Multilevel maximum likelihood estimation with application to covariance matrices

Marie Turčičová, Jan Mandel & Kryštof Eben

Published online: 23 Jan 2018.

Submit your article to this journal ↗

Article views: 138

View Crossmark data ↗

Taylor & Francis
Taylor & Francis Group

Check for updates

Multilevel maximum likelihood estimation with application to covariance matrices

Marie Turčičová ®[a], Jan Mandel ®[b], and Kryštof Eben ®[c]

[a]Institute of Computer Science, Academy of Sciences of the Czech Republic Pod Vodárenskou věží 271/2, 182 07 Praha 8, Czech Republic, and Charles University in Prague, Faculty of Mathematics and Physics, Sokolovská 83, Prague 8, 186 75, Czech Republic; [b]University of Colorado Denver, Denver, CO 80217-3364, USA, and Institute of Computer Science, Academy of Sciences of the Czech Republic Pod Vodárenskou věží 271/2, 182 07 Praha 8, Czech Republic; [c]Institute of Computer Science, Academy of Sciences of the Czech Republic Pod Vodárenskou věží 271/2, 182 07 Praha 8, Czech Republic

ABSTRACT

The asymptotic variance of the maximum likelihood estimate is proved to decrease when the maximization is restricted to a subspace that contains the true parameter value. Maximum likelihood estimation allows a systematic fitting of covariance models to the sample, which is important in data assimilation. The hierarchical maximum likelihood approach is applied to the spectral diagonal covariance model with different parameterizations of eigenvalue decay, and to the sparse inverse covariance model with specified parameter values on different sets of nonzero entries. It is shown computationally that using smaller sets of parameters can decrease the sampling noise in high dimension substantially.

1. Introduction

Estimation of large covariance matrices from small samples is an important problem in many fields, including spatial statistics, genomics, and ensemble filtering. One of the prominent applications is data assimilation in meteorology and oceanography, where the dimension of state vector describing the atmosphere or ocean is in order of millions or larger. Every practically available sample is a small sample in this context, since a reasonable approximation of the full covariance can be obtained only with sample size of the order of the dimension of the problem (Vershynin 2012). In practice, the sample covariance[1] is singular and polluted by spurious correlations. Nevertheless, it carries useful information (e.g. on covariances present in the actual atmospheric flow) and different techniques can be applied in order to improve the covariance model and its practical performance.

One common technique is shrinkage, that is, a linear combination of sample covariance and a positive definite target matrix, which prevents the covariance from being singular. The target matrix embodies some prior information about the covariance; it can be, e.g., unit diagonal or, more generally, positive diagonal (Ledoit and Wolf 2004). See, e.g., Schäfer and

---

CONTACT Jan Mandel ✉ jan.mandel@gmail.com ▣ Mathematical and Statistical Sciences, University of Colorado Denver, PO Box 173364, Denver 80217-3364, United States.

[1] In this paper, by sample covariance we mean the maximum likelihood estimate of covariance matrix using the norming constant N as opposed to the unbiased estimate with norming constant $(N-1)$.

Strimmer (2005) for a survey of such shrinkage approaches. Shrinkage of sample covariance towards a fixed covariance matrix based on a specific model and estimated from historical data (called background covariance) was used successfully in meteorology (Hamill and Snyder 2000; Wang et al. 2008). This approach is justified as one which combines actual (called flow-dependent) and long-term average (called climatologic) information on spatial covariances present in the 3D meteorological fields.

Another approach to improving on the sample covariance matrix is localization by suppressing long-range spurious correlations, which is commonly done by multiplying the sample covariance matrix term by term by a gradual cutoff matrix (Buehner and Charron 2007; Furrer and Bengtsson 2007) to suppress off-diagonal entries. The extreme case, when only the diagonal is left, is particularly advantageous in the spectral domain, as the covariance of a random field in Fourier space is diagonal if and only if the random field in cartesian geometry is second order stationary, i.e., the covariance between the values at two points depends only on their distance vector. Alternatively, diagonal covariance in a wavelet basis provides spatial variability as well (Pannekoucke, Berre, and Desroziers 2007). Spectral diagonal covariance models were successfully used in operational statistical interpolation in meteorology in spherical geometry (Parrish and Derber 1992), and versions of Ensemble Kalman Filter (EnKF) were developed which construct diagonal covariance in Fourier or wavelet space in every update step of the filter at low cost, and can operate successfully with small ensembles (Mandel, Beezley, and Kondratenko 2010; Beezley, Mandel, and Cobb 2011; Kasanický, Mandel and Vejmelka 2015).

Sparse covariance models, such as the spectral diagonal, allow a compromise between realistic assumptions and cheap computations. Another covariance model taking advantage of sparsity is a Gauss-Markov Random Field (GMRF), based on the fact that conditional independence of variables implies zero corresponding elements in the inverse of the covariance matrix (Rue and Held 2005), which leads to modeling the covariance as the inverse of a sparse matrix.

However, both spectral diagonal and sparse inverse covariance models have a large number of parameters, namely all terms of the sparse matrix (up to symmetry) which are allowed to attain nonzero values. This results in overfitting and significant sampling noise for small samples. Therefore, it is of interest to reduce the number of parameters by adopting additional, problem-dependent assumptions on the true parameter values.

The principal result of this paper is the observation that if parameters are fitted as the Maximum Likelihood Estimator (MLE) and the additional assumptions are satisfied by the true parameters, then the estimate using fewer parameters is asymptotically more accurate, and often very significantly so even for small samples.

The paper is organized as follows. In Section 2, we provide a brief statement of MLE and its asymptotic variance. In Section 3, we use the theory of maximum likelihood estimation to prove that for any two nested subspaces of the parametric space containing the true parameter, the asymptotic covariance matrix of the MLE is smaller for the smaller parameter space. These results hold for a general parameter and, in the special case of MLE for covariance matrices we do not need any invertibility assumption. The applications to estimation of covariance matrices by spectral diagonal and GMRF are presented in Section 4, and Section 5 contains computational illustrations. A comparison of the performance of MLE for parametric models and of related shrinkage estimators is in Section 6.

2. Asymptotic variance of the maximum likelihood estimator

First, we briefly review some standard results for reference. Suppose $\mathbb{X}_N = [\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N]$ 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 $\Theta \subset \mathbb{R}^p$. 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

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

We adopt the usual assumptions that (i) the true parameter $\boldsymbol{\theta}^0$ lies in the interior of Θ, (ii) 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.s. if and only if $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$, and (iii) $f(\boldsymbol{x}, \boldsymbol{\theta})$ is a sufficiently smooth function of $\boldsymbol{x}$ and $\boldsymbol{\theta}$. Then the error of the estimate is asymptotically normal,

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}^0) \xrightarrow{d} \mathcal{N}_p(\boldsymbol{0}, Q_{\boldsymbol{\theta}^0}), \text{ as } N \to \infty, \tag{2}$$

where

$$Q_{\boldsymbol{\theta}^0} = J_{\boldsymbol{\theta}^0}^{-1}, \quad J_{\boldsymbol{\theta}^0} = \mathrm{E}(\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^0|\boldsymbol{X})^\top \nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^0|\boldsymbol{X})), \quad \boldsymbol{X} \sim f(\boldsymbol{x}, \boldsymbol{\theta}^0). \tag{3}$$

The matrix $J_{\boldsymbol{\theta}^0}$ is called the Fisher information matrix for the parameterization $\boldsymbol{\theta}^0$. 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 Jacobi matrices below. The mean value in (3) is taken with respect to $\boldsymbol{X}$, which is the only random quantity in (3). Cf., e.g., (Lehmann and Casella 1998, Theorem 5.1) for details.

