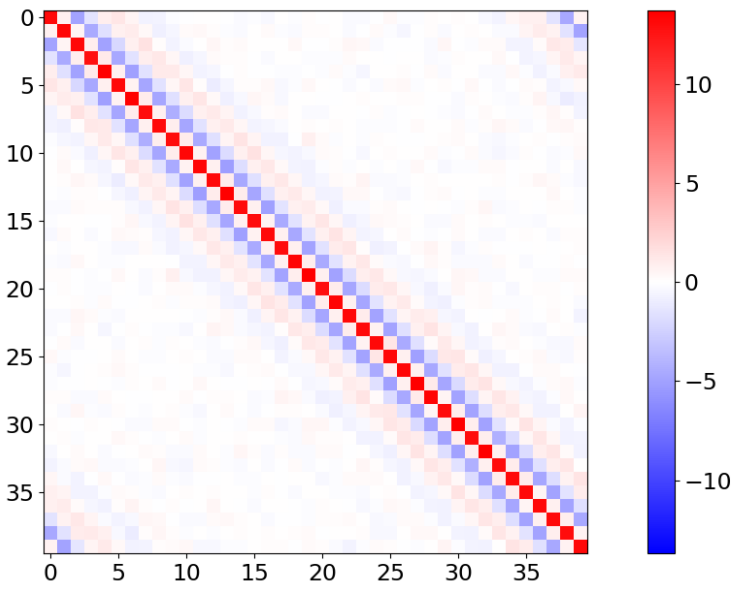


Figure 7.4: Lorenz 96 (simulation): Inversion of the sample covariance matrix computed from 5000 vectors of length $n = 40$ starting from the uniform distribution on $\left[-\frac{1}{2}, \frac{1}{2}\right]$ at each point and advanced by Lorenz 96 (with $F = 8$) for 1000 time steps.

Backward selection of design matrices

The form $\sum_{k=1}^{r} \beta_k A_k$ of the precision matrix model does not guarantee its score matching estimate to be positive definite. Sometimes, the SME based on an ensemble $\mathbb{X}_N = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N\}$ may overestimate the model fit as it corresponds to the minimum of the objective function $\mathcal{S}_N(\boldsymbol{\mu}, \boldsymbol{\beta}|\mathbb{X}_N)$ over the entire space $L = \{(\boldsymbol{\mu}, \boldsymbol{\beta}) \,|\boldsymbol{\mu} \in \mathbb{R}^n, \boldsymbol{\beta} \in \mathbb{R}^r\}$ rather than over

$$\tilde{\boldsymbol{\Theta}} = \left\{ (\boldsymbol{\mu}, \boldsymbol{\beta}) \,\middle|\, \boldsymbol{\mu} \in \mathbb{R}^n, \sum_{k=1}^{r} \beta_k A_k \text{ is positive definite} \right\} \subset L.$$

This problem occurs mainly for small ensemble sizes. For larger ensemble sizes, the estimate tends to be positive definite due to the consistency of SME, as stated in Lemma 7. In order to make the score matching filters practically applicable even for small ensemble sizes, we need a method for finding acceptable value of $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\beta}})$ in $\tilde{\boldsymbol{\Theta}}$ in the situation when the minimum of $\mathcal{S}_N(\boldsymbol{\mu}, \boldsymbol{\beta}|\mathbb{X}_N)$ lies in $L \setminus \tilde{\boldsymbol{\Theta}}$. Since there is no restriction on $\boldsymbol{\mu}$, we keep $\hat{\boldsymbol{\mu}}$ associated with the minimum of $\mathcal{S}_N(\boldsymbol{\mu}, \boldsymbol{\beta}|\mathbb{X}_N)$ and focus on adjusting $\hat{\boldsymbol{\beta}}$.

The main idea is to model Σ^{-1} only by means of a subset $\mathcal{A}_*$ of $\mathcal{A}$ that leads to a positive definite score matching estimate and contributes significantly to the objective function (4.14).

The set $\mathcal{A}_0$ of design matrices spanning the diagonal of Σ^{-1} has to be involved in $\mathcal{A}_*$ in any case. Denote by $\boldsymbol{\beta}_0$ the set of parameters corresponding to matrices in $\mathcal{A}_0$. For each design matrix $A_{jk} \in \mathcal{A} \setminus \mathcal{A}_0$, which is associated with an off-diagonal element of Σ^{-1}, we compute the optimal value of the objective function (4.14) associated with the SME $\hat{\boldsymbol{\beta}}_{jk}$ (given by (4.40)) of the parameter $\boldsymbol{\beta}_{jk} = (\boldsymbol{\beta}_0^\top, \beta_{jk})^\top$. The optimal value of $\mathcal{S}_N(\boldsymbol{\mu}, \boldsymbol{\beta}|\mathbb{X}_N)$ (up to constants $1/N$ and $c_N^*(\mathbb{X}_N)$, which do

not depend on parameters) is equal to

$$\frac{1}{N}\mathcal{S}_N(\bar{\boldsymbol{X}}, \hat{\boldsymbol{\beta}}_{jk}|\mathbb{X}_N) - \frac{1}{N}c_N^*(\mathbb{X}_N) = -\frac{1}{2}\Big(\operatorname{tr}(A_{11}), \ldots, \operatorname{tr}(A_{nn}), \operatorname{tr}(A_{jk})\Big)\hat{\boldsymbol{\beta}}_{jk}. \quad (7.13)$$

The exact computation of (7.13) is provided in Appendix A.1. Afterwards, all matrices $A_{jk} \in \mathcal{A} \setminus \mathcal{A}_0$ are ordered in ascending manner according to their value of (7.13). Thus the A_{jk} yielding the largest contribution to the objective function when added to the set which spans the diagonal, are ranked first in the list. Then, design matrices from the opposite end of the list are successively discarded until we reach a positive definite matrix. A treshold for the minimal eigenvalue of the estimated covariance can also be set in this way.

