P6.5 A BAYESIAN TECHNIQUE FOR ESTIMATING COVARIANCE PARAMETERS IN LARGE SCALE STATISTICAL OBJECTIVE ANALYSIS

David F. Parrish

Environmental Modeling Center, NCEP
Camp Springs, MD

R. James Purser

General Sciences Corporation
Beltsville, MD

1. INTRODUCTION

As operational variational analysis schemes evolve to accommodate more adaptive representations of the estimated background error covariance, including inhomogeneities and anisotropies, there is a corresponding greater need for objective statistical methods to establish the parameters of the covariances involved on a case-to-case basis. In their traditional forms, methods for maximum-likelihood and Bayesian estimation, while statistically 'efficient', are prohibitively expensive to apply directly when the measurement datasets are as large as those typical of a modern meteorological assimilation system. However, the Monte-Carlo method of randomized trace estimation, proposed in another context by Girard (1989, 1991), which sidesteps the exorbitant cost of directly estimating the trace of a large symmetric matrix, can be exploited to eliminate the computational bottle-neck of the Bayesian estimation problem. This makes it possible to extract objective real-time estimates of several covariance parameters simultaneously from the observation data. An outline of the method is given here (a more complete account is available in Purser and Parrish 2000) together with a discussion of its applicability to practical data analysis schemes.

2. LIKELIHOOD AND BAYES' THEOREM

The principles of statistical parameter estimation that we intend to use may be explained in general terms by way of the following idealized example. Let a vector of statistical parameters, $\boldsymbol{\lambda}$, be realized with a prior (probability) density, $p(\boldsymbol{\lambda})$. Given a particular realization of $\boldsymbol{\lambda}$, let the conditional density, for a vector of measurable events, $\boldsymbol{y}$, be $p(\boldsymbol{y}|\boldsymbol{\lambda})$. Then, according to elementary probability theory, the joint density for $\boldsymbol{\lambda}$ *and* $\boldsymbol{y}$ is,

$$p(\boldsymbol{y},\boldsymbol{\lambda}) = p(\boldsymbol{y}|\boldsymbol{\lambda})p(\boldsymbol{\lambda}). \tag{1}$$

Corresponding author address: David F. Parrish, W/NP2 RM 207, WWBG, 5200 Auth Road, Camp Springs, MD 20746-4304

Equally, we may express the joint density as the product of the conditional density of $\boldsymbol{\lambda}$ given $\boldsymbol{y}$ and the unconditional density $p(\boldsymbol{y})$ of $\boldsymbol{y}$:

$$p(\boldsymbol{y},\boldsymbol{\lambda}) = p(\boldsymbol{\lambda}|\boldsymbol{y})p(\boldsymbol{y}). \tag{2}$$

Combining these, we obtain the result known as Bayes' theorem:

$$p(\boldsymbol{\lambda}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{\lambda})p(\boldsymbol{\lambda})}{p(\boldsymbol{y})} \tag{3}$$

or, since $p(\boldsymbol{y}) = \int p(\boldsymbol{y}|\boldsymbol{\lambda})p(\boldsymbol{\lambda})d\lambda$, where $d\lambda$ is the volume measure in λ-space,

$$p(\boldsymbol{\lambda}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{\lambda})p(\boldsymbol{\lambda})}{\int p(\boldsymbol{y}|\boldsymbol{\lambda}')p(\boldsymbol{\lambda}')d\lambda'}. \tag{4}$$

When the objective is an inference based on the conditional probability of $\boldsymbol{\lambda}$, we refer to $p(\boldsymbol{\lambda})$ as being the 'prior' (density), $p(\boldsymbol{\lambda}|\boldsymbol{y})$ as being the 'posterior' (density). Note that the role of $p(\boldsymbol{y})$, which is *not* a function of $\boldsymbol{\lambda}$, is to normalize the posterior. The function of $\boldsymbol{\lambda}$ which modulates the prior to obtain the posterior evaluates numerically to the conditional, $p(\boldsymbol{y}|\boldsymbol{\lambda})$, but, in the context in which the measurable vector $\boldsymbol{y}$ is known and parameter $\boldsymbol{\lambda}$ is regarded as the variable, this function is known as the 'likelihood'. To summarize Bayes' rule: the posterior is, apart from a normalizing constant, the product of the prior and the likelihood.

