U.S. Department of Commerce
National Oceanic and Atmospheric Administration
National Weather Service
National Centers for Environmental Prediction
5200 Auth Road
Camp Springs, MD 20746-4304

**Office Note 429**

# A BAYESIAN TECHNIQUE FOR ESTIMATING CONTINUOUSLY VARYING STATISTICAL PARAMETERS OF A VARIATIONAL ASSIMILATION

R. James Purser*

David F. Parrish†
Environmental Modeling Center

May 2000



* General Sciences Corporation, Beltsville, Maryland; e-mail: jpurser@ncep.noaa.gov
† National Centers for Environmental Prediction, Camp Springs, Maryland; e-mail: dparrish@ncep.noaa.gov

**Abstract**

This note addresses the challenging problem of inferring, from observed meteorological data, a set of continuous parameters defining the error covariances used to analyze these data in a variational assimilation scheme. The method we propose is a Bayesian extension of the "maximum-likelihood" technique, which means that prior information about the parameters is brought into play. The method uses a stochastic approximation in the computation of some of the required terms, which are difficult and costly to evaluate by other, more standard methods. One important advantage of the proposed Bayesian approach is that it makes it possible to estimate objectively a spatially dependent but smoothly varying set of parameters in a consistent manner, provided the scale over which the variations occur are sufficiently large. This ability is illustrated in the idealized tests presented here.

## 1. Introduction

In three-dimensional variational assimilation (3D-VAR) of meteorological data, it has been customary to represent the covariances of background error as spatially homogeneous and isotropic functions. These assumptions have been dictated by practical computational restrictions rather than by any belief that the actual covariances of error behave in such a simple way. Increasingly, it is now being realized that a better model of background error allows the statistical parameters defining the background covariances to be spatially adaptive in response to a combination of recent data density and quality, to the ambient climatic type (e.g., cyclonic or anticyclonic), to the local deformation of the synoptic-scale flow, or to the local predictability as indicated by, for example, the degree of dispersion of an ensemble of short-range forecasts. It is a challenging goal for present and future research to identify the extents to which these various factors contribute to the specification of an effective model of the covariances that incorporates spatial inhomogeneity and, ultimately, variations in aspect ratio and even in the shape of the covariance profiles. Recent experiments in applying adaptive anisotropic covariances in variational data assimilation, using a variety of local diagnostics of the background field, have been reported by Desroziers (1997), Riishøjgaard (1998), and by Swinbank et al. (2000).

In the case where the covariances may be defined globally by a small number (two or three per analysis variable) of statistical parameters, such as the variance (the diagonal entries in the covariance matrix), the characteristic spatial scale of "coherence", or a parameter describing some feature of the covariance's shape, then there are several ways to estimate the values of these parameters. Some methods require relatively little statistical complication; others rather more. Some recently proposed methods rely on simplified or approximated forms of Monte Carlo Kalman filters (Evensen 1994; Fisher 1998; Mitchell and Houtekamer 2000) to obtain the adaptive covariances directly, but such methods are clearly expensive and involve a complicated supporting structure. Among the relatively less complicated approaches, Thiébaux (1976), Thiébaux et al. (1986), use a parameterized curve fitting technique applied to the sample correlations plotted against separation. This is probably among the most robust methods for estimating global parameters. Hollingsworth and Lönnberg (1986) use the sample covariances of

the observation-minus-background pairs directly to adjust the fit of parameterized covariances. Both of the aforementioned simple methods work well when the available data are numerous. In that case, the statistical efficiency, which may not be particularly high for these methods, does not even become an issue. The concept of statistical efficiency relates the extent to which the actual expected improvement to the precision of the estimated parameter compares against the expected improvement one would achieve from one of the best possible statistical estimation procedures. On the other hand, it is unlikely that the methods described above would be the most effective in cases where there are only a small number of relevant data to infer each set of statistical parameters. In such cases, more efficient statistical methods begin to become more attractive options. The method of "Generalized Cross-Validation" (GCV) has been proposed as a general framework for the estimation of a few smoothing parameters from independent data (Craven and Wahba 1979) and has been applied to meteorological analysis problems by Wahba and her colleagues (e.g., Wahba 1990, Wahba and Wendelberger 1980). Conceptually, the method is equivalent to minimizing, with respect to the statistical parameters, the accumulation of certain weighted squared-residuals (the GCV function). Each component residual is obtained as the difference between the particular validating observation and the objective statistical estimate of that value based on all the *other* observations. The summation of the squared residual over all possible choices of validating data taken singly is essentially the *Ordinary* Cross-Validation function; in GCV, the weighting is chosen to guarantee algebraic elegance, making the evaluation of the GCV function the usual residual sum of squares divided by the square of the trace of a matrix intrinsic to the estimation problem. The method has been demonstrated to be robust and to be reasonably (though not perfectly) efficient in the usual statistical sense. However, the need to compute the trace of a matrix of an order equal to the size of the estimation problem (essentially the number of observations going into the assimilation) was, for a long time, a serious impediment to the application of this method to real meteorological data assimilation problems. In recent years, this situation has changed as a result of work by Girard (1989, 1991). He demonstrates that the *exact* calculation of the matrix trace can be replaced in practice, with acceptably small degradation of statistical efficiency and robustness, by a *stochastic* or "Monte Carlo" estimate. For large observational datasets, this estimate is considerably cheaper to apply. The statistical parameters of three- and four-dimensional assimilation have been estimated using this technique, by Wahba et al. (1995); and a variety of statistical and physical parameters in an idealized numerical prediction model were estimated simultaneously with this method in the study of Gong et al. (1998).