Even though we do not have any optimality result to justify this approach, it worked well in practice.

Simulation results

The simulation was carried out as follows. At the beginning, we took a random vector sampled from the uniform distribution on $[-\frac{1}{2}, \frac{1}{2}]$ and performed 1000 steps of free run (the so called *spin-up*), so as to let the system to catch the attractor. After the spin-up, we generate a vector representing the truth and an initial ensemble of N vectors by adding white noise to each component of the spin-up vector. Then we start the assimilation process. We choose a linear observation operator H selecting every second component of $\boldsymbol{X}$ and we observe in every time step with the observation error matrix $R = 0.5 \cdot I_m$. Beside the EnKF (Algorithm 1), SMF-GR (Algorithm 3) and the SMEF (Algorithm 4), we used also the diagonal EnKF (Algorithm 2). The RMSE (7.9) of the analysis ensemble mean for all these filters is plotted in Figure 7.5. The averaged RMSE from all time steps is recorded in Table 7.2.

	ensemble size N		
	10	30	80
EnKF	4.6679	4.5796	**0.2570**
SMF-GR	4.6650	1.9357	0.4940
SMEF	**0.7008**	**0.4705**	0.4317
diag EnKF	1.3748	1.4754	1.7292
free run	4.9194	5.1320	4.8785

Table 7.2: Lorenz 96 (simulation): RMSE of different filtering algorithms averaged from 500 time steps. Minimum in each column displayed in bold font.

We see in Figure 7.5 that for small ensembles, the SMEF has constantly the lowest RMSE, even though the structure of the regularized precision matrix was derived under the assumption of normality. The diagonal EnKF performs better than EnKF but not as well as SMEF. This confirms the fact that estimating off-diagonal elements of the forecast covariance is beneficial in small samples. The performance of EnKF is similar to the free run. Since the system state is not Gaussian, the SMF-GR (which preserves normality by resampling) performs poorly, however, especially for $N = 30$, it is still better than the EnKF.

When the ensemble size significantly exceeds the dimension of the state, the sample covariance matrix becomes the best available estimate of the true covariance, which results in the excellent performance of EnKF (cf., the very last picture in Figure 7.5).

Remark 8. In this simulation study we did not employ any of the heuristic techniques of Section 2.1 or inflation of the ensemble. Apart from the simulations presented here, we performed a number of experiments where some of these techniques were employed. The performance of both traditional and proposed filters improved. Also, hybrid approach combining regularization with resampling was successful in some cases. These considerations are beyond the scope of this thesis and will be treated elsewhere.