3. Nested maximum likelihood estimators

Now suppose that we have an additional information that the true parameter $\boldsymbol{\theta}^0$ lies in a subspace of Θ, which is parameterized by $k \leq p$ parameters $(\varphi_1, \ldots, \varphi_k)^\top = \boldsymbol{\varphi}$. Denote by $\nabla_{\boldsymbol{\varphi}}\boldsymbol{\theta}(\boldsymbol{\varphi})$ the $p \times k$ Jacobi matrix with entries $\frac{\partial \theta_i}{\partial \varphi_j}$. In the next theorem, we derive the asymptotic covariance of the maximum likelihood estimator for $\boldsymbol{\varphi}$,

$$\hat{\boldsymbol{\varphi}}_N = \arg\max_{\boldsymbol{\varphi}} \ell(\boldsymbol{\varphi}|\mathbb{X}_N), \quad \ell(\boldsymbol{\varphi}|\mathbb{X}_N) = \sum_{i=1}^N \ell(\boldsymbol{\varphi}|\boldsymbol{X}_i), \quad \ell(\boldsymbol{\varphi}|\boldsymbol{x}) = \log f(\boldsymbol{x}, \boldsymbol{\theta}(\boldsymbol{\varphi})), \tag{4}$$

based on the asymptotic covariance of $\boldsymbol{\theta}$ in (2).

Theorem 1. *Assume that the map $\boldsymbol{\varphi} \mapsto \boldsymbol{\theta}(\boldsymbol{\varphi})$ is one-to-one from $\Phi \subset \mathbb{R}^k$ to Θ, the map $\boldsymbol{\varphi} \mapsto \boldsymbol{\theta}(\boldsymbol{\varphi})$ is continuously differentiable, $\nabla_{\boldsymbol{\varphi}}\boldsymbol{\theta}(\boldsymbol{\varphi})$ is full rank for all $\boldsymbol{\varphi} \in \Phi$, and $\boldsymbol{\theta}^0 = \boldsymbol{\theta}(\boldsymbol{\varphi}^0)$ with $\boldsymbol{\varphi}^0$ in the interior of Φ. Then,*

$$\sqrt{N}(\hat{\boldsymbol{\varphi}}_N - \boldsymbol{\varphi}^0) \xrightarrow{d} \mathcal{N}_k(\boldsymbol{0}, Q_{\boldsymbol{\varphi}^0}) \text{ as } N \to \infty, \tag{5}$$

where $Q_{\boldsymbol{\varphi}^0} = J_{\boldsymbol{\varphi}^0}^{-1}$, with $J_{\boldsymbol{\varphi}^0}$ the Fisher information matrix of the parameterization $\boldsymbol{\varphi}$ given by

$$J_{\boldsymbol{\varphi}^0} = \nabla_{\boldsymbol{\varphi}}\boldsymbol{\theta}(\boldsymbol{\varphi}^0)^\top J_{\boldsymbol{\theta}^0} \nabla_{\boldsymbol{\varphi}}\boldsymbol{\theta}(\boldsymbol{\varphi}^0). \tag{6}$$

Proof. From (3) and the chain rule

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

we have

$$\begin{aligned}
J_{\boldsymbol{\varphi}^0} &= \mathrm{E}(\nabla_{\boldsymbol{\varphi}}\ell(\boldsymbol{\varphi}^0|\boldsymbol{X})^{\top}\nabla_{\boldsymbol{\varphi}}\ell(\boldsymbol{\varphi}^0|\boldsymbol{X})) \\
&= \nabla_{\boldsymbol{\varphi}}\boldsymbol{\theta}(\boldsymbol{\varphi}^0)^{\top}\,\mathrm{E}(\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^0|\boldsymbol{X})^{\top}\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^0|\boldsymbol{X}))\nabla_{\boldsymbol{\varphi}}\boldsymbol{\theta}(\boldsymbol{\varphi}^0) \\
&= \nabla_{\boldsymbol{\varphi}}\boldsymbol{\theta}(\boldsymbol{\varphi}^0)^{\top}J_{\boldsymbol{\theta}^0}\nabla_{\boldsymbol{\varphi}}\boldsymbol{\theta}(\boldsymbol{\varphi}^0).
\end{aligned}$$

The asymptotic distribution (5) is now (2) applied to $\boldsymbol{\varphi}$. $\qquad\square$

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 2. *Under the assumptions of Theorem 1,*

$$\sqrt{N}(\boldsymbol{\theta}(\hat{\boldsymbol{\varphi}}_N) - \boldsymbol{\theta}^0) \overset{d}{\to} \mathcal{N}_p(\mathbf{0}, Q_{\boldsymbol{\theta}(\boldsymbol{\varphi}^0)}) \quad \text{as } N \to \infty, \tag{7}$$

where

$$Q_{\boldsymbol{\theta}(\boldsymbol{\varphi}^0)} = \nabla_{\boldsymbol{\varphi}}\boldsymbol{\theta}(\boldsymbol{\varphi}^0)J_{\boldsymbol{\varphi}^0}^{-1}\nabla_{\boldsymbol{\varphi}}\boldsymbol{\theta}(\boldsymbol{\varphi}^0)^{\top} = \nabla_{\boldsymbol{\varphi}}\boldsymbol{\theta}(\boldsymbol{\varphi}^0)(\nabla_{\boldsymbol{\varphi}}\boldsymbol{\theta}(\boldsymbol{\varphi}^0)^{\top}J_{\boldsymbol{\theta}^0}\nabla_{\boldsymbol{\varphi}}\boldsymbol{\theta}(\boldsymbol{\varphi}^0))^{-1}\nabla_{\boldsymbol{\varphi}}\boldsymbol{\theta}(\boldsymbol{\varphi}^0)^{\top}. \tag{8}$$

Proof. The lemma follows from (5) by the delta method (Rao 1973, p. 387), since the map $\boldsymbol{\varphi} \mapsto \boldsymbol{\theta}(\boldsymbol{\varphi})$ is continuously differentiable. $\qquad\square$

Remark 3. The matrix $Q_{\boldsymbol{\theta}(\boldsymbol{\varphi}^0)}$ is singular, so it cannot be written as the inverse of another matrix, but it can be understood as the inverse $J_{\boldsymbol{\theta}(\boldsymbol{\varphi}^0)}^{-1}$ of the Fisher information matrix for $\boldsymbol{\varphi}$, embedded in the larger parameter space Θ.

