A SIMPLE SCHEME FOR TUNING FORECAST ERROR COVARIANCE PARAMETERS

D. P. Dee
Delft Hydraulics
Emmeloord, The Netherlands

We present a simple scheme by which parameterized covariance models can be tuned to the data they describe. The method is appropriate when the number of tunable parameters is much less than the number of measurements n. The computational cost is roughly $O(n^3)$ per parameter. We work out examples that provide some insight into the reliability of the parameter estimates, and briefly discuss potential applications of the method in the context of statistical data assimilation. These include the design of adaptive filtering schemes, and validation of covariance models used for the prediction of forecast error statistics.

1. INTRODUCTION

The primary purpose of meteorological data assimilation is to provide the initial values for a numerical weather prediction model, in such a way that it will lead to optimal predictions. Modern data assimilation methods attempt to achieve this by combining information from observational systems with information about atmospheric processes which is implicit in the model equations.

Which approach should be used in solving the data assimilation problem is still a subject of some debate. The idea behind the so-called variational method is to determine a model hindcast which is closest to the available observational data in a least-squares sense, and then to use that hindcast to generate initial values for a prediction (Talagrand and Courtier, 1987). The variational method is conceptually equivalent to least-squares fitting of a parameterized function to a set of measurements. The special difficulties that arise in applying it to the meteorological data assimilation problem are strictly technical. Due to the large number of parameters (namely, the entire field of initial values) a gradient-based optimization algorithm must be used in order to solve the least-squares problem. The required gradients (sensitivities of the least-squares criterion with respect to the unknown parameters) can be obtained efficiently by using the adjoint of the prediction model.

Statistical data assimilation methods, on the other hand, are based on a probabilistic approach. Uncertainties are associated with measurements as well as with the prediction model itself. If these

uncertainties can be quantified, then it is possible to produce an estimate of the atmospheric state which is as accurate as possible in a statistical sense. This can then be used to initialize a prediction (Cohn, 1982; Ghil *et al.*, 1981). The statistical approach is conceptually no more complex than the variational approach. However, when attempting to apply it to the meteorological data assimilation problem one is faced with an obstacle of a more fundamental nature. Very little is known about statistical properties of the uncertainties, particularly those associated with the prediction model. Yet a complete specification of these properties is required in order to implement a statistical data assimilation scheme.

Dismissing the problems associated with implementing either of the two approaches for the moment, it makes little sense to argue that one of them is conceptually superior to the other. The crucial difference between the two lies in the criterion upon which they are based. The statistical approach aims at optimal accuracy; the variational approach emphasizes optimal dynamical balancing. There is a growing consensus that future data assimilation methods should combine both of these features.

Regardless of the data assimilation method being used, it will be useful to compute estimates of the accuracy of model predictions. This is why forecast error evolution in prediction models is an important subject in itself, and should not be regarded merely as a necessary step toward the development of sophisticated statistical data assimilation methods.

Considering the extent of our present understanding of model error, it is safe to say that the study of forecast error evolution is still at an early stage of development. Given a statistical description of model error, techniques are available by which the forecast error covariance evolution can be approximately computed. However, the available data on model error are insufficient to implement these techniques in an operational context, and therefore they have so far been used as theoretical devices in idealized studies only. It is a mistake to think that the high cost of computation is the primary impediment to forecast error covariance prediction in practice, since no amount of computing power can produce meaningful results if the required input data (i.e., model error statistics) are largely unavailable.

Thus there is a need for simple conceptual models that describe the gross features of forecast error evolution, and for schemes that allow these models to be tested against the available data. The fundamental problem remains the specification of model error statistics, and it must be accepted that, at best, any specification will be crude. The total number of parameters required to completely describe model error statistics is simply too large, relative to the amount of observational data at our

disposal. In practice this means that one will have to rely on idealized representations of model error statistics, typically based on strong assumptions about underlying probability distributions. Possibly such representations depend on a small number of free parameters, which one could then attempt to estimate on the basis of actual measurements.