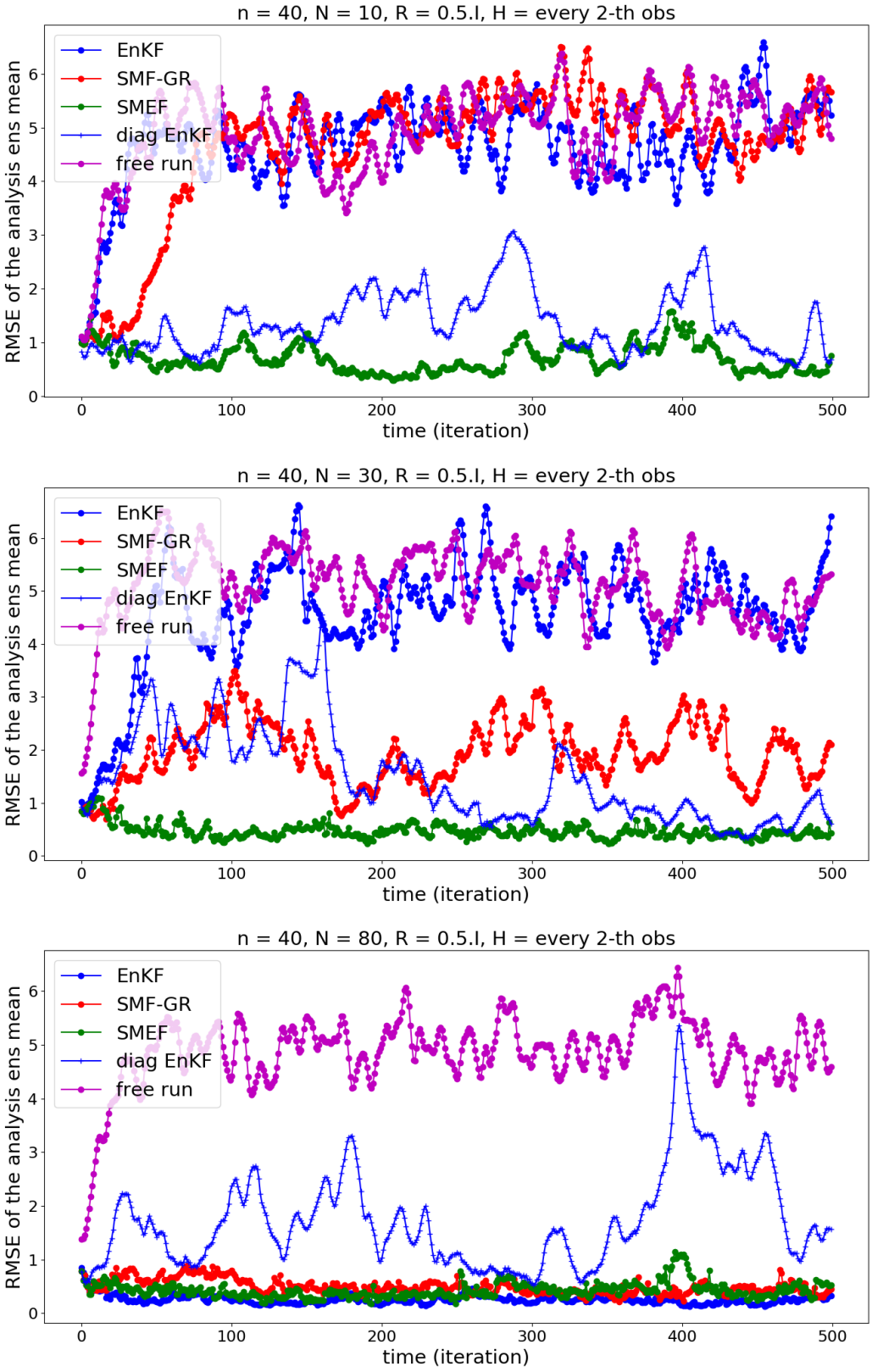


Figure 7.5: Lorenz 96 (simulation): Comparison of SMF-GR (Algorithm 3), SMEF (Algorithm 4), EnKF (Algorithm 1) and diagonal EnKF (Algorithm 2). Spin-up was 1000 steps. The state vector had dimension $n = 40$ and the ensemble size was $N = 10, 30, 80$. Observations were available for every second variable and the observation error had covariance $R = 0.5 \cdot I_m$.

Conclusion

Many applied problems require an estimate of a covariance matrix or its inverse. In such problems, the matrix dimension can be large compared to the sample size. In the first part of this thesis, we have provided an overview of several estimating methods with focus on their use in data assimilation.

After summarizing basic estimating techniques, that are usually based on element-wise transformation of the sample covariance matrix, we shifted our attention on parametric models for the covariance matrix or its inverse. The associated parameters were estimated by the maximum likelihood or the score matching method. Both of these techniques were supplemented by several new results. We have shown that asymptotic covariance matrices of nested M-estimators have a hierarchical structure, which, in particular, applies to the maximum likelihood and score matching estimators. The hierarchical comparison was in terms of traces of asymptotic covariance matrices of two nested parametrizations after a specific transform. For the maximum likelihood estimators, we have derived a stronger result that compares the whole asymptotic covariance matrices of two nested parametrizations in terms of positive definiteness. Moreover, we derived explicit formulas for parameter estimators for two particular covariance models. First, we computed maximum likelihood estimators of parameters in models intended for the decay of eigenvalues of a covariance matrix of weakly stationary random field. Second, we computed score matching estimators for parameters of a linear model for the precision matrix of a Gaussian Markov random field. These covariance models allow a compromise between realistic assumptions and relatively cheap computations.

The second part of the thesis deals with filtering algorithms used in data assimilation. The performance of these filtering algorithms is highly influenced by the quality of the covariance estimate. We proposed three filtering algorithms based on the score matching estimator of a covariance model for a Gaussian Markov random field. We proved that the Score matching filter with Gaussian resampling provides consistent estimates of the mean and covariance matrix of the true forecast distribution in every time step. The key component in the proof is the continuity of score matching estimators to random perturbations, which we have also shown. The second proposed filter, called the Score matching ensemble filter, is directly based on the well-known Ensemble Kalman filter and it seems to work well even for a large class of dynamical systems (even without normality or Markov property). However, its limit properties for large ensembles are not studied in this thesis and they are left as a subject of further research. One problem in using a linear model for a precision matrix in filtering algorithms is that positive definiteness of the resulting estimator is not guaranteed and has to be addressed separately. In our algorithm, we proposed a method of covariance selection. However, an optimal way of making the estimated covariance positive definite is a non-trivial open problem.

The main contribution of this thesis is contained in Chapters 3, 4, 5 and 7. All these results together with other outcomes from related topics that we have dealt with during my PhD studies, have been published in papers listed in the "List of publications" at the end of this thesis.

Bibliography

[1] J. L. Anderson. An ensemble Adjustment Kalman filter for data assimilation. *Monthly Weather Review*, 129:2884–2903, 2001. doi: 10.1175/1520-0493(2001)129⟨2884:AEAKFF⟩2.0.CO;2.

[2] K. Atkinson. Convergence rates for approximate eigenvalues of compact integral operators. *SIAM Journal on Numerical Analysis*, 12(2):213–222, 1975. doi: 10.1137/0712020.

[3] K. Atkinson and W. Han. *Theoretical numerical analysis: A functional analysis framework*, volume 39 of *Texts in Applied Mathematics*. Springer, Dordrecht, third edition, 2009. doi: 10.1007/978-1-4419-0458-4.

[4] K. E. Atkinson. The numerical solutions of the eigenvalue problem for compact integral operators. *Transactions of the American Mathematical Society*, 129:458–465, 1967. doi: 10.2307/1994601.

[5] R. N. Bannister. A review of forecast error covariance statistics in atmospheric variational data assimilation II: Modelling the forecast error covariance statistics. *Quarterly Journal of the Royal Meteorological Society*, 134 (637):1971–1996, 2008. doi: 10.1002/qj.340.