Suppose that $\boldsymbol{\psi}$ is another parameterization which satisfies the same assumption as $\boldsymbol{\varphi}$ in Theorem 1: the map $\boldsymbol{\psi} \mapsto \boldsymbol{\theta}(\boldsymbol{\psi})$ is one-to-one from $\Psi \subset \mathbb{R}^m$, $k \leq m \leq p$, to Θ, $\boldsymbol{\psi} \mapsto \boldsymbol{\theta}(\boldsymbol{\psi})$ is continuously differentiable, $\nabla_{\boldsymbol{\psi}}\boldsymbol{\theta}(\boldsymbol{\psi})$ is full rank for all $\boldsymbol{\psi} \in \Psi$, and $\boldsymbol{\theta}^0 = \boldsymbol{\theta}(\boldsymbol{\psi}^0)$, where $\boldsymbol{\psi}^0$ is in the interior of Ψ. Then, similarly as in (7), we have also

$$\sqrt{N}(\boldsymbol{\theta}(\hat{\boldsymbol{\psi}}_N) - \boldsymbol{\theta}^0) \overset{d}{\to} \mathcal{N}_p(\mathbf{0}, Q_{\boldsymbol{\theta}(\boldsymbol{\psi}^0)}) \quad \text{as } N \to \infty, \tag{9}$$

where, as in (8),

$$Q_{\boldsymbol{\theta}(\boldsymbol{\psi}^0)} = \nabla_{\boldsymbol{\psi}}\boldsymbol{\theta}(\boldsymbol{\psi}^0)J_{\boldsymbol{\psi}^0}^{-1}\nabla_{\boldsymbol{\psi}}\boldsymbol{\theta}(\boldsymbol{\psi}^0)^{\top} = \nabla_{\boldsymbol{\psi}}\boldsymbol{\theta}(\boldsymbol{\psi}^0)(\nabla_{\boldsymbol{\psi}}\boldsymbol{\theta}(\boldsymbol{\psi}^0)^{\top}J_{\boldsymbol{\theta}^0}\nabla_{\boldsymbol{\psi}}\boldsymbol{\theta}(\boldsymbol{\psi}^0))^{-1}\nabla_{\boldsymbol{\psi}}\boldsymbol{\theta}(\boldsymbol{\psi}^0)^{\top}. \tag{10}$$

The next theorem shows that when we have two parameterizations $\boldsymbol{\varphi}$ and $\boldsymbol{\psi}$ which are nested, then the smaller parameterization has smaller or equal asymptotic covariance than the larger one. For symmetric matrices A and B, $A \leq B$ means that $A - B$ is positive semidefinite.

Theorem 4. *Suppose that $\boldsymbol{\varphi}$ and $\boldsymbol{\psi}$ satisfy the assumptions in Theorem 1, and there exists a differentiable mapping $\boldsymbol{\varphi} \mapsto \boldsymbol{\psi}$ from Φ to Ψ, such that $\boldsymbol{\varphi}^0 \mapsto \boldsymbol{\psi}^0$. Then,*

$$Q_{\boldsymbol{\theta}(\boldsymbol{\varphi}^0)} \leq Q_{\boldsymbol{\theta}(\boldsymbol{\psi}^0)}. \tag{11}$$

In addition, if $U \sim \mathcal{N}_p(\mathbf{0}, Q_{\boldsymbol{\theta}(\boldsymbol{\varphi}^0)})$ and $V \sim \mathcal{N}_p(\mathbf{0}, Q_{\boldsymbol{\theta}(\boldsymbol{\psi}^0)})$ are random vectors with the asymptotic distributions of the estimates $\boldsymbol{\theta}(\hat{\boldsymbol{\varphi}}_N)$ and $\boldsymbol{\theta}(\hat{\boldsymbol{\psi}}_N)$, then

$$\mathrm{E}\,|U|^2 = \frac{1}{N}\,\mathrm{Tr}\,Q_{\boldsymbol{\theta}(\boldsymbol{\varphi}^0)} \leq \frac{1}{N}\,\mathrm{Tr}\,Q_{\boldsymbol{\theta}(\boldsymbol{\psi}^0)} = \mathrm{E}\,|V|^2, \tag{12}$$

where $|V| = (V^{\top}V)^{1/2}$ is the standard Euclidean norm in $\mathbb{R}^p$.

Proof. Denote $A = J_{\theta^0}$, $B = \nabla_\varphi \theta(\varphi^0)$, $C = \nabla_\psi \theta(\psi^0)$. From the chain rule,

$$\nabla_\varphi \theta \left(\varphi^0\right) = \nabla_\psi \theta \left(\psi^0\right) \nabla_\varphi \psi \left(\varphi^0\right),$$

we have that $B = C\nabla_\varphi \psi(\varphi^0)$, and, consequently, Range $B \subset$ Range C. Define

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

The matrices P_B and P_C are symmetric and idempotent, hence they are orthogonal projections. In addition,

$$\text{Range } P_B = \text{Range } A^{1/2}B \subset \text{Range } A^{1/2}C = \text{Range } P_C.$$

Consequently, $P_B \leq P_C$ holds from standard properties of orthogonal projections, and (11) follows.

To prove (12), note that for random vector X with $\text{E}\, X = 0$ and finite second moment, $\text{E}\,|X|^2 = \text{Tr Cov}\, X$ from Karhunen-Loève decomposition and Parseval identity. The proof is concluded by using the fact that for symmetric matrices, $A \leq B$ implies $\text{Tr}\, A \leq \text{Tr}\, B$, cf. e.g., Carlen (2010). □

Remark 5. In the practically interesting cases when there is a large difference in the dimensions of the parameters φ and ψ, many eigenvalues in the covariance of the estimation error become zero. The computational tests in Section 5 show that the resulting decrease of the estimation error can be significant.

4. Application: Nested covariance models

Models of covariance (e.g., of the state vector in a numerical weather prediction model) and the quality of the estimated covariance are one of the key components of data assimilation algorithms. High dimension of the problem usually prohibits working with the covariance matrix explicitly. In ensemble filtering methods, this difficulty may be circumvented by working directly with the original small sample like in the classical Ensemble Kalman filter. This, however, effectively means using the sample covariance matrix with its rank deficiency and spurious correlations. Current filtering methods use shrinkage and localization as noted above, and ad hoc techniques for dimension reduction.

A reliable way towards effective filtering methods lies in introducing sparsity into covariance matrices or their inverses by means of suitable covariance models. The results of previous section suggest that it is beneficial to choose parsimonious models, and indeed, in practical application we often encounter models with a surprisingly low number of parameters.

A large class of covariance models which encompass sparsity in an efficient manner arises from Graphical models (Lauritzen 1996) and Gaussian Markov Random Fields (GMRF), (Rue and Held 2005), where a special structure of inverse covariance is assumed. In the area of GMRF, nested covariance models arise naturally. If, for instance, we consider a GMRF on a rectangular mesh, each gridpoint may have 4, 8, 12, 20 etc. neighbouring points which have nonzero corresponding element in the inverse covariance matrix. Thus, a block band-diagonal structure in the inverse covariance arises (Ueno and Tsuchiya 2009). The results of Section 3 apply for this case and we shall illustrate them in the simulation study of Section 5.

Finally, variational assimilation methods, which dominate today's practice of meteorological services, usually employ a covariance model based on a series of transformations leading to

independence of variables (Bannister 2008; Michel and Auligné 2010). At the end, this results in an estimation problem for normal distribution with a diagonal covariance matrix.

For both ensemble and variational methods, any additional knowledge can be used to improve the estimate of covariance. Second-order stationarity leads to diagonality in spectral space, diagonality in wavelet space is often a legitimate assumption (e.g., Pannekoucke, Berre, and Desroziers 2007) and we shall treat the diagonal case in more detail. Suppose

$$X \sim \mathcal{N}_n(\mathbf{0}, D), \tag{13}$$

where X denotes the random field after the appropriate transform and D is a diagonal matrix. It is clear that estimating D by the full sample covariance matrix (what would be the case when using the classical EnKF) is ineffective in this situation and it is natural to use only the diagonal part of the sample covariance. In practice, the resulting diagonal matrix may still turn out to be noisy (Kasanický, Mandel and Vejmelka 2015), and further assumptions like a certain type of decay of the diagonal entries may be realistic.