In this paper we present a simple scheme by which parameterized covariance models can be tuned to the data they describe. The method is based on a parameter estimation technique designed for time series analysis by Burg *et al.*, (1982), and is appropriate when the number of tunable parameters is much less than the number of measurements n. The computational cost of the method is roughly $O(n^3)$ per parameter.

We describe the method in Section 2, and work out some examples in Section 3. For a simple analytical example we show that the standard deviation of a parameter estimate, obtained on the basis of n measurement data, is proportional to $n^{-\frac{1}{2}}$. We briefly discuss possible applications of the method in the context of statistical data assimilation in Section 4.

2. SINGLE-SAMPLE ESTIMATION OF COVARIANCE PARAMETERS

Let $\boldsymbol{v}$ denote a vector of jointly distributed random variables, whose covariance we wish to estimate. If the dimension of $\boldsymbol{v}$ is n, say, then $n\times(n-1)$ parameters are required to completely specify its covariance. The availability of measurements of $\boldsymbol{v}$ will determine how many of these covariance parameters can actually be estimated, and how reliably this can be done. In general it can be said that, independently of the estimation scheme being used, the number of measurements must be much larger than the number of parameters being estimated.

Particularly in real-time situations where the number of available observations of $\boldsymbol{v}$ may be $O(n)$ or less, it will be necessary to supplement the measurements with a theoretical model for the covariance of $\boldsymbol{v}$. Assumptions must be made about the probability distribution of $\boldsymbol{v}$ which result in a prescription for the covariance that depends on a few ($\ll n$) parameters only. These parameters can then be estimated on the basis of the available measurements, even if they consist of only one sample of $\boldsymbol{v}$.

2.1 Maximum likelihood parameter estimation

Suppose that the elements of $\boldsymbol{v}$ are jointly distributed with

$$E[\boldsymbol{v}] = \mathbf{o} \tag{1}$$

$$E[\boldsymbol{v}\boldsymbol{v}^T] = \boldsymbol{P}(\boldsymbol{\alpha}) \tag{2}$$

where $\boldsymbol{\alpha}$ is a vector of unknown parameters with

$$\dim \boldsymbol{\alpha} \ll \dim \boldsymbol{v} \tag{3}$$

and $E[\cdot]$ denotes the expectation operator. The parameterized $n \times n$ matrix $\boldsymbol{P}(\boldsymbol{\alpha})$ represents our covariance model for $\boldsymbol{v}$.

We propose to estimate the unknown parameters $\boldsymbol{\alpha}$ on the basis of a single vector-sample of $\boldsymbol{v}$ — that is, one measurement of each element of $\boldsymbol{v}$. This can be done by maximizing the conditional probability density function $p(\boldsymbol{v}|\boldsymbol{\alpha})$ of $\boldsymbol{v}$ given $\boldsymbol{\alpha}$, which is assumed to be Gaussian:

$$p(\boldsymbol{v}|\boldsymbol{\alpha}) = (2\pi)^{-\frac{n}{2}} (\det \boldsymbol{P})^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\boldsymbol{v}^T \boldsymbol{P}^{-1} \boldsymbol{v})\right] \tag{4}$$

Thinking of $\boldsymbol{P}(\boldsymbol{\alpha})$ as a family of candidate covariance matrices, parameterized by $\boldsymbol{\alpha}$, maximizing the conditional density function (4) is a way of picking the most plausible member $\boldsymbol{P}(\boldsymbol{\alpha}_{ml})$ on the basis of one sample of $\boldsymbol{v}$.

The maximum-likelihood parameter estimate of $\boldsymbol{\alpha}$ is given by

$$\boldsymbol{\alpha}_{ml} = \arg\max_{\boldsymbol{\alpha}} p(\boldsymbol{v}|\boldsymbol{\alpha}) \tag{5}$$

$$= \arg\min_{\boldsymbol{\alpha}} f(\boldsymbol{\alpha}) \tag{6}$$

with

$$f(\boldsymbol{\alpha}) = \log\det \boldsymbol{P}(\boldsymbol{\alpha}) + \boldsymbol{v}^T \boldsymbol{P}^{-1}(\boldsymbol{\alpha}) \boldsymbol{v} \tag{7}$$

equal to the negative of the logarithm of the Gaussian density function (4).