[6] O. Barndorff-Nielsen. *Information and Exponential Families in Statistical Theory*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Ltd., Chichester, 1978.

[7] O. E. Barndorff-Nielsen and D. R. Cox. *Inference and Asymptotics*. Monographs on Statistics and Applied Probability. Springer Science+Business Media, 1994.

[8] J. Berner, G. J. Shutts, M. Leutbecher, and T. N. Palmer. A spectral stochastic kinetic energy backscatter scheme and its impact on flow-dependent predictability in the ECMWF ensemble prediction system. *Journal of the Atmospheric Sciences*, 66(3):603–626, 2009. doi: 10.1175/2008JAS2677.1.

[9] P. Bickel and E. Levina. Covariance regularization by thresholding. *Annals of Statistics*, 36(6):2577–2604, 12 2008. doi: 10.1214/08-AOS600.

[10] C. H. Bishop, B. J. Etherton, and S. J. Majumdar. Adaptive sampling with the ensemble transform Kalman filter. Part I: Theoretical aspects. *Monthly Weather Review*, 129:420–436, 2001. doi: 10.1175/1520-0493(2001)129⟨0420:ASWTET⟩2.0.CO;2.

[11] P. Brockwell and R. Davis. *Time Series: Theory and Methods*. Springer Science + Business Media, LLC, New York, 2006.

[12] M. Buehner and M. Charron. Spectral and spatial localization of background-error correlations for data assimilation. *Quarterly Journal of the Royal Meteorological Society*, 133(624):615–630, 2007. doi: 10.1002/qj.50.

[13] G. Burgers, P. J. van Leeuwen, and G. Evensen. Analysis scheme in the ensemble Kalman filter. *Monthly Weather Review*, 126:1719–1724, 1998.

[14] C. Burrus, R. Gopinath, and H. Guo. *Introduction to Wavelets and Wavelet Transforms: A Primer*. Prentice Hall, New Jersey, 1998.

[15] E. Carlen. Trace inequalities and quantum entropy: An introductory course. In *Entropy and the quantum*, volume 529 of *Contemp. Math.*, pages 73–140. Amer. Math. Soc., Providence, RI, 2010. doi: 10.1090/conm/529/10428.

[16] P. Courtier, E. Andersson, W. Heckley, D. Vasiljevic, M. Hamrud, A. Hollingsworth, F. Rabier, M. Fisher, and J. Pailleux. The ECMWF implementation of three-dimensional variational assimilation (3D-Var). I: Formulation. *Quarterly Journal of the Royal Meteorological Society*, 124(550): 1783–1807, 1998. doi: 10.1002/qj.49712455002.

[17] I. Daubechies. Orthonormal bases of compactly supported wavelets. *Communications on Pure and Applied Mathematics*, 41(7):909–996, 1988. doi: 10.1002/cpa.3160410705.

[18] I. Daubechies. *Ten Lectures on Wavelets*, volume 61 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1992. doi: 10.1137/1.9781611970104.

[19] A. P. Dempster. Covariance selection. *Biometrics*, 28(1):157 – 175, 1972. doi: 10.2307/2528966.

[20] A. Doucet, N. de Freitas, and N. Gordon, editors. *Sequential Monte Carlo in Practice*. Springer, 2001.

[21] Y. Dwivedi and S. S. Rao. A test for second-order stationarity of a time series based on the discrete fourier transform. *Journal of Time Series Analysis*, 32 (1):68–91, September 2010. doi: 10.1111/j.1467-9892.2010.00685.x.

[22] G. Evensen. *Data Assimilation: The Ensemble Kalman Filter*. Springer, second edition, 2009. doi: 10.1007/978-3-642-03711-5.

[23] P. G. M. Forbes and S. Lauritzen. Linear estimating equations for exponential families with application to Gaussian linear concentration models. *Linear Algebra and its Applications*, 473:261–283, 2015. doi: https://doi.org/10.1016/j.laa.2014.08.015. Special issue on Statistics.

[24] R. Furrer and T. Bengtsson. Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants. *Journal of Multivariate Analysis*, 98(2):227 – 255, 2007. doi: https://doi.org/10.1016/j.jmva.2006.08.003.

[25] G. Gaspari and S. E. Cohn. Construction of correlation functions in two and three dimensions. *Quarterly Journal of the Royal Meteorological Society*, 125 (554):723–757, 1999. doi: 10.1002/qj.49712555417.

[26] G. Gaspari, S. E. Cohn, J. Guo, and S. Pawson. Construction and application of covariance functions with variable length-fields. *Quarterly Journal of the Royal Meteorological Society*, 132(619):1815–1838, 2006. doi: 10.1256/qj.05.08.

[27] P. Giudici and P. J. Green. Decomposable graphical gaussian model determination. *Biometrika*, 86(4):785–801, 1999. doi: https://doi.org/10.1093/biomet/86.4.785.

[28] T. Hamill, J. Whitaker, and C. Snyder. Distance-dependent filtering of background error covariance estimates in an ensemble Kalman filter. *Monthly Weather Review*, 129(11):2776–2790, 2001. doi: 10.1175/1520-0493(2001)129⟨2776:DDFOBE⟩2.0.CO;2.

[29] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Second edition. Cambridge University Press, 2013.

[30] P. Houtekamer and H. Mitchell. A sequential ensemble Kalman filter for atmospheric data assimilation. *Monthly Weather Review*, 129(1):123–137, 2001. doi: 10.1175/1520-0493(2001)129⟨0123:ASEKFF⟩2.0.CO;2.