In practice, it is almost always convenient to refer to the logarithms of these quantities, thereby converting the multiplicative relationship into an additive one. The negative log-likelihood,

$$l_y(\boldsymbol{\lambda}) \equiv -\log p(\boldsymbol{y}|\boldsymbol{\lambda}) \tag{5}$$

together with the negative log-prior and negative log-posterior allow many Bayesian inference problems to be expressed in their algebraically simplest forms.

For a more specific example of meteorological relevance, let us adopt some of notation suggested by Ide et al. (1997) and replace the generic "parameters", $\boldsymbol{\lambda}$, by the gridded values, $\boldsymbol{x}$, of an objective analysis of atmospheric fields. Let the $\boldsymbol{y}^o$, be noisy measurements of $\boldsymbol{Hx}$ where, for simplicity, we assume $\boldsymbol{H}$ to be a linear operator. If we now assume unbiased normal statistics, with a covariance matrix $\boldsymbol{B}$ for the n errors of the gridded background field $\boldsymbol{x}^b$, that is:

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{n}{2}}|\boldsymbol{B}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x}^b - \boldsymbol{x})^T \boldsymbol{B}^{-1}(\boldsymbol{x}^b - \boldsymbol{x})\right) \tag{6}$$

and with a covariance matrix $\boldsymbol{R}$ for the m measurement errors, the negative log-prior for this problem is,

$$-\log p(\boldsymbol{x}) = \frac{1}{2}\log[(2\pi)^n|\boldsymbol{B}|] + \frac{1}{2}(\boldsymbol{x}^b - \boldsymbol{x})^T \boldsymbol{B}^{-1}(\boldsymbol{x}^b - \boldsymbol{x}) \tag{7}$$

and, similarly, the negative log-likelihood function is,

$$l_y(\boldsymbol{x}) = \frac{1}{2}\log[(2\pi)^m|\boldsymbol{R}|] + \frac{1}{2}(\boldsymbol{y}^o - \boldsymbol{Hx})^T \boldsymbol{R}^{-1}(\boldsymbol{y}^o - \boldsymbol{Hx}). \tag{8}$$

Since the determinants, $|\boldsymbol{B}|$ and $|\boldsymbol{R}|$, are not dependent upon the values $\boldsymbol{x}$, the posterior probablility is maximized when we minimize the quadratic form:

$$\mathcal{L}(\boldsymbol{x}) = \mathcal{L}_a(\boldsymbol{x}) + \mathcal{L}_y(\boldsymbol{x}) \tag{9}$$

with

$$2\mathcal{L}_a(\boldsymbol{x}) = (\boldsymbol{x}^b - \boldsymbol{x})^T \boldsymbol{B}^{-1}(\boldsymbol{x}^b - \boldsymbol{x}) \tag{10}$$

and

$$2\mathcal{L}_y(\boldsymbol{x}) = (\boldsymbol{y}^o - \boldsymbol{Hx})^T \boldsymbol{R}^{-1}(\boldsymbol{y}^o - \boldsymbol{Hx}). \tag{11}$$

The expression above is, of course, the usual "cost function" of a variational analysis, but we have derived it here from explicitly Bayesian principles. The minimization leads to a linear problem, though typically one of a nontrivially large size, since the number of data (m) tends to be several hundreds or thousands and the number of gridded variables (n) can be considerably larger still.

The solution vector, $\boldsymbol{x}^a$, that minimizes this $\mathcal{L}(\boldsymbol{x})$ is the optimal variational analysis state expressed by either of the two equivalent forms:

$$\boldsymbol{x}^a = \boldsymbol{x}^b + \boldsymbol{BH}^T \boldsymbol{f}, \tag{12}$$

$$\boldsymbol{x}^a = \boldsymbol{x}^b + \boldsymbol{P}^a \boldsymbol{H}^T \boldsymbol{R}^{-1} \boldsymbol{d}, \tag{13}$$

with the vector, $\boldsymbol{f}$, of analysis forcing components given by the solution of the auxiliary linear problem of size m:

$$\boldsymbol{Qf} = \boldsymbol{d}, \tag{14}$$

where,

$$\boldsymbol{d} \equiv \boldsymbol{y}^o - \boldsymbol{y}^b \equiv \boldsymbol{y}^o - \boldsymbol{Hx}^b, \tag{15}$$

