SCORE MATCHING FILTERS FOR GAUSSIAN MARKOV RANDOM FIELDS WITH A LINEAR MODEL OF THE PRECISION MATRIX

MARIE TURČIČOVÁ*,1,2
1Department of Statistical Modelling
Institute of Computer Science
Pod Vodárenskou věží 271/2
182 07, Prague, Czech Republic

2Department of Probability and Mathematical Statistics
Faculty of Mathematics and Physics, Charles University
Sokolovská 83
186 75, Prague, Czech Republic

JAN MANDEL
Department of Mathematical and Statistical Sciences
University of Colorado Denver
PO Box 173364
Denver, CO 80217-3364, USA

KRYŠTOF EBEN
Department of Complex Systems
Institute of Computer Science
Pod Vodárenskou věží 271/2
182 07, Prague, Czech Republic

(Communicated by Kody Law)

ABSTRACT. We present an ensemble filtering method based on a linear model for the precision matrix (the inverse of the covariance) with the parameters determined by Score Matching Estimation. The method provides a rigorous covariance regularization when the underlying random field is Gaussian Markov. The parameters are found by solving a system of linear equations. The analysis step uses the inverse formulation of the Kalman update. Several filter versions, differing in the construction of the analysis ensemble, are proposed, as well as a Score matching version of the Extended Kalman Filter.

1. Introduction. Filters based on Monte-Carlo approximation of the traditional Kalman filter are a popular group of methods for data assimilation in high-dimensional systems. Examples of these filters are the Ensemble Kalman filter (EnKF) [4] and different versions of Ensemble Square-root filter [29]. If such a filter uses a linear update step, in its core, there is always a model of covariance matrix, its

2020 Mathematics Subject Classification. 62F12, 62H12, 62P12.

Key words and phrases. Score matching, ensemble filter, Gaussian Markov random field, covariance modelling.

This research was partially supported by TA CR grant TL01000238, NSF grant ICER-1664175 and by the project TO01000219 funded by Norway Grants and TA CR.

* Corresponding author: Marie Turčičová.
square root, or its inverse. In the prevalent situation when the ensemble size is small, the sample covariance suffers from a low rank and from the presence of spurious covariances, to mention the most common issues, and methods such as tapering and shrinkage are widely used for regularization of the covariance estimate, e.g., [7, 22]. Often, a sparsity structure is imposed on the covariance matrix; in the extreme (but often well motivated) case, a diagonal approximation in physical, spectral or wavelet space may be used, e.g., [23, 11].

We consider a regularization based on a spatial Markov property of the underlying random field, i.e., the value at each point of the field is conditionally independent on the values at all the other points of the field, given a separating neighbourhood of the point [13, 24]. Markov random fields occur naturally in various situations. They are ubiquitous in spatial statistics; e.g., in meteorology, many fields can be modelled as Markov. Gaussian Markov random fields offer elegant and effective approaches for modelling based on their precision matrix (inverse of the covariance matrix) [18, 27, 36], which has zero entries outside of the separating neighbourhoods. Modelling of covariance by the inverse of a discrete diffusion operator [35] can be understood as a special case. Also, the sparsity structure of the precision matrix can be utilized in a direct way, e.g., in covariance selection techniques [5], which have come to be known as Gaussian graphical models [13].

The precision matrix can be approximated as a linear combination of sparse “design matrices”. Such linear model approach goes back to [1] for covariance matrices, and it has been used for the precision matrix of meteorological fields [32]. These papers employ the Maximum Likelihood method for the estimation of the unknown parameters, which requires the use of numerical optimization algorithms.

In this paper, we focus on parameter estimation by the Score Matching estimation method [9, 6] and we use it for construction of a new ensemble filter, which combines score matching estimation and the linear model for the precision matrix under the assumption of normality. In this case, a closed formula is available for estimators of the parameters of the precision matrix (Section 3). We generalize this approach and estimate both the unknown state and the precision matrix from the forecast ensemble and the available observations. As a result, we obtain an explicit update formula for the parameters based on an ensemble of realizations of a Gaussian Markov random field with a precision matrix obeying a (sparse) linear model. The state-vector component of the joint score matching estimator (SME) is the sample mean. In geoscience, the update step of a filter is often called “analysis” and several versions of SME-based filters may be constructed, depending on the formation of the analysis ensemble. After the analysis step, the ensemble is then propagated through the model dynamics.

If the analysis ensemble is formed by sampling from the estimated Gaussian distribution, a filter with Gaussian resampling (SMEF-GR) arises. We prove several theoretical properties of this filter, in particular the convergence of the estimated parameters to their true values in each step of the filter, if the ensemble size goes to infinity. These estimators are based on new results on asymptotic consistence of SMEs for a triangular array of samples from an exponential family of distributions with random parameters (Section 2.5), which may be of independent interest.

Instead of resampling, one can use the estimated precision matrix in the inverse formulation of the update step of the Kalman filter directly, i.e., use the update step of the information formulation of the Kalman filter for all ensemble members. This filter seems to be more robust than the previous one in cases when the dynamics
is nonlinear and chaotic and the distribution of the forecast ensemble is far from Gaussian. We call this filter the Score Matching Ensemble Filter (SMEF).

A version of Ensemble transform Kalman filter that utilizes the information contained in the precision matrix is also developed. This filter (SMETKF) uses the subspace spanned by the principal components corresponding to the smallest eigenvalues of the estimated precision matrix in the construction of the transform matrix.

Finally, a non-ensemble SME-based filter is proposed as a SME version of the Extended Kalman Filter (SMExKF). This filter propagates the parameters of a linear precision matrix model with the help of the tangent linear dynamics.

The score matching method has been already used for estimating parameters in von Mises – Fisher filter on a unit sphere [3, 30]. However, its usage in ensemble filtering algorithms for Gaussian Markov random fields seems to be new.

The paper is organized as follows: In Section 2, we review the score matching method and complement the joint estimation of the parameters of a normal distribution by new results. In Section 3, we derive results for the estimation of Gaussian Markov random fields, which are used for the formulation and analysis of filtering algorithms in Section 4. Computational results are in Section 5, and Section 6 is the conclusion.

2. Score matching estimation.

2.1. Notation. The Euclidean norm of a vector \( \mathbf{x} = (x_1, x_2, \ldots, x_n)^\top \) from \( \mathbb{R}^n \) is denoted by \( \| \mathbf{x} \|_{\mathbb{R}^n} \) and the inner product of \( \mathbf{x}, \mathbf{v} \in \mathbb{R}^n \) by \( \langle \mathbf{x}, \mathbf{v} \rangle_{\mathbb{R}^n} \). An \( m \times n \) matrix with elements \( a_{ij} \) is written as \( [a_{ij}]_{m,n} \) or \( [a_{ij}]_{i,j=1}^n \) when \( m = n \); \( [\mathbf{v}_1, \ldots, \mathbf{v}_k] \) stands for a matrix with columns \( \mathbf{v}_1, \ldots, \mathbf{v}_k \) and \( I_n \) for an identity matrix of type \( n \times n \).

For a scalar function \( g \) of vector argument \( \mathbf{x} \in \mathbb{R}^n \), \( \Delta_{\mathbf{x}} \) stands for the Laplacian and \( \nabla_{\mathbf{x}} \) for the gradient,

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

If the gradient is needed as a column vector, we denote \( \nabla_{\mathbf{x}}^\top g(\mathbf{x}) = (\nabla_{\mathbf{x}} g(\mathbf{x}))^\top \).

The Jacobian matrix of a vector function \( \mathbf{h} = (h_1, \ldots, h_m)^\top : \mathbb{R}^n \to \mathbb{R}^m \) is an \( m \times n \) matrix

\[
\nabla_{\mathbf{x}} \mathbf{h}(\mathbf{x}) = \left[ \frac{\partial h_i}{\partial x_j}(\mathbf{x}) \right]_{i,j=1}^m \cdot \cdot \cdot \cdot \cdot \cdot \cdot \n
\]

We denote the expected value of \( h(\mathbf{X}) \) with respect to a probability density \( p \) on \( \mathcal{X} \subset \mathbb{R}^n \) by

\[
\mathbb{E}_{\mathbf{X} \sim p} [h(\mathbf{X})] = \int_{\mathcal{X}} h(\mathbf{x}) p(\mathbf{x}) d\mathbf{x}.
\]

The density \( p(\mathbf{x}) \) is modelled by \( f(\mathbf{x} ; \theta) \), where \( \theta \) is a parameter.

2.2. Motivation. When estimating the parameters of a distribution, the maximum likelihood method is usually the first choice. However, the likelihood of a structured precision matrix is complicated even for normal distribution and only numerical maximization is possible [32]. Nevertheless, in our case of estimating the parameters \( (\mu_t^t, (Q_t^t)^{-1}) \) of normal distribution (and, more generally, for distributions belonging to the exponential family), the score matching estimation method can provide linear estimators in a closed form.
2.3. Overview of SME. In this section, we provide a brief summary of the score matching estimation method, following [6, 9, 10].

Consider a random vector $X$ with values in a set $\mathcal{X} \subset \mathbb{R}^n$ and an unknown probability density function $p(x)$, and a parametrized density model of the form

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

where $\theta$ varies over $\Theta$, which is an open set in a finite dimensional vector space $L$. The idea of score matching estimation is to make inference about the parameter $\theta$ using the gradient of the log-density, called a score function, instead of the density itself. The principal observation is that

$$\nabla_x \log f(x|\theta) = \nabla_x (\log q(x|\theta) - \log Z(\theta)) = \nabla_x \log q(x|\theta),$$

thus the score function $\nabla_x \log f(x|\theta)$ does not depend on $Z(\theta)$. The score matching estimator (SME) of the parameter $\theta$ is then obtained by minimizing the mean square distance, i.e.,

$$\hat{\theta} = \arg \min_{\theta \in \Theta} J(\theta), \quad J(\theta) = \int_{\mathcal{X}} \| \nabla_x^T \log q(x|\theta) - \nabla_x^T \log p(x) \|_{\mathbb{R}^n}^2 p(x) dx. \quad (2)$$

SME relies on the following assumptions on the densities, which is satisfied for a large class of probability distributions, in particular the normal distribution.

**Assumption 2.1.** Assume that

(a) $p(x)$ and $\nabla_x \log q(x|\theta)$ are differentiable on $\mathcal{X}$,

(b) $\int_{\mathcal{X}} \| \nabla_x^T \log q(x|\theta) \|_{\mathbb{R}^n}^2 p(x) dx < \infty$ for all $\theta \in \Theta$,

(c) $\int_{\mathcal{X}} \| \nabla_x^T \log p(x) \|_{\mathbb{R}^n}^2 p(x) dx < \infty$, and

(d) function $g(x|\theta) = \log q(x|\theta) : \mathbb{R}^n \rightarrow \mathbb{R}$ satisfies that $p(x) \nabla_x^T g(x|\theta) \rightarrow 0$ for any $\theta \in \Theta$ when $x \rightarrow \partial \mathcal{X}$, and the boundary $\partial \mathcal{X}$ of $\mathcal{X}$ is sufficiently regular for integration by parts, in particular

$$\int_{\mathcal{X}} \langle \nabla_x^T p(x), \nabla_x^T g(x|\theta) \rangle_{\mathbb{R}^n} dx = - \int_{\mathcal{X}} p(x) \Delta_x g(x|\theta) dx, \quad \forall \theta \in \Theta.$$

It follows from Assumption 2.1 that the mean value in the definition of the objective function (2) exists. When Assumption 2.1 holds, it can be shown by integration by parts [9] that the objective function (2) equals to

$$J(\theta) = E_{X \sim p} \left[ \frac{1}{2} \| \nabla_x^T \log q(X|\theta) \|_{\mathbb{R}^n}^2 + \Delta_x \log q(X|\theta) \right] + c,$$

where $c = E_{X \sim p} \left[ \| \nabla_x^T \log p(X) \|_{\mathbb{R}^n}^2 \right]$ does not depend on $\theta$. 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 non-normalized model density $q(x|\theta)$.