[31] C. Huang, H. Guo, and Z. Zhang. A spectral collocation method for eigenvalue problems of compact integral operators. *Journal of Integral Equations and Applications*, 25(1):79–101, 2013. doi: 10.1216/JIE-2013-25-1-79.

[32] B. R. Hunt, E. J. Kostelich, and I. Szunyogh. Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter. *Physica D: Nonlinear Phenomena*, 230:112–126, 2007. doi: 10.1016/j.physd.2006.11.008.

[33] A. Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6:695–709, 2005.

[34] A. Hyvärinen. Some extensions of score matching. *Computational Statistics & Data Analysis*, 51(5):2499–2512, 2007. doi: 10.1016/j.csda.2006.09.003.

[35] R. E. Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME – Journal of Basic Engineering, Series D*, 82: 35–45, 1960. doi: 10.1115/1.3662552.

[36] I. Kasanický. *Ensemble Kalman filter on high and infinite dimensional spaces*. PhD thesis, Charles University, Faculty of Mathematics and Physics, 2016.

[37] I. Kasanický, J. Mandel, and M. Vejmelka. Spectral diagonal ensemble Kalman filters. *Nonlinear Processes in Geophysics*, 22(4):485 – 497, 2015. doi: 10.5194/npg-22-485-2015.

[38] M. Katzfuss, J. R. Stroud, and C. K. Wikle. Understanding the ensemble Kalman filter. *The American Statistician*, 70(4):350–357, 2016. doi: 10.1080/00031305.2016.1141709.

[39] H. König. *Eigenvalue Distribution of Compact Operators*. Operator Theory: Advances and Applications. Springer Basel AG, 1986. doi: 10.1007/978-3-0348-6278-3.

[40] K. Law, A. Stuart, and K. Zygalakis. *Data assimilation: A mathematical introduction*, volume 62 of *Texts in Applied Mathematics*. Springer, Cham, 2015. doi: 10.1007/978-3-319-20325-6.

[41] F. Le Gland, V. Monbet, and V.-D. Tran. Large sample asymptotics for the ensemble Kalman filter. In D. Crisan and B. Rozovskiĭ, editors, *The Oxford Handbook of Nonlinear Filtering*, pages 598–631. Oxford University Press, 2011.

[42] O. Ledoit and M. Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004. doi: 10.1016/S0047-259X(03)00096-4.

[43] E. L. Lehmann and G. Casella. *Theory of Point Estimation*. Springer Texts in Statistics. Springer-Verlag, New York, second edition, 1998. doi: 10.1007/b98854.

[44] E. L. Lehmann and J. P. Romano. *Testing Statistical Hypotheses*. Springer Texts in Statistics. Springer, New York, third edition, 2005.

[45] L. Lin, M. Drton, and A. Shojaie. Estimation of high-dimensional graphical models using regularized score matching. *Electronic Journal of Statistics*, 10 (1):806–854, 2016. doi: 10.1214/16-ejs1126.

[46] F. Lindgren, H. Rue, and J. Lindström. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498, 2011. doi: 10.1111/j.1467-9868.2011.00777.x.

[47] D. M. Livings, S. L. Dance, and N. K. Nichols. Unbiased ensemble square root filters. *Phys. D*, 237(8):1021–1028, 2008. doi: 10.1016/j.physd.2008.01.005.

[48] E. N. Lorenz. Predictability - a problem partly solved. In T. Palmer and R. Hagendorn, editors, *Predictability of Weather and Climate*, pages 40–58. Cambridge University Press, 2006. doi: https://doi.org/10.1017/CBO9780511617652.004.

[49] J. R. Magnus and H. Neudecker. *Matrix Differential Calculus with Applications in Statistics and Econometrics*. John Wiley, third edition, 2007.

[50] J. Mandel, J. D. Beezley, J. L. Coen, and M. Kim. Data assimilation for wildland fires: Ensemble Kalman filters in coupled atmosphere-surface models. *IEEE Control Systems Magazine*, 29(3):47–65, June 2009. doi: 10.1109/MCS.2009.932224.

[51] J. Mandel, L. Cobb, and J. D. Beezley. On the convergence of the ensemble Kalman filter. arXiv:0901.2951, January 2009.

[52] J. Mandel, L. Cobb, and J. D. Beezley. On the convergence of the ensemble Kalman filter. *Applications of Mathematics*, 56:533–541, 2011. doi: 10.1007/s10492-011-0031-2.

[53] T. Matsuo, D. W. Nychka, and D. Paul. Nonstationary covariance modeling for incomplete data: Monte Carlo EM approach. *Computational Statistics & Data Analysis*, 55(6):2059–2073, 2011. doi: 10.1016/j.csda.2010.12.002.

[54] Y. Michel and T. Auligné. Inhomogeneous Background Error Modeling and Estimation over Antarctica. *Monthly Weather Review*, 138(6):2229–2252, 2010. doi: 10.1175/2009mwr3139.1.

[55] I. Mirouze and A. T. Weaver. Representation of correlation functions in variational assimilation using an implicit diffusion operator. *Quarterly Journal of the Royal Meteorological Society*, 136:1421–1443, 2010. doi: 10.1002/qj.643.

[56] R. Muirhead. Developments in eigenvalue estimation. In A. K. Gupta, editor, *Advances in Multivariate Statistical Analysis*, pages 277–288. D. Reidel Publishing Company, 1987.

[57] R. Muirhead. *Aspects of Multivariate Statistical Theory*. Wiley Series in Probability and Statistics. Wiley, 2005. doi: 10.1002/9780470316559.