Notice that $f(\boldsymbol{\alpha})$ is a scalar function of N variables ($N = \dim \boldsymbol{\alpha}$) whose definition depends on the n random elements of $\boldsymbol{v}$. Once $\boldsymbol{v}$ is measured, this function can be minimized using a standard optimization technique. Depending on the choice of technique, $f(\boldsymbol{\alpha})$ will have to be evaluated repeatedly

for different values of $\boldsymbol{\alpha}$. It may be worthwhile to employ a gradient-based optimization algorithm, in which case the gradient of f with respect to $\boldsymbol{\alpha}$ will have to be evaluated as well.

Evaluating $f(\boldsymbol{\alpha})$ or its gradient will be expensive when n is large. However, in many cases the special structure of $\boldsymbol{P}(\boldsymbol{\alpha})$ and its derivatives with respect to $\boldsymbol{\alpha}$ can be exploited in order to simplify the computations. We show here how Cholesky factorizations might be used to arrive at an efficient optimization scheme.

2.2 Evaluation of f

For a given value of $\boldsymbol{\alpha}$, let

$$\boldsymbol{P}(\boldsymbol{\alpha}) = \boldsymbol{G}\boldsymbol{G}^T \tag{8}$$

$$\boldsymbol{s} = \boldsymbol{G}^{-1}\boldsymbol{v} \tag{9}$$

where $\boldsymbol{G}$ is the lower triangular Cholesky factor of $\boldsymbol{P}$ (Golub and Van Loan, 1983). Then it is easy to show that

$$f(\boldsymbol{\alpha}) = \log \prod_{i=1}^{n} g_{ii}^2 + ||\boldsymbol{s}||_2^2 \tag{10}$$

where the g_{ii} are the diagonal elements of $\boldsymbol{G}$. The total operation count for this algorithm is $\frac{1}{6}n^3 + \frac{1}{2}n^2 + O(n)$ flops.

2.3 Evaluation of the gradient of f.

In the Appendix it is shown that

$$\frac{\partial f}{\partial \alpha_i} = \text{trace}\left[\left(\boldsymbol{P}^{-1} - \boldsymbol{P}^{-1}\boldsymbol{v}\boldsymbol{v}^T\boldsymbol{P}^{-1}\right)\frac{\partial \boldsymbol{P}}{\partial \alpha_i}\right] \tag{11}$$

where $\frac{\partial}{\partial \alpha_i}\boldsymbol{P}$ denotes the matrix obtained by differentiating each element of $\boldsymbol{P}(\boldsymbol{\alpha})$ with respect to α_i, the i^{th} element of $\boldsymbol{\alpha}$. Computation of (11) costs $O(Nn^3)$ flops. Efficient algorithms will employ matrix factorizations. For example, if each $\frac{\partial}{\partial \alpha_i}\boldsymbol{P}$ is positive definite, then it admits a Cholesky factorization

$$\frac{\partial \boldsymbol{P}}{\partial \alpha_i} = \boldsymbol{G}_i\boldsymbol{G}_i^T \tag{12}$$

Then, as shown in the Appendix,

$$\frac{\partial f}{\partial \alpha_i} = ||(\boldsymbol{G}^{-1}\boldsymbol{G}_i)||_F^2 - ||(\boldsymbol{G}^{-1}\boldsymbol{G}_i)\boldsymbol{s}||_2^2 \tag{13}$$

where $||.||_F$ denotes the Frobenius norm. The operation count for computing the gradient of f in this way is $N(2n^3/3 + 2n^2 + O(n))$.

3. EXAMPLES

We will work out two simple examples of single-sample covariance estimation. The first can be solved analytically, while the second requires some computation. The main purpose of these examples is to gain insight into the reliability of the estimator.