Another statistically efficient way of estimating parameters from the available data is the classical method of "maximum likelihood" estimation. Here, the probability model of the distribution of errors must be made explicit. In practice, it is assumed that the distribution of difference between observation values and collocated background values, which is equivalent to the distribution of the differences of the corresponding *errors*, is normal or "Gaussian". (In principle, the likelihood method allows much greater generality than this, but the main problem then shifts to one of handling the descriptive complexity of the covariance model.) Dee (1995), Dee and da Silva (1999), have applied the maximum-likelihood principle to the optimal estimation of statistical parameters in the meteorological context. For the assumption of normal statistics, the log-likelihood function is comprised of two components: a quadratic form in the innovation vector, which is relatively straight-forward to compute; and the log-

determinant of the system matrix (whose inverse is the kernel of the aforementioned quadratic form), which is extremely costly to compute for large problems. Thus, like the direct calculation of the GCV function, the practical calculation of the log-likelihood function seems at first sight to be inherently restricted, in a practical sense, to datasets of only a moderate size (a few hundreds, but not thousands). This is an unfortunate restriction because the likelihood estimation method is applicable to cases where, owing to correlations in the observational errors, the direct application of GCV is no longer valid. Moreover, the likelihood estimation method is, virtually by definition, asymptotically efficient at estimating the unknown parameters as the number of data, or independent samples on which the inference is based, increases.

Evidently, the stochastic trace estimation procedure has liberated the GCV method from the realm of merely "moderate" sized problems. It would be desirable if, similarly, a stochastic estimation procedure could be applied to sidestep the costly evaluation of the log-determinant or its gradients in the case of likelihood estimation. It is the purpose of this paper to investigate just such a possibility and to lay out a proposal for a Bayesian, or "penalized" extension of the maximum-likelihood method. This will permit the robust estimation of spatially varying statistical parameters at an affordable computational cost. Section 2 reviews the statistical concepts of likelihood and Bayesian estimation. Sections 3 and 4 show how Girard's methods of stochastic trace estimation can be adapted to the problem of estimating the terms in derivatives of the log-likelihood function, with Section 4 dealing with the case in which comparisons are made among completely different constructions of the statistical covariance models. In Section 5 we introduce techniques for handling the cases in which statistical parameters smoothly vary. Section 6 describes some simple idealized experiments used to illustrate our proposed methods and Section 7 discusses some of the implications of this work in the context of operational data assimilation.

## 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, $\mathbf{y}$, be $p(\mathbf{y}|\boldsymbol{\lambda})$. Then, according to elementary probability theory, the joint density for $\boldsymbol{\lambda}$ *and* $\mathbf{y}$ is,

$$p(\mathbf{y}, \boldsymbol{\lambda}) = p(\mathbf{y}|\boldsymbol{\lambda})p(\boldsymbol{\lambda}). \tag{2.1}$$

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

$$p(\mathbf{y}, \boldsymbol{\lambda}) = p(\boldsymbol{\lambda}|\mathbf{y})p(\mathbf{y}). \tag{2.2}$$

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

$$p(\boldsymbol{\lambda}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\lambda})p(\boldsymbol{\lambda})}{p(\mathbf{y})} \tag{2.3}$$

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

$$p(\boldsymbol{\lambda}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\lambda})p(\boldsymbol{\lambda})}{\int p(\mathbf{y}|\boldsymbol{\lambda}')p(\boldsymbol{\lambda}')d\lambda'}. \tag{2.4}$$

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), $p(\boldsymbol{\lambda}|\mathbf{y})$ as being the 'posterior' (density). Note that the role of $p(\mathbf{y})$, 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, $p(\mathbf{y}|\boldsymbol{\lambda})$, but, in the context in which the measurable vector $\mathbf{y}$ 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,

$$l_y(\boldsymbol{\lambda}) \equiv -\log p(\mathbf{y}|\boldsymbol{\lambda}) \tag{2.5}$$

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, $\mathbf{x}$, of an objective analysis of atmospheric fields. Let the $\mathbf{y}^o$, be noisy measurements of $\mathbf{Hx}$ where, for simplicity, we assume $\mathbf{H}$ to be a linear operator. If we now assume unbiased normal statistics, with a covariance matrix $\mathbf{B}$ for the $n$ errors of the gridded background field $\mathbf{x}^b$, that is:

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{n}{2}}|\mathbf{B}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}^b - \mathbf{x})^T\mathbf{B}^{-1}(\mathbf{x}^b - \mathbf{x})\right) \tag{2.6}$$

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

$$-\log p(\mathbf{x}) = \frac{1}{2}\log[(2\pi)^n|\mathbf{B}|] + \frac{1}{2}(\mathbf{x}^b - \mathbf{x})^T\mathbf{B}^{-1}(\mathbf{x}^b - \mathbf{x}) \tag{2.7}$$

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

$$l_y(\mathbf{x}) = \frac{1}{2}\log[(2\pi)^m|\mathbf{R}|] + \frac{1}{2}(\mathbf{y}^o - \mathbf{Hx})^T\mathbf{R}^{-1}(\mathbf{y}^o - \mathbf{Hx}). \tag{2.8}$$

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

$$\mathcal{L}(\mathbf{x}) = \frac{1}{2}(\mathbf{x}^b - \mathbf{x})^T\mathbf{B}^{-1}(\mathbf{x}^b - \mathbf{x}) + \frac{1}{2}(\mathbf{y}^o - \mathbf{Hx})^T\mathbf{R}^{-1}(\mathbf{y}^o - \mathbf{Hx}). \tag{2.9}$$

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

$$\mathbf{x}^a = \mathbf{x}^b + \mathbf{B}\mathbf{H}^T\mathbf{f}, \tag{2.10}$$

$$\mathbf{x}^a = \mathbf{x}^b + \mathbf{P}^a\mathbf{H}^T\mathbf{R}^{-1}\mathbf{d}, \tag{2.11}$$

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

$$\mathbf{Q}\mathbf{f} = \mathbf{d}, \tag{2.12}$$

where,

$$\mathbf{d} \equiv \mathbf{y}^o - \mathbf{y}^b \equiv \mathbf{y}^o - \mathbf{H}\mathbf{x}^b, \tag{2.13}$$

is the "innovation" vector, where

$$\mathbf{Q} = \mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R} = \langle \mathbf{d}\mathbf{d}^T \rangle, \tag{2.14}$$

is the autocovariance of the innovation vector, and

$$\mathbf{P}^a = (\mathbf{B}^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})^{-1} \equiv \mathbf{B} - \mathbf{B}\mathbf{H}^T\mathbf{Q}^{-1}\mathbf{H}\mathbf{B} \tag{2.15}$$

is the covariance of error in the resulting analysis, $\mathbf{x}^a$. Eq. (2.12) can be solved with moderate effort, even for large $m$, by the method of preconditioned conjugate gradients or by an appropriate version of a quasi-Newton method. Several such methods are described in detail in Gill et al. (1981). This step, combined with the solution of (2.10), is referred to as the "measurement space" form of the variational analysis. Alternatively, the "grid space" form of the analysis equation, (2.11), may be solved by similar methods, and with an approximately equivalent degree of difficulty (e.g., Courtier 1997).