is the "innovation" vector, where

$$\boldsymbol{Q} = \boldsymbol{HBH}^T + \boldsymbol{R} = \langle \boldsymbol{dd}^T \rangle, \tag{16}$$

is the autocovariance of the innovation vector, and

$$\boldsymbol{P}^a = (\boldsymbol{B}^{-1} + \boldsymbol{H}^T \boldsymbol{R}^{-1} \boldsymbol{H})^{-1} \equiv \boldsymbol{B} - \boldsymbol{BH}^T \boldsymbol{Q}^{-1} \boldsymbol{HB} \tag{17}$$

is the covariance of error in the resulting analysis, $\boldsymbol{x}^a$.

Now consider another example where the objective is to estimate a vector of κ parameters, $\boldsymbol{\lambda}$, which define certain qualities of the covariance $\boldsymbol{B}$ itself. Let us suppose that the parameterization by $\boldsymbol{\lambda}$ is constructed such that the prior estimate is $\boldsymbol{\lambda} = \boldsymbol{0}$ and the prior autocovariance of $\boldsymbol{\lambda}$ is simply the identity, $\langle \boldsymbol{\lambda\lambda}^T \rangle = \boldsymbol{I}$. Adopting the normal model for the distribution of $\boldsymbol{\lambda}$, the Bayesian solution, obtained as the maximization of the posterior probability density of $\boldsymbol{\lambda}$, leads to the problem of minimizing the functional,

$$\mathcal{L}(\boldsymbol{\lambda}) = \frac{1}{2}\boldsymbol{\lambda}^T \boldsymbol{\lambda} + l_1(\boldsymbol{\lambda}) + l_2(\boldsymbol{\lambda}), \tag{18}$$

where l_1 and l_2 are the two terms of the negative log-likelihood in this case:

$$l_1(\boldsymbol{\lambda}) = \frac{1}{2}\log|\boldsymbol{Q}|, \tag{19}$$

$$l_2(\boldsymbol{\lambda}) = \frac{1}{2}\boldsymbol{d}^T \boldsymbol{Q}^{-1} \boldsymbol{d}. \tag{20}$$

If the quadratic 'prior' term, $\boldsymbol{\lambda}^T\boldsymbol{\lambda}$, is omitted from (16) the solution, if it exists, is the maximum-likelihood solution which is discussed in Dee (1995) and in Dee and da Silva (1999). The main difficulty encountered in putting the estimation procedure into practice is the evaluation of the log-determinant term, $\log|\boldsymbol{Q}|$, or at least its derivatives. The next section will focus on this problem and a possible solution in terms of stochastic estimation methods.

3. STOCHASTIC TRACE ESTIMATION FOR LIKELIHOOD CALCULATIONS

If we apply an orthogonal transformation of the vector basis to diagonalize the matrix $\boldsymbol{Q}$ we can extend the logarithmic function in the obvious way to matrix arguments and show that the log-determinant has an equivalent expression in terms of the matrix trace:

$$\log|\boldsymbol{Q}| = \mathrm{trace}(\log \boldsymbol{Q}). \tag{21}$$

Unfortunately, even with an efficient procedure to evaluate the trace, the above substitution would not be directly helpful in practice because the diagonalization involved in constructing $\log \boldsymbol{Q}$ would itself be at least as expensive as the direct evaluation of the determinant. In order to make practical use of the likelihood function we must make some approximations based on reasonable assumptions. We start by explicitly recognizing the fact that it is only the *relative* differences in the log-likelihood function that are ever meaningful. Then the fundamental assumption we make is that the matrix quantities $\boldsymbol{Q}$, whose associated likelihood functions we compare, are never too dissimilar numerically within the plausible range of the statistical parameters that we explore. For a *homomorphic* set of covariances, by which we mean a set taken from the same continuously parameterized family, the optimal parameters can be taken as those locally maximizing the log-posterior density, which involves evaluating the *derivatives* of the log-likelihood with respect to these parameters. In this context,the identity,

$$d \log |\boldsymbol{Q}| = \text{trace}\left(\boldsymbol{Q}^{-1}\, d\boldsymbol{Q}\right) \tag{22}$$