3.1 An analytical example

Suppose

$$\boldsymbol{P}(\alpha) = \alpha \boldsymbol{P}_0 \tag{14}$$

with $\boldsymbol{P}_0$ known. We wish to estimate the coefficient α on the basis of a single vector of measurements $\boldsymbol{v}$. In this case we can solve the optimization problem analytically, since

$$f(\alpha) = \log\det(\alpha \boldsymbol{P}_0) + \boldsymbol{v}^T(\alpha \boldsymbol{P}_0)^{-1}\boldsymbol{v} \tag{15}$$

$$= n\log\alpha + \log\det \boldsymbol{P}_0 + \frac{1}{\alpha}\boldsymbol{v}^T\boldsymbol{P}_0^{-1}\boldsymbol{v} \tag{16}$$

so that

$$\frac{df}{d\alpha} = \frac{n}{\alpha} - \frac{1}{\alpha^2}\boldsymbol{v}^T\boldsymbol{P}_0^{-1}\boldsymbol{v} \tag{17}$$

The maximum likelihood estimate of α is therefore given by

$$\alpha_{ml} = \frac{1}{n}\boldsymbol{v}^T\boldsymbol{P}_0^{-1}\boldsymbol{v} \tag{18}$$

$$= \frac{1}{n}||\boldsymbol{v}||_{\boldsymbol{P}_0}^2 \tag{19}$$

Clearly this solution insures that the scaling of $\boldsymbol{P}$ is consistent with that which is observed.

It is useful to work out this example in some more detail. An important issue in connection with parameter estimation is the reliability of the estimates as a function of, for example, the number of measurements employed. Particularly when a parameter estimate is based on a single sample only, this issue needs to be addressed.

Suppose that $\boldsymbol{v}$ is indeed normally distributed, with

$$E[\boldsymbol{v}] = \mathbf{o} \tag{20}$$

$$E[\boldsymbol{v}\boldsymbol{v}^T] = \bar{\alpha}\boldsymbol{P}_0 \tag{21}$$

for some $\bar{\alpha}$. Using the Cholesky decomposition

$$\boldsymbol{P}_0 = \boldsymbol{G}_0\boldsymbol{G}_0^T \tag{22}$$

let

$$\boldsymbol{s} = \boldsymbol{G}_0^{-1}\boldsymbol{v} \tag{23}$$

Then $\boldsymbol{s}$ is normally distributed with

$$E[\boldsymbol{s}] = \mathbf{o} \tag{24}$$

$$E[\boldsymbol{s}\boldsymbol{s}^T] = \bar{\alpha}\boldsymbol{I} \tag{25}$$

Formula (18) can be rewritten as

$$\alpha_{ml} = \frac{1}{n}\boldsymbol{s}^T\boldsymbol{s} \tag{26}$$

This shows that the estimate α_{ml} is a sum of squares of independent, normally distributed random variables. Therefore α_{ml} is itself a random variable with a gamma distribution. It is easy to show that, in fact,

$$\alpha_{ml} \sim \Gamma\left(\frac{n}{2\bar{\alpha}}, \frac{n}{2}\right) \tag{27}$$

and therefore

$$E[\alpha_{ml}] = \bar{\alpha} \tag{28}$$

$$E[|\alpha_{ml} - \bar{\alpha}|] = \bar{\alpha}\sqrt{\frac{2}{n}} \tag{29}$$

Thus, the maximum-likelihood estimate is unbiased, and its standard deviation is proportional to $n^{-\frac{1}{2}}$. Based on this simple example one might conjecture about the number of measurements that are required in order to obtain reasonably reliable single-sample parameter estimates in more complicated situations.

3.2 A numerical example

Suppose now that the covariance can be described by two parameters, as follows:

$$\boldsymbol{P}(\boldsymbol{\alpha}) = (p_{ij}) = (\sigma_i \sigma_j \rho_{ij}) \tag{30}$$

where

$$\sigma_j = 1 + \alpha_1 \sin\frac{2\pi j}{n} \tag{31}$$

$$\rho_{ij} = \exp\left(-\alpha_2 \left|\frac{i-j}{n}\right|\right) \tag{32}$$

The parameter α_1 relates to the variances of the elements of $\boldsymbol{v}$, while α_2 is a correlation-length parameter. Given a vector of measurements $\boldsymbol{v}$, the two unknown parameters can be estimated by optimizing (7) with $\boldsymbol{P}(\boldsymbol{\alpha})$ given by (30–32).