With a sample $\mathbb{X}_N = \{X_1, \ldots, X_N\}$ from the density $p$, the expected value in (3) can be approximated by the sample mean,

$$J_N(\theta|\mathbb{X}_N) = \frac{1}{N} \sum_{i=1}^{N} \left( \frac{1}{2} \| \nabla_x^T \log q(X_i|\theta) \|_{\mathbb{R}^n}^2 + \Delta_x \log q(X_i|\theta) \right) + c_N (\mathbb{X}_N), \quad (4)$$

where $c_N (\mathbb{X}_N) = \frac{1}{N} \sum_{i=1}^{N} \| \nabla_x^T \log p(X_i) \|_{\mathbb{R}^n}$ does not depend on $\theta$. 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) is attained. Thus, we have the empirical estimator

$$\hat{\Theta}_N = \arg\min_{\hat{\Theta}} \sum_{i=1}^{N} \left( \frac{1}{2} \| \nabla_{\Theta} \log q(X_i|\Theta) \|^2_{\mathbb{R}^n} + \Delta_x \log q(X_i|\Theta) \right).$$

(5)

Now suppose that the vector space $L \ni \Theta$ is equipped with inner product $(\cdot , \cdot)_L$ and, in addition, the density model (1) belongs to the exponential family, i.e.,

$$\log q(x|\theta) = (\theta, T(x))_L - a(\theta) + b(x).$$

(6)

Further assume that $\Theta$ is such that $Z(\Theta) < \infty$ for all $\theta \in \Theta$. Function $T(x)$ is
called the canonical sufficient statistic and $\theta$ is called the canonical parameter. For a density from the exponential family, the terms in the objective function (3)

$$\nabla_x \log q(x|\theta) = \nabla_x (\theta, T(x))_L + \nabla_x b(x),$$

$$\Delta_x \log q(x|\theta) = \Delta_x (\theta, T(x))_L + \Delta_x b(x),$$

are linear functions of $\theta$. Substituting into (3), we obtain that $J(\theta)$ is quadratic,

$$J(\theta) = \mathbb{E}_{X \sim p} \left[ \frac{1}{2} \| \nabla_{\theta} (\theta, T(X))_L + \nabla_{\theta} b(X) \|^2_{\mathbb{R}^n} + \Delta_x (\theta, T(X))_L + \right.$$

$$\left. + \Delta_x b(X) \right] + c_1$$

$$= \mathbb{E}_{X \sim p} \left[ \frac{1}{2} \| \nabla_{\theta} (\theta, T(X))_L \|^2_{\mathbb{R}^n} + \langle \nabla_{\theta} (\theta, T(X))_L, \nabla_{\theta} b(X) \rangle_{\mathbb{R}^n} + \right.$$

$$\left. + \Delta_x (\theta, T(X))_L \right] + c_2,$$

(7)

where $c_1$ and $c_2$ do not depend on $\theta$. It follows from Assumption 2.1 that the mean value exists.

For a fixed $x$, define linear operator $D(x)$ by

$$D(x): L \to \mathbb{R}^n, \quad D(x)\theta = \nabla_{\theta} (\theta, T(x))_L,$$

(8)

its adjoint operator $D^*(x)$ by

$$D^*(x): \mathbb{R}^n \to L, \quad (\theta, D^*(x)v)_L = \langle D(x)\theta, v \rangle_{\mathbb{R}^n} \forall \theta \in L \forall v \in \mathbb{R}^n,$$

(9)

and the Laplacian vector $\Delta_x T(x)$ by

$$\Delta_x T(x) \in L: \quad (\theta, \Delta_x T(x))_L = \Delta_x (\theta, T(x))_L \forall \theta \in L.$$  

(10)

Then, (7) becomes

$$J(\theta) = \mathbb{E}_{X \sim p} \left[ \frac{1}{2} \| D(X)\theta \|^2_{\mathbb{R}^n} + \langle \theta, D^*(X)\nabla_{\theta} b(X) \rangle_L + \langle \theta, \Delta_x T(X) \rangle_L \right] + c_2.$$  

(11)

With a random sample $X_1, \ldots, X_N$, (11) results in the SME

$$\hat{\theta}_N = \arg\min_{\theta \in L} \frac{1}{N} \sum_{i=1}^{N} \left( \frac{1}{2} \| D(X_i)\theta \|^2_{\mathbb{R}^n} + \langle \theta, D^*(X_i)\nabla_{\theta} b(X_i) \rangle_L + \langle \theta, \Delta_x T(X_i) \rangle_L \right)$$

$$= - \left[ \sum_{i=1}^{N} D^*(X_i)D(X_i) \right]^{-1} \sum_{i=1}^{N} (D^*(X_i)\nabla_{\theta} b(X_i) + \Delta_x T(X_i)),$$

(12)

if the inverse exists.
In the case when \( L = \mathbb{R}^s \) and \((\cdot, \cdot)_{\mathbb{R}^s}\) is the usual Euclidean inner product,

\[
\nabla_x T(x) = \left[ \frac{\partial T}{\partial x_1}, \ldots, \frac{\partial T}{\partial x_n} \right] = \begin{bmatrix}
\frac{\partial T_1}{\partial x_1} & \cdots & \frac{\partial T_1}{\partial x_n} \\
\vdots & \ddots & \vdots \\
\frac{\partial T_n}{\partial x_1} & \cdots & \frac{\partial T_n}{\partial x_n}
\end{bmatrix}
\]

is the Jacobian matrix of \( T \), so \( D(x) \) is the Jacobian transposed,

\[
D(x) = (\nabla_x T(x))^\top,
\]

the adjoint operator in (9) becomes simply the Jacobian itself,

\[
D^*(x) = \nabla_x T(x),
\]

(10) becomes the Laplacian applied to \( T \) entry by entry,

\[
\Delta_x T(x) = \begin{pmatrix}
\Delta_x T_1(x) \\
\vdots \\
\Delta_x T_n(x)
\end{pmatrix}
\]

and the estimator (12) becomes

\[
\hat{\theta}_N = -\left[ \sum_{i=1}^{N} \nabla_x T(X_i) \nabla_x^\top T(X_i) \right]^{-1} \sum_{i=1}^{N} \left( \nabla_x T(X_i) \nabla_x^\top b(X_i) + \Delta_x T(X_i) \right),
\]

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

In the rest of the paper, we will assume that the true density \( p(x) \) equals \( f(x|\theta^0) \) for a unique \( \theta^0 \), where \( f(x|\theta) \) belongs to the exponential family. In addition to Assumption 2.1, we need the exponential family distribution to satisfy the following.

**Assumption 2.2.** Assume that

\( (a) \) \( \int_{\mathcal{X}} \|D(x)\|_{op}^2 f(x|\theta) \, dx < \infty \) for all \( \theta \in \Theta \),

\( (b) \) \( \int_{\mathcal{X}} \|\nabla_x b(x)\|_{\mathbb{R}^n}^2 f(x|\theta) \, dx < \infty \) for all \( \theta \in \Theta \),

\( (c) \) \( \int_{\mathcal{X}} \|\Delta_x T(x)\|_{L} f(x|\theta) \, dx < \infty \) for all \( \theta \in \Theta \), and

\( (d) \) Assumption 2.1(d) holds with \( g(x|\theta) = \langle T(x), \theta \rangle_L \) in place of \( g \),

where \( \|D(x)\|_{op}^2 = \sup \left\{ \|D(x)\theta\|_{\mathbb{R}^n}^2 : \theta \in L, \|\theta\|_L \leq 1 \right\} \).

**Remark 1.** For distributions with \( b(x) = 0 \), Assumption 2.2(b) is satisfied automatically and Assumption 2.2(d) coincides with Assumption 2.1(d). Since \( p(x) = f(x|\theta^0) \), Assumption 2.1(c) follows from Assumption 2.1(b). Moreover, Assumption 2.1(b) follows from Assumption 2.2(b) and Assumption 2.2(c), because

\[
\mathbb{E}_{X \sim p} \|\nabla_x \log g(X|\theta)\|_{\mathbb{R}^n}^2 = \mathbb{E}_{X \sim p} \|D(X)\theta + \nabla_x^\top b(x)\|_{\mathbb{R}^n}^2
\leq 2 \mathbb{E}_{X \sim p} \|D(X)\|_{op}^2 \|\theta\|_L^2 + 2 \mathbb{E}_{X \sim p} \|\nabla_x^\top b(x)\|_{\mathbb{R}^n}^2.
\]

Under these assumptions, the SME (5) can be shown to be consistent.
Theorem 2.3 (Consistency of SME, [6, Proposition 1]). Assume that \( f(\mathbf{x}|\theta) \) is a density of exponential family (6) that satisfies Assumptions 2.1(a,d) and 2.2, \( \mathbf{X}_1, \ldots, \mathbf{X}_N \) is a sample from \( f(\mathbf{x}|\theta^0) \), and the inverse of \( \mathbb{E}_{\mathbf{X} \sim f(\cdot|\theta^0)}(D^*(\mathbf{X}) D(\mathbf{X})) \) exists. Then, the score matching estimator \( \hat{\theta}_N \) obtained by (12) is consistent, i.e., \( \hat{\theta}_N \overset{P}{\to} \theta^0 \) as \( N \to \infty \).

2.4. SME of parameters of normal distribution. The \( n \)-dimensional normal distribution \( N_n(\mu, \Sigma) \) belongs to the exponential family with

\[
\mathcal{X} = \mathbb{R}^n, \quad T(\mathbf{x}) = \left[ -\frac{1}{2} [x_ix_j]_{i,j=1}^n \right], \quad \theta = \left[ \Sigma^{-1}\mu \Sigma^{-1} \right],
\]

\[
L = \left\{ \left[ \begin{array}{cc} \mathbf{x} \\ A \end{array} \right] : \mathbf{x} \in \mathbb{R}^n, A \in \mathbb{R}^{n \times n}, A = A^\top \right\},
\]

\[
\left\langle \left[ \begin{array}{c} \mathbf{x} \\ A \end{array} \right], \left[ \begin{array}{c} \mathbf{y} \\ B \end{array} \right] \right\rangle_L = \mathbf{x}^\top \mathbf{y} + \text{tr}(A^\top B),
\]

\[
\Theta = \left\{ \left[ \begin{array}{c} \mathbf{x} \\ A \end{array} \right] \in L : A \text{ is symmetric positive definite} \right\}.
\]

When \( L = \mathbb{R}^s \) and \( \langle \cdot, \cdot \rangle_{\mathbb{R}^s} \) is the usual Euclidean inner product, the SME simplifies further, analogously as in (14). In the notation here, which is from [6], the canonical density (6) becomes

\[
\log q(\mathbf{x}|\theta) = \sum_{k=1}^s \theta_k T_k(\mathbf{x}) - a(\theta) + b(\mathbf{x}),
\]

the linear operator \( D \) in (8) becomes

\[
D(\mathbf{x}) \theta = \left[ \frac{\partial}{\partial x_j} \sum_{k=1}^s \theta_k T_k(\mathbf{x}) \right]_{j=1}^n = \left[ \sum_{k=1}^s \frac{\partial T_k(\mathbf{x})}{\partial x_j} \theta_k \right]_{j=1}^n = (\nabla_x T(\mathbf{x}))^\top \theta.
\]

See also [10, eq. (34)] for the case \( b = 0 \).

2.5. Continuity of SME. To demonstrate the consistency of our SME filter, we need to prove the convergence of SME for a triangular array of samples from an exponential family of distributions with random parameters, which can be viewed as a continuity of SME with respect to random perturbations.

Lemma 2.4. Suppose that \( f(\mathbf{x}|\theta) \) is a parametric probability density on \( \mathcal{X} \subset \mathbb{R}^n \) of the exponential family (6) with parameter \( \theta \in \Theta \subset L, \Theta \) open. Further, suppose that \( \mathcal{W} \) is a finite-dimensional normed space, \( h : \mathbb{R}^n \to \mathcal{W} \) is Borel measurable and such that

\[
\int_{\mathcal{X}} \| h(\mathbf{x}) \|_W^2 f(\mathbf{x}|\theta) \, d\mathbf{x} < \infty \quad \text{for all } \theta \in \Theta. \tag{15}
\]

Finally, suppose that \( \theta_N \) are random parameters such that \( \theta_N \overset{P}{\to} \theta^0 \) as \( N \to \infty \), and, for each \( N \), \( \{ \mathbf{X}_N : i = 1, \ldots, N \} \) is a sample from \( f(\mathbf{x}|\theta_N) \). Then,

\[
\frac{1}{N} \sum_{i=1}^N h(\mathbf{X}_{Ni}) \overset{P}{\to} \mathbb{E}_{\mathbf{X} \sim f(\cdot|\theta^0)} h(\mathbf{X}) \quad \text{as } N \to \infty.
\]

For the proof see Appendix A. The next lemma will be used to show the existence of the SME.
Lemma 2.5. Suppose that \( \{B_N\}_{N \in \mathbb{N}} \) are random linear operators on a finite dimensional normed space. Assume that \( B_N \xrightarrow{P} A \) as \( N \to \infty \) and that \( A^{-1} \) exists. Then, \( P(B_N^{-1} \text{ exists}) \to 1 \) and \( B_N^{-1} \xrightarrow{P} A^{-1} \) as \( N \to \infty \).