looks more promising for practical manipulation than (21). We shall assume that the construction of the covariance $\boldsymbol{B}$ is such that there exists an explicitly known operator $\boldsymbol{C}$, (which, in matrix terms, must be rectangular) into which $\boldsymbol{Q}$ formally factorizes:

$$\boldsymbol{Q} = \boldsymbol{C}\boldsymbol{C}^T. \tag{23}$$

We use a zero subscript to signify a $\boldsymbol{\lambda}$-dependent operator or vector at the initial default value, $\boldsymbol{\lambda} = \boldsymbol{0}$. For example,

$$\boldsymbol{f}_0 = \boldsymbol{Q}_0^{-1}\boldsymbol{d}. \tag{24}$$

Now we can apply the method of Girard (1989, 1991) to the estimation of the trace in (22) by realizing a Gaussian random white-noise vector $\boldsymbol{\epsilon}$ and using it to form:

$$\boldsymbol{q} = \boldsymbol{C}_0\boldsymbol{\epsilon}, \tag{25}$$

$$\boldsymbol{r} = \boldsymbol{Q}_0^{-1}\boldsymbol{q}. \tag{26}$$

Then each $\boldsymbol{q}$ approximately shares the statistical properties of the innovation vector $\boldsymbol{d}$ in the sense that,

$$\langle \boldsymbol{q} \rangle = \boldsymbol{0}, \tag{27}$$

$$\langle \boldsymbol{q}\boldsymbol{q}^T \rangle = \boldsymbol{Q}_0, \tag{28}$$

while each $\boldsymbol{r}$ obeys the statistics:

$$\langle \boldsymbol{r} \rangle = \boldsymbol{0}, \tag{29}$$

$$\langle \boldsymbol{r}\boldsymbol{r}^T \rangle = \boldsymbol{Q}_0^{-1}. \tag{30}$$

An infinitesimal change of l_1 at $\boldsymbol{\lambda} = \boldsymbol{0}$ satisfies,

$$dl_1 = \frac{1}{2}\text{trace}\left(\boldsymbol{Q}_0^{-1}\, d\boldsymbol{Q}\right) \equiv \frac{1}{2}\langle \boldsymbol{r}^T\, d\boldsymbol{Q}\boldsymbol{r} \rangle. \tag{31}$$

The principle of stochastic trace estimation allows us to assume that the result obtained by replacing the expectation operator in (31) by the sample average, denoted by an overbar, provides a consistent and reasonably accurate estimate for dl_1:

$$\frac{1}{2}\langle \boldsymbol{r}^T\, d\boldsymbol{Q}\boldsymbol{r} \rangle \approx \frac{1}{2}\overline{\boldsymbol{r}^T\, d\boldsymbol{Q}\boldsymbol{r}}. \tag{32}$$

Thus, combining this result with the exact derivative of l_2 leads to estimates for all the gradient components of the complete negative log-likelihood function at $\boldsymbol{\lambda} = \boldsymbol{0}$:

$$l_\alpha = \left.\frac{\partial l(\boldsymbol{\lambda})}{\partial \lambda_\alpha}\right|_{\lambda=0} \approx \frac{1}{2}\left(\overline{\boldsymbol{r}^T\boldsymbol{Q}_\alpha\boldsymbol{r}} - \boldsymbol{f}_0^T\boldsymbol{Q}_\alpha\boldsymbol{f}_0\right), \tag{33}$$

where

$$\boldsymbol{Q}_\alpha = \frac{\partial \boldsymbol{Q}}{\partial \lambda_\alpha}. \tag{34}$$

The cost of applying either $\boldsymbol{C}$ or $\boldsymbol{C}^T$ to vectors, and hence the cost of applying $\boldsymbol{Q}$, is relatively insignificant compared to the cost of performing a linear inversion such as is implied by (26). Thus, the gradient of the log-likelihood is estimated for the equivalent cost of an inversion of (24), which we need to do anyway in order to obtain the optimal analysis, plus the cost of an auxiliary linear inversion of (26) for each one of the p independent random realizations of $\boldsymbol{\epsilon}$ (and hence of $\boldsymbol{q}$ and of $\boldsymbol{r}$). But typically, it suffices to use only a single realization (i.e., $p = 1$) for large problems, as was noted by Girard (1989).