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

4.1. Sample covariance

The top-level parameter space Θ consists of all symmetric positive definite matrices, resulting in the parameterization Σ with $\frac{n(n+1)}{2}$ independent parameters. The likelihood of a sample $\mathbb{X}_N = [X^{(1)}, \ldots, X^{(N)}]$ from $\mathcal{N}_n(\mathbf{0}, \Sigma)$ is

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

If $N \geq n$, it is well known (e.g. Muirhead (2005), p. 83) that the likelihood is maximized at what we call here sample covariance matrix

$$\hat{\Sigma}_N = \frac{1}{N} \sum_{i=1}^N X^{(i)} (X^{(i)})^\top. \tag{14}$$

The Fisher information matrix of the sample covariance estimator is (Magnus and Neudecker 2007, p. 356)

$$J^{(0)}(\operatorname{vec}(\Sigma)) = \frac{1}{2} \Sigma^{-1} \otimes \Sigma^{-1},$$

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

$$\tilde{\Sigma}_N = \begin{bmatrix} \tilde{\Sigma}_{11} & 0 \\ 0 & 0 \end{bmatrix}, \quad \tilde{\Sigma}_{11} > 0.$$

Then $\lim_{a\to 0^+} \mathbb{X}_N^\top (\tilde{\Sigma}_N + aI)^{-1} \mathbb{X}_N$ exists, but $\lim_{a\to 0^+} \det(\tilde{\Sigma}_N + aI) = 0$, thus

$$\lim_{a\to 0^+} L(\tilde{\Sigma}_N + aI | \mathbb{X}_N) = \infty.$$

Note that when the likelihood is redefined in terms of the subspace $\text{span}(\mathbb{X}_N)$ only, the sample covariance can be obtained by maximization on the subspace (Rao 1973, p. 527).

When the true covariance is diagonal ($\Sigma \equiv D$, cf. (13)), a significant improvement can be achieved by setting the off-diagonal terms of sample covariance to zero,

$$\hat{D}_N^{(0)} = \text{diag}\left(\hat{\Sigma}_N\right). \tag{15}$$

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

$$\hat{\Sigma}_N^u = \frac{1}{N-1} \sum_{i=1}^N X^{(i)} \left(X^{(i)}\right)^\top$$

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

$$\left| \hat{D}_N^{(0)} - D \right|_F \leq \left| \hat{\Sigma}_N^u - D \right|_F \tag{16}$$

cf. Furrer and Bengtsson (2007) for the case when the mean is assumed to be known like here, and Kasanický, Mandel and Vejmelka (2015) for the unbiased sample covariance and unknown mean.

4.2. Diagonal covariance

The parameter space Θ_1 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 $D = \text{diag}(d_1, \ldots, d_n)$ with a given random sample $\mathbb{X}_N = [X^{(1)}, \ldots, X^{(N)}]$ from $\mathcal{N}_n(\mathbf{0}, D)$ is

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

and has its maximum at

$$\hat{d}_j = \frac{1}{N} \sum_{k=1}^N \left(X_j^{(k)}\right)^2, \quad j = 1, \ldots, n, \tag{17}$$

where $X_j^{(k)}$ denotes the j-th entry of $X^{(k)}$. The sum of squares $S_j^2 = \sum_{k=1}^N (X_j^{(k)})^2$ is a sufficient statistic for the variance d_j. Thus, we get the maximum likelihood estimator

$$\hat{D}_N^{(1)} = \frac{1}{N} \text{diag}\left(S_1^2, \ldots, S_n^2\right). \tag{18}$$

It is easy to compute the Fisher information matrix explicitly,

$$J_{D^{(1)}} = \text{diag}\left(\frac{1}{2d_1^2}, \ldots, \frac{1}{2d_n^2}\right). \tag{19}$$

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

$$\frac{1}{N} Q_{D^{(1)}} = \frac{1}{N} J_{D^{(1)}}^{-1} = \frac{1}{N} \text{diag}\left(2d_1^2, \ldots, 2d_n^2\right)$$

from (2).

4.3. Diagonal covariance with prescribed decay by 3 parameters

A more specific situation appears when we have an additional information that the matrix D 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 $d_i = ((c_1 + c_2 h_i) f_i(\alpha))^{-1}, i = 1, \ldots, n$, where c_1, c_2 and α are unknown parameters, $h_1, \ldots, h_n$ are known positive numbers, and $f_1, \ldots, f_n$ are known differentiable functions. For easier computation it is useful to work with $\tau_i = \frac{1}{d_i} = (c_1 + c_2 h_i) f_i(\alpha)$. Maximum likelihood estimators for c_1, c_2, and α can be computed effectively from the likelihood

$$\ell(D|\mathbb{X}_N) = -\frac{N}{2} n \log(2\pi) + \frac{N}{2} \sum_{i=1}^{n} \log \tau_i - \frac{1}{2} \sum_{i=1}^{n} \tau_i S_i^2 \qquad (20)$$

by using the chain rule. It holds that

$$\frac{\partial \ell}{\partial c_1} = \sum_{i=1}^{n} \frac{\partial \ell}{\partial \tau_i} \frac{\partial \tau_i}{\partial c_1} = \sum_{i=1}^{n} \left(\frac{N}{2\tau_i} - \frac{S_i^2}{2} \right) \frac{\partial \tau_i}{\partial c_1}$$

$$= \frac{N}{2} \sum_{i=1}^{n} \left(\frac{1}{(c_1 + c_2 h_i) f_i(\alpha)} - \frac{1}{N} S_i^2 \right) f_i(\alpha).$$

Setting this derivative equal to zero we get

$$\sum_{i=1}^{n} \left(\frac{1}{c_1 + c_2 h_i} - \frac{1}{N} S_i^2 f_i(\alpha) \right) = 0. \qquad (21)$$

Analogously,

$$\frac{\partial \ell}{\partial c_2} = \sum_{i=1}^{n} \frac{\partial \ell}{\partial \tau_i} \frac{\partial \tau_i}{\partial c_2} = \frac{N}{2} \sum_{i=1}^{n} \left(\frac{1}{(c_1 + c_2 h_i) f_i(\alpha)} - \frac{1}{N} S_i^2 \right) h_i f_i(\alpha),$$

so the equation for estimating the parameter c_2 is

$$\sum_{i=1}^{n} \left(\frac{h_i}{c_1 + c_2 h_i} - \frac{1}{N} S_i^2 h_i f_i(\alpha) \right) = 0. \qquad (22)$$

Similarly,

$$\frac{\partial \ell}{\partial \alpha} = \sum_{i=1}^{n} \frac{\partial \ell}{\partial \tau_i} \frac{\partial \tau_i}{\partial \alpha} = \frac{N}{2} \sum_{i=1}^{n} \left(\frac{1}{(c_1 + c_2 h_i) f_i(\alpha)} - \frac{1}{N} S_i^2 \right) (c_1 + c_2 h_i) \frac{\partial f_i(\alpha)}{\partial \alpha}$$

$$= \frac{N}{2} \sum_{i=1}^{n} \left(\frac{1}{f_i(\alpha)} - \frac{1}{N} S_i^2 (c_1 + c_2 h_i) \right) \frac{\partial f_i(\alpha)}{\partial \alpha}$$

and setting the derivative to zero, we get

$$\sum_{i=1}^{n} \left(\frac{1}{f_i(\alpha)} \frac{\partial f_i(\alpha)}{\partial \alpha} - \frac{1}{N} S_i^2 (c_1 + c_2 h_i) \frac{\partial f_i(\alpha)}{\partial \alpha} \right) = 0. \qquad (23)$$

The maximum likelihood estimator for D is then given by

$$\hat{D}^{(3)} = \text{diag}\{((\hat{c}_1 + \hat{c}_2 h_i) f_i(\hat{\alpha}))^{-1}, i = 1, \ldots, n\}, \qquad (24)$$

where $(\hat{c}_1, \hat{c}_2, \hat{\alpha})$ is the solution of the system (21), (22), (23). This expression corresponds to searching a maximum likelihood estimator of D in the subspace $\Theta_3 \subset \Theta_1 \subset \Theta$ formed by diagonal matrices $\mathrm{diag}\{((c_1 + c_2 h_i) f_i(\alpha))^{-1}, i = 1, \ldots, n\}$.