Proof. This statement follows from the continuity of the mapping \( A \to A^{-1} \), the fact that the set of all invertible mappings is open, and the continuous mapping theorem.

To prove the convergence of SME, we also need a stronger version of Assumptions 2.2(a)-(c). These are also satisfied by the normal distribution.

Assumption 2.6. Assume that
(a) \( \mathbb{E}_{X \sim f(\cdot|\theta)} \|D(X)\|_2^{op} < \infty \) for all \( \theta \in \Theta \),
(b) \( \mathbb{E}_{X \sim f(\cdot|\theta)} \|\nabla_x b(X)\|_R < \infty \) for all \( \theta \in \Theta \), and
(c) \( \mathbb{E}_{X \sim f(\cdot|\theta)} \|\Delta_x T(X)\|_L^2 < \infty \) for all \( \theta \in \Theta \).

Theorem 2.7. Suppose that \( p(x) = f(x|\theta^0) \) for a unique \( \theta^0 \in \Theta \), where \( f(x|\theta) = q(x|\theta)/Z(\theta) \) is a parametric density on \( \mathcal{X} \) from an exponential family (6) that satisfies Assumptions 2.1(a,d), 2.2(d), 2.6(a)-(c). Let \( \theta_N, N = 1,2,\ldots, \) be random parameters with values in an open set \( \Theta \subset \mathbb{R} \) and such that \( \theta_N \xrightarrow{P_{N \to \infty}} \theta^0 \) in \( \Theta \). Assume that the inverse of \( \mathbb{E}_{X \sim f(\cdot|\theta^0)} (D^*(X)D(X)) \) exists. For each \( N \in \mathbb{N} \), denote by \( \hat{\theta}_N \) the SME of \( \theta_N \) computed by (12) using a sample \( X_{N1}, \ldots, X_{NN} \) from \( f(\cdot|\theta_N) \). Then \( \hat{\theta}_N \xrightarrow{P_{N \to \infty}} \theta^0 \).

Proof. From Lemma 2.4 with \( h(x) = D^*(x)D(x) \) and Assumption 2.6(a), it follows that
\[
\frac{1}{N} \sum_{i=1}^{N} D^*(X_{Ni})D(X_{Ni}) \xrightarrow{P_{N \to \infty}} \mathbb{E}_{X \sim f(\cdot|\theta^0)} (D^*(X)D(X)).
\] (16)

Next, we apply Lemma 2.4 with \( h(x) = D^*(x)\nabla_x b(x) + \Delta_x T(x) \). By the definition of the operator norm, Cauchy inequality, and Assumption 2.6(a,b),
\[
\left( \int_{\mathcal{X}} \|D^*(x)\nabla_x b(x)\|_L f(x|\theta) \, dx \right)^2 \leq \left( \int_{\mathcal{X}} \|D^*(x)\|_2^{op} \|\nabla_x b(x)\|_R f(x|\theta) \, dx \right)^2 \leq \int_{\mathcal{X}} \|D^*(x)\|_2^{op} f(x|\theta) \, dx \int_{\mathcal{X}} \|\nabla_x b(x)\|_R f(x|\theta) \, dx < \infty \quad \text{for all } \theta \in \Theta,
\]
which, together with Assumption 2.6(c), yields (15). Thus, by Lemma 2.4,
\[
\frac{1}{N} \sum_{i=1}^{N} D^*(X_{Ni})\nabla_x b(X_{Ni}) + \Delta_x T(X_{Ni}) \xrightarrow{P_{N \to \infty}} \mathbb{E}_{X \sim f(\cdot|\theta^0)} (D^*(X)\nabla_x b(X) + \Delta_x T(X)).
\] (17)
Since \( \nabla_x J(\theta^0) = 0 \), it follows from (11) that
\[
\mathbb{E}_{X \sim f(\cdot|\theta^0)} (D^*(X)D(X)) \theta^0 + \mathbb{E}_{X \sim f(\cdot|\theta^0)} (D^*(X)\nabla_x b(X) + \Delta_x T(X)) = 0.
\] (18)
Thus, from (16, 17, 18), and the continuous mapping theorem,
\[ \left( \sum_{i=1}^{N} D^*(X_{Ni}) D(X_{Ni}) \right)^{-1} \sum_{i=1}^{N} \left( D^*(X_{Ni}) \nabla_a b(X_{Ni}) + \Delta_a T(X_{Ni}) \right) \xrightarrow{P} \left( \mathbb{E}_{X \sim f(\cdot|\theta_0)} (D^*(X) D(X)) \right)^{-1} \mathbb{E}_{X \sim f(\cdot|\theta_0)} \left( D^*(X) \nabla_a b(X) + T(X) \right) = \theta_0. \]

Existence of the inverse on the left-hand side follows from Lemma 2.5 with
\[ B_N = \frac{1}{N} \sum_{i=1}^{N} D^*(X_{Ni}) D(X_{Ni}) \text{ and } A = \mathbb{E}_{X \sim f(\cdot|\theta_0)} (D^*(X) D(X)). \]

3. SME for parameters of a Gaussian Markov random field. Following [25], consider a random field understood as a vector of \( n \) random variables \( X = (X_1, \ldots, X_n)^\top \) defined on a probability space \( (\Omega, \mathcal{S}, P) \). For a subset of indices \( A \subset \{1, \ldots, n\} \), \( X_A \) denotes the subfield \( (X_i : i \in A) \) and \( X_{-A} \) stands for \( X \{1, \ldots, n\} \setminus A \). The random field \( X \) is equipped with an adjacency structure by an undirected graph, and for each \( X_k \), \( \mathcal{N}_{X_k} \subset X_{-k} \) is the set of neighbours of \( X_k \). The random field \( X \) is said to have the spatial Markov property if for every \( k \in \{1, \ldots, n\} \), the conditional distribution of \( X_k \) depends only on the neighbourhood \( \mathcal{N}_{X_k} \), i.e., for every \( k \) and every Borel set \( B \),
\[ P(X_k \in B|X_{-k}) = P(X_k \in B|\mathcal{N}_{X_k}). \]

That is, each variable \( X_k \) of a Markov random field \( X \) is conditionally independent on variables \( X_j \notin \mathcal{N}_{X_k}, j \neq k \), given \( \mathcal{N}_{X_k} \). If the field \( X \) is Gaussian with covariance matrix \( \Sigma \), then the precision matrix \( Q = \Sigma^{-1} = [q_{jk}]_{j,k=1}^n \) is sparse [25, Theorem 2.2]; for \( j \neq k \),
\[ q_{jk} = 0 \quad \text{if} \quad X_j \notin \mathcal{N}_{X_k}. \]

This observation gives us a powerful tool for a sparse representation of the covariance matrix of a Markov random field, since its inverse (the precision matrix) is a sparse, band matrix. For illustration, in Figure 1, we provide three examples of a neighbourhood of a point in a two-dimensional Gaussian Markov random field and the corresponding precision matrices.

![Figure 1](image-url)  
**Figure 1.** Block band-diagonal structure of inverse covariance matrix for two-dimensional Markov fields of type 10 × 10 points. There are assumed 4, 8, 12 neighbours of any gridpoint.

A linear model \( Q = \sum_{k=1}^r \beta_k A_k \) for the precision matrix \( \Sigma^{-1} \equiv Q \) was introduced in [32], where the matrices \( A_k, k = 1, \ldots, r \), are called “design matrices”, which
Theorem 3.1. Let $N$ since the computation is done for one fixed $N$ and parameters of linear model for the precision matrix of normal distribution $X$ value $E$ and estimate its parameters by the score matching method together with the mean matching in a general setting was studied in [6]. Also been considered in [28]. Linear estimation of the precision matrix by score maximization of the likelihood function. However, the present method is incomparably faster, because it avoids numerical but not the same results as the maximum likelihood method described in [32].

In this paper, we adopt the linear model for the precision matrix of the forecast matrices. Denote $\Sigma$ for the precision matrix $A$ will be adopted in this paper as well. The design matrices $802$ MARIE TURČIČOVÁ, JAN MANDEL AND KRYŠTOF EBEN

Regularization of the covariance matrix by means of the Markov property has usually chosen as sparse matrices, whose diagonals and subdiagonals can effectively provide a regularization by constraining the precision matrix to the span of the design matrices. In [32], the parameters were estimated by the maximum likelihood method, which, however, does not reduce to a linear problem, and has to be done numerically. We have found out that score matching estimator gives very similar results as the maximum likelihood method described in [32]. However, the present method is incomparably faster, because it avoids numerical maximization of the likelihood function.

Regularization of the covariance matrix by means of the Markov property has also been considered in [28]. Linear estimation of the precision matrix by score matching in a general setting was studied in [6].

In this paper, we adopt the linear model for the precision matrix of the forecast and estimate its parameters by the score matching method together with the mean value $E X$. Then, by conditioning, we derive the analysis step of a new filter, with regularization provided by the design matrices.

The following theorem provides explicit formulas for SME of the mean value estimator $(\beta_0, \Sigma^0)$, which are defined as maximizers of an objective function

$$\psi(X, \theta) = \sum_{i=1}^N \psi_i(X, \theta),$$

where $\psi_i(X, \theta)$ is called the inverse Godambe information. For the score matching estimator (14), $\psi(X, \theta) = -D^*(X)D(X)\theta - D^*(X)\nabla_\theta b(X) - \Delta_2 T(X)$, and the assumptions of [33, Theorem 5.41] reduce to the condition that the expected value $E \left\| \psi(X, \theta^0) \right\|^2_{\mathbb{R}^n}$ is finite and the expectation of the Jacobian matrix $\nabla_\theta (\psi(X, \theta))$ will be adopted in this paper as well. The design matrices $A_k, k = 1, \ldots, r$, are usually chosen as sparse matrices, whose diagonals and subdiagonals can effectively model the appropriate precision matrix (Figure 1). Such a linear model can provide a regularization by constraining the precision matrix to the span of the design matrices. In [32], the parameters were estimated by the maximum likelihood method, which, however, does not reduce to a linear problem, and has to be done numerically. We have found out that score matching estimator gives very similar results as the maximum likelihood method described in [32]. However, the present method is incomparably faster, because it avoids numerical maximization of the likelihood function.

Regularization of the covariance matrix by means of the Markov property has also been considered in [28]. Linear estimation of the precision matrix by score matching in a general setting was studied in [6].

In this paper, we adopt the linear model for the precision matrix of the forecast and estimate its parameters by the score matching method together with the mean value $E X$. Then, by conditioning, we derive the analysis step of a new filter, with regularization provided by the design matrices.

The following theorem provides explicit formulas for SME of the mean value estimator $(\beta_0, \Sigma^0)$, which are defined as maximizers of an objective function

$$\psi(X, \theta) = \sum_{i=1}^N \psi_i(X, \theta),$$

where $\psi_i(X, \theta)$ is called the inverse Godambe information. For the score matching estimator (14), $\psi(X, \theta) = -D^*(X)D(X)\theta - D^*(X)\nabla_\theta b(X) - \Delta_2 T(X)$, and the assumptions of [33, Theorem 5.41] reduce to the condition that the expected value $E \left\| \psi(X, \theta^0) \right\|^2_{\mathbb{R}^n}$ is finite and the expectation of the Jacobian matrix $\nabla_\theta (\psi(X, \theta))$...
exists and is non-singular at $\theta^0$. For the SME (21), this condition is satisfied when the matrix $[\text{tr}(A_k\Sigma)_{k=1}^r]_{k,l=1}^r$ is regular [31]. Hence, the rate of weak convergence of a consistent SME $\hat{\theta}_N$ to the true value is $1/\sqrt{N}$.

The design matrices $A_k$ need to be selected in a way that the inverse in (21) exists.

**Theorem 3.2 (Existence of SME).** Suppose $A_k \in \mathbb{R}^{n \times n}$, $k = 1, \ldots, r$, are symmetric, and $X_1, \ldots, X_N \in \mathbb{R}^n$. Denote the matrix of “anomalies”

$$Z = [Z_1, \ldots, Z_N] = [X_1 - \bar{X}, \ldots, X_N - \bar{X}].$$

so that the sample covariance is $S = ZZ^T/N$. Then the following conditions are equivalent:

1. $[\text{tr}(SA_kA_l)]_{k,l=1}^r$ is invertible.
2. The set $\{A_kZ : k = 1, \ldots, r\} \subset \mathbb{R}^{n \times N}$ is linearly independent.
3. The set $\{A_kS : k = 1, \ldots, r\} \subset \mathbb{R}^{n \times n}$ is linearly independent.