Now consider another example where the objective is to estimate a vector of $\kappa$ parameters, $\boldsymbol{\lambda}$, which define certain qualities of the covariance $\mathbf{B}$ itself. Such parameters might be immediately identifiable qualities of the covariance, such as the amplitude (variance) and characteristic spatial scale, or they may consist of more subtle parameters that might, for example, control the degree to which gradients in the background field modulate the anisotropic stretching and transverse shrinking of the local covariance function. We do not need to consider specific details of such a parameterization here (a subject that merits substantial research by itself) since our focus at this time is on general principles of the estimation problem. Although in the previous example, we were able to ignore the contributions of the determinants to the estimation problem, this time, we cannot ignore them (but we can still drop the powers of $2\pi$). We shall leave implicit the fact that $\mathbf{B}$, and hence $\mathbf{Q}$, are actually functions of $\boldsymbol{\lambda}$. Let us suppose that the parameterization by $\boldsymbol{\lambda}$ is constructed such that the prior estimate is $\boldsymbol{\lambda} = \mathbf{0}$ and the prior autocovariance of $\boldsymbol{\lambda}$ is simply the identity, $\langle \boldsymbol{\lambda}\boldsymbol{\lambda}^T \rangle = \mathbf{I}$. 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,

$$\mathcal{L}(\boldsymbol{\lambda}) = \frac{1}{2}\boldsymbol{\lambda}^T\boldsymbol{\lambda} + \frac{1}{2}\log|\mathbf{Q}| + \frac{1}{2}\mathbf{d}^T\mathbf{Q}^{-1}\mathbf{d}. \tag{2.16}$$

If the quadratic 'prior' term, $\boldsymbol{\lambda}^T\boldsymbol{\lambda}$, is omitted from (2.16) the solution, if it exists, is the maximum-likelihood solution which is discussed in Dee (1995) and in Dee and da Silva (1999). For data vastly more numerous than the parameters being estimated, there may be no difficulty in principle in using the maximum-likelihood estimator. However, when the number of parameters themselves becomes large, the maximum-likelihood procedure has a tendency to become severely ill-conditioned, even when the data are plentiful. It is for this reason that, in general, it is desirable to incorporate the prior information if at all possible. The presence of the prior term, provided that the covariance matrix is sensibly proportioned (it is the identity, $\mathbf{I}$, in the present case), has a stabilizing influence on the solution and a healthy conditioning effect on the optimization problem; this is essential in practical applications. The traditional objection to the use of a prior is that it requires that the inference problem be contaminated by an unpalatable *subjective* element. In the context of meteorological assimilation, where parameter estimations might be repeated regularly, it is fair to assume that we *do* possess prior information about the parameters, based on all the previous experience with their estimation from earlier occasions. Assuming that we shall retain the prior and make the estimation procedure a strictly 'Bayesian' one, the remaining difficulty encountered in putting the estimation procedure into practice is the evaluation of the log-determinant term, $\log|\mathbf{Q}|$, or at least its derivatives. The next section will focus on this problem and a possible solution in terms of stochastic estimation methods.

## 3. Application of stochastic trace estimation for likelihood calculations

If we apply an orthogonal transformation of the vector basis to diagonalize the matrix $\mathbf{Q}$ 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:

$$\log|\mathbf{Q}| = \mathrm{trace}(\log\mathbf{Q}). \tag{3.1}$$

A brief derivation of this identity is given in Appendix A. 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 $\log\mathbf{Q}$ 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 $\mathbf{Q}$, 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,

$$d\log|\mathbf{Q}| = \mathrm{trace}\left(\mathbf{Q}^{-1}\,\mathbf{dQ}\right) \tag{3.2}$$

looks more promising for practical manipulation than (3.1). However, there will be some occasions when we *do* need to compare directly the log-likelihoods corresponding to *heteromorphic*

pairs of covariances, in which the respective parameter vectors refer to completely different prescriptions for constructing the covariances. In this context, we must still assume that the two innovation covariances involved in the comparison, say $\mathbf{Q}_0$ and $\mathbf{Q}_1$, are sufficiently similar to make the combination, $\mathbf{S}_1 = \mathbf{Q}_0^{-1}\mathbf{Q}_1$, whose log-determinants we need to approximate, 'small' perturbations from the identity, $\mathbf{I}$. This assumption justifies approximations of the logarithm function as polynomials:

$$\log \mathbf{S}_1 \approx \sum_{k=0}^{n} \gamma_k^{(n)} \mathbf{S}_1^k \tag{3.3}$$

If the approximation is based upon truncated Taylor expansions centered about the identity state, $\mathbf{S} = \mathbf{I}$, we find that the coefficients $\gamma_k^{(n)}$ are essentially equivalent to those of an $n$th-order one-sided difference operator (also used in the "stiffly-stable" multistep numerical integration schemes of Gear, 1971). We list a few of these coefficient sequences in Table 1 and show the quality of the fit of these approximating polynomials to the logarithmic function in figure 1. We shall return to the heteromorphic comparisons in Section 4 and describe ways to estimate the trace of each power of $\mathbf{S}$. But first, we develop the method required to treat the homomorphic case.