For completeness, the asymptotic covariance of the estimation error about

$$D^{(3)} = \mathrm{diag}\{d_i(c_1, c_2, \alpha), i = 1, \ldots, n\},$$

contained in $\mathbb{X}_N$ is

$$\frac{1}{N} Q_{D^{(3)}} = \frac{1}{N} \nabla \boldsymbol{d}(c_1, c_2, \alpha) J^{-1}_{c_1, c_2, \alpha} \nabla \boldsymbol{d}(c_1, c_2, \alpha)^\top \tag{25}$$

from (8), where the Fisher information matrix $J_{c_1, c_2, \alpha}$ is the 3×3 matrix

$$
J_{c_1, c_2, \alpha}
$$
$$
= \begin{bmatrix}
\frac{1}{2} \sum_{i=1}^{n} \frac{1}{(c_1 + c_2 h_i)^2} & \frac{1}{2} \sum_{i=1}^{n} \frac{h_i}{(c_1 + c_2 h_i)^2} & \frac{1}{2} \sum_{i=1}^{n} \frac{1}{(c_1 + c_2 h_i) f_i(\alpha)} \frac{\partial f_i(\alpha)}{\partial \alpha} \\
\frac{1}{2} \sum_{i=1}^{n} \frac{h_i}{(c_1 + c_2 h_i)^2} & \frac{1}{2} \sum_{i=1}^{n} \frac{h_i^2}{(c_1 + c_2 h_i)^2} & \frac{1}{2} \sum_{i=1}^{n} \frac{h_i}{(c_1 + c_2 h_i) f_i(\alpha)} \frac{\partial f_i(\alpha)}{\partial \alpha} \\
\frac{1}{2} \sum_{i=1}^{n} \frac{1}{(c_1 + c_2 h_i) f_i(\alpha)} \frac{\partial f_i(\alpha)}{\partial \alpha} & \frac{1}{2} \sum_{i=1}^{n} \frac{h_i}{(c_1 + c_2 h_i) f_i(\alpha)} \frac{\partial f_i(\alpha)}{\partial \alpha} & \frac{1}{2} \sum_{i=1}^{n} \frac{1}{f_i^2(\alpha)} \left(\frac{\partial f_i(\alpha)}{\partial \alpha} \right)^2
\end{bmatrix}
$$

and

$$
\begin{aligned}
\boldsymbol{d}(c_1, c_2, \alpha) &= [d_1(c_1, c_2, \alpha), \ldots, d_n(c_1, c_2, \alpha)]^\top \\
&= \left[((c_1 + c_2 h_1) f_1(\alpha))^{-1}, \ldots, ((c_1 + c_2 h_n) f_n(\alpha))^{-1} \right]^\top.
\end{aligned}
$$

4.4. Diagonal covariance with prescribed decay by 2 parameters

We may consider a more specific model for diagonal elements with two parameters: $d_i = (c f_i(\alpha))^{-1}$, i.e. $\tau_i = c f_i(\alpha), i = 1, \ldots, n$, where c and α are unknown parameters. Maximum likelihood estimators for c and α can be computed similarly as in the previous case. The estimating equations have the form

$$\frac{1}{c} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{N} S_i^2 f_i(\alpha)$$

$$\frac{1}{c} \sum_{i=1}^{n} \frac{1}{f_i(\alpha)} \frac{\partial f_i(\alpha)}{\partial \alpha} = \sum_{i=1}^{n} \frac{1}{N} S_i^2 \frac{\partial f_i(\alpha)}{\partial \alpha},$$

which can be rearranged to

$$\frac{1}{c} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{N} S_i^2 f_i(\alpha) \tag{26}$$

$$0 = \sum_{i=1}^{n} S_i^2 f_i(\alpha) \left(\frac{1}{f_i(\alpha)} \frac{\partial f_i(\alpha)}{\partial \alpha} - \frac{1}{n} \sum_{j=1}^{n} \frac{1}{f_j(\alpha)} \frac{\partial f_j(\alpha)}{\partial \alpha} \right). \tag{27}$$

Equation (27) is an implicit formula for estimating α. Its result can be used for estimating c through (26). The maximum likelihood estimator for D is then given by

$$\hat{D}^{(2)} = \mathrm{diag}\left((\hat{c} f_1(\hat{\alpha}))^{-1}, \ldots, (\hat{c} f_n(\hat{\alpha}))^{-1} \right), \tag{28}$$

where $\hat{c}$ and $\hat{\alpha}$ are MLEs of c and α. It corresponds to searching a maximum likelihood estimator of D in the subspace $\Theta_2 \subset \Theta_3 \subset \Theta_1 \subset \Theta$ formed by diagonal matrices

diag$\{(cf_i(\alpha))^{-1}, i = 1, \ldots, n\}$. Of course, the estimator $\hat{D}^{(2)}$ does not have "larger" variance than $\hat{D}^{(3)}$.

The covariance of the asymptotic distribution of the parameters $d_1, \ldots, d_n$ is

$$\frac{1}{N} Q_{D^{(2)}} = \frac{1}{N} \nabla \boldsymbol{d}(c, \alpha) J_{c,\alpha}^{-1} \nabla \boldsymbol{d}(c, \alpha)^\top, \tag{29}$$

from (8), where Fisher information matrix at $D = \text{diag}\{d_i(c, \alpha), i = 1, \ldots, n\}$ is the 2×2 matrix

$$J_{c,\alpha} = \begin{bmatrix} \frac{n}{2c^2} & \frac{1}{2c} \sum_{i=1}^n \frac{1}{f_i(\alpha)} \frac{\partial f_i(\alpha)}{\partial \alpha} \\ \frac{1}{2c} \sum_{i=1}^n \frac{1}{f_i(\alpha)} \frac{\partial f_i(\alpha)}{\partial \alpha} & \frac{1}{2} \sum_{i=1}^n \frac{1}{f_i^2(\alpha)} \left(\frac{\partial f_i(\alpha)}{\partial \alpha} \right)^2 \end{bmatrix}$$

and $\boldsymbol{d}(c, \alpha) = [d_1(c, \alpha), \ldots, d_n(c, \alpha)]^\top = [(cf_1(\alpha))^{-1}, \ldots, (cf_n(\alpha))^{-1}]^\top$.

4.5. Sparse inverse covariance and GMRF

In the GMRF method for fields on a rectangular mesh, we assume that a variable on a grid-point is conditionally independent on the rest of the gridpoints, given values on neighboring gridpoints. It follows that nonzero entries in the inverse of the covariance matrix can be only between neighbor gridpoints. We start with 4 neighbors (up, down, right, left), and adding neighbors 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.

The inverse covariance model fitted by MLE was introduced by Ueno and Tsuchiya (2009) and applied on data from oceanography. The corresponding Fisher information matrix may be found as the negative of the Hessian (Ueno and Tsuchiya 2009, eq. (C17)).