[58] O. Pannekoucke, L. Berre, and G. Desroziers. Filtering properties of wavelets for local background-error correlations. *Quarterly Journal of the Royal Meteorological Society*, 133(623, Part B):363–379, 2007. doi: 10.1002/qj.33.

[59] D. F. Parrish and J. C. Derber. The National Meteorological Center's spectral statistical-interpolation analysis system. *Monthly Weather Review*, 120 (8):1747–1763, 1992. doi: 10.1175/1520-0493(1992)120⟨1747:TNMCSS⟩2.0.CO;2.

[60] M. Pourahmadi. Covariance estimation: the GLM and regularization perspectives. *Statistical Science. A Review Journal of the Institute of Mathematical Statistics*, 26(3):369–387, 2011. doi: 10.1214/11-STS358.

[61] P. N. Raanes, A. Carrassi, and L. Bertino. Extending the square root method to account for additive forecast noise in ensemble methods. *Monthly Weather Review*, 143(10):3857–3873, 2015. doi: 10.1175/MWR-D-14-00375.1.

[62] C. R. Rao. *Linear Statistical Inference and its Applications*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York-London-Sydney, second edition, 1973. doi: 10.1002/9780470316436.

[63] H. Rue and L. Held. *Gaussian Markov Random Fields: Theory and Applications*. Monographs on Statistics and Applied Probability. Chapman & Hall/CRC, 2005.

[64] D. Simpson, F. Lindgren, and H. Rue. Think continuous: Markovian Gaussian models in spatial statistics. *Spatial Statistics*, 1:16–29, 2012. doi: 10.1016/j.spasta.2012.02.003.

[65] A. Spantini, R. Baptista, and Y. Marzouk. Coupling techniques for nonlinear ensemble filtering. arXiv:1907.00389, 2019.

[66] M. Turčičová, J. Mandel, and K. Eben. Score matching filters for Gaussian Markov fields with a linear model of the precision matrix. Paper submitted.

[67] M. Turčičová, J. Mandel, and K. Eben. Multilevel maximum likelihood estimation with application to covariance matrices. *Communications in Statistics. Theory and Methods*, 48(4):909–925, 2019. doi: 10.1080/03610926.2017.1422755.

[68] G. Ueno and T. Tsuchiya. Covariance regularization in inverse space. *Quarterly Journal of the Royal Meteorological Society*, 135(642):1133–1156, 2009. doi: 10.1002/qj.445.

[69] A. W. Van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 2000.

[70] J. Whittaker. *Graphical Models in Applied Multivariate Statistics*. Wiley Series in Probability and Statistics. Chichester: John Wiley & Sons, Ltd., 1990.

[71] J.-H. Won and S.-J. Kim. Maximum likelihood covariance estimation with a condition number constraint. In M. B. Matthews, editor, *In Proc. 40th Asilomar Conf. Signals, Systems and Computers, Pacific Grove, Oct. 29th–Nov. 1st*, pages 1445 – 1449. New York: Institute of Electrical and Electronics Engineers, 2006.

[72] J.-H. Won, J. Lim, S.-J. Kim, and B. Rajaratnam. Condition-number-regularized covariance estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(3):427–450, 2013. doi: 10.1111/j.1467-9868.2012.01049.x.

[73] A. Yaglom. *Correlation Theory of Stationary and Related Random Functions: Volume I: Basic Results*. Springer Series in Statistics. Springer New York, 1987.

[74] S. Yu, M. Drton, and A. Shojaie. Generalized score matching for non-negative data. *Journal of Machine Learning Research (JMLR)*, 20:1–70, 2019.

List of Figures

List of Tables

List of Abbreviations

Abbreviations

EnKF Ensemble Kalman filter.

GMRF Gaussian Markov random fields.

MLE Maximum Likelihood Estimator.

SME Score Matching Estimator.

SMEF Score matching ensemble filter.

SMF-GR Score matching filter with Gaussian resampling.

Nomenclature

$A \circ B$ Schur product of matrices A and B.

$A \otimes B$ the Kronecker product of matrices A and B.

I_n identity matrix of type $n \times n$.

N size of a sample or ensemble.

$[v_{ij}]_{i,j=1}^{m,n}$ $m \times n$ matrix of elements v_{ij}.

$[v_j]_{j=1}^{m}$ column vector $(v_1, \ldots, v_n)^\top$.

$\boldsymbol{X}$ random vector (usually the state vector of a dynamical system).

$\|A\|_F$ Frobenius norm of a matrix A.

$\|T\|_{op}$ operator norm of a linear operator T.

$\|\boldsymbol{v}\|_{\mathcal{V}}$ norm of an element $\boldsymbol{v}$ from a vector space $\mathcal{V}$.

$\|\boldsymbol{v}\|_n$ Euclidean norm of a vector $\boldsymbol{v} \in \mathbb{R}^n$.

$\mathbf{1}_{[B]}$ indicator function of an event B.

$\nabla_{\boldsymbol{u}} h(\boldsymbol{u})$ gradient of a scalar function h with respect to $\boldsymbol{u}$ (row vector).

$\det(A)$ determinant of the matrix A.

$\operatorname{tr}(A)$ trace of a matrix A.

$\xrightarrow{P}$ convergence in probability.

$\xrightarrow{d}$ convergence in distribution.

n lenght of the random vector $\boldsymbol{X}$.