For a direct application of the Newton Raphson method to the problem of finding a zero in the estimated gradient of either the negative log-posterior (Bayesian case) or just the negative log-likelihood (maximum-likelihood case), one needs to estimate the Hessian of the functional minimized. Purser and Parrish (2000) discuss various ways of extending the stochastic estimation procedure so that, for a relatively small dimensionality of the parameter vector, $\boldsymbol{\lambda}$, all the Hessian components may be stochastically estimated and appropriate modifications made to the Newton Raphson formula to make due allowance for the "fuzziness" introduced into the problem as a consequence of adopting a stochastic estimation procedure. Alternatively, a steepest descent algorithm which employs crude finite differences in $\boldsymbol{\lambda}$-space to estimate each step size, avoids the extra computational burden of complete Hessian estimation.

4. DISCUSSION

We have provided an outline of a procedure to identify the statistical parameters of a variational assimilation scheme. The technical discussion has focused on 3D-VAR but the technique is equally applicable within the framework of 4D-VAR where, as discussed by Fisher and Courtier (1995), there is still a need to estimate the forecast error covariances. The incorporation of a Bayesian prior into the estimation has the effect of stabilizing the procedure (Verter and Dee 2000, and personal communication) but risks imposing an undue imprint of the prior guess on the set of parameters obtained. Even when data are numerous, there is one well-known potential danger of the application of the pure maximum-likelihood method (without a prior), and even the Bayesian method to some extent. This is the lack of statistical 'robustness' of the estimates that can occur when there is an insufficient resemblance between the parameterized statistical model assumed for the innovation vector and the actual statistical behavior of it. This is a recurring problem in meteorological data analysis, not just because of inadequacies in the modeling of the spatial structure of the background errors themselves, but because of the complexity inherent in the structure of biases and correlations in the observational data that are very difficult to account for in a satisfactory way. The evidence from attempts to apply the methods described here to actual data strongly suggest the need to address the problems of bias and correlation in the data *before* any substantial progress can be made towards the refinement of background error covariances. Of course, there is no reason in principle to restrict the estimation methods to parameters that relate only to the background statistics and, in fact, the inclusion of statistical parameters that relate to the observational error distributions is just as valid.

For cases where the number of parameters to be estimated is very large, the Bayesian prior becomes a practical necessity. Conversely, having a Bayesian prior in place means that a very large number of parameters can be accommodated. One context where this could prove very valuable is in the estimation and tuning of parameters for the diagnostic relationships by which features of a background might control the amplitude and shape of anisotropic and spatially adaptive covariances, which are expected to come into operational use at NCEP in the near future.

5. ACKNOWLEDGMENTS

We are grateful to Drs. Fran Verter and Dick Dee for helpful discussions. This work was partially supported by the NSF/NOAA Joint Grants Program of the US Weather Research Program. This research is also in response to requirements and funding by the Federal Aviation Administration (FAA). The views expressed are those of the authors and do not necessarily represent the official policy or position of the FAA.

REFERENCES

Dee, D., 1995: On-line estimation of error covariance parameters for atmospheric data assimilation. *Mon. Wea. Rev.*, **123,** 1128–1145.

——, and A. M. da Silva, 1999: Maximum-likelihood estimation of forecast and observation error covariance parameters. Part I: Methodology. *Mon. Wea. Rev.*, **127**, 1822–1834.

Fisher, M., and P. Courtier, 1995: Estimating the covariance matrices of analysis and forecast error in variational data assimilation. *ECMWF Technical Memorandum* 220, 27pp.

Girard, D., 1989: A fast 'Monte-Carlo cross-validation' procedure for large least squares problems with noisy data. *Numer. Math.*, **56,** 1–23.

——, 1991: Asymptotic optimality of the fast randomized versions of GCV and C_L in ridge regression and regularization. *Ann. Stat.*, **19,** 1950–1963.

Ide, K., P. Courtier, M. Ghil, and A. C. Lorenc, 1997: Unified notation for data assimilation: operational, sequential and variational. *J. Meteor. Soc. Japan*, **75,** 181–189.

Purser, R. J., and D. F. Parrish, 2000: A Bayesian technique for estimating continuously varying statistical parameters of a variational assimilation. NOAA/NCEP Office Note 429. 28 pp.

Verter, F., and D. P. Dee, 2000: Bayesian error estimates. (M/s available at http://dao.gsfc.nasa.gov/monitoring/errest-IT/Bayes/).