5. Computational study

In Section 3, we have 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, we illustrate this behavior by means of simulations.

5.1. Simulation of simple GMRF

We first show that in the case of GMRF with four neighbors per gridpoint, adding dependencies (parameters) which are not present brings a loss of precision of the MLE. Using the sample covariance in this case causes a substantial error.

We have generated an ensemble of realizations of a GMRF with dimensions 10×10 (resulting in $n = 100$) and inverse covariance structure as in Figure 1. The values on the diagonals of the covariance matrix have been set to constant, since we assume the correlation with left and right neighbor to be identical, as well as the correlation with upper and lower neighbor (by symmetry of the covariance matrix and isotropy in both directions of the field, but different correlation in each direction). This leads to a model with 3 parameters for 4 neighbors, 5 parameters for 8 neighbors and 7 parameters for 12 neighbors,

The covariance structure of Σ^{-1} with 4 neighbors was set as "truth" and random samples were generated from $\mathcal{N}_n(\boldsymbol{0}, \Sigma)$ with sample sizes $N = 10, 15, 20, \ldots, 55$. The values on first, second and tenth diagonal have been set as 5, -0.2 and 0.5. For each sample, we computed
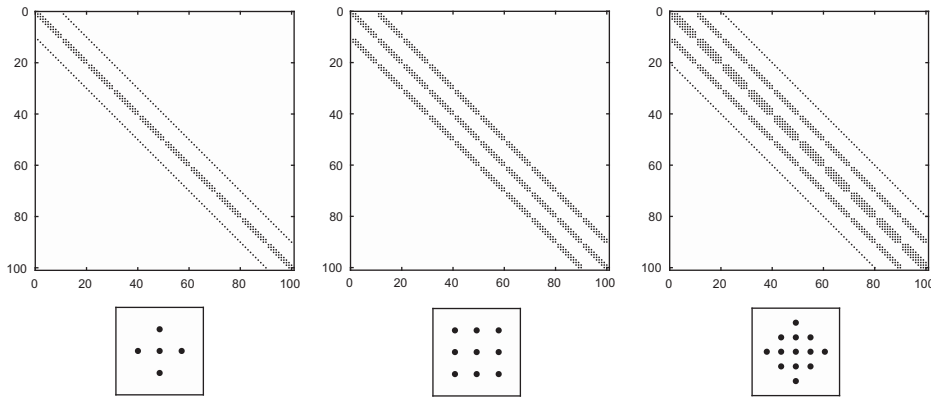
Figure 1. Block band-diagonal structure of inverse covariance matrix. 10 columns of dimension 10, stacked vertically. 4, 8, 12 neighbors of any gridpoint.

successively the MLE with 3, 5 and 7 unknown parameters numerically by Newton's method, as described in Ueno and Tsuchiya (2009).

The difference of each estimator from the true matrix Σ was measured in the Frobenius norm, which is the same as the Euclidean norm of a matrix written as one long vector. 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 2.

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.

5.2. Simulation of fields with diagonal covariance

The simulation for spectral diagonal covariance was carried out in a similar way. First, a diagonal matrix D was prepared, whose diagonal entries decay according to the model $d_i = \frac{1}{c} e^{\alpha \lambda_i}$, $i = 1, \ldots, n$, where c and α are parameters and λ_i are the eigenvalues of Laplace operator in two dimensions on 10×10 nodes (so again $n = 100$). Such models are useful in modeling smooth random fields, e.g., in meteorology. Then, random samples were generated from $\mathcal{N}_n(\mathbf{0}, D)$ with sample sizes $N = 5, \ldots, 20$. For each sample, four covariance matrix estimators were computed:
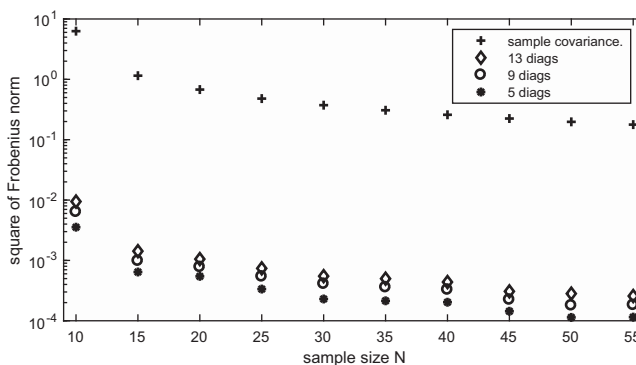
Figure 2. Error of the MLE in Frobenius norm for sample covariance and models with 4, 8, 12 neighbors, i.e. 5, 9, 13 nonzero diagonals in the inverse covariance matrix.

- sample covariance matrix $\hat{\Sigma}_N$, cf. (14)
- diagonal part $\hat{D}^{(0)}$ of the sample covariance matrix, cf. (15)
- MLE $\hat{D}^{(1)}$ in the space of diagonal matrices, cf. (18)
- MLE $\hat{D}^{(3)} = \text{diag}\{(\hat{c}_1 - \hat{c}_2\lambda_i)^{-1}e^{\hat{\alpha}\lambda_i}, i = 1, \ldots, n\}$ with 3 parameters c_1, c_2 and α, cf. (24).
- MLE $\hat{D}^{(2)} = \text{diag}\{\hat{c}^{-1}e^{\hat{\alpha}\lambda_i}, i = 1, \ldots, n\}$ with 2 parameters c and α, cf. (28).

Let us briefly discuss the choice of the covariance model $d_i = \frac{1}{c}e^{\alpha\lambda_i}$. We decided to carry out the simulation with a second-order stationary random field, whose covariance can be diagonalized by the Fourier transform. This transform is formed by the eigenvectors of the Laplace operator. Hence, it is reasonable to model the diagonal terms of this covariance matrix (i.e. the covariance eigenvalues) by some function of eigenvalues of the Laplace operator. This function needs to have a sufficiently fast decay in order to fulfil the necessary condition for the proper covariance (the so-called trace class property, e.g., Kuo (1975)). Exponential decay is used, e.g., in Mirouze and Weaver (2010). 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 $-\lambda_i$, $i = 1, \ldots, n$, e.g., Berner et al. (2009); Gaspari et al. (2006); Simpson, Lindgren, and Rue (2012).

The difference of each estimator from the true matrix D was measured in the Frobenius norm again. 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.

For the diagonal MLE, given by (18), (24), and (28), we can expect from (12) that these estimators should satisfy asymptotically

$$E\left(\left|\hat{D}_N^{(k)} - D\right|_F^2\right) \approx \frac{1}{N}\text{Tr}\left(J_{D^{(k)}}^{-1}\right), \quad k = 1, 2, 3, \tag{30}$$

even if convergence in distribution does not imply convergence of moments without additional assumptions. This conjecture can be supported by a comparison of Figures 5 and 4, where we observe the same decay. From the nesting, we know that

$$\text{Tr}\left(J_{D^{(2)}}^{-1}\right) \leq \text{Tr}\left(J_{D^{(3)}}^{-1}\right) \leq \text{Tr}\left(J_{D^{(1)}}^{-1}\right) \tag{31}$$

and we can expect that the Frobenius norm should decrease for more restrictive models, that is,

$$E\left|\hat{D}_N^{(2)} - D\right|_F^2 \leq E\left|\hat{D}_N^{(3)} - D\right|_F^2 \leq E\left|\hat{D}_N^{(1)} - D\right|_F^2, \tag{32}$$