**Proof.** From the properties of trace,

$$[\text{tr}(SA_kA_l)]_{k,l=1}^r = \left[ \text{tr}\left( \frac{1}{N} \sum_{i=1}^N Z_iZ_i^T A_kA_l \right) \right]_{k,l=1}^r = \frac{1}{N} \left[ \sum_{i=1}^N Z_i^T A_k^r A_l^r Z_i \right]_{k,l=1}^r,$$

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

$$A_kZ = [A_kZ_1, \ldots, A_kZ_N] \in \mathbb{R}^{n \times N}, \quad k = 1, \ldots, r,$$

of $\mathbb{R}^{n \times N}$ equipped with the Frobenius inner product

$$\langle [U_1, \ldots, U_N], [V_1, \ldots, V_N] \rangle_F = \sum_{i=1}^N U_i^T V_i.$$

Since the Gram matrix of a set of elements of a vector space is nonsingular if and only if the set is linearly independent (e.g., [8, Section 3.3.10]), the matrix $[\text{tr}(SA_kA_l)]_{k,l=1}^r$ is invertible if and only if the set of matrices $\{A_kZ : k = 1, \ldots, r\} \subset \mathbb{R}^{n \times N}$ is linearly independent, which is equivalent to the linear independence of $\{A_kS : k = 1, \ldots, r\} \subset \mathbb{R}^{n \times n}$, since $S = ZZ^T/N$.

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

**Corollary 2.** If the matrix $[\text{tr}(SA_kA_l)]_{k,l=1}^r$ is invertible and $J$ is a nonempty subset of $\{1, \ldots, r\}$, then $[\text{tr}(SA_kA_l)]_{k,l\in J}^r$ is also invertible.

We can now apply the consistency of SME to Gaussian Markov random fields. 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 $X_N$ instead of $X$. See (45) in Appendix B for a form of $D^*(X)D(X)$ specific to a linear model of the precision matrix and normal distribution.

**Theorem 3.3.** Assume that all the assumptions of Theorem 3.1 are satisfied and denote by $\beta_N$ the SME of $\beta^0$ from (21) based on a sample of size $N$, if the inverse of $[\text{tr}(SA_kA_l)]_{k,l=1}^r$ exists. Further assume that $E_{X \sim f(\mu^0, \theta)}(D^*(X)D(X))$ is invertible. Then, $(\hat{X}_N, \hat{\beta}_N)$ is a consistent estimator of $(\mu^0, \beta^0)$, and, in particular, $Q(\hat{\beta}_N)$ is a consistent estimator of $Q(\beta^0)$.
Proof. Denote $\beta^0 = [\beta^0_k]_{k=1}^r$, and apply Theorem 2.3 (whose assumptions can be easily verified) to the parameter

$$\eta = \begin{pmatrix} Q(\beta)\mu \\ \beta \end{pmatrix},$$

which provides

$$\hat{\eta}_N \xrightarrow{P} \eta^0 = \begin{pmatrix} Q(\beta^0)\mu^0 \\ \beta^0 \end{pmatrix},$$

where $P(\hat{\eta}_N \text{ exists}) \to 1$ as $N \to \infty$ (Lemma 2.5). Then, the convergence

$$(\bar{X}_N, \hat{\beta}_N) \xrightarrow{P} \eta^0 = (\mu^0, \beta^0)$$

results from the continuous mapping theorem applied to the mapping of $\eta$ to $(\mu, \beta)$, which is continuous by Lemma B.1 (Appendix B).

Corollary 3. Under the assumptions of Theorem 3.3, $\hat{\Sigma}_N = Q(\hat{\beta}_N)^{-1}$ (when the inverse exists) is a consistent estimator of $\Sigma^0$.

Proof. From Theorem 3.3,

$$Q(\hat{\beta}_N) \xrightarrow{P} Q(\beta^0).$$

Since $Q(\beta^0)$ exists by assumption, the consistency of $\hat{\Sigma}_N$ follows from Lemma 2.5.

4. Filtering.

4.1. State space model and filtering. We use standard Kalman filter setting and terminology following [12, eq. (6,7)]. Consider a discrete-time stochastic linear dynamical system and observation model

$$X_t = MX_{t-1} + e^X_t, \quad t \in \mathbb{N} \quad (24)$$

$$Y_t = HX_t + e^Y_t, \quad t \in \mathbb{N}, \quad (25)$$
with initial condition $X_0 \sim \mathcal{N}_n(\mu_0, \Sigma_0)$. Here, $X_t$ is an unobservable system state and the linear operator $M \in \mathbb{R}^{n \times n}$ represents the system dynamics, $H \in \mathbb{R}^{m \times n}$ is the observation operator. The additive random perturbations $e_t^X \sim \mathcal{N}_n(0, \Psi)$ and the additive observation errors $e_t^Y \sim \mathcal{N}_m(0, R)$ are independent mutually and also among different time instants $t \in \mathbb{N}$.

The forecast $X_t^f = (X_t|Y_1 = y_1, \ldots, Y_{t-1} = y_{t-1})$ thus has a normal distribution $\mathcal{N}_n(\mu_t^f, \Sigma_t^f)$ for some $\mu_t^f$ and $\Sigma_t^f$. Assume that $\Sigma_t^f$ is invertible and denote by $Q_t^f = (\Sigma_t^f)^{-1}$ the forecast precision matrix.

The initial state covariance $\Sigma_0$ and the covariances $R$ and $\Psi$ of the observations and model error are assumed to be known. Then the forecast covariance $\Sigma_t^f$ follows by recursion, cf., (28, 29) below, but due to the high dimension of the problem, $\Sigma_t^f$ is estimated from an ensemble. Different regularization methods are used to produce practically useful estimates.

Our objective is to build a filter leveraging the Markov property of the field $X_t$, defining a (sparse) linear model for precision matrices. Using the score matching estimation explained above, we get explicit formulas for parameter estimators. This approach presents a viable alternative to regularization of sample covariance, avoiding heuristic methods like tapering.

Since we model precision matrices, we will use in the update step (analysis) the inverse form of Kalman update [14, Section 4.1], giving

$$X_t^a = (X_t|y_1, \ldots, y_t) \sim \mathcal{N}_n(\mu_t^a, \Sigma_t^a)$$  (26)

with parameters fulfilling

$$Q_t^a \mu_t^a = Q_t^f \mu_t^f + H^T R^{-1} y_t$$  (27)

$$\Sigma_t^a = (Q_t^a)^{-1},$$  (28)

where

$$Q_t^a = Q_t^f + H^T R^{-1} H$$

is the analysis precision matrix. Such filters are usually called inverse filters or information filters in engineering literature, where often the forecast step is adjusted accordingly. In geosciences, the model usually works in “grid space” and we cannot use transformed state like $Q_t^f \mu_t^f$ directly. We nevertheless use the term “inverse filter” when using the inverse form of update.

### 4.2. Score matching filters
Consider the discrete-time stochastic linear dynamical system (24, 25), and assume that $X_0$ has the Markov property so that the inverse of $\Sigma_0$ is sparse. Further, assume that the matrix $M$, representing the dynamics, has sparse inverse so that $X_t^f = X_t = M X_0 + e_t^X$ is again a GMRF. For a short assimilation time step, this may be a realistic assumption, e.g. in case the variable at one location is influenced only by points from its immediate neighbourhood. Assuming $H$ is sparse too, the distribution of the analysis $X_t^a = X_t|y_1$ has a sparse precision matrix and therefore, $X_t^a$ is also a GMRF. Sparsity of $H$ means that observations are available only at a small number of locations, which is often the case in applications. By induction, we can consider the forecast $X_t^f = (X_t|y_1, \ldots, y_{t-1})$ and the analysis $X_t^a = (X_t|y_1, \ldots, y_t)$ to be Gaussian Markov random fields for all $t \in \mathbb{N}$. 


If the dynamical model $M$ is nonlinear or available as machine code only, the information of state covariance is usually carried forward in form of an ensemble (unless a tangent linear model is available). Many ensemble methods use the standard linear update in this case too, effectively inserting an estimate of forecast covariance, based on the ensemble, into the update formulas. Here, we propose three versions of an ensemble filter based on score matching estimate of forecast state and precision matrix

$$\mu_t^f = M\mu_{t-1}^f, \quad Q_t^f = (M\Sigma_{t-1}^a M^\top + \Psi)^{-1}, \quad \text{with } \mu_0^a = \mu_0, \Sigma_0^a = \Sigma_0,$$

(29) namely an inverse filter with Gaussian resampling, an inverse filter with perturbed observations and an inverse filter of the ETKF type.

We start with the following Algorithm 1 named as the Score matching ensemble filter with Gaussian resampling (SMEF-GR). It approximates the parameters (29) and under linearity and Gaussianity, the theory derived in the preceding sections allows us to prove the consistency of estimators of the analysis distribution (26) in every time step (Theorem 4.1).

**Algorithm 1 (SMEF-GR).** Choose a set of design matrices $\{A_k\}_{k=1}^r$. The initial ensemble $X_{01}, \ldots, X_{0N}^a$ is sampled from a given initial distribution $\mathcal{N}_n(\mu_0, \Sigma_0)$. Then, for $t \geq 1$:

1. **Forecast step:** The ensemble is pushed forward in time by applying the model $M$ to each ensemble member independently, to create the forecast ensemble at time $t$:

$$X_{ti}^f = MX_{t-1,i}^a + V_{ti}, \quad V_{ti} \sim \mathcal{N}_n(0, \Psi), \quad i = 1, \ldots, N.$$  

2. **Analysis step:** Compute the sample mean $\bar{X}_t^f$ and sample covariance $S_t$ of the forecast ensemble and then compute the SME ($\bar{X}_t^f, \hat{Q}_t^f$) from Theorem 3.1 as

$$\hat{Q}_t^f = \sum_{k=1}^r \hat{\beta}_{tk} A_k,$$

with $\hat{\beta}_t = (\hat{\beta}_{t1}, \ldots, \hat{\beta}_{tr})^\top$ being a solution of the (sparse) linear system

$$\left(\left[\text{tr} (S_t A_k)\right]_{k,l=1}^r\right) \beta = \left[\text{tr} A_k\right]_{k=1}^r.$$

For a given data vector $y_t$, the analysis $(\hat{\mu}_t^a, \hat{Q}_t^a)$ is found according to the formulas (27, 28) for the conditional mean and covariance as a solution of

$$\hat{Q}_t^a \hat{\mu}_t^a = \hat{Q}_t^f \bar{X}_t^f + H^\top R^{-1} y_t,$$

(30) where

$$\hat{Q}_t^f = \hat{Q}_t^f + H^\top R^{-1} H.$$  

(31)

3. **Resampling:** The ensemble $X_{t1}^a, \ldots, X_{tN}^a$ is sampled from $\mathcal{N}_n(\hat{\mu}_t^a, (\hat{Q}_t^a)^{-1})$.

Note that the inverse $(\hat{Q}_t^a)^{-1}$ does not need to be computed explicitly. From (31), the analysis precision matrix $\hat{Q}_t^a$ is a linear combination of sparse matrices and thus can be expected to be sparse as well. Sampling $X_{ti}^a$ from $\mathcal{N}_n(\hat{\mu}_t^a, (\hat{Q}_t^a)^{-1})$ can be then implemented efficiently by sampling $\zeta_i \sim \mathcal{N}_a(0, I_n)$, computing a square root $L, LL^\top = \hat{Q}_t^a$, solving a system with multiple right-hand sides $L^\top Z_i = \zeta_i$, and taking $X_{ti}^a = Z_i + \hat{\mu}_t^a$, $i = 1, \ldots, N$.

In the standard EnKF, the mean and covariance matrix are estimated by the ensemble mean and sample covariance, which converge in the limit for large
ensembles to the mean and covariance of the true filtering distribution in $L^p$ for all $p \in [1, \infty)$ \cite{21, 16}. Hence, in every time step, EnKF (with linear dynamics) provides consistent estimators of the true mean and covariance matrix of the analysis and forecast distribution. The following theorem proves this type of consistency for the SMEF-GR filter.

The set of design matrices $\{A_k : k = 1, \ldots, r\}$ is assumed to be fixed in every time step $t$, however, some of $\beta_{t1}, \ldots, \beta_{tr}$ can be zero. Therefore, the set of “active” design matrices may differ over time, if the structure of $Q_t$ is not stable.