$J_{\boldsymbol{u}}(h(\boldsymbol{u})) = \left[\frac{\partial h_i(\boldsymbol{u})}{\partial u_j}\right]_{i,j=1}^{m,n}$ Jacobian matrix of a vector function $h\colon \mathbb{R}^n \to \mathbb{R}^m$ with respect to $\boldsymbol{u}$ (matrix of type $m \times n$).

$[\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_K]$ matrix consisting of columns $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_K$.

$\langle \cdot, \cdot \rangle_{\mathcal{V}}$ inner product in a vector space $\mathcal{V}$.

$\mathrm{diag}(d_1, \ldots, d_m), \mathrm{diag}\{d_k, k = 1, \ldots, m\}$ diagonal matrix with elements $d_1, ..., d_m$ on its diagonal.

List of publications

(i) M. Turčičová, J. Mandel, and K. Eben. Score matching filters for Gaussian Markov fields with a linear model of the precision matrix. Paper submitted.

(ii) M. Turčičová and K. Eben: Odhad varianční matice ve vysoké dimenzi (in Czech). *Informační bulletin České statistické společnosti*, 31(4):24–38, 2020. doi: 10.5300/IB.

(iii) M. Turčičová, J. Mandel and K. Eben: Multilevel maximum likelihood estimation with application to covariance matrices. *Communications in Statistics - Theory and Methods*, 48(4): 909–925, 2019. doi: 10.1080/03610926.2017.1422755.

(iv) M. Turčičová, J. Mandel and K. Eben: Stability of the Spectral EnKF under nested covariance estimators. *Proceedings of the 20th European Young Statisticians Meeting*, p. 137–142, 2017.

(v) M. Turčičová, J. Mandel and K. Eben: Maximum likelihood estimation of a diagonal covariance matrix. *Institute of Computer Science, Technical report No. V-1228*, 2016.

(vi) M. Turčičová, J. Mandel and K. Eben: Covariance Modeling by Means of Eigenfunctions of Laplace Operator. *In JSM Proceedings, Section on Statistics and the Environment.* Alexandria, VA: American Statistical Association. p. 3454–3461, 2015.

A. Appendix

A.1 Computing the optimal value (7.13) of the score matching objective function

For a sample $\mathbb{X}_N = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N\}$, the precision matrix model selection in Section 7.4.2 is based on the sample version (4.14) of the score for normal distribution, where $\nabla_{\boldsymbol{x}}^\top b(\boldsymbol{X}) = \mathbf{0}$,

$$\mathcal{S}_N(\boldsymbol{\eta}|\mathbb{X}_N) = \sum_{i=1}^{N} \left(\frac{1}{2} \boldsymbol{\eta}^\top D^*(\boldsymbol{X}_i) D(\boldsymbol{X}_i) \boldsymbol{\eta} + \boldsymbol{\eta}^\top \Delta_{\boldsymbol{x}} T(\boldsymbol{X}_i) \right) + c_N^*(\mathbb{X}_N), \tag{A.1}$$

where the constant $c_N^*(\mathbb{X}_N)$ does not depend on parameter and $\boldsymbol{\eta}$ is defined in (4.43) as

$$\boldsymbol{\eta} = \begin{pmatrix} \boldsymbol{\eta}_1 \\ \boldsymbol{\eta}_2 \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^{r} \beta_k A_k \boldsymbol{\mu} \\ \boldsymbol{\beta} \end{pmatrix}. \tag{A.2}$$

By substituting for $D^*(\boldsymbol{X}_i) D(\boldsymbol{X}_i)$ from (4.45), we get

$$\begin{aligned} \boldsymbol{\eta}^\top D^*(\boldsymbol{X}_i) D(\boldsymbol{X}_i) \boldsymbol{\eta} = \boldsymbol{\eta}_1^\top \boldsymbol{\eta}_1 - \sum_{j=1}^{r} \eta_{2j} \boldsymbol{\eta}_1^\top A_j \boldsymbol{X}_i - \sum_{j=1}^{r} \eta_{2j} \boldsymbol{X}_i^\top A_j \boldsymbol{\eta}_1 + \\ + \sum_{j=1}^{r} \sum_{k=1}^{r} \eta_{2j} \eta_{2k} \boldsymbol{X}_i^\top A_j A_k \boldsymbol{X}_i. \end{aligned} \tag{A.3}$$

By substituting for $\boldsymbol{\eta}$ from (A.2), the terms in (A.3) are

$$\boldsymbol{\eta}_1^\top \boldsymbol{\eta}_1 = \sum_{k=1}^{r} \sum_{j=1}^{r} \beta_k \beta_j \boldsymbol{\mu}^\top A_k A_j \boldsymbol{\mu}, \tag{A.4}$$

$$\sum_{j=1}^{r} \eta_{2j} \boldsymbol{\eta}_1^\top A_j \boldsymbol{X}_i = \sum_{k=1}^{r} \sum_{j=1}^{r} \beta_k \beta_j \boldsymbol{\mu}^\top A_k A_j \boldsymbol{X}_i, \tag{A.5}$$

$$\sum_{j=1}^{r} \eta_{2j} \boldsymbol{X}_i^\top A_j \boldsymbol{\eta}_1 = \sum_{k=1}^{r} \sum_{j=1}^{r} \beta_j \beta_k \boldsymbol{X}_i^\top A_j A_k \boldsymbol{\mu}, \tag{A.6}$$

$$\sum_{j=1}^{r} \sum_{k=1}^{r} \eta_{2j} \eta_{2k} \boldsymbol{X}_i^\top A_j A_k \boldsymbol{X}_i = \sum_{j=1}^{r} \sum_{k=1}^{r} \beta_j \beta_k \boldsymbol{X}_i^\top A_j A_k \boldsymbol{X}_i. \tag{A.7}$$