which is confirmed by the simulations (see Figure 4, resp. 5).
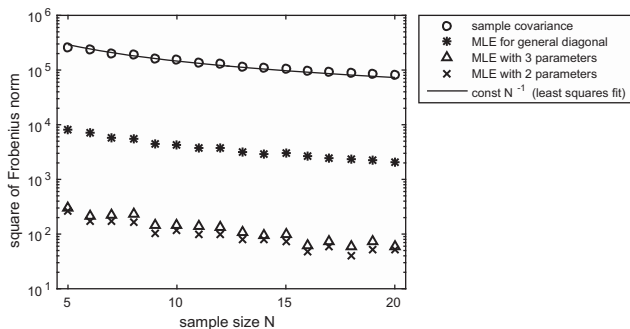
Figure 3. Comparison of the error matrix $\hat{D} - D$ in the Frobenius norm. The field had dimension $n = 10 \times 10$. Exponential decay of eigenvalues (i.e. $\tau_i = ce^{\alpha\lambda_i}$, $i = 1, \ldots, n$) was used with parameters $c = 1/30$ and $\alpha = 0.002$. The full line is the order of convergence $\text{const}(N^{-1})$ fitted to the error of the sample covariance.
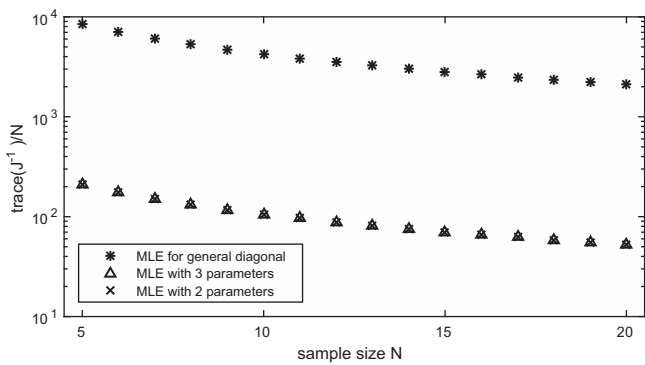
Figure 4. $\frac{1}{N} \operatorname{Tr}(J_D^{-1})$ for the three parameterizations.

The comparisons (32) of the Frobenius norm of the error in the mean squared complement the pointwise comparison (16) 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, $\hat{\Sigma}_N$ is singular and cannot be cast as MLE with an accompanying Fisher information matrix, cf. Remark 6. 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 $N < n$.

6. Comparison with regularization methods

In the previous sections, we pointed out the advantages of using low-parametric models for estimating a covariance matrix using a small sample. As mentioned in the Introduction, there is another large class of estimating methods for high-dimensional covariance matrices: shrinkage estimators. The principle of these methods is to move the sample covariance towards a target matrix that possesses some desired properties (e.g., full rank, proper structure). This can be seen as a convex combination of the sample covariance matrix $\hat{\Sigma}_N$ and the so called target matrix T:

$$\hat{\Sigma}_S = \gamma \hat{\Sigma}_N + (1 - \gamma)T, \qquad \text{for } \gamma \in [0, 1]. \tag{33}$$

One of the simplest shrinkage estimators has the form of (33) with the target matrix equal to identity, which results in shrinking all sample eigenvalues with the same intensity towards
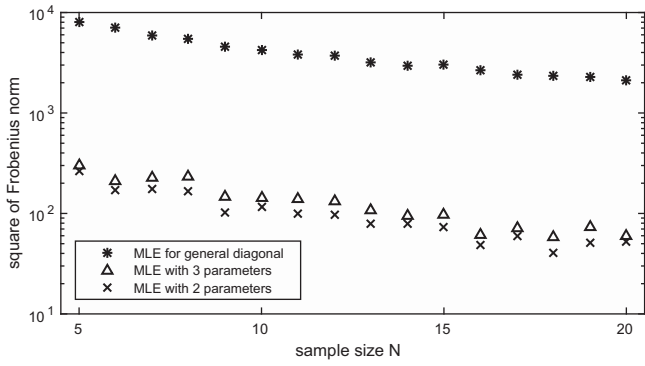

Figure 5. Mean of $|\hat{D}_N^{(k)} - D|_F^2$ based on 50 replications.

their mean value. Ledoit and Wolf (2004) derived the optimal shrinkage parameter γ to minimize the squared Frobenius loss

$$\min_{\gamma} \mathrm{E} \, ||\hat{\Sigma}_S - D||_F^2. \tag{34}$$

The comparison of this estimator with the maximum likelihood estimator $\hat{D}^{(2)}$ was accomplished by a simulation with identical setting as in Section 5. The results are shown in Figure 6. For reference, the sample covariance $\hat{\Sigma}_N$ and its diagonal $\hat{D}^{(0)}$ are also added.

Another regularization method is described in Won et al. (2013). They consider a type of covariance estimator, where the regularization effect is achieved by bounding the condition number of the estimate by a regularization parameter κ_{max}. 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 maximum likelihood estimator restricted on the subspace of matrices with condition number bounded by κ_{max}, i.e.

$$\max_{\Sigma} \ell(\Sigma) \quad \text{subject to } \frac{\lambda_{max}(\Sigma)}{\lambda_{min}(\Sigma)} \leq \kappa_{max}, \tag{35}$$

where $\lambda_{max}(\Sigma)$, resp. $\lambda_{min}(\Sigma)$, is the largest, resp. the smallest, eigenvalue of the covariance matrix Σ. An optimal κ_{max} is selected by maximization of the expected likelihood, which is approximated by using K-fold cross-validation. The authors proved that κ_{max} selected in this way is a consistent estimator for the true condition number (i.e. the condition number of D). 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. The form of the resulting covariance estimator together with the details of the computational process is provided in Won et al. (2013). In Figure 6, we can see the performance of this estimator (denoted as `cond-num-regularization`) in comparison of other methods.

The shrinkage estimator $\hat{\Sigma}_S$ and the condition-number-regularized estimator result in non-diagonal matrices, which in our case predetermines them to perform worse than the diagonal estimator $\hat{D}^{(0)}$. However, we have to note that performance of these methods strongly depends on the particular form of the true covariance matrix D. In the case when the decrease
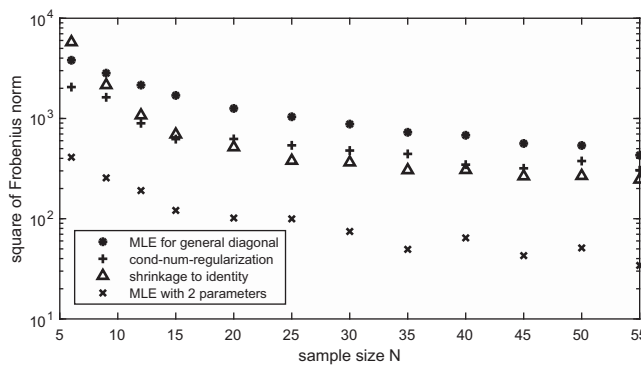
Figure 6. Comparison of regularization estimators with maximum likelihood estimators. The error matrices $\hat{\Sigma} - D$ are compared in the Frobenius norm. The simulation setting was identical with the Section 5.

of the true eigenvalues is less rapid, both methods may provide better results than the diagonal of sample covariance. The performance of $\hat{\Sigma}_S$ could be possibly improved by choosing a different target matrix that is closer to reality but such a study is out of the scope of this paper.