TABLE 1. COEFFICIENTS OF POLYNOMIAL APPROXIMATIONS TO THE LOGARITHM FUNCTION

| Order of accuracy, $n$ | $\beta$ | $\beta\gamma_0^{(n)}$ | $\beta\gamma_1^{(n)}$ | $\beta\gamma_2^{(n)}$ | $\beta\gamma_3^{(n)}$ | $\beta\gamma_4^{(n)}$ | $\beta\gamma_5^{(n)}$ | $\beta\gamma_6^{(n)}$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | -1 | 1 | | | | | |
| 2 | 2 | -3 | 4 | -1 | | | | |
| 3 | 6 | -11 | 18 | -9 | 2 | | | |
| 4 | 12 | -25 | 48 | -36 | 16 | -3 | | |
| 5 | 60 | -137 | 300 | -300 | 200 | -75 | 12 | |
| 6 | 60 | -147 | 360 | -450 | 400 | -225 | 72 | -10 |

As we saw in the previous Section, the efficient solution of the variational assimilation can be accomplished via the solution of a symmetric linear problem whose system matrix is $\mathbf{Q}$. It is therefore fair to assume that the machinery exists for this task. Let us make the additional assumption that the machinery by which we apply the operator $\mathbf{Q}$ to a vector can be expressed in algebraic terms through the symmetrical factorization of $\mathbf{Q}$ into mutually adjoint rectangular matrix factors:

$$\mathbf{Q} \equiv \mathbf{C}\mathbf{C}^T. \tag{3.4}$$

The background error covariance can often be synthesized via a succession of simpler filtering operations. For example, as an additive combination of weighted Gaussian smoothers, each smoother expressible as a self-adjoint pairing, $\mathbf{G}_k\mathbf{G}_k^T$, corresponding to a particular characteristic scale, and each pair modulated via right- and left-multiplication by spatially dependent weight functions, represented algebraically by diagonal operators $\mathbf{w}_k$, for $k = 1, \ldots, p$, say. The observation error covariance contribution, $\mathbf{R}$, is diagonal, or at least block-diagonal, in most implementations, so its square-root can be taken without difficulty. Thus, the operator, $\mathbf{C}$, can be given the block representation:

$$\mathbf{C} = \left[ \; \mathbf{R}^{1/2}; \quad \mathbf{H}\mathbf{w}_1\mathbf{G}_1; \quad \ldots \quad ; \mathbf{H}\mathbf{w}_p\mathbf{G}_p \; \right]. \tag{3.5}$$

It is convenient to let $\mathbf{C}_0$ denote the operator $\mathbf{C}$ when $\boldsymbol{\lambda} = \mathbf{0}$, and likewise:

$$\mathbf{Q}_0 = \mathbf{C}_0\mathbf{C}_0^T, \tag{3.6}$$
$$\mathbf{f}_0 = \mathbf{Q}_0^{-1}\mathbf{d}. \tag{3.7}$$

Suppose we form one or several independent realizations of a "white noise" vector, $\boldsymbol{\epsilon}$, each consisting of as many independent unit-variance Gaussian random numbers as there are columns of $\mathbf{C}$, and form the associated random vectors,

$$\mathbf{q} = \mathbf{C}_0\boldsymbol{\epsilon}, \tag{3.8}$$
$$\mathbf{r} = \mathbf{Q}_0^{-1}\mathbf{q}. \tag{3.9}$$

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

$$\langle \mathbf{q} \rangle = \mathbf{0}, \tag{3.10}$$
$$\langle \mathbf{q}\mathbf{q}^T \rangle = \mathbf{Q}_0. \tag{3.11}$$

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

$$\langle \mathbf{r} \rangle = \mathbf{0}, \tag{3.12}$$
$$\langle \mathbf{r}\mathbf{r}^T \rangle = \mathbf{Q}_0^{-1}. \tag{3.13}$$

When we differentiate a generic quantity, $\psi$, at $\boldsymbol{\lambda} = \mathbf{0}$, with respect to one of the parameters, $\lambda_\alpha$, we shall write the result as $\psi_\alpha$. It is also useful to name the two terms of the likelihood function for $\boldsymbol{\lambda}$:

$$l_1(\boldsymbol{\lambda}) = \frac{1}{2}\log|\mathbf{Q}|, \tag{3.14}$$
$$l_2(\boldsymbol{\lambda}) = \frac{1}{2}\mathbf{d}^T\mathbf{Q}^{-1}\mathbf{d}, \tag{3.15}$$

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

$$dl_1 = \frac{1}{2}\text{trace}\left(\mathbf{Q}_0^{-1}\,\mathbf{dQ}\right) \equiv \frac{1}{2}\langle \mathbf{r}^T\,\mathbf{dQ}\,\mathbf{r}\rangle. \tag{3.16}$$

The principle of stochastic trace estimation allows us to assume that the result obtained by replacing the expectation operator in (3.16) by the sample average, denoted by an overbar, provides a consistent and reasonably accurate estimate for $dl_1$:

$$\frac{1}{2}\langle \mathbf{r}^T\,\mathbf{dQ}\,\mathbf{r}\rangle \approx \frac{1}{2}\overline{\mathbf{r}^T\,\mathbf{dQ}\,\mathbf{r}}. \tag{3.17}$$

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 $\boldsymbol{\lambda} = \mathbf{0}$:

$$l_\alpha = \left.\frac{\partial l(\boldsymbol{\lambda})}{\partial \lambda_\alpha}\right|_{\boldsymbol{\lambda}=0} \approx \frac{1}{2}\left(\overline{\mathbf{r}^T\mathbf{Q}_\alpha\mathbf{r}} - \mathbf{f}_0^T\mathbf{Q}_\alpha\mathbf{f}_0\right). \tag{3.18}$$

The cost of applying either $\mathbf{C}$ or $\mathbf{C}^T$ to vectors, and hence the cost of applying $\mathbf{Q}$, is relatively insignificant compared to the cost of performing a linear inversion such as is implied by (3.7) or (3.9). Thus, the gradient of the log-likelihood is estimated for the equivalent cost of an inversion of (3.7), which we need to do anyway in order to obtain the optimal analysis, plus the cost of an auxiliary linear inversion of (3.9) for each one of the $p$ independent random realizations of $\boldsymbol{\epsilon}$ (and hence of $\mathbf{q}$ and of $\mathbf{r}$). But typically, it suffices to use only a single realization (i.e., $p = 1$) for large problems, as was noted by Girard (1989).