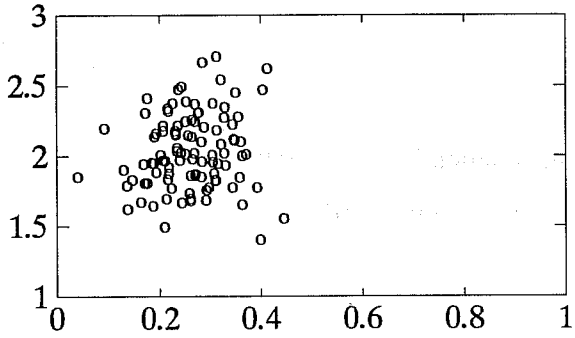

Fig. 1 Scatter diagram of parameter estimates

The resulting estimates will of course be random, since they depend on the measurements. In order to illustrate the performance of the estimator, we have used a random number generator to repeatedly generate measurement samples v, and we then computed the maximum-likelihood parameter estimates based on each single sample. In Figure 1 we show the results in a scatter diagram. Each point in the diagram corresponds to one estimate based on one sample. The coordinates of each point are the two estimated parameter values. The true parameter values, used to generate the measurements, are $\alpha_1 = 0.25$, $\alpha_2 = 2$.

In the experiments shown in Figure 1, the number of measurement points — that is, the length of the vector v — was taken to be $n = 128$. Clearly the spread of the parameter estimates around their true values depends on n. In order to investigate this dependence, the experiments should be repeated for other values of n. In this particular example, the relative standard deviations of the estimates were computed on the basis on 100 experiments, and found to be equal to 29% for α_1, 13% for α_2.

4. APPLICATION TO STATISTICAL DATA ASSIMILATION

The central problem in statistical data assimilation is to adequately represent the error covariance associated with a model forecast. The forecast error is affected by errors in the initial data that are used for the forecast (*forecast error propagation*), and by the accumulation of errors due to the discrepancy between the forecast model and the actual atmosphere (*model error forcing*).

How to compute the propagation of a known error covariance function is conceptually well-understood. The cost of such a computation, however, is very large unless substantial approximations are introduced. More importantly, data on model error forcing are lacking in practice, so that extremely costly computations of forecast error covariances are actually not justifiable except for research purposes.

Any practical statistical data assimilation scheme will therefore have to rely on a multitude of assumptions about the nature of model errors in order to compensate for the lack of data. Once such assumptions are made, it is natural to attempt to simplify the computation of forecast error covariance propagation as well (Dee, 1991). The resulting scheme will, at best, provide rough representations of the actual error statistics for a given forecast.

The estimation method outlined in the Section 2 can be used to validate such approximate representations of forecast error statistics against available measurement data. If a statistical data assimilation

scheme is based on a description of forecast error which contains unknown parameters (e.g., variance characteristics, correlation length scales), then these parameters can be tuned by matching the theoretically predicted covariance with observations.

4.1 Relation between measurements and forecast error covariance

Let $\boldsymbol{x}_k^f$ denote the model forecast at time k, then

$$\boldsymbol{P}_k^f \equiv E[(\boldsymbol{x}_k - \boldsymbol{x}_k^f)(\boldsymbol{x}_k - \boldsymbol{x}_k^f)^T] \tag{33}$$

is the forecast error covariance. Suppose that measurements $\boldsymbol{y}_k$ at time k are of the form

$$\boldsymbol{y}_k = \boldsymbol{H}_k \boldsymbol{x}_k + \boldsymbol{e}_k^o \tag{34}$$

where $\boldsymbol{H}_k$ is a known matrix and $\boldsymbol{e}_k^o$ is a vector of random measurement errors. The *forecast-minus-observed* residual is defined by

$$\begin{aligned} \boldsymbol{v}_k &= \boldsymbol{y}_k - \boldsymbol{H}_k \boldsymbol{x}_k^f & (35) \\ &= \boldsymbol{H}_k(\boldsymbol{x}_k - \boldsymbol{x}_k^f) + \boldsymbol{e}_k^o & (36) \end{aligned}$$