By using (4.46) for $\Delta_{\boldsymbol{x}} T(\boldsymbol{X}_i)$, the second term of the sum in (A.1) is

$$\boldsymbol{\eta}^\top \Delta_{\boldsymbol{x}} T(\boldsymbol{X}) = (\boldsymbol{\eta}_1^\top, \boldsymbol{\eta}_2^\top) \begin{pmatrix} \mathbf{0} \\ [-\operatorname{tr} A_k]_{k=1}^{r} \end{pmatrix} = -\sum_{k=1}^{r} \beta_k \operatorname{tr} A_k. \tag{A.8}$$

Now, we can rewrite the objective function $\mathcal{S}_N(\boldsymbol{\eta}|\mathbb{X}_N)$ in terms of $(\boldsymbol{\mu}, \boldsymbol{\beta})$ by substituting (A.4)-(A.8) into (A.1):

$$\begin{aligned}
\mathcal{S}_N(\boldsymbol{\mu}, \boldsymbol{\beta}|\mathbb{X}_N) - c_N^*(\mathbb{X}_N) &= \sum_{i=1}^{N} \frac{1}{2} \Big[\sum_{k=1}^{r} \sum_{j=1}^{r} \beta_k \beta_j \boldsymbol{\mu}^\top A_k A_j (\boldsymbol{\mu} - \boldsymbol{X}_i) - \\
&\qquad - \sum_{k=1}^{r} \sum_{j=1}^{r} \beta_k \beta_j \boldsymbol{X}_i^\top A_k A_j (\boldsymbol{\mu} - \boldsymbol{X}_i) \Big] - N \sum_{k=1}^{r} \beta_k \operatorname{tr}(A_k) \\
&= \sum_{i=1}^{N} \frac{1}{2} \sum_{k=1}^{r} \sum_{j=1}^{r} \beta_k \beta_j (\boldsymbol{\mu}^\top - \boldsymbol{X}_i^\top) A_k A_j (\boldsymbol{\mu} - \boldsymbol{X}_i) - N \sum_{k=1}^{r} \beta_k \operatorname{tr}(A_k) \\
&= \sum_{i=1}^{N} \frac{1}{2} \sum_{k=1}^{r} \sum_{j=1}^{r} \beta_k \operatorname{tr}\left(A_k A_j (\boldsymbol{\mu} - \boldsymbol{X}_i)(\boldsymbol{\mu} - \boldsymbol{X}_i)^\top \right) \beta_j - N \sum_{k=1}^{r} \beta_k \operatorname{tr}(A_k) \\
&= \frac{N}{2} \boldsymbol{\beta}^\top \left[\operatorname{tr}\left(A_k A_j \frac{1}{N} \sum_{i=1}^{N} (\boldsymbol{\mu} - \boldsymbol{X}_i)(\boldsymbol{\mu} - \boldsymbol{X}_i)^\top \right) \right]_{k,j=1}^{r} \boldsymbol{\beta} - N \boldsymbol{\beta}^\top \left[\operatorname{tr}(A_k) \right]_{k=1}^{r}.
\end{aligned}$$

By evaluating $\mathcal{S}_N(\boldsymbol{\mu}, \boldsymbol{\beta}|\mathbb{X}_N)$ at $(\bar{\mathbf{X}}, \hat{\boldsymbol{\beta}})$, which represents arguments of its maximum, we get its optimal value

$$\mathcal{S}_N\left(\bar{\boldsymbol{X}}, \hat{\boldsymbol{\beta}}|\mathbb{X}_N\right) = \frac{N}{2} \hat{\boldsymbol{\beta}}^\top \left[\operatorname{tr}\left(A_k A_j S \right) \right]_{k,j=1}^{r} \hat{\boldsymbol{\beta}} - N \hat{\boldsymbol{\beta}}^\top \left[\operatorname{tr}(A_k) \right]_{k=1}^{r} + c_N^*(\mathbb{X}_N),$$

where S denotes the sample covariance matrix computed from $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N$.

From (4.40),

$$\left[\operatorname{tr}\left(S A_k A_l \right) \right]_{k,l=1}^{r} \hat{\boldsymbol{\beta}} = \left[\operatorname{tr}(A_k) \right]_{k=1}^{r},$$

and therefore,

$$\begin{aligned}
\frac{1}{N} \mathcal{S}_N\left(\bar{\boldsymbol{X}}, \hat{\boldsymbol{\beta}}|\mathbb{X}_N\right) - \frac{1}{N} c_N^*(\mathbb{X}_N) &= \frac{1}{2} \operatorname{tr}\left(\hat{\boldsymbol{\beta}}^\top \left[\operatorname{tr}(A_k) \right]_{k=1}^{r} \right) - \hat{\boldsymbol{\beta}}^\top \left[\operatorname{tr}(A_k) \right]_{k=1}^{r} \\
&= \frac{1}{2} \hat{\boldsymbol{\beta}}^\top \left[\operatorname{tr}(A_k) \right]_{k=1}^{r} - \hat{\boldsymbol{\beta}}^\top \left[\operatorname{tr}(A_k) \right]_{k=1}^{r} \\
&= -\frac{1}{2} \hat{\boldsymbol{\beta}}^\top \left[\operatorname{tr}(A_k) \right]_{k=1}^{r}.
\end{aligned}$$