It is seen from Figure 6 that the condition-number-regularized estimator provides more precise estimates than the sample covariance $\hat{\Sigma}_N$, as expected. This is in accordance with the preceding theory and empirical findings about the higher precision of estimators from a smaller parametric subspace (the corresponding parametric subspace consists of matrices with the condition number smaller or equal to κ_{max}). If, however, the theoretical condition number is very large as in our case, the method has a problem in estimating this number and its performance is limited.

Both regularization estimators perform well against sample covariance, but the setting of our simulation is less favourable for them. Neither of them can compete with the maximum likelihood estimator found in the true small subspace of diagonal matrices with proper decay.

7. Conclusions

Our main aim was to point out the significant advantage resulting from computing the MLE of the covariance matrix in a proper parameter subspace, especially in the high-dimensional setting, when the available sample has small size relative to the dimension of the problem. This subspace can be formed, e.g., by a parametric model for covariance eigenvalues or for a diagonal matrix resulting from a suitable set of transformations.

We provided theoretical results on asymptotic comparison of covariance matrices of each estimator for multivariate normal distribution, where we can lean on the well-developed maximum likelihood theory. The situation for small samples was illustrated by means of a simulation. We consider two-parametric models for the covariance eigenvalues based on the eigenvalues of Laplace operator. In practice, the proper model/subspace can be inferred from historical data.

Using a properly specified model, one can reach a significant improvement in performance, which can have a positive impact on the subsequent tasks like data assimilation and prediction.

ORCID

Marie Turčičová   http://orcid.org/0000-0002-7468-0817
Jan Mandel   http://orcid.org/0000-0002-8489-5766
Kryštof Eben   http://orcid.org/0000-0003-4602-748X

References

Bannister, R. N. 2008. 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):1971–96. doi:10.1002/qj.340.

Beezley, J. D., J. Mandel, and L. Cobb. 2011. Wavelet ensemble Kalman filters. In *Proceedings of IEEE IDAACS'2011, Prague, September 2011*, vol. 2, 514–8. IEEE, doi:10.1109/IDAACS.2011.6072819.

Berner, J., G. J. Shutts, M. Leutbecher, and T. N. Palmer. 2009. 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):603–26. doi:10.1175/2008JAS2677.1.

Buehner, M., and M. Charron. 2007. Spectral and spatial localization of background-error correlations for data assimilation. *Quarterly Journal of the Royal Meteorological Society* 133 (624):615–30. doi:10.1002/qj.50.

Carlen, E. 2010. Trace inequalities and quantum entropy: an introductory course. In *Entropy and the quantum*, vol. 529, of *Contemp. Math.*, 73–140. Amer. Math. Soc., Providence, RI, doi:10.1090/conm/529/10428.

Furrer, R., and T. Bengtsson. 2007. Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants. *Journal of Multivariate Analysis*, 98 (2):227–55. doi:10.1016/j.jmva.2006.08.003.

Gaspari, G., S. E. Cohn, J. Guo, and S. Pawson. 2006. Construction and application of covariance functions with variable length-fields. *Quarterly Journal of the Royal Meteorological Society* 132 (619):1815–38. doi:10.1256/qj.05.08.

Hamill, T. M., and C. Snyder. 2000. A hybrid ensemble Kalman filter–3D variational analysis scheme. *Monthly Weather Review* 128 (8):2905–19. doi:10.1175/1520-0493(2000)128 <2905:AHEKFV>2.0.CO;2.

Kasanický, I., J. Mandel, and M. Vejmelka. 2015. Spectral diagonal ensemble Kalman filters. *Nonlinear Processes in Geophysics* 22 (4):485–97. doi:10.5194/npg-22-485-2015.

Kuo, H. H. 1975. *Gaussian measures in Banach spaces*. Lecture Notes in Mathematics, vol. 463. Springer-Verlag, Berlin, doi:10.1007/BFb0082008.

Lauritzen, S. L. 1996. *Graphical models*, vol. 17. Clarendon Press.

Ledoit, O., and M. Wolf. 2004. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88 (2):365–411, doi:10.1016/S0047-259X(03)00096-4.

Lehmann, E. L., and G. Casella. 1998. *Theory of point estimation*. Springer Texts in Statistics. New York: Springer-Verlag, second edition., doi:10.1007/b98854.

Magnus, J. R., and H. Neudecker. 2007. *Matrix Differential Calculus with Applications in Statistics and Econometrics*. John Wiley, third edition.

Mandel, J., J. D. Beezley, and V. Y. Kondratenko. 2010. Fast Fourier transform ensemble Kalman filter with application to a coupled atmosphere-wildland fire model. In A. M. Gil-Lafuente and J. M. Merigo, editors, *Computational Intelligence in Business and Economics, Proceedings of MS'10*, 777–784. World Scientific, doi:10.1142/9789814324441_0089.

Michel, Y., and T. Auligné 2010. Inhomogeneous Background Error Modeling and Estimation over Antarctica. *Monthly Weather Review* 138 (6):2229–52. doi:10.1175/2009mwr3139.1.

Mirouze, I., and A. T. Weaver. 2010. Representation of correlation functions in variational assimilation using an implicit diffusion operator. *Quarterly Journal of the Royal Meteorological Society* 136: 1421–1443, doi:10.1002/qj.643.

Muirhead, R. J. 2005. *Aspects of Multivariate Statistical Theory*. Wiley Series in Probability and Statistics. Wiley, doi: 10.1002/9780470316559.

Pannekoucke, O., L. Berre, and G. Desroziers. 2007. Filtering properties of wavelets for local background-error correlations. *Quarterly Journal of the Royal Meteorological Society*, 133 (623, Part B):363–79, doi:10.1002/qj.33.

Parrish, D. F., and J. C. Derber. 1992. The National Meteorological Center's spectral statistical-interpolation analysis system. *Monthly Weather Review* 120 (8):1747–63. doi:10.1175/1520-0493 (1992)120<1747:TNMCSS>2.0.CO;2.

Rao, C. R. 1973. *Linear statistical inference and its applications*. New York-London-Sydney: John Wiley & Sons, 2nd ed. Wiley Series in Probability and Mathematical Statistics, doi:10.1002/9780470316436.

Rue, H., and L. Held. 2005. *Gaussian Markov random fields: theory and applications*. CRC Press.

Schäfer, J., and S. Korbinian. 2005. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology* 4:(32). doi:10.2202/1544-6115.1175.

Simpson, D., F. Lindgren, and H. Rue. 2012. Think continuous: Markovian gaussian models in spatial statistics. *Spatial Statistics* 1:16–29, doi:10.1016/j.spasta.2012.02.003.

Ueno, G., and T. Tsuchiya. 2009. Covariance regularization in inverse space. *Quarterly Journal of the Royal Meteorological Society* 135 (642):1133–56, doi:10.1002/qj.445.

Vershynin, R. 2012. How close is the sample covariance matrix to the actual covariance matrix?. *Journal of Theoretical Probability* 25 (3):655–86, doi:10.1007/s10959-010-0338-z.

Wang, X., D. M. Barker, C. Snyder, and T. M. Hamill. 2008. A hybrid ETKF–3DVAR data assimilation scheme for the WRF model. Part I: Observing system simulation experiment. *Monthly Weather Review* 136 (12):5116–31, doi:10.1175/2008MWR2444.1.

Won, J.-H., J. Lim, S.-J. Kim, and B. Rajaratnam. 2013. Condition-number-regularized covariance estimation. *Journal of the Royal Statistical Society. Series B, Statistical Methodology* 75 (3):427–50. ISSN 1369-7412. doi:10.1111/j.1467-9868.2012.01049.x.