The $\boldsymbol{\lambda}$-derivatives of the products of the linear operator $\mathbf{Q}$ may be carried out by finite differencing on an appropriate stencil. Appendix B describes a way of choosing a centered quasi-isotropic stencil in $n$ dimensions suitable for the construction of the gradient components. While the gradient of the likelihood function provides a local direction along which we can change the statistical parameters and expect an improvement, unless the log-posterior is overwhelmingly dominated by the contribution from the prior, the appropriate change to the parameters required to achieve close to optimal analysis performance would seem generally to require us to estimate the *Hessian* matrix of *second derivatives* of the log-posterior, so that we may apply at least one step of a Newton-Raphson iteration towards this optimal $\boldsymbol{\lambda}$. Further differentiation of the log-likelihood incorporating the stochastic trace estimation leads to:

$$l_{\alpha\beta} = \left. \frac{\partial^2 l(\boldsymbol{\lambda})}{\partial \lambda_\alpha \partial \lambda_\beta} \right|_{\boldsymbol{\lambda}=\mathbf{0}} \approx \frac{1}{2} \left[ \overline{\mathbf{r}^T \mathbf{Q}_{\alpha\beta} \mathbf{r}} - \overline{\mathbf{p}_\alpha^T \mathbf{u}_\beta} - \mathbf{f}_0^T \mathbf{Q}_{\alpha\beta} \mathbf{f}_0 + 2 \mathbf{g}_\alpha^T \mathbf{h}_\beta \right], \tag{3.19}$$

where we define:

$$\mathbf{p}_\alpha \equiv \mathbf{Q}_\alpha \mathbf{r}, \tag{3.20}$$

$$\mathbf{g}_\alpha \equiv \mathbf{Q}_\alpha \mathbf{f}_0, \tag{3.21}$$

and where each $\mathbf{u}_\alpha$ and each $\mathbf{h}_\alpha$ are defined by solving the additional linear inversion problems,

$$\mathbf{Q}_0 \mathbf{u}_\alpha = \mathbf{p}_\alpha, \tag{3.22}$$

$$\mathbf{Q}_0 \mathbf{h}_\alpha = \mathbf{g}_\alpha. \tag{3.23}$$

This procedure for estimating the Hessian therefore requires $(p+1)\kappa$ further linear inversions for $p$ realizations of the stochastic vectors, which, even for $p = 1$, might be regarded as unacceptably extravagant. The expense is especially questionable when we bear in mind that a Newton-Raphson correction made with the aid of a Hessian, even when the latter is locally exact, provides only an imperfect approximation to the maximum-likelihood solution (or to its Bayesian counterpart when this Hessian is suitably augmented). As a first step to economizing, we note that, for those cases where the autocovariances of $\mathbf{d}$ and $\mathbf{q}$ are both already known to be closely approximated by $\mathbf{Q}_0$, approximate algebraic cancellations in (3.19) lead to simpler stochastic approximations to $l_{\alpha\beta}$, such as the alternatives:

$$l_{\alpha\beta} \approx \frac{1}{2} \mathbf{g}_\alpha^T \mathbf{h}_\beta, \tag{3.24}$$

$$\approx \frac{1}{2} \mathbf{p}_\alpha^T \mathbf{u}_\beta, \tag{3.25}$$

either of which cuts the number of additional linear inversions needed to only $\kappa$. However, if we restrict the search for an improved parameter estimate to the line in parameter space of steepest descent, which is provided by (3.18), then a further computational economy is effected by estimating only the component of the Hessian along this single direction — an additional expense of as little as one linear inversion. In those Bayesian scenarios where accumulated previous experience leads to a relatively dominant prior, or where the number, $\kappa$, of parameters estimated is quite large, then the restricted line search is the practical option.

The approximate maximum likelihood solution, $\hat{\boldsymbol{\lambda}}^{(M.L.)}$, consistent with the one-step Newton-Raphson refinement,

$$\sum_{\beta=1}^{\kappa} l_{\alpha\beta}\hat{\lambda}_{\beta}^{(M.L.)} = -l_{\alpha} \tag{3.26}$$

would ordinarily be accorded a precision weight matrix, $W_{\alpha\beta} = l_{\alpha\beta}$, but it is evident from the form of (3.18) that *both* contributing terms inject a random element into the resulting approximation. In the case of a single stochastic realization for the first term on the right (i.e., for $p$=1), then, provided the prior estimate, $\mathbf{Q}_0$, for $\mathbf{Q}$ is not much in error, it is clear from the symmetric form of (3.18) that, in this case, there must be an approximate *doubling* of all components of the covariance of this estimate of $l_{\alpha}$, compared to the corresponding non-stochastic evaluation, where only the second term would contribute any randomness to the parameter-vector estimate. In which case, the effective weight accorded this estimate presumably should be halved. But more generally, when a total of $p$ independent random realizations of $\boldsymbol{\epsilon}$, and hence of $\mathbf{q}$ and $\mathbf{r}$, are employed, the inherent randomness of the parameter-vector estimate, as measured by the covariance of error in $\hat{\boldsymbol{\lambda}}^{(M.L.)}$, must be expected to be enhanced by a factor of $(p+1)/p$ times the non-stochastic estimate. With these considerations taken into account, the appropriate *effective* weight, $\hat{\mathbf{W}}$, that we should accord to the stochastic maximum-likelihood estimator is,

$$\hat{\mathbf{W}} = \frac{p}{p+1}\mathbf{W}, \tag{3.27}$$

provided it is legitimate to disregard the sampling error contributions involved in the stochastic estimation of the matrix $l_{\alpha\beta}$ itself in (3.26). Adopting this admittedly questionable assumption, we would formulate the corresponding stochastically corrected Bayesian estimate, $\boldsymbol{\lambda}^{(B)}$, by solving,

$$(\hat{\mathbf{W}} + \mathbf{I})\boldsymbol{\lambda}^{(B)} = \hat{\mathbf{W}}\hat{\boldsymbol{\lambda}}^{(M.L.)} \equiv \frac{-p}{p+1}\frac{\partial l}{\partial \boldsymbol{\lambda}}, \tag{3.28}$$

in which the identity operator is the precision matrix of the Bayesian prior for $\boldsymbol{\lambda}$ implied by (2.16). The effective posterior covariance for $\boldsymbol{\lambda}$ consistent with this stochastically modified Bayesian estimate is,

$$\boldsymbol{\Lambda} = (\hat{\mathbf{W}} + \mathbf{I})^{-1}. \tag{3.29}$$

Note that, even in the degenerate limiting cases where $\mathbf{W}$ becomes singular and where, consequently, the maximum-likelihood estimate $\hat{\boldsymbol{\lambda}}^{(M.L.)}$ becomes undefined, Eq.(3.28) allows us to bypass the evaluation of $\hat{\boldsymbol{\lambda}}^{(M.L.)}$ itself and obtain the regularized estimate, $\boldsymbol{\lambda}^{(B)}$, instead without difficulty.

Regardless of the choice of method used to estimate the correction $\hat{\boldsymbol{\lambda}}$ to the statistical parameters vector, if the quantities $\mathbf{h}_\alpha$ used in the Hessian calculations are retained, then,

$$\hat{\mathbf{f}} = \mathbf{f}_0 - \sum_{\alpha=1}^{\kappa} \mathbf{h}_\alpha \hat{\lambda}_\alpha \tag{3.30}$$

provides a first-order correction to the analysis, accounting for the new parameter estimates and at negligible additional cost. In practice, it may be more desirable to treat this corrected analysis as a good "first guess" in a continuing process of improvement of both the parameters and the gridded analysis in situations where nonlinear elements in the data assimilation come into play.

We have implicitly assumed the adoption of the "observation space" form of the variational analysis throughout this Section but, in view of the well-known duality of solutions (Courtier 1997), we should expect that a corresponding set of manipulations will pertain to the estimation of parameters in the case of a variational analysis formulated in "model space". The model space form of the estimation problem can be treated in the manner indicated in Appendix C.

## 4. Heteromorphic statistical models

In the scenario we consider here, two or more different families of methods for constructing the statistics are assumed to have been conditionally optimized within the scopes of their respective parameterizations, resulting in two or more contending innovation covariances, $\mathbf{Q}_0$, $\mathbf{Q}_1, \ldots$. Considering just the pair, $\mathbf{Q}_0$ and $\mathbf{Q}_1$, the difference between the respective first terms, $l_1$, of the log-likelihood is the quantity,

$$\Delta l_1 = \frac{1}{2}\mathrm{trace}(\log \mathbf{S}_1), \tag{4.1}$$

$$\approx \frac{1}{2}\mathrm{trace}\left(\sum_{k=0}^{n} \gamma_k^{(n)} \mathbf{S}_1^k\right) \tag{4.2}$$

where we have invoked the approximation of (3.3) with the coefficients, $\gamma_k^{(n)}$ given in Table 1. To apply this approximation with stochastic trace estimation, we need a procedure for estimating the trace of each power of $\mathbf{S}_1$. The stochastic trace method applied to $\mathbf{S}_1$ itself gives,

$$\mathrm{trace}(\mathbf{S}_1) \approx \overline{\mathbf{r}^T \mathbf{Q}_1 \mathbf{r}} \equiv \overline{\mathbf{r}^T \mathbf{s}_1}, \tag{4.3}$$

where $\mathbf{s}_1$ is defined:

$$\mathbf{s}_1 = \mathbf{Q}_1 \mathbf{r}. \tag{4.4}$$

We may establish a connection with the stochastic approximations of Section 3 by expressing $\mathbf{Q}_1$ as the linear perturbation to $\mathbf{Q}_0$:

$$\mathbf{Q}_1 = \mathbf{Q}_0 + \sum_{\alpha} \mathbf{Q}_\alpha \, \delta\lambda_\alpha \tag{4.5}$$

which implies that,

$$\mathbf{s}_1 = \mathbf{q} + \sum_{\alpha} \mathbf{p}_\alpha \, \delta\lambda_\alpha \tag{4.6}$$

and,

$$\mathrm{trace}(\mathbf{S}_1) \approx \overline{\mathbf{r}^T\mathbf{q}} + \sum_\alpha \overline{\mathbf{r}^T\mathbf{p}_\alpha}\,\delta\lambda_\alpha. \tag{4.7}$$

To first order in $\delta\lambda$, the trace estimate of $\mathbf{S}_1$ is therefore consistent with that used in (3.18).

For higher powers of $\mathbf{S}_1$, we proceed by defining $\mathbf{t}_1$ by means of a further linear inversion:

$$\mathbf{Q}_0\mathbf{t}_1 = \mathbf{s}_1, \tag{4.8}$$

whereupon, the stochastic estimates of the next two powers of $\mathbf{S}_1$ are found:

$$\mathrm{trace}(\mathbf{S}_1^2) \approx \overline{\mathbf{t}_1^T\mathbf{s}_1}, \tag{4.9}$$

$$\mathrm{trace}(\mathbf{S}_1^3) \approx \overline{\mathbf{t}_1^T\mathbf{Q}_1\mathbf{t}_1}, \tag{4.10}$$

and from which the general pattern for the extension to arbitrary powers is by now evident. The Hessian derived from the second degree polynomial approximation (4.2) and the stochastic approximations to $\mathrm{trace}(\mathbf{S}_1)$ and $\mathrm{trace}(\mathbf{S}_1^2)$, when $\mathbf{Q}_1$ is allowed to vary continuously about $\mathbf{Q}_0$, can be verified to be exactly consistent with the previous Hessian approximation, (3.19). But the heteromorphic comparisons can be carried out to higher degrees of approximation, if desired, for *finite* departures of $\mathbf{Q}_1$ from $\mathbf{Q}_0$.

We notice one defect of the above procedure, and that is that the method in general fails to deliver the correct value, $\mathrm{trace}(\mathbf{S}_0^k) = m$, for the case in which $\mathbf{Q}_1$ is replaced by $\mathbf{Q}_0$. This inconsistency is purely a consequence of the "sampling error" inherent in the stochastic input, vector $\boldsymbol{\epsilon}$, from which are derived the $\mathbf{q}$, $\mathbf{r}$, and so on. A partial compensation of this defect is achieved by explicitly evaluating the corresponding stochastic trace estimates for the powers of "$\mathbf{S}_0$" and subtracting their contributions to the trace of each term of the power series (3.3). However, if this additional work is to be done, then a more symmetrical comparison, with all the compared $\mathbf{Q}$ being put on an equal basis, would be achieved by replacing the inverse operator factors, $\mathbf{Q}_0^{-1}$, in the trace estimates by the inverse of the *average* of both, or all, of the $\mathbf{Q}$-operators involved in such a heteromorphic comparison.

The heteromorphic comparisons are relevant mainly to global parameter estimations and, since the contributions of the log-prior cancel in each pair-wise comparison, the method essentially reduces to a "likelihood-ratio" test in each case. But, within a homomorphic family of statistical models for a variational data assimilation covering a broad geographical area, it is quite unlikely that the vector of parameters describing the covariances locally in one region will have the same optimal values as those optimized to a distant geographical region. In the next section, we return to the homomorphic case, but extend it to accommodate a gradual geographical variation in the parameter values.

## 5. Homomorphic optimization with spatially varying parameter fields

Recalling that the computational cost is dominated by the occasions when it is required to perform one of the large-scale linear inversion steps, it is evident that the computational cost of estimating the gradient of the log-likelihood function is approximately double that of computing the analysis itself. Likewise, in order to compute the matrix of Hessian components, the cost essentially becomes multiplied by at least the number of parameters to be estimated. Thus, as

formally described, our method will not be viable when the parameters are themselves permitted the freedom to vary geographically, which is unfortunately what we would like them to do in an adaptive assimilation scheme. We shall suppose a moderate number of such parameters are required at each geographical location, but allow that their values may change smoothly in space over a large horizontal scale (larger than the typical coherence scale of the background error covariances $\mathbf{B}$ that these parameters are intended to prescribe, for example). Note that, by "parameters", we do not have to mean the obvious amplitude or scale features of a covariance model, but perhaps the more subtle parameters alluded to in Section 2, coupling diagnostics of the background field to local distortions of the covariance. The smooth, and purely large-scale spatial variation of such parameters therefore does not preclude a small-scale modulation of the actual covariance functions in response to abrupt changes of, say, the thermal gradient diagnosed in the background field. By analogy with the more familiar theory of data analysis, we should expect that the formal treatment of such a situation could be obtained by substituting for the prior term in (2.16) a quadratic form whose kernel represents the inverse of a spatially smooth covariance for the distributed $\boldsymbol{\lambda}$. It will provide some insight into the problem to at least lay out the requisite formalism and to point out the connection between the more rigorous development and our proposed approximation to it.

We shall assume the negative log-likelihood now becomes a *functional* of the parameter distribution $\boldsymbol{\lambda}(z)$ where $z$ denotes the spatial location (the number of spatial dimensions is not important here). We generalize the previous formalism by assuming that, in the vicinity of the parameter state $\boldsymbol{\lambda} = 0$, we may expand the negative log-likelihood as a functional Taylor series:

$$l(\boldsymbol{\lambda}) = l_0 + \int \sum_{\alpha} l_\alpha(z)\lambda_\alpha(z)\, dz + \frac{1}{2}\int\int \sum_{\alpha,\beta} l_{\alpha\beta}(z,z')\lambda_\alpha(z)\lambda_\beta(z')\, dz\, dz' + \ldots \tag{5.1}$$

or, in abbreviated notation:

$$l(\boldsymbol{\lambda}) = l_0 + (\mathbf{l'}_0, \boldsymbol{\lambda}) + \frac{1}{2}(\boldsymbol{\lambda}, \mathbf{l''}_0 \circ \boldsymbol{\lambda}) + \ldots \tag{5.2}$$

where $(,)$ denotes the inner product used in the second term of (5.1) and the functional operator, $\circ$ , generalizes a spatial convolution and is defined:

$$(\mathbf{l''}_0 \circ \boldsymbol{\lambda})_\alpha(z) \equiv \int \sum_{\beta} l''_{\alpha\beta}(z,z')\lambda_\beta(z')\, dz'.$$

The functional derivative of $l$ satisfies:

$$\mathbf{l'} \equiv \frac{\partial l}{\partial \lambda_\alpha}(z) = l_\alpha(z) + \int \sum_{\beta} l_{\alpha\beta}(z,z')\lambda_\beta(z')\, dz' + \ldots \tag{5.3}$$

or,

$$\mathbf{l'}(\boldsymbol{\lambda}) = \mathbf{l'}_0 + \mathbf{l''}_0 \circ \boldsymbol{\lambda}. \tag{5.4}$$

The negative log-prior for normal statistics is in the form:

$$\begin{aligned} -\log p(\boldsymbol{\lambda}) &\equiv \frac{1}{2}\int\int \sum_{\alpha\beta} \lambda_\alpha^T L^{-1}_{\alpha\beta}(z,z')\lambda_\beta(z')\, dz\, dz' \\ &\equiv \frac{1}{2}(\boldsymbol{\lambda}, \mathbf{L}^{-1} \circ \boldsymbol{\lambda}) \end{aligned} \tag{5.5}$$

where we do not exclude the possibility that the kernel, $\mathbf{L}^{-1}(z,z')$, of the functional operator appearing in (5.5) involves a generalized function. Carrying out the standard variational procedures to maximize the posterior density, we obtain a formal statement of the condition to optimize the parameter distribution, $\boldsymbol{\lambda}(z)$:

$$\boldsymbol{\lambda} + \mathbf{L} \circ \mathbf{l}'(\boldsymbol{\lambda}) = 0, \tag{5.6}$$

where the kernel of this operator is the covariance,

$$L_{\alpha\beta}(z,z') \equiv \langle \lambda_\alpha(z) \lambda_\beta(z') \rangle. \tag{5.7}$$

By the assumption of the smoothness of this covariance, $\mathbf{L}$, we may establish at least an approximate connection between this more rigorous formalism and the practical modification which extends the theory of Section 3 to smoothly continuous parameters. The main approximation we make is expressed in the following, which uses the substitution (5.4) in (5.6):

$$\begin{aligned} \boldsymbol{\lambda}(z) &= -(\mathbf{L} \circ \mathbf{l}'_0)(z) - \left(\mathbf{L} \circ (\mathbf{l}''_0 \circ \boldsymbol{\lambda})\right)(z) + \ldots, & (5.8) \\ &\approx -(\mathbf{L} \circ \mathbf{l}'_0)(z) - (\mathbf{L} \circ \mathbf{l}''_0)(z) \boldsymbol{\lambda}(z). & (5.9) \end{aligned}$$

It is this approximation, (5.9), that alows us to obtain the practical generalization of the Bayesian solution described in Section 3 since, with it, we now obtain:

$$\sum_\beta \left[ \mathbf{I}_{\alpha\beta} + (\mathbf{L} \circ \mathbf{l}''_0)_{\alpha\beta} \right] \lambda_\beta(z) \approx -(\mathbf{L} \circ \mathbf{l}'_0)_\alpha(z). \tag{5.10}$$

But, as in Section 3, we may identify the effective weight provided by the data:

$$W_{\alpha\beta}(z) \equiv (\mathbf{L} \circ \mathbf{l}''_0)_{\alpha\beta}(z), \tag{5.11}$$

and an approximation to the continuously varying maximum-likelihood estimate, $\hat{\boldsymbol{\lambda}}^{(M.L.)}(z)$ that satisfies:

$$\mathbf{W}(z) \hat{\boldsymbol{\lambda}}^{(M.L.)}(z) \equiv -(\mathbf{L} \circ \mathbf{l}'_0)(z). \tag{5.12}$$

To be approximately consistent with the presence of additional sampling noise when stochastic estimation procedures are adopted for this problem, we again reduce the effective weight, $\hat{\mathbf{W}}(z) = (p/(p+1))\mathbf{W}(z)$, so that, corresponding to (3.28), we obtain a spatially adaptive form of the Bayesian estimation:

$$(\hat{\mathbf{W}}(z) + \mathbf{I}) \boldsymbol{\lambda}^{(B)}(z) = \frac{-p}{p+1} (\mathbf{L} \circ \mathbf{l}'_0)(z). \tag{5.13}$$

As written above, the application in a practical context of the large-scale smoothing operator, $\mathbf{L}\circ$, has been left rather vague We notice that, in the estimates of the *global* log-likelihood function, every term has the form of an inner product over the space of measurements. In a related context of formulating estimates of effective information densities in a data network, where the global quantities are also expressed formally as measurement-space summations, Purser and Huang (1993) and Huang and Purser (1996) were able to show that serviceable

*local* estimates of these information densities could be consistently derived by replacing each formal summation by a smoothly-weighted regional summation, through the intervention of a spatial smoothing filter applied to the terms to be summed. We expect that a similar device should apply in the present context, giving us, in effect, a local log-likelihood-density, provided there exists a reliable method of projecting measurement-space quantities onto the analysis grid. (Efficient filters require a regular grid to operate on.)

The prerequisites for the spatially adaptive procedure are: firstly, the availability of a positivity-preserving large-scale smoother; secondly, a way of projecting a quantity associated with each measurement into the grid domain in a way that preserves the integrated magnitude of the quantity. For discrete data, this latter requirement is not too hard to meet, but may require more careful thought in the case of data which are themselves distributed measurements. The idea, then, is to project the contributions of the vector inner products of Eq. (3.18) for the gradient and the contributions of Eq. (3.19) for the Hessian into the analysis grid domain and smooth them using the operator $L\circ$. Consider the example of the term $\mathbf{f}_0^T\mathbf{Q}_\alpha\mathbf{f}_0$ of (3.18) obtained by finite differencing, in parameter space, the terms, $\mathbf{f}_0^T\mathbf{Q}\mathbf{f}_0$. The latter may be expressed symmetrically as the vector inner product, $\mathbf{v}^T\mathbf{v}$, where,

$$\mathbf{v} = \mathbf{C}^T\mathbf{f}_0. \tag{5.14}$$

Although some of the contributing components, $v_i^2$ are in measurement space and need to be interpolated, somehow, into the spatial domain, once this has been done it is fairly obvious how the global summation of such components may be replaced by a smoothly modulated local accumulation, essentially through the application of a smoother whose profile mimics that of the covariance $\mathbf{L}$:

$$(\mathbf{f}_0^T\mathbf{Q}\mathbf{f})_i = \sum_j L_{ij}(v_j)^2. \tag{5.15}$$

Finite differencing of these fields with respect to simultaneous *global* changes of the basic parameters, $\lambda_\alpha$, produce the requisite spatially varying and smooth components, $(\mathbf{f}_0^T\mathbf{Q}_\alpha\mathbf{f}_0)_i$ at all locations. Analogous methods can be used in *all* the terms of (3.18) and (3.19) in order to extend the respective quantities in a symmetric way to spatