Maximum Likelihood Estimation of Diagonal Covariance Matrix

Turčičová, Marie
2016

Institute of Computer Science
The Czech Academy of Sciences

Maximum likelihood estimation of a diagonal covariance matrix

Marie Turčičová, Jan Mandel, Kryštof Eben

Technical report No. V-1228

January 2016

Pod Vodárenskou věží 2, 182 07 Prague 8, phone: +420 266 052 089, fax: +420 286 585 789,
e-mail:turcicova@cs.cas.cz

Institute of Computer Science
The Czech Academy of Sciences

Maximum likelihood estimation of a diagonal covariance matrix

Marie Turčičová[1], Jan Mandel[2], Kryštof Eben[3]

Technical report No. V-1228

January 2016

Abstract:

Sparse approximation of a covariance in spectral space is an actual topic in data assimilation because it significantly reduces the computational cost arising from the big dimension of data. In this paper we focused on a situation when the covariance matrix in the spectral space is diagonal and we estimated its diagonal entries by the maximum likelihood method. Further we assumed a special parametric structure of the diagonal and computed the maximum likelihood estimator for this situation. The bottom line of this paper is a computation and comparison of variances of these two estimators. In accordance with our intuition, the estimate computed under the correct parametric assumption has smaller variance than the estimate computed without any additional assumptions.

Keywords:
maximum likelihood estimator, parametric model, Fisher information, delta method

[1]Institute of Computer Science, The Czech Academy of Sciences, Pod Vodárenskou věží 2, 182 07, Prague 8, Czech Republic, E-mail: turcicova@cs.cas.cz
[2]Department of Mathematical and Statistical Sciences, University of Colorado Denver, Denver, CO 80217-3364, E-mail: Jan.Mandel@ucdenver.edu
[3]Institute of Computer Science, The Czech Academy of Sciences, Pod Vodárenskou věží 2, 182 07, Prague 8, Czech Republic, E-mail: eben@cs.cas.cz

1 Introduction

Sparse approximation of covariance in spectral domain have been used in data assimilation for some time. One motivation is the fact that a random field in cartesian geometry is second order stationary (that is, the covariance between the values at two points depends only on their distance vector) if and only if its covariance in the Fourier space is diagonal e.g., [11]. On a sphere, an isotropic random field has diagonal covariance in the basis of spherical harmonics [3]. For wavelets, the effect of the diagonal spectral approximation is equivalent to a weighted spatial averaging of local covariance functions [11].

Thus, diagonal, and, more generally, sparse covariance models in a spectral basis are useful in applications. The optimal statistical interpolation system from [12] was based on a diagonal approximation in spherical harmonics. already used as horizontal basis functions in the model, and a change of state variables into physically balanced analysis variables. The ECMWF 3DVAR system [5] also used diagonal covariance in spherical harmonics. Diagonal approximation in the Fourier space for homogeneous 2D error fields, with physically balanced crosscovariances, was proposed in [2]. The Fourier diagonalization approach was extended by [11] to sparse wavelet representation of the background covariance, and into a combined spatial and spectral localization by [4]. Restricting sample covariance to its diagnal in spectral space was in the ensemble Kalman filter has an effect similar to localization and it allows the successful use of very small ensembles [1, 7, 9].

Therefore, error estimates for spectral diagonal models are of interest. In [7], it was proved that if the covariance is diagonal in the spectral space, then setting all off-diagonal terms of the sample covariance in the spectral basis to zero improves the sampling error of the covariance in the Frobenius norm, generalizing a similar estimate in [6]. However, the diagonal still has considerable sampling error, which can be reduced by exploiting prior knowledge about the behavior of the diagonal entries, such as the known rate of decay. Therefore, it was suggested in [13] to fit the diagonal entries of the covariance to a model with a small number of parameters determined by maximum likelihood.

In this paper, we use the theory of maximum likelihood estimation to prove that if the true covariance is in fact in the form assumed by the model, then maximum likelihood estimation will improve the estimation asymptotically. The abstract result can be used hierarchically: sample covariance is the maximum likelihood estimate when nothing is known; the diagonal covariance is the maximum from the maximum likelihood estimate in the space of diagonal matrices; and diagonal covariance with the diagonal entries dependent on a small parameter is obtained from maximum likelihood estimation in the space of such parametric representation. In each case, the more specific model results in a smaller variance, if the covariance is indeed of the assumed form. The theory applies to the more general case of nested parametric models or nested sparse covariance models as well.

2 Maximum likelihood estimates of diagonal entries

Assume $\mathbf{e}^{(k)} \sim \mathcal{N}_n(\mathbf{0}, \mathbf{D})$ for $k = 1, \ldots, N$, where $\mathbf{D} = \operatorname{diag}(d_1, \ldots, d_n) = \operatorname{diag}(\frac{1}{\varphi_1}, \ldots, \frac{1}{\varphi_n})$. Denote by g the density of $\mathcal{N}_n(\mathbf{0}, \mathbf{D})$ distribution. The likelihood function for $\mathbf{D}$ with a given random sample $\mathbf{E} = \left[\mathbf{e}^{(1)}, \ldots, \mathbf{e}^{(N)}\right]$ has the form

$$\begin{aligned} L(\mathbf{D}|\mathbf{E}) &= \prod_{k=1}^{N} g(\mathbf{e}^{(k)}) \\ &= [(2\pi)^n |\mathbf{D}|]^{-N/2} \exp\left(-\frac{1}{2}\sum_{k=1}^{N}\left(\mathbf{e}^{(k)}\right)^\top \mathbf{D}^{-1}\mathbf{e}^{(k)}\right) \\ &= \left[(2\pi)^n \prod_{i=1}^{n}\frac{1}{\varphi_i}\right]^{-N/2} \exp\left(-\frac{1}{2}\sum_{k=1}^{N}\sum_{i=1}^{n}\left(e_i^{(k)}\right)^2 \varphi_i\right) \\ &= \left[(2\pi)^n \prod_{i=1}^{n}\frac{1}{\varphi_i}\right]^{-N/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\varphi_i S_i^2\right), \end{aligned}$$

where $S_i^2 = \sum_{k=1}^{N}\left(e_i^{(k)}\right)^2$ is a sufficient statistic for variance.

The log-likelihood function is equal to

$$\begin{aligned} \ell(\mathbf{D}|\mathbf{E}) &= \log L(\mathbf{D}|\mathbf{e}^{(1)}, \ldots, \mathbf{e}^{(N)}) \\ &= -\frac{N}{2} n \log(2\pi) + \frac{N}{2}\sum_{i=1}^{n}\log\varphi_i - \frac{1}{2}\sum_{i=1}^{n}\varphi_i S_i^2. \end{aligned} \tag{1}$$

Partial derivatives of ℓ with respect to φ_j are

$$\frac{\partial \ell}{\partial \varphi_j} = \frac{N}{2\varphi_j} - \frac{S_j^2}{2}.$$

Imposing them equal to zero we get the maximum likelihood estimates for eigenvalues

$$\frac{1}{\hat{\varphi}_j} = \hat{d}_j = \frac{1}{N} S_j^2, \quad j = 1, \ldots, n,$$

i.e. maximum likelihood estimates of variances.

3 Error of the MLE in Frobenius norm

The Frobenius norm of the error matrix is

$$\begin{aligned} \mathbb{E}||\hat{\mathbf{D}} - \mathbf{D}||_F^2 &= \sum_{i=1}^{n}\mathbb{E}(\hat{d}_i - d_i)^2 = \sum_{i=1}^{n}\left(\mathbb{E}(\hat{d}_i)^2 - 2d_i\mathbb{E}(\hat{d}_i) + d_i^2\right) \\ &= \sum_{i=1}^{n}\left(\frac{1}{N^2}\mathbb{E}(S_i^2)^2 - \frac{2}{N}d_i\mathbb{E}(S_i^2) + d_i^2\right) \\ &= \sum_{i=1}^{n}\left(\frac{2}{N}d_i^2 + d_i^2 - 2d_i^2 + d_i^2\right) = \frac{2}{N}\sum_{i=1}^{n} d_i^2 \end{aligned}$$

where we used the fact that $e_i^{(k)}/\sqrt{d_i} \sim \mathcal{N}(0,1)$, so $S_i^2 \sim d_i\chi_N^2$ and therefore

- $\mathbb{E}(S_i^2) = d_i N$
- $\mathbb{E}(S_i^2)^2 = \operatorname{Var} S_i^2 + (\mathbb{E}(S_i^2))^2 = d_i^2 2N + N^2 d_i^2$.

The error of the diagonal of sample covariance matrix has the Frobenius norm equal to $\frac{2}{N-1}\sum_{j=1}^n d_j^2$, so it is evident that the maximum likelihood estimate of the diagonal terms has smaller error (provided that the Frobenius norm is the criterion of optimality).

4 Parametric model of diagonal elements

Now we assume that the diagonan entries of the covariance are of the form

$$\varphi_j = c f_j(\alpha)$$

(e.g. $ce^{\alpha\lambda_j}$, $\frac{c}{\alpha+j^{1.5}}$ etc.) and compute the maximum likelihood estimates for this case.

Partial derivatives of φ_j with respect to the parameters (c, α) are

$$\frac{\partial \varphi_j}{\partial c} = f_j(\alpha), \qquad \frac{\partial \varphi_j}{\partial \alpha} = c\frac{\partial f_j}{\partial \alpha}.$$

Using the chain rule, we get

$$\begin{aligned}\frac{\partial \ell}{\partial c} &= \sum_{i=1}^n \frac{\partial \ell}{\partial \varphi_i}\frac{\partial \varphi_i}{\partial c} = \sum_{i=1}^n \left(\frac{N}{2\varphi_i} - \frac{S_i^2}{2}\right)\frac{\partial \varphi_i}{\partial c} \\ &= \frac{N}{2}\sum_{i=1}^n \left(\frac{1}{c f_i(\alpha)} - \frac{1}{N}S_i^2\right) f_i(\alpha).\end{aligned}$$

Setting this derivative equal to zero we get

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

Similarly,

$$\begin{aligned}\frac{\partial \ell}{\partial \alpha} &= \sum_{i=1}^n \frac{\partial \ell}{\partial \varphi_i}\frac{\partial \varphi_i}{\partial \alpha} = \sum_{i=1}^n \left(\frac{N}{2\varphi_i} - \frac{S_i^2}{2}\right)\frac{\partial \varphi_i}{\partial \alpha} \\ &= \frac{N}{2}\sum_{i=1}^n \left(\frac{1}{c f_i(\alpha)} - \frac{1}{N}S_i^2\right) c\frac{\partial f_i}{\partial \alpha} \\ &= \frac{N}{2}\sum_{i=1}^n \left(\frac{1}{f_i(\alpha)} - \frac{1}{N}S_i^2 c\right)\frac{\partial f_i}{\partial \alpha}.\end{aligned}$$

and setting the deriative to zero, we get

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

Using (2) it implies

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

which can be rearranged to

$$\sum_{i=1}^{n} S_i^2 f_i^2(\alpha) \left(\frac{1}{f_i(\alpha)} \frac{\partial f_i}{\partial \alpha} - \frac{1}{n} \sum_{j=1}^{n} \frac{1}{f_j(\alpha)} \frac{\partial f_j}{\partial \alpha} \right) = 0. \tag{3}$$

The expression (3) is an implicit formula for estimating α.

5 Comparison of variances

The asymptotic distribution of the maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ is normal with mean equal to the estimating parameter $\boldsymbol{\theta}$ and variance equal to the inverse of the Fisher information matrix [10, p. 2146, Theorem 3.3],

$$\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}\left(0, J(\boldsymbol{\theta})^{-1}\right). \tag{4}$$

Denote $\boldsymbol{\varphi} = (\varphi_1, \ldots, \varphi_n)^\top$. According to this law, the asymptotic distribution of maximum likelihood estimate $\hat{\boldsymbol{\varphi}} = (\hat{\varphi}_1, \ldots, \hat{\varphi}_n)^\top$ is

$$\sqrt{N}(\hat{\boldsymbol{\varphi}} - \boldsymbol{\varphi}) \xrightarrow{d} \mathcal{N}\left(0, J(\boldsymbol{\varphi})^{-1}\right),$$

where $J(\boldsymbol{\varphi}) = \operatorname{diag}\left\{ \frac{N}{2} \frac{1}{\varphi_j^2} \right\}$, i.e. it is the Fisher information matrix for general $\boldsymbol{\varphi}$ without assuming any particular structure.

This fact can be verified by a straightforward computation. Recall the definition of the Fisher information matrix

$$J_{kl}(\boldsymbol{\varphi}) = \mathbb{E}_{\boldsymbol{\varphi}} \left[-\frac{\partial^2 \ell(\boldsymbol{\varphi}|\mathbb{e})}{\partial \varphi_k \partial \varphi_l} \right], \qquad k, l = 1, \ldots, n.$$

Providing (1), it is evident that for $k \neq l$ it holds $J_{kl} = 0$. For $k = l$, we have

$$\begin{aligned} J_{kk}(\boldsymbol{\varphi}) &= \mathbb{E}_{\boldsymbol{\varphi}} \left[-\frac{\partial^2 \ell(\boldsymbol{\varphi}|\mathbb{e})}{\partial \varphi_k^2} \right] = \mathbb{E}_{\boldsymbol{\varphi}} \left[-\frac{\partial}{\partial \varphi_k} \left(\frac{N}{2} \frac{1}{\varphi_k} - \frac{1}{2} S_k^2 \right) \right] \\ &= \mathbb{E}_{\boldsymbol{\varphi}} \left[\frac{N}{2} \frac{1}{\varphi_k^2} \right] = \frac{N}{2} \frac{1}{\varphi_k^2}. \end{aligned}$$

If φ_j, $j = 1, \ldots, n$, are functions of the parameter $(c, \alpha)^\top$ (i.e. $\varphi_j = f_j(c, \alpha)$) then according to the behaviour of the Fisher information matrix under reparametrization [8, p. 125, eq. (6.16)], the Fisher information matrix for the parameter $(c, \alpha)^\top$ is

$$J(c, \alpha) = (\nabla \boldsymbol{\varphi}(c, \alpha))^\top J(\boldsymbol{\varphi}) \nabla \boldsymbol{\varphi}(c, \alpha). \tag{5}$$

From (4) we know that

$$\sqrt{N} \left(\begin{pmatrix} \hat{c} \\ \hat{\alpha} \end{pmatrix} - \begin{pmatrix} c \\ \alpha \end{pmatrix} \right) \xrightarrow{d} \mathcal{N}_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, J(c, \alpha)^{-1} \right). \tag{6}$$

How this distribution change if we compute a function of $(c, \alpha)^\top$? This question can be answered by the Delta method.

Theorem 1 (Multivariate delta method [8, p. 61, Theorem 8.22]) *Suppose that* $\mathbf{Y}_N = (Y_{N1}, \ldots, Y_{Nk})^\top$*, is a random vector. If* $g : \mathbb{R}^k \to \mathbb{R}^l$ *is a real-valued function which is continuously differentiable in a neighbourhood of the parameter point* a *and which has a non-singular derivative* $\nabla g(a)$ *at* $a \in \mathbb{R}^k$*, where*

$$\nabla g(\mathbf{a}) = \begin{pmatrix} \frac{\partial g_1}{\partial a_1} & \cdots & \frac{\partial g_1}{\partial a_l} \\ \vdots & \vdots & \\ \frac{\partial g_k}{\partial a_1} & \cdots & \frac{\partial g_k}{\partial a_l} \end{pmatrix}$$

and if it holds

$$\sqrt{N}(\mathbf{Y}_N - \mathbf{a}) \xrightarrow{d} \mathcal{N}_k(\mathbf{0}, \Sigma),$$

then

$$\sqrt{N}(g(\mathbf{Y}_N) - g(\mathbf{a})) \xrightarrow{d} \mathcal{N}_l\left(\mathbf{0}, \nabla g(\mathbf{a}) \Sigma \left(\nabla g(\mathbf{a})\right)^\top\right). \tag{7}$$

Assume that $\nabla\varphi(c, \alpha)$ exists. By using the Multivariate delta method (7) on relation (6) we get

$$\sqrt{N}(\boldsymbol{\varphi}(\hat{c}, \hat{\alpha}) - \boldsymbol{\varphi}(c, \alpha)) \xrightarrow{d} \mathcal{N}_n\left(\mathbf{0}, \nabla\boldsymbol{\varphi}(c, \alpha) J(c, \alpha)^{-1} (\nabla\boldsymbol{\varphi}(c, \alpha))^\top\right). \tag{8}$$

The covariance matrix in (8) can be denoted by $J(\varphi(c, \alpha))^{-1}$ in order to refer to its dependence on c and α. Moreover, according to the Zehna's invariance principle [14], $\varphi(\hat{c}, \hat{\alpha})$ is a maximum likelihood estimate of $\varphi(c, \alpha)$. Hence, its covariance matrix should be an inversion of some Fisher information matrix, which justifies the use of the letter J here.

Replacing the matrix $J(c, \alpha)$ in (8) by (5) gives us the relationship between the covariance matrix for the estimates of inverse entries of the diagonal matrix D with some general structure and the covariance matrix of the estimates in the case that these entries are assumed to be dependent on the parameter $(c, \alpha)^\top$. This relationship has the form

$$J(\boldsymbol{\varphi}(c, \alpha))^{-1} = \nabla\boldsymbol{\varphi}(c, \alpha) \left((\nabla\boldsymbol{\varphi}(c, \alpha))^\top J(\boldsymbol{\varphi}) \nabla\boldsymbol{\varphi}(c, \alpha)\right)^{-1} (\nabla\boldsymbol{\varphi}(c, \alpha))^\top.$$

It can be expected that covariance matrix for $\varphi(\hat{c}, \hat{\alpha})$ is smaller than the covariance matrix of $\hat{\boldsymbol{\varphi}}$ because it corresponds to an estimate found in a smaller (but correct) subspace. We prove that it is indeed the case. The inequality is meant in the way that the difference matrix is non-negative definite.

Theorem 2 *It holds that*

$$\nabla\boldsymbol{\varphi}(c, \alpha) \left((\nabla\boldsymbol{\varphi}(c, \alpha))^\top J(\boldsymbol{\varphi}) \nabla\boldsymbol{\varphi}(c, \alpha)\right)^{-1} (\nabla\boldsymbol{\varphi}(c, \alpha))^\top \leq J(\boldsymbol{\varphi})^{-1}. \tag{9}$$

Proof. Denote $A := J(\varphi)$ and $B := \nabla\varphi(c, \alpha)$. Then the equation has the form

$$B(B^\top A B)^{-1} B^\top \leq A^{-1}.$$

After multiplying this equation by $A^{1/2}$ from the right and from the left we get

$$A^{1/2} B (B^\top A B)^{-1} B^\top A^{1/2} \leq I.$$

Matrix $P = A^{1/2}B(B^\top AB)^{-1}B^\top A^{1/2}$ is a projection since $P^2 = P$. Moreover, it is an orthogonal projection. To see this just recall that matrix $B(B^\top AB)^{-1}B^\top$ is a covariance matrix, so it is symmetric, and matrix $A^{1/2}$ is diagonal. Therefore, matrix P is symmetric as well, which implies P to be an orthogonal projection.

It remains to show that $I - P$ is a non-negative definite matrix, i.e.

$$v^\top(I - P)v \geq 0 \quad \forall v \in \mathbb{R}^n.$$

This can be shown easily by the following computation

$$\begin{aligned} v^\top(I - P)v &= v^\top(I - P)(I - P)v = v^\top(I - P)^\top(I - P)v \\ &= \langle(I - P)v, (I - P)v\rangle = ||(I - P)v||^2 \geq 0. \end{aligned}$$

where we used the fact that $I - P$ is also a projection and that it is also symmetric. Note that $\langle\cdot,\cdot\rangle$ is the standard inner product in R^n and $||\cdot||$ is the associated (Euclidean) norm. ■

Therefore, a maximum likelihood estimate computed in the correct subspace has smaller variance than an estimate computed in the whole space. Further notice, that in the whole derivation we have not used the assumption of normal distribution, so this statements is valid for general maximum likelihood estimates. Moreover, it can be easily extended to higher-dimensional parameters.

Acknowledgements

This work was supported by the grant 13-34856 of the Czech Science Foundation (GACR) and the grant No. 1216481 of NSF (USA).

References

[1] Jonathan D. Beezley, Jan Mandel, and Loren Cobb. Wavelet ensemble Kalman filters. In *Proceedings of IEEE IDAACS'2011, Prague, September 2011*, volume 2, pages 514–518. IEEE, 2011.

[2] Loïk Berre. Estimation of synoptic and mesoscale forecast error covariances in a limited-area model. *Monthly Weather Review*, 128(3):644–667, 2000.

[3] G. J. Boer. Homogeneous and isotropic turbulence on the sphere. *Journal of the Atmospheric Sciences*, 40(1):154–163, 1983.

[4] Mark Buehner and Martin Charron. Spectral and spatial localization of background-error correlations for data assimilation. *Quarterly Journal of the Royal Meteorological Society*, 133(624):615–630, 2007.

[5] P. Courtier, E. Andersson, W. Heckley, D. Vasiljevic, M. Hamrud, A. Hollingsworth, F. Rabier, M. Fisher, and J. Pailleux. The ECMWF implementation of three-dimensional variational assimilation (3D-Var). I: Formulation. *Quarterly Journal of the Royal Meteorological Society*, 124(550):1783–1807, 1998.

[6] Reinhard Furrer and Thomas Bengtsson. Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants. *J. Multivariate Anal.*, 98(2):227–255, 2007.

[7] I. Kasanický, J. Mandel, and M. Vejmelka. Spectral diagonal ensemble Kalman filters. *Nonlinear Processes in Geophysics*, 22(4):485 – 497, 2015.

[8] E. L. Lehmann and George Casella. *Theory of point estimation*. Springer Texts in Statistics. Springer-Verlag, New York, second edition, 1998.

[9] Jan Mandel, Jonathan D. Beezley, and Volodymyr Y. Kondratenko. 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*, pages 777–784. World Scientific, 2010.

[10] Whitney K. Newey and Daniel McFadden. Large sample estimation and hypothesis testing. In *Handbook of econometrics, Vol. IV*, volume 2 of *Handbooks in Econom.*, pages 2111–2245. North-Holland, Amsterdam, 1994.

[11] Olivier Pannekoucke, Loïk Berre, and Gerald Desroziers. Filtering properties of wavelets for local background-error correlations. *Quarterly Journal of the Royal Meteorological Society*, 133(623, Part B):363–379, 2007.

[12] David F. Parrish and John C. Derber. The National Meteorological Center's spectral statistical-interpolation analysis system. *Monthly Weather Review*, 120(8):1747–1763, 1992.

[13] M. Turčičová, J. Mandel, and K. Eben. Covariance modeling by means of eigenfunctions of Laplace operator. In *JSM Proceedings*, pages 3454–3461, 2015. Joint Statistical Meetings 2015 Seattle, Section on Statistics and the Environment.

[14] Peter W. Zehna. Invariance of maximum likelihood estimators. *Ann. Math. Statist.*, 37:744, 1966.