**Theorem 4.1 (SMEF-GR).** Assume the discrete-time stochastic linear dynamical system \((24, 25)\) with $X_0 \sim N_n(\mu_0, \Sigma_0)$ being a GMRF and with $M$ and $H$ sparse. Assume also that the forecast and analysis ensemble are generated by Algorithm 1, and, at every time $t$, the covariance matrix $\Sigma_t^f$ of the forecast is regular and its inverse $Q_t^f$ is in the span of the design matrices at time $t$, i.e.,

$$Q_t^f = \sum_{k=1}^r \beta_{tk} A_k,$$

for some $|\beta_{tk}|_{k=1}^r \in \mathbb{R}^r$. Suppose that the matrix $E_{x_t^f \sim f(\cdot | \mu^o, \Psi)} (D^*(X_t^f)D(X_t^f))$ is invertible for all $t \geq 1$. For $X_{t1}^f, \ldots, X_{tN}^f$, let $\hat{X}_{tN}^f$ and $\hat{\beta}_{tN} = [\hat{\beta}_{tNk}]_{k=1}^r$ be the SME computed from (21). Assume that $\hat{Q}_{tN}^f = \sum_{k=1}^r \hat{\beta}_{tNk} A_k$ is invertible and denote $\hat{\Sigma}_{tN}^f = (\hat{Q}_{tN}^f)^{-1}$. Then, for all $t \geq 1$,

$$(\hat{X}_{tN}^f, \hat{\Sigma}_{tN}^f) \overset{P}{\rightarrow} (\mu_t^f, \Sigma_t^f) \text{ as } N \rightarrow \infty.$$ \hfill (33)

**Proof.** At $t = 1$, Theorem 3.3 provides

$$(\hat{X}_{1N}^f, \hat{\beta}_{1N}) \overset{p}{\rightarrow} (\mu_1^f, \beta_1)$$

and

$$(\hat{X}_{1N}^f, \hat{Q}_{1N}^f) \overset{p}{\rightarrow} (\mu_1^f, Q_1^f),$$

since $Q_t^f$ is assumed to be of the form (32). Hence, (33) for $t = 1$ follows from Lemma 2.5.

Suppose now that (33) holds with $t-1$ in place of $t$ for some $t \geq 2$. Then, by the continuous mapping theorem applied to the mapping $(\mu_{t-1}^f, \Sigma_{t-1}^f) \mapsto (\mu_{t-1}^a, \Sigma_{t-1}^a)$ defined by (30, 31), we obtain the convergence

$$(\hat{\mu}_{t-1,N}^a, \hat{\Sigma}_{t-1,N}^a) \overset{p}{\rightarrow} (\mu_{t-1}^a, \Sigma_{t-1}^a).$$

The forecast ensemble at time $t$ is then a sample from $N_n(M \hat{\mu}_{t-1,N}^a, M \hat{\Sigma}_{t-1,N}^a M^T + \Psi)$ and due to the sparsity of $M$ and $H$, each its member is a GMRF. Consider $(\hat{X}_{tN}^f, \hat{\beta}_{tN})$ computed from the forecast ensemble by using the expression (21). Then, it follows from Theorem 3.4 that $\hat{X}_{tN}^f$ and $\hat{\beta}_{tN}$ converge, i.e.,

$$(\hat{X}_{tN}^f, \hat{\beta}_{tN}) \overset{p}{\rightarrow} (\mu_t^f, \beta_t),$$

and therefore,

$$(\hat{X}_{tN}^f, (\hat{Q}_{tN}^f)^{-1}) \overset{p}{\rightarrow} (\mu_t^f, \Sigma_t^f),$$

as $N \rightarrow \infty$. \hfill (34)
where the inverse exists with probability approaching one as $N \to \infty$, cf., Lemma 2.5.

A common practice in applied tasks is to plug in some regularized covariance estimate into the Kalman update formula. Using a linear model for precision matrix and SME from Theorem 3.1 in the inverse update formula results in a filter that we call the **Score matching ensemble filter (SMEF)**.

**Algorithm 2 (SMEF).** Choose a set of design matrices $\{A_k\}_{k=1}^r$. The initial ensemble $X_{01}, \ldots, X_{0N}$ is sampled from $\mathcal{N}_n(\mu_0, \Sigma_0)$. For $t \geq 1$:

1. **Forecast step:** The analysis ensemble is pushed forward in time by the (imperfect) model $M$:
   \[
   X_{t_i}^f = M X_{t-1,i}^a + V_{t_i}, \quad i = 1, \ldots, N,
   \]
   with $V_{t_i} \sim \mathcal{N}_n(0, \Psi)$ being the model error. Compute the sample mean $\bar{X}_t^f$ and sample covariance $S_t$ from the forecast ensemble and then compute the regularized estimate of $Q_{t_i}^f$ by
   \[
   \hat{Q}_{t_i}^f = \sum_{k=1}^r \hat{\beta}_{tk} A_k,
   \]
   where $\hat{\beta}_t = (\hat{\beta}_{t1}, \ldots, \hat{\beta}_{tr})^\top$ is the SME from Theorem 3.1, which can be computed by solving the system
   \[
   \left( \text{tr} \left( S_t A_k A_l \right) \right)_{k,l=1}^r \beta = \left( \text{tr} A_k \right)_{k=1}^r.
   \]

2. **Analysis step:** For a given observation $y_t$, the $i$-th member of the analysis ensemble is a solution of the system
   \[
   \left( \hat{Q}_{t_i}^f + H^\top R^{-1} H \right) X_{t_i}^a = \hat{Q}_{t_i}^f X_{t_i}^f + H^\top R^{-1} (y_t + W_{t_i}),
   \]
   where $W_{t_i} \sim \mathcal{N}_m(0, R), i = 1, \ldots, N$, are random perturbations.

In the simulation study of Section 5.1, the SMEF from Algorithm 2 performs well and outperforms the standard EnKF (without regularization) when the ensemble size is small or the variance of observation error is large. However, unlike the SMEF-GR, in this filter, the ensemble members are not normally distributed even for $t = 1$ and, therefore, the score matching estimator of $(\bar{\mu}_t^f, \bar{\beta}_t)$ cannot be shown to be consistent by the same method as in Theorem 4.1. Analysis along the lines of [21, 15, 16] would require an a-priori $L^p$ bound on the estimated covariance $(\hat{Q}_t^f)^{-1}$, which is not available. Hence, convergence in large ensemble asymptotics is left to future research.

4.3. **A filter of square-root type.** For a small ensemble size, it is usually beneficial to use a square-root filter and avoid sampling errors which influence the performance of filters using perturbed observations. Thus, we construct a variant of the Ensemble transform Kalman filter (ETKF, [19, 34]), where we modify the transform matrix, utilizing the information contained in the Markov structure of the precision matrix.

Recall the ensemble of anomalies defined in (22),
\[
Z = [X_1 - \bar{X}, \ldots, X_N - \bar{X}]^\top.
\]
In ETKF, the forecast state is updated by the standard Kalman update, which we may perform as in (27). The analysis ensemble is formed by adding transformed anomalies $Z_T$ to the mean, for a suitable transform matrix $T$. One possible choice of the transform matrix is [19, eq. (27)]

$$T = C(I_n + \Gamma)^{-1/2}C^\top,$$

resulting from the eigenvalue decomposition

$$\frac{1}{N-1} Z^\top H^\top R^{-1} H Z = C C^\top.$$

Consider the matrix $U_N$ consisting of $N$ eigenvectors belonging to the smallest eigenvalues in the decomposition of our estimated precision matrix,

$$\hat{Q}^f_t = U_A U^\top = [U_{n-N} \quad U_N] \begin{bmatrix} \Lambda_{n-N} & 0 \\ 0 & \Lambda_N \end{bmatrix} \begin{bmatrix} U^\top_n \\ U^\top_N \end{bmatrix}.$$

These eigenvectors belong to the largest $N$ eigenvalues of covariance and span the subspace, where we expect the largest anomalies. It seems natural to incorporate this information into the construction of the transform matrix. Let $\tilde{Z} = U_N U_N^\top Z$ be the anomalies projected onto the subspace of the first principal components of the estimated covariance matrix. We replace $Z$ by $\tilde{Z}$ in (35), construct the matrices $C$ and $\Gamma$ from the eigendecomposition

$$\frac{1}{N-1} \tilde{Z}^\top H^\top R^{-1} H \tilde{Z} = C C^\top$$

and put these $C$ and $\Gamma$ into (34). We refer to this algorithm as SMETKF.

### 4.4. Score Matching Extended Kalman Filter

If the tangent linear model and its adjoint are available, (21) allows a straightforward way for pushing forward the parameters of the linear model for precision matrix. A filter of the type of Extended Kalman filter may be set up, rephrasing the standard Extended Kalman filter algorithm in terms of the precision matrix and its linear model.

At $t = 1$, let $\hat{\mu}^f_t$ and $\hat{\beta}_1$ be the score matching estimates of $\mu^f_1$ and $\beta_1$ from Theorem 3.1 (either constructed from an initial ensemble or given by prior knowledge). For $t \geq 1$, we shall derive the forward model for $\beta_{t+1}$ from the explicit form of the score matching estimator.

Thus, assume that in time $t$, we have estimates $\hat{\mu}^f_t$ and $\hat{\beta}_t$. Then, the estimates of analysis mean and precision matrix are given by

$$\hat{Q}^a_t \hat{\mu}^a_t = \hat{Q}^f_t \hat{\mu}^f_t + H^\top R^{-1} y_t$$

$$\hat{Q}^a_t = \hat{Q}^f_t + H^\top R^{-1} H,$$

where

$$\hat{Q}^f_t = \sum_{j=1}^r \hat{\beta}_{tj} A_j.$$

The forecast mean and covariance matrix at $t + 1$ are

$$\hat{\mu}^f_{t+1} = M \hat{\mu}^a_t$$

$$\hat{\Sigma}^f_{t+1} = M \left( \hat{Q}^a_t \right)^{-1} M^\top + \Psi.$$
The estimate (36) is then used in (21) in place of the sample covariance $S$. Hence, the estimate of $\beta_{t+1}$ is

$$
\hat{\beta}_{t+1} = \left( \left[ \text{tr} \left( \left( M \left( \hat{Q}_t^{-1} M^T + \Psi \right) A_k A_l \right) \right)_{k,l=1}^r \right]^{-1} \cdot \left[ \text{tr}(A_k) \right]_{k=1}^r \right)
$$

using the invariance of trace under cyclic permutation and the symmetry of $A_k$. We refer to this filter as Score matching extended Kalman filter (SMExKF).

Algorithm 3 (SMExKF). Assume we have a set of initial parameters $\hat{\mu}_f^1$ and $\hat{\beta}_1$, viewed as a forecast. For $t \geq 1$:

1. **Analysis step:** The updated state vector $\hat{\mu}_a^t$ is obtained from $\hat{\mu}_f^t$ according to the inverse update (27) with $\hat{Q}_t^a = \sum_{j=1}^r \hat{\beta}_{tj} A_j + H^T R^{-1} H$.

2. **Forecast step:** Put $\hat{\mu}_f^{t+1} = M \hat{\mu}_a^t$. For a small model with $M$ available as a matrix, the construction of the forecast covariance matrix at $t+1$ is straightforward:

$$
\hat{\Sigma}_f^{t+1} = M \left( \hat{Q}_t^a \right)^{-1} M^T + \Psi.
$$

If the (tangent linear) model $M$ and its adjoint are available in form of a code only, we have to propagate the covariance in form of an artificial ensemble and use an SVD of $\hat{Q}_t^a$ similarly as in SMETKF. For a suitable $N$, let $U_N$ consist of $N$ eigenvectors belonging to the smallest $N$ eigenvalues in the decomposition of our estimated precision matrix $\hat{Q}_t^a = U \Lambda U^T = [U_{n-N} \ U_N] \begin{bmatrix} \Lambda_{n-N} & 0 \\ 0 & \Lambda_N \end{bmatrix} \begin{bmatrix} U_{n-N}^T \\ U_N^T \end{bmatrix}$.

Now, push forward the “ensemble” of columns of $\Lambda_N^{1/2} U_N$ one time step ahead and form an approximation of forecast covariance from these columns:

$$
\hat{\Sigma}_f^{t+1} = (M \Lambda_N^{1/2} U_N)(M \Lambda_N^{1/2} U_N)^T + \Psi.
$$

Finally, compute $\hat{\beta}_{t+1}$ from (21) with the forecast covariance $\hat{\Sigma}_f^{t+1}$ in place of the sample covariance $S$.

**Remark 3.** The consistence of the estimated analysis mean $\hat{\mu}_a^t$ and covariance $\hat{\Sigma}_a^t := \left( \sum_{k=1}^r \hat{\beta}_{tk} A_k + H^T R^{-1} H \right)^{-1}$ follows from the consistence of $(\hat{\mu}_f^t, \hat{\beta}_1)$ and the continuous mapping theorem.