Its covariance is given by

$$E[\boldsymbol{v}_k \boldsymbol{v}_k^T] = \boldsymbol{H}_k \boldsymbol{P}_k^f \boldsymbol{H}_k^T + \boldsymbol{R}_k \tag{37}$$

where

$$\boldsymbol{R}_k = E[\boldsymbol{e}_k^o (\boldsymbol{e}_k^o)^T] \tag{38}$$

In deriving (37) cross-correlations among forecast errors and measurement errors are neglected, i.e., it is assumed that

$$E[\boldsymbol{H}_k(\boldsymbol{x}_k - \boldsymbol{x}_k^f)(\boldsymbol{e}_k^o)^T] = 0 \tag{39}$$

Equation (37) relates measurement information, represented by $\boldsymbol{v}_k$, to the forecast error covariance matrix $\boldsymbol{P}_k^f$. If a statistical data assimilation scheme is based on a parameterized model for $\boldsymbol{P}_k^f$, then the parameters can be tuned by using the estimation procedure outlined in the Section 2.

4.2 Application to Kalman filtering

Forecast error covariance evolution is computed by the (extended) Kalman filter according to

$$\boldsymbol{P}_k^f = \boldsymbol{A}(k,k-1)\boldsymbol{P}_{k-1}^f\boldsymbol{A}^T(k,k-1) + \boldsymbol{Q}_k \tag{40}$$

where $\boldsymbol{A}(k,k-1)$ denotes the tangent linear model operator associated with a forecast (or some approximation of it) and $\boldsymbol{Q}_k$ represents model error covariance. If $\boldsymbol{Q}_k$ contains unknown parameters, say $\boldsymbol{Q}_k = \boldsymbol{Q}(\boldsymbol{\alpha})$, then (37) becomes

$$E[\boldsymbol{v}_k\boldsymbol{v}_k^T] = \boldsymbol{H}_k\left[\boldsymbol{A}(k,k-1)\boldsymbol{P}_{k-1}^f\boldsymbol{A}^T(k,k-1) + \boldsymbol{Q}(\boldsymbol{\alpha})\right]\boldsymbol{H}_k^T + \boldsymbol{R}_k \tag{41}$$

This equation is a specific instance of (2), and so the estimation procedure outlined in Section 2 can be used to tune the parameters $\boldsymbol{\alpha}$.

Combining the Kalman filter with on-line estimation of covariance parameters makes the filter *adaptive*. We believe that the simple adaptive scheme proposed here will be more effective in an operational context than the method described in Dee (1983) and Dee *et al.* (1985). The present scheme does not rely on assumptions about error dynamics, and can therefore be expected to be more robust. Particularly when model error statistics are dependent on prevailing atmospheric conditions, as can be expected, the idea of single-sample parameter estimation appears attractive.

A numerical example of this approach to adaptive Kalman filtering was reported for a linear one-dimensional shallow water system in Dee (1990).

4.3 Some practical remarks

We emphasize that it should at all times be kept in mind that predicted forecast error covariances are subject to many approximations. This is true whether covariance evolution is computed by means of an extended Kalman filter, or by another method such as Optimal Interpolation (OI). It is therefore perfectly permissible to introduce further approximations into the parameter estimation scheme described in this paper as well.

For example, it may not be worthwhile to expend a great deal of computing power in order to obtain parameter estimates for which the maximum-likelihood function (7) is optimized *exactly*. Reducing

the value of the objective function (7) is equivalent to improving the consistency between predicted covariance and measurements, and it may be sufficient to simply evaluate the function for a few parameter values and then to choose the 'best' ones.

Similarly, only a subset of all available measurement data could be employed for the purpose of covariance tuning. Using more than, say, 10^2 data per parameter does not make much sense if the accuracy of the underlying covariance model is dubious to begin with. For example, if the assumed shape of a correlation function is erroneous, then a perfect tuning of the correlation length scale to the data will not help improve matters very much.

If more data are available than required in order to produce reasonably accurate parameter estimates, then it is possible to cross-validate the underlying covariance model. This is done by dividing the data into batches, and then producing parameter estimates based on each batch separately. If the covariance model is valid, then parameter estimates obtained from different batches of data should be consistent. Such an approach could be used, for example, to test hypotheses about model error statistics.

If used properly, single-sample estimation of covariance parameters provides a simple way of improving covariance predictions, by rendering them more responsive to prevailing atmospheric conditions. The analytical example worked out in the previous section, for example, could serve immediately to improve operational OI schemes. After all, such schemes explicitly provide a certain representation $\boldsymbol{P}_0$ of forecast error covariance at a particular instance. In the example it was shown that a relatively simple calculation will ensure that the average variance implied by this representation becomes consistent with measurements. More involved computations would allow OI parameters, such as average error growth rates and correlation scales, to be tuned continuously on the basis of real-time measurements.

5. APPENDIX

5.1 Derivation of the gradient of f

We first show that

$$\frac{d}{d\alpha}\log\det \boldsymbol{P} = \text{trace}\left[\boldsymbol{P}^{-1}\frac{d\boldsymbol{P}}{d\alpha}\right] \tag{42}$$

Expanding the determinant of $\boldsymbol{P}$ in terms of cofactors of the j^{th} column, we have

$$\det \boldsymbol{P} = \sum_i p_{ij}P_{ij} \tag{43}$$

where P_{ij} denotes the cofactor associated with the element p_{ij} of $\boldsymbol{P}$. Equation (43) shows that the partial derivative of the determinant of a matrix with respect to the ij^{th} element of that matrix is equal to the ij^{th} cofactor. Applying the chain rule for differentiation we get

$$\begin{aligned}
\frac{d}{d\alpha}\det \boldsymbol{P} &= \sum_{ij}\frac{\partial(\det \boldsymbol{P})}{\partial p_{ij}}\frac{dp_{ij}}{d\alpha} && (44)\\
&= \sum_{ij}P_{ij}\frac{dp_{ij}}{d\alpha} && (45)\\
&= \sum_i\left[\sum_j P_{ij}\frac{dp_{ij}}{d\alpha}\right] && (46)\\
&= \sum_i\left[(P_{ij})^T\frac{d\boldsymbol{P}}{d\alpha}\right]_{ii} && (47)\\
&= \text{trace}\left[(P_{ij})^T\frac{d\boldsymbol{P}}{d\alpha}\right] && (48)
\end{aligned}$$

where (P_{ij}) denotes the matrix whose elements are the cofactors of $\boldsymbol{P}$. Substituting the familiar formula

$$\boldsymbol{P}^{-1} = \frac{1}{\det \boldsymbol{P}}(P_{ij})^T \tag{49}$$

for the inverse of a matrix, we obtain

$$\frac{d}{d\alpha}\det \boldsymbol{P} = (\det \boldsymbol{P})\,\text{trace}\left[\boldsymbol{P}^{-1}\frac{d\boldsymbol{P}}{d\alpha}\right] \tag{50}$$

which proves (42). This gives us the derivative of the first term of f, see Eq. (7).

For the derivative of the second term we use

$$\frac{d\boldsymbol{P}^{-1}}{d\alpha} = -\boldsymbol{P}^{-1}\frac{d\boldsymbol{P}}{d\alpha}\boldsymbol{P}^{-1} \tag{51}$$

which follows readily by taking the derivative of $\boldsymbol{P}\boldsymbol{P}^{-1} = \boldsymbol{I}$. Then we obtain

$$\begin{aligned}
\frac{d}{d\alpha}\boldsymbol{v}^T\boldsymbol{P}^{-1}\boldsymbol{v} &= -\boldsymbol{v}^T\boldsymbol{P}^{-1}\frac{d\boldsymbol{P}}{d\alpha}\boldsymbol{P}^{-1}\boldsymbol{v} && (52)\\
&= \text{trace}\left[-\boldsymbol{v}^T\boldsymbol{P}^{-1}\frac{d\boldsymbol{P}}{d\alpha}\boldsymbol{P}^{-1}\boldsymbol{v}\right] && (53)\\
&= \text{trace}\left[-\boldsymbol{P}^{-1}\boldsymbol{v}\boldsymbol{v}^T\boldsymbol{P}^{-1}\frac{d\boldsymbol{P}}{d\alpha}\right] && (54)
\end{aligned}$$

where we repeatedly used trace$\boldsymbol{AB}$ = trace$\boldsymbol{BA}$. Combining (42) and (54) proves (11).

5.2 Computation of the gradient of f

We substitute the Cholesky decompositions of $\boldsymbol{P}$ and $\frac{\partial}{\partial\alpha_i}\boldsymbol{P}$, as well as (9) into (11):

$$\begin{aligned}
\frac{\partial f}{\partial\alpha_i} &= \text{trace}\left[\left(\boldsymbol{G}^{-T}\boldsymbol{G}^{-1} - \boldsymbol{G}^{-T}\boldsymbol{s}\boldsymbol{s}^T\boldsymbol{G}^{-1}\right)\boldsymbol{G}_i\boldsymbol{G}_i^T\right] && (55)\\
&= \text{trace}\left[(\boldsymbol{G}^{-1}\boldsymbol{G}_i)^T(\boldsymbol{G}^{-1}\boldsymbol{G}_i)\right] - \text{trace}\left[(\boldsymbol{G}^{-1}\boldsymbol{G}_i)^T\boldsymbol{s}\boldsymbol{s}^T(\boldsymbol{G}^{-1}\boldsymbol{G}_i)\right] && (56)\\
&= ||\boldsymbol{G}^{-1}\boldsymbol{G}_i||_F^2 - ||(\boldsymbol{G}^{-1}\boldsymbol{G}_i)^T s||_2^2 && (57)
\end{aligned}$$

where we used some well-known properties of the trace operator. The operation count for a Cholesky factorization is $n^3/6$ flops. Computation of $(\boldsymbol{G}^{-1}\boldsymbol{G}_i)$ costs $n^3/2$ flops. Computation of the first term of (57) therefore costs $2n^3/3 + n^2$ flops. Computing the second term of (57) costs an additional $n^2 + n$ flops.

6. REFERENCES

Burg, J. P., D. G. Luenberger, D. L. Wenger, 1982: Estimation of structured covariance matrices. Proc. IEEE, 70, 963–974.

Cohn, S. E., 1982: Methods of sequential estimation for determining initial data in numerical weather prediction, PhD thesis, Courant Institute of Mathematical Sciences report no. CI-6-82, New York University, New York.

Dee, D. P., 1983: Computational aspects of adaptive filtering and applications to numerical weather prediction, PhD thesis, Courant Institute of Mathematical Sciences report no. CI-6-83, New York University, New York.

Dee, D. P., 1990: Simplified adaptive Kalman filtering for large-scale geophysical models. In: Realization and modeling in systems theory: Proceedings of the International Symposium MTNS-89, Volume 1, eds. M. A. Kaashoek, A. C. M. Ran, and J. H. van Schuppen, Birkhäuser Boston, New York, 567–574.

Dee, D. P., 1991: Simplification of the Kalman filter for meteorological data assimilation. Q. J. R. Meteorol. Soc., 117, 365–384.

Dee, D. P., S. E. Cohn, A. Dalcher, and M. Ghil, 1985: An efficient algortithm for estimating noise covariances in distributed systems. IEEE Trans. Autom. Control, AC-30, 1057–1065.

Ghil, M., S. E. Cohn, J. Tavantzis, K. Bube, and E. Isaacson, 1981: Applications of estimation theory to numerical weather prediction. In: Dynamic meteorology: Data assimilation methods, eds. L. Bengtsson, M. Ghil, and E. Källén, Springer Verlag, New York, 139–224.

Talagrand, O., and P. Courtier, 1987: Variational assimilation of meteorological observations with the adjoint vorticity equation. I: Theory. Q. J. R. Meteorol. Soc., 113, 1311–1328.