5. **Practical aspects and a simulation study.** In the previous sections, we proposed several versions of filters based on the score matching estimation method. As a preliminary evaluation, we carried out a simulation study and compared their performance with reference EnKF and ETKF filters for the Lorenz 96 model.
The performance of the filters is measured by the root-mean-square-error of the analysis ensemble mean, given by

$$\text{RMSE}_t = \sqrt{\frac{1}{n} \| \mathbf{X}_t - \bar{\mathbf{X}}_t \|^2},$$

(37)

for every time step $t$. Recall that $n$ is the state vector dimension, $\mathbf{X}_t$ denotes the true system state and $\bar{\mathbf{X}}_t$ is the analysis ensemble mean produced by the filtering algorithm.

5.1. Lorenz 96. This model was published in [20] as a candidate for a one-dimensional atmospheric model. The system state is represented by a random vector $\mathbf{X}_t = (X_{t,1}, \ldots, X_{t,n})^\top$ defined on a discrete circle with $n = 40$ points. The time evolution of each point $j \in \{1, \ldots, n\}$ is governed by the equations

$$X_{t+1,j} = X_{t,j} + \frac{dX_{t,j}}{dt}, \quad t \in \mathbb{N},$$

(38)

and

$$\frac{dX_{t,j}}{dt} = (X_{t,j+1} - X_{t,j-2})X_{t,j-1} - X_{t,j} + F,$$

(39)

where $X_{t,-1} = X_{t,n-1}$, $X_{t,0} = X_{t,n}$, $X_{t,n+1} = X_{t,1}$. The forcing term $F$ was set to 8, which is a known value that causes chaotic behaviour. For simplicity and comparison with earlier literature, we used the “perfect model” setup with $\Psi = 0$. The empirical distribution of Euclidean distance of two randomly selected states, computed from our pool, had a mean of approximately 5.11. Based on this fact, we classified any filter simulation as divergent if the distance of analysis from the truth exceeded a threshold of 6, which roughly corresponds to the upper 96% quantile of the empirical distribution.

5.1.1. Choice of design matrices for Lorenz 96 model. The state of the Lorenz system does not have the spatial Markov property in the traditional sense of (19), however, equation (39) foreshadows some kind of relationship between each point and its neighbours. We use an approximate Gaussian Markov model for the precision matrix $Q$ with $q_{ij}$ allowed to be nonzero if $(i,j) \in \mathcal{I}$, where

$$\mathcal{I} = \{(i,j) : i, j \in \{1, \ldots, n\}, \ |i - j + kn| \leq 4, \ \text{for some} \ k \in \{-1, 0, 1\}\}.$$

That is, the separation neighbourhood of every node in the Markov property (19) consists of 4 nodes on each side. The nonzeros of the matrix $Q$ are thus constrained to the diagonal, 4 co-diagonals below and 4 above, and the lower left and the upper right corners arising from periodicity. We choose the set of design matrices $A$ so that the matrices $A_{ii}$, $i = 1, \ldots, n$, corresponding to the diagonal entries of the precision matrix, are placed first. Then the first $n$ diagonal entries of $\tilde{D}$ are

$$\tilde{D}_{(i,i),(i,i)} = s_{ii},$$

(40)
and the remaining diagonal entries have the form
\[ \tilde{D}_{(i,j),(i,j)} = \text{tr}(SA_{ij}A_{ij}) = s_{ii} + s_{jj}, \quad i \neq j. \] (41)

The sparsity structure of the matrix \( \tilde{D} \) with such ordering and \( n = 40 \) is depicted in Figure 2, together with the structure of the precision matrix \( Q \).

The form \( Q = \sum_{k=1}^{r} \beta_k A_k \) of the precision matrix model generally does not guarantee that the score matching estimate \( \hat{Q} \) is positive definite. With the choice of the design matrices here, boosting of the diagonal of \( \hat{D} \) has a similar effect as boosting the diagonal of the sample covariance \( S \) due to (40, 41), which tends to push \( \hat{Q} \) towards being positive definite as well. In the simulation study below, it was sufficient to add the same constant 0.2 (for \( N = 10 \)) or 0.1 (for \( N = 20, 30 \)) to all diagonal entries of the matrix \( \tilde{D} \) if every second element of the state vector was observed. In the case when every fourth element was observed, this constant was increased to 0.35.

Another option to push \( \hat{Q} \) towards being positive definite is \( l_1 \) penalization of the coefficients \( \{\beta_k\}_{k=1}^{r} \) corresponding to off-diagonal entries of the precision matrix. This can be implemented easily by regression LASSO with weights.

The design matrices considered above can be viewed as “elementary” but there is a large flexibility in their choice. As an example consider a case when we have some information on smoothness of the field so that the variances and covariances in neighbouring gridpoints cannot jump abruptly. To reflect this fact, the precision matrix can be further parametrized by e.g. a B-spline expansion, choosing an appropriate degree of the polynomials. The basis is given analytically and for each B-spline basis vector we can construct a corresponding design matrix consisting of values of the basis vector on the gridpoints. Then the expansion, discretized on the gridpoints, is performed by the linear model (20). In this way, the number of unknown parameters can be further reduced. This approach has been tested in [31] for a real meteorological field transformed to the wavelet space.

5.1.2. Simulation results for Lorenz 96 model where every second element of the state vector is observed. Small ensemble size. The settings of our study took some inspiration from [26], where the case \( N = 35, R = I_n \) was considered and a complete state was observed. We focus on smaller ensemble sizes and larger error variance
Table 1. Lorenz 96: RMSE of different filtering algorithms averaged from 1000 time steps and 50 replications of the run, initialized from different randomly selected states. The “div” columns show the number of divergent replications.

| N  | ρ  | EnKF RMSE | ETKF RMSE | SMEF RMSE | SMETKF RMSE | SMEF-GR RMSE |
|----|----|-----------|-----------|-----------|-------------|--------------|
| 10 | 1  | 4.31 50   | 3.54 50   | 1.00 1    | 0.55 0      | 0.80 0       |
| 10 | 2  | 4.30 50   | 3.63 50   | 0.89 0    | 0.84 0      | 1.18 13      |
| 10 | 3  | 4.27 50   | 3.85 50   | 1.52 2    | 1.42 0      | 1.59 35      |
| 20 | 1  | 2.79 48   | 0.45 0    | 0.43 0    | 0.40 0      | 0.61 0       |
| 20 | 2  | 3.55 48   | 0.74 0    | 0.62 1    | 0.55 0      | 0.87 0       |
| 20 | 3  | 3.37 48   | 0.97 0    | 1.15 8    | 0.82 2      | 1.16 3       |
| 30 | 1  | 0.44 0    | 0.45 0    | 0.41 0    | 0.40 0      | 0.56 0       |
| 30 | 2  | 0.69 0    | 0.68 0    | 0.55 0    | 0.55 0      | 0.75 0       |
| 30 | 3  | 0.95 0    | 0.88 0    | 0.74 0    | 0.72 2      | 0.94 0       |

in observations and thus we use larger inflation than [26]. At the beginning, we generate a vector representing the truth and an ensemble of $N$ vectors resulting from a random $N(0,1)$ perturbation of elements of the true state vector. Both the initial ensemble and the true state were forwarded two time steps, then assimilation started. First 200 steps of assimilation were reserved for the spin-up, then another 1000 steps were run. The observations have been selected as every second element of the state vector, i.e. 20 equidistant elements. The ensemble size was $N=10, 20$ and 30, the matrix $R$ was diagonal, $R = \rho I_m$ with $\rho = 1, 2, \text{and} 3$. Each simulation was replicated 50 times with the true state vector selected at random from the pool of states. If, after the spin-up, the RMSE of a replication exceeded the value of 6 (as discussed above), the replication was stopped and considered divergent. The value of RMSE reported in graphs is a mean of available replications (those which did not diverge up to current time). If the number of divergent replications is 50, the corresponding graph ends before step 1200.

We compared the RMSE in the update step (analysis) of five filters: EnKF with inflation 1.12, ETKF with inflation 1.12, SMETKF (Section 4.3), SMEF (Algorithm 2) and SMEF-GR (Algorithm 1) with the true state. Figure 3 is an illustration of the behaviour of the RMSE (37) of the filters for different combinations of ensemble size $N$ and observation error $\rho$. In the upper plot which corresponds to the least favourable case $N = 10, \rho = 3$, both EnKF and ETKF diverge whereas the regularized SME-based filters are relatively stable, with higher risk of instability in the resampling filter. In the “average” situation, the EnKF still diverges whereas ETKF and SMEF-GR perform about the same and the remaining filters slightly better. In the lower plot, where the ensemble size is relatively high and observation error smaller, the EnKF performs better than SMEF-GR and the other filters are comparable. A comparison of overall error (computed as mean in time steps $\geq 200$) for all combinations is recorded in Table 1.

From Table 1, we may infer that the SME-based filters show a better performance and are less prone to diverge. The resampling filter is heavily dependent on Gaussianity; its virtues will become more evident in the case when every fourth element of the state vector is observed (see the next section).
Figure 3. Performance illustration of EnKF, ETKF, SMEF (Algorithm 2), SMETKF (Section 4.3) and SMEF-GR (Algorithm 1) for the Lorenz 96 model, for selected combinations of ensemble size and observation error. $N = 10, \rho = 3$ (a), $N = 20, \rho = 2$ (b), $N = 30, \rho = 1$ (c). Observation step 2.

For completeness, we include simulation results for the SMExKF filter (Algorithm 3). In a small model like Lorenz 96, its application is straightforward. If, however, the tangent model and its adjoint are available as a code only, the forecast covariance matrix has to be approximated.

We rather want to illustrate here that there exists a tradeoff between the degree of regularization (as provided e.g. by SME) and overall performance measured
by RMSE. A regularized filter ensures a lower risk of gross errors and higher filter stability but cannot possibly achieve the highest performance. This issue is documented in Figure 4 showing averaged RMSE as before and Figure 5 where, for $\rho = 2$, all 50 replications of the filtering process are depicted. Although overall performance of ExKF is better for $\rho = 2$, more “excurses” and higher maximal errors are present in ExKF than in SMExKF. Finally for $\rho = 3$, regularization is necessary to keep stability of the filter. A summary of this simulation is in Table 2.

\begin{table}[h]
\centering
\begin{tabular}{c|cc|cc}
\hline
Obs. step & $\rho$ & ExKF & & SMExKF & \\
\hline
\hline
2 & 1 & 0.34 & 0 & 0.54 & 0 \\
2 & 2 & 0.53 & 1 & 0.71 & 0 \\
2 & 3 & 1.71 & 16 & 0.86 & 0 \\
\hline
\end{tabular}
\caption{Lorenz 96: RMSE of the two non-ensemble filtering algorithms averaged from 1000 time steps and 50 replications of the run, initialized from different randomly selected states. The “div” columns show the number of divergent replications.}
\end{table}

5.1.3. Simulation results for Lorenz 96 model where every fourth element of the state vector is observed. Larger ensemble size. In the preceding subsection, we have not
dealt with larger ensemble sizes, since, for $\rho = 1, 2$ or 3, enlarging the ensemble size does not improve the performance of the filters. On the other hand, for ensemble sizes 10, 20, or 30 and observation step 4, all filters diverge (not shown). It is of some interest, however, to tackle the situation with observation step 4, larger ensemble size and larger $\rho$. All filters perform at a lower level of accuracy. For SMEF and SMETKF as well as for EnKF and ETKF, there is a large number of divergent replications. This case is apparently closer to Gaussianity and the SMEF-GR is far more robust than its competitors and outperforms them, although those replications of SMEF that do not diverge give a slightly better result. The results are summarised in Table 3 and a sample plot of RMSE is in Figure 6.

6. Discussion and conclusion. The score matching filters present an alternative method of covariance regularization. They also avoid numeric optimization of the likelihood and despite being “inverse”, they are computationally feasible, as they require a solution of a sparse linear system (or possibly extraction of a small number of eigenvectors of the estimated precision matrix). This is nowadays achievable in many practical situations. They are particularly suited for the situation where the underlying random field is Gaussian Markov, in which case they have nice asymptotic properties.

The filters show a good performance for the Lorenz 96 model and a reasonable robustness in situations where the sample size is small. It is seen that if there is not
Table 3. Lorenz 96: RMSE of different filtering algorithms averaged from 1000 time steps and 50 replications of the run, initialized from different randomly selected states. The “div” columns show the number of divergent replications. Observation step 4.

| N  | r | EnKF RMSE | EnKF div | ETKF RMSE | ETKF div | SMEF RMSE | SMEF div | SMETKF RMSE | SMETKF div | SMEF-GR RMSE | SMEF-GR div |
|----|---|-----------|----------|-----------|---------|-----------|---------|-------------|-----------|-------------|-----------|
| 30 | 3.0 | 3.46 | 46 | 3.35 | 40 | 2.49 | 30 | 3.51 | 34 | 2.53 | 1 |
| 30 | 4.0 | 3.56 | 40 | 3.50 | 41 | 2.63 | 45 | 3.66 | 43 | 2.67 | 0 |
| 40 | 3.0 | 3.55 | 36 | 3.32 | 24 | 2.05 | 23 | 3.35 | 39 | 2.34 | 0 |
| 40 | 4.0 | 3.75 | 30 | 3.58 | 23 | 2.42 | 47 | 3.56 | 42 | 2.49 | 0 |
| 50 | 3.0 | 3.85 | 33 | 3.53 | 15 | 2.02 | 19 | 3.25 | 35 | 2.18 | 0 |
| 50 | 4.0 | 3.98 | 39 | 3.75 | 28 | 2.22 | 41 | 3.39 | 44 | 2.36 | 1 |

Figure 6. RMSE, averaged over 1000 replications, $N = 50$, observation step 4, $\rho = 3$ (a), $\rho = 4$ (b).

enough information on the true covariance structure, a regularized approximation by a Gaussian Markov model may be beneficial. Some preliminary investigation of the score matching estimation of precision matrix in real problems has been performed in [31], where regularization of background covariance matrix for the numeric weather prediction model WRF has been tackled.
Acknowledgments. Turčičová was partially supported by the project TL01000238 funded by Technology Agency of the Czech Republic. Mandel was partially supported by the U.S. National Science Foundation under grant ICER-1664175. Eben was supported by the project TO01000219 funded by Norway Grants and Technology Agency of the Czech Republic.

Appendix A. Proof of Lemma 2.4. Recall that for a random vector $X$ with a probability density $p$ on $\mathcal{X} \subset \mathbb{R}^n$, we denote

$$E_{X \sim p} X = \int_{\mathcal{X}} x p(x) dx.$$ 

When $X$ has finite second moments, we will further denote

$$\text{Cov}_{X \sim p}(X) = \int_{\mathcal{X}} (x - E_{X \sim p} X)(x - E_{X \sim p} X)^\top p(x) dx.$$ 

**Lemma A.1.** Suppose that $f(x|\theta)$ is a parametric probability density on a measurable set $\mathcal{X} \subset \mathbb{R}^n$ with parameter $\theta \in \Theta \subset \mathcal{L}$, $\Theta$ open, such that for each $\theta$, $C(\theta) \equiv \text{Cov}_{X \sim f(\cdot|\theta)}(X)$ exists and is continuous from $\Theta$ to $\mathbb{R}^{n \times n}$. Suppose that $\theta_N$ are random parameters such that $\theta_N \xrightarrow{p} \theta^0 \in \Theta$ as $N \to \infty$, and, for each $N$, $\{X_{N_i} : i = 1, \ldots, N\}$ is a sample from $f(\cdot|\theta_N)$. Then,

$$\frac{1}{N} \sum_{i=1}^{N} X_{N_i} - E_{X_{N_1} \sim f(\cdot|\theta_N)}(X_{N_1}) \xrightarrow{p} 0 \text{ as } N \to \infty.$$ 

**Proof.** Since $\text{Cov}_{X_{N_1} \sim f(\cdot|\theta_N)}(X_{N_1})$ exists by assumption, $E_{X_{N_1} \sim f(\cdot|\theta_N)}(X_{N_1})$ exists. In order to simplify the notation, we will denote

$$E_{X_{N_1} \sim f(\cdot|\theta_N)}(X_{N_1}) = E(X_{N_1}|\theta_N)$$

$$\text{Cov}_{X_{N_1} \sim f(\cdot|\theta_N)}(X_{N_1}) = \text{Cov}(X_{N_1}|\theta_N) = C(\theta_N).$$

Further, denote

$$W_{N_i} = X_{N_i} - E(X_{N_1}|\theta_N), \quad \bar{W}_N = \frac{1}{N} \sum_{i=1}^{N} W_{N_i}.$$ 

We need to show that $\bar{W}_N \xrightarrow{p} 0$ as $N \to \infty$. Fix $N$ and $\theta_N$. Then,

$$E(W_{N_1}|\theta_N) = 0,$$

$$\text{Cov}(W_{N_1}|\theta_N) = \text{Cov}(X_{N_1}|\theta_N) = C(\theta_N)$$

and, since $W_{N_i}$ are uncorrelated, following the standard $L^2$ law of large numbers argument,

$$E\left(\|\bar{W}_N\|_{\mathbb{R}^n}^2 | \theta_N\right) = \frac{1}{N^2} \sum_{i=1}^{N} E\left(\|W_{N_i}\|_{\mathbb{R}^n}^2 | \theta_N\right) = \frac{1}{N} E\left(\|W_{N_1}\|_{\mathbb{R}^n}^2 | \theta_N\right)$$

$$= \frac{1}{N} \text{tr} \left(\text{Cov}(W_{N_1} | \theta_N)\right) = \frac{1}{N} \text{tr} (C(\theta_N)).$$

However, since $\theta_N$ is not constant, we cannot conclude convergence immediately and need to modify the standard argument. Let $\varepsilon > 0$ and $\delta > 0$. Using the Markov inequality, we have

$$P\left(\|\bar{W}_N\|_{\mathbb{R}^n}^2 \geq \varepsilon^2 | \theta_N\right) \leq E\left(\|\bar{W}_N\|_{\mathbb{R}^n}^2 | \theta_N\right) \leq \frac{\text{tr} (C(\theta_N))}{\varepsilon^2}.$$
Since $C(\theta)$ is a continuous function of $\theta$ by assumption, there exists $\eta > 0$ such that
\[ \text{tr}(C(\theta)) < \text{tr}(C(\theta^0)) + 1 \quad \text{if} \quad \|\theta - \theta^0\|_L < \eta. \]
Since $\theta_N \xrightarrow{P} \theta^0$ for $N \to \infty$, there exists $N_1$ such that
\[ P(\|\theta_N - \theta^0\|_L \geq \eta) < \frac{\delta}{2} \quad \text{if} \quad N \geq N_1. \]
Then, by the law of total probability,
\[
P(\|W_N\|_{\mathbb{R}^n} \geq \varepsilon) = \sum_{\varepsilon_0} \varepsilon_0 P(\|W_N\|_{\mathbb{R}^n} \geq \varepsilon_0 | \|\theta_N - \theta^0\|_L \geq \eta) \cdot P(\|\theta_N - \theta^0\|_L \geq \eta)
+ \sum_{\varepsilon_1} \varepsilon_1 P(\|W_N\|_{\mathbb{R}^n} \geq \varepsilon_1 | \|\theta_N - \theta^0\|_L < \eta) \cdot P(\|\theta_N - \theta^0\|_L < \eta)
\]
\[ < \frac{\delta}{2} + \frac{\text{tr}(C(\theta^0)) + 1}{\varepsilon^2 N} \quad \text{if} \quad N \geq N_1. \]
There exists $N_2$ such that
\[ \frac{\text{tr}(C(\theta^0)) + 1}{\varepsilon^2 N} < \frac{\delta}{2} \quad \text{if} \quad N \geq N_2, \]
and, finally,
\[ P(\|W_N\|_{\mathbb{R}^n} \geq \varepsilon) < \frac{\delta}{2} + \frac{\delta}{2} = \delta \quad \text{if} \quad N > \max\{N_1, N_2\}, \]
which proves that $W_N \xrightarrow{P} 0$ as $N \to \infty$. \hfill \Box

**Proof of Lemma 2.4.** We use Lemma A.1 with $h(X_{N_i})$ in place of $X_{N_i}$. Since all norms on a finite dimensional space are equivalent, we can assume without loss of generality that $W = \mathbb{R}^n$. From (15), for all $\theta \in \Theta$, the integrals
\[
M(\theta) = \int_X h(x) h(x)^T f(x|\theta) dx = e^{-a(\theta)} \int_X e^{(T(x), \theta)} h(x) h(x)^T e^{b(x)} dx
\]
\[
m(\theta) = \int_X h(x) f(x|\theta) dx = e^{-a(\theta)} \int_X e^{(T(x), \theta)} h(x) e^{b(x)} dx
\]
exist. Since $M(\theta)$ and $m(\theta)$ are Fourier-Laplace transforms of the sufficient statistic $T(x)$, they are analytic in $\Theta$ [17, Theorem 2.7.1][2, Theorem 7.2] and, in particular, continuous. Thus,
\[ C(\theta) = \text{Cov}_{X \sim f(\cdot|\theta)}(h(X)) = M(\theta) - m(\theta)m(\theta)^T \]
exists and is continuous from $\Theta$ to $\mathbb{R}^{n \times n}$. Now, from Lemma A.1,
\[ \frac{1}{N} \sum_{i=1}^N h(X_{N_i}) - E(h(X_{N})|\theta_N) \xrightarrow{P_{N \to \infty}} 0, \]
and since $m(\theta)$ is a continuous function of $\theta$,
\[ E(h(X_{N})|\theta_N) = m(\theta_N) \xrightarrow{P_{N \to \infty}} m(\theta^0) = E_{X \sim f(\cdot|\theta^0)} h(X), \]
by the continuous mapping theorem. \hfill \Box
Appendix B. Proof of Theorem 3.1.

Proof of Theorem 3.1. Normal distribution belongs to the exponential family of distributions, as specified in Section 2.4. So we can compute SME from (12) for the special case of normal distribution with a linear model for precision matrix. Dimension of the parameter space is \( s = n + r \). The density of \( \mathcal{N}_n(\mu, \Sigma) \) with \( \Sigma^{-1} = \sum_{k=1}^r \beta_k A_k \) is

\[
p(x|\mu, \beta) = \frac{q(x|\mu, \beta)}{\int_{\mathbb{R}^n} q(x|\mu, \beta)}, \tag{42}
\]

where

\[
\log q(x|\mu, \beta) = -\frac{1}{2} (x - \mu)^\top \sum_{k=1}^r \beta_k A_k (x - \mu)
\]

\[
= \left( \begin{array}{c} \sum_{k=1}^r \beta_k A_k \mu \\ \beta \end{array} \right)^\top \left( \begin{array}{c} x \\ -\frac{1}{2} [x^\top A_1 x, \ldots, x^\top A_r x]^\top \end{array} \right) \bigg|_L
\]

\[
- \frac{1}{2} \left( \begin{array}{c} \mu \\ \sum_{k=1}^r \beta_k A_k \mu \end{array} \right)^\top_L = (\eta, T(x))_L - a(\eta).
\]

Thus, we can identify the canonical parameter of the exponential family

\[
\eta = \left( \begin{array}{c} \eta_1 \\ \eta_2 \end{array} \right) = \left( \begin{array}{c} \sum_{k=1}^r \beta_k A_k \mu \\ \beta \end{array} \right) = \left( \begin{array}{c} Q(\beta) \mu \\ \beta \end{array} \right), \tag{43}
\]

the canonical sufficient statistic

\[
T(x) = \left( \begin{array}{c} x \\ -\frac{1}{2} [x^\top A_1 x, \ldots, x^\top A_r x]^\top \end{array} \right), \tag{44}
\]

and

\[
a(\eta) = \frac{1}{2} \langle \mu, Q(\beta) \mu \rangle_L, \quad b(x) = 0.
\]

The original parameters \((\mu, \beta)\) define probability density by (42) if and only if \((\mu, \beta) \in \tilde{\Theta} = \{(\mu, \beta) | \mu \in \mathbb{R}^n, Q(\beta) \text{ is positive definite} \} \).

Define

\[
Q(\eta_2) = \sum_{k=1}^r \eta_2 k A_k
\]

and

\[
\Theta = \left\{ \eta \in \mathbb{R}^{n+r} | \eta_1 \in \mathbb{R}^n, Q(\eta_2) \text{ is positive definite} \right\},
\]

where \( \eta_2 = [\eta_{2k}]_{k=1}^r \). Lemma B.1 (below) shows that (43) defines a one-to-one correspondence between \( \eta \in \Theta \) and \((\mu, \beta) \in \tilde{\Theta}\).

We are now ready to evaluate the estimate (12). From (13),

\[
D^*(x) = \nabla_x T(x) = \nabla_x \left[ \begin{array}{c} x \\ -\frac{1}{2} [x^\top A_1 x, \ldots, x^\top A_r x]^\top \end{array} \right] = \left[ \begin{array}{c} I_n \\ -[A_1 x, \ldots, A_r x]^\top \end{array} \right]
\]

and then \( D \) is the transpose,

\[
D(x) = [I_n, -[A_1 x, \ldots, A_r x]].
\]
\[ D^*(x)D(x) = \begin{bmatrix} I_n \\ -[A_1 x, \ldots, A_r x]^\top \end{bmatrix} [I_n, -[A_1 x, \ldots, A_r x]] \]
\[ = \begin{bmatrix} I_n \\ -[A_1 x, \ldots, A_r x]^\top \end{bmatrix} - [A_1 x, \ldots, A_r x] [x^\top A_k A_l x]_{k,l=1}^{r}. \tag{45} \]

By a direct computation, the Laplacian of the sufficient statistic from (44) is
\[ \Delta x T(x) = \begin{bmatrix} 0 \\ -[\text{tr}(A_1), \ldots, \text{tr}(A_r)]^\top \end{bmatrix}. \]

With a random sample \( X_1, \ldots, X_N \), we have the SME (12)
\[ \hat{\eta} = -\left( \frac{1}{N} \sum_{i=1}^N D^*(X_i)D(X_i) \right)^{-1} \frac{1}{N} \sum_{i=1}^N \Delta x T(X_i) \tag{46} \]
\[ = \begin{bmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ [\text{tr}(A_1), \ldots, \text{tr}(A_r)]^\top \end{bmatrix}, \]

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

Using the formula for the inverse of 2 \times 2 block matrix,
\[ \begin{bmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{bmatrix}^{-1} = \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix} = \begin{bmatrix} E_{11}^{-1} (I_n + E_{12} M_{22} E_{21} E_{11}^{-1}) - E_{11}^{-1} E_{12} M_{22} \\ -M_{22} E_{21} E_{11}^{-1} M_{22} \end{bmatrix}, \]

where
\[ M_{22}^{-1} = E_{22} - E_{21} E_{11}^{-1} E_{12} \]
\[ = \left( \frac{1}{N} \sum_{i=1}^N X_i^\top A_k A_l X_i \right)_{k,l=1}^r - \frac{1}{N} \left( \sum_{i=1}^N A_1 X_i, \ldots, \sum_{i=1}^N A_r X_i \right)^\top \frac{1}{N} \left[ \sum_{i=1}^N A_1 X_i, \ldots, \sum_{i=1}^N A_r X_i \right] \]
\[ = \left( \frac{1}{N} \sum_{i=1}^N X_i^\top A_k A_l X_i - \frac{1}{N} \sum_{i=1}^N X_i^\top A_k \frac{1}{N} \sum_{j=1}^N A_j X_j \right)_{k,l=1}^r \]
\[ = \text{tr} \left( \left( \frac{1}{N} \sum_{i=1}^N X_i X_i^\top - \frac{1}{N} \sum_{j=1}^N X_j \frac{1}{N} \sum_{i=1}^N X_i^\top \right) A_k A_l \right)_{k,l=1}^r \]
\[ = \text{tr} (SA_k A_l)_{k,l=1}^r. \]
Since $M_{22}$ exists by assumption, the inverse in (46) exists, and
\[
\hat{\eta} = -\begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix} \begin{bmatrix} 0 \\ (\text{tr}(A_1), \ldots, \text{tr}(A_r))^\top \end{bmatrix} \\
= \begin{bmatrix} -E_{11}^{-1}E_{12}M_{22} \\ M_{22} \end{bmatrix} (\text{tr}(A_1), \ldots, \text{tr}(A_r))^\top \\
= \begin{bmatrix} -E_{11}^{-1}E_{12} \\ I_n \end{bmatrix} M_{22} (\text{tr}(A_1), \ldots, \text{tr}(A_r))^\top,
\]
which gives
\[
\hat{\eta}_2 = \hat{\beta} = \left(\left[\text{tr} \left( S A_k A_{\ell} \right) \right]_{k,\ell=1}^r \right)^{-1} (\text{tr}(A_1), \ldots, \text{tr}(A_r))^\top
\]
and
\[
\hat{\eta}_1 = -E_{11}^{-1}E_{12}\hat{\eta}_2 = \frac{1}{N} \sum_{i=1}^N \left[ A_1 X_i, \ldots, A_r X_i \right] \hat{\beta}
\]
\[
= \frac{1}{N} \sum_{i=1}^N \sum_{k=1}^r \beta_k A_k X_i = \sum_{k=1}^r \beta_k A_k \bar{X}.
\]
By (43), $\eta_1 = \sum_{k=1}^r \beta_k A_k \mu$, and since the mapping of the parameters $(\eta_1, \eta_2)$ and the original parameters $(\mu, \beta)$ is one-to-one by Lemma B.1 below, it follows that $\hat{\mu} = \bar{X}$.

**Lemma B.1.** Let all the assumptions of Theorem 3.1 hold and denote
\[
\tilde{\Theta} = \left\{ (\mu, \beta) \mid \mu \in \mathbb{R}^n, \sum_{k=1}^r \beta_k A_k \text{ is positive definite} \right\}
\]
\[
\Theta = \left\{ \eta = \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} \mid \eta_1 \in \mathbb{R}^n, \sum_{k=1}^r \eta_2 A_k \text{ is positive definite} \right\}.
\]
Then
\[
\eta = \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} = \begin{bmatrix} \sum_{k=1}^r \beta_k A_k \mu \\ \beta \end{bmatrix}
\]
(47) defines a homeomorphism between $\eta \in \Theta$ and $(\mu, \beta) \in \tilde{\Theta}$.

**Proof.** Evaluating (47), $(\mu, \beta) \in \tilde{\Theta}$ gives a unique $\eta \in \Theta$. In the opposite direction, if $\eta \in \Theta$, then $\eta_2 = \beta$ and, since $\sum_{k=1}^r \beta_k A_k$ is nonsingular, $\mu = \left(\sum_{k=1}^r \beta_k A_k\right)^{-1} \eta_2$.

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

**REFERENCES**

[1] T. W. Anderson, Asymptotically efficient estimation of covariance matrices with linear structure, *Ann. Statist.*, 1 (1973), 135–141.

[2] O. Barndorff-Nielsen, *Information and Exponential Families in Statistical Theory*, Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons, Ltd., Chichester, 1978.

[3] M. Bukal, I. Marković and I. Petrović, Score matching based assumed density filtering with the von Mises-Fisher distribution, 20th International Conference on Information Fusion (Fusion), Xi’an, China, 2017.
[4] G. Burgers, P. J. van Leeuwen and G. Evensen, Analysis scheme in the ensemble Kalman filter, *Monthly Weather Review*, 126 (1998), 1719–1724.

[5] A. P. Dempster, Covariance selection, *Biometrics*, 28 (1972), 157–175.

[6] P. G. M. Forbes and S. Lauritzen, Linear estimating equations for exponential families with application to Gaussian linear concentration models, *Linear Algebra Appl.*, 473 (2015), 261–283.

[7] R. Furrer and T. Bengtsson, Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants, *J. Multivariate Anal.*, 98 (2007), 227–255.

[8] J. E. Gentle, *Matrix Algebra. Theory, Computations and Applications in Statistics*, 2nd edition, Springer Texts in Statistics, Springer, Cham, 2017.

[9] A. Hyvärinen, Estimation of non-normalized statistical models by score matching, *J. Mach. Learn. Res.*, 6 (2005), 695–709.

[10] A. Hyvärinen, Some extensions of score matching, *Comput. Statist. Data Anal.*, 51 (2007), 2499–2512.

[11] I. Kasanický, J. Mandel and M. Vejmelka, Spectral diagonal ensemble Kalman filters, *Nonlin. Processes Geophys.*, 22 (2015), 485–497.

[12] M. Katzfuss, J. R. Stroud and C. K. Wikle, Understanding the ensemble Kalman filter, *Amer. Statist.*, 70 (2016), 350–357.

[13] S. L. Lauritzen, *Graphical Models*, Oxford Statistical Science Series, 17, Oxford Science Publications, The Clarendon Press, Oxford University Press, New York, 1996.

[14] K. Law, A. Stuart and K. Zygalakis, *Data assimilation. A Mathematical Introduction*, Texts in Applied Mathematics, 62, Springer, Cham, 2015.

[15] K. Law, H. Tembine and R. Tempone, Deterministic mean-field ensemble Kalman filtering, *SIAM J. Multivariate Anal.*, 38 (2016), A1251–A1279.

[16] F. Le Gland, V. Monbet and V.-D. Tran, Large sample asymptotics for the ensemble Kalman filter, in *The Oxford Handbook of Nonlinear Filtering*, Oxford Univ. Press, Oxford, 2011, 598–631.

[17] E. L. Lehmann and J. P. Romano, *Testing Statistical Hypotheses*, 3rd edition, Springer Texts in Statistics, Springer, New York, 2005.

[18] F. Lindgren, H. Rue and J. Lindström, An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach, *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 73 (2011), 423–498.

[19] D. M. Livings, S. L. Dance and N. K. Nichols, Unbiased ensemble square root filters, *Phys. D.*, 237 (2008), 1021–1028.

[20] E. N. Lorenz, Predictability - A problem partly solved, in *Predictability of Weather and Climate*, Cambridge University Press, 2006, 40–58.

[21] J. Mandel, L. Cobb and J. D. Beezley, On the convergence of the ensemble Kalman filter, *Appl. Math.*, 56 (2011), 533–541.

[22] E. D. Nino-Ruiz and A. Sandu, Ensemble Kalman filter implementations based on shrinkage covariance matrix estimation, *Ocean Dynamics*, 65 (2015), 1423–1439.

[23] O. Pannekoucke, L. Berre and G. Desroziers, Filtering properties of wavelets for local background-error correlations, *Quart. J. Roy. Meteorol. Soc.*, 133 (2007), 363–379.

[24] J. A. Rozanov, Markov random fields, and stochastic partial differential equations, *Mat. Sb. (N.S.)*, 103 (1977), 590–613.

[25] H. Rue and L. Held, *Gaussian Markov Random Fields. Theory and Applications*, Monographs on Statistics and Applied Probability, 104, Chapman & Hall/CRC, Boca Raton, FL, 2005.

[26] P. Sakov and P. Oke, A deterministic formulation of the ensemble Kalman filter: An alternative to ensemble square root filters, *Tellus A*, 60 (2008), 361–371.

[27] D. Simpson, F. Lindgren and H. Rue, Think continuous: Markovian Gaussian models in spatial statistics, *Statistical Science*, 1 (2012), 16–29.

[28] A. Spantini, R. Baptista and Y. Marzouk, Coupling techniques for nonlinear ensemble filtering, preprint, arXiv:1907.00389.

[29] M. K. Tippett, J. L. Anderson, C. H. Bishop, T. M. Hamill and J. S. Whitaker, Ensemble square root filters, *Monthly Weather Review*, 131 (2003), 1485–1490.

[30] F. Tronarp, R. Hostettler and S. Särkkä, Continuous-discrete von Mises-Fisher filtering on S2 for reference vector tracking, 21st International Conference on Information Fusion (FUSION), Cambridge, UK, 2018.

[31] M. Turčíčová, *Covariance Estimation for Filtering in High Dimension*, Ph.D thesis, Charles University, 2021. Available from: http://hdl.handle.net/20.500.11956/136400.
[32] G. Ueno and T. Tsuchiya, Covariance regularization in inverse space, Quart. J. Roy. Meterol. Soc., 135 (2009), 1133–1156.

[33] A. W. van der Vaart, Asymptotic Statistics, Cambridge Series in Statistical and Probabilistic Mathematics, 3, Cambridge University Press, Cambridge, 1998.

[34] X. Wang, C. H. Bishop and S. J. Julier, Which is better, an ensemble of positive–negative pairs or a centered spherical simplex ensemble?, Monthly Weather Review, 132 (2004), 1590–1605.

[35] A. Weaver and P. Courtier, Correlation modelling on the sphere using a generalized diffusion equation, Quart. J. Roy. Meterol. Soc., 127 (2001), 1815–1846.

[36] A. T. Weaver and I. Mirouze, On the diffusion equation and its application to isotropic and anisotropic correlation modelling in variational assimilation, Quart. J. Roy. Meterol. Soc., 139 (2013), 242–260.

Received for publication May 2021; early access November 2021.

E-mail address: turcicova@cs.cas.cz
E-mail address: jan.mandel@ucdenver.edu
E-mail address: eben@cs.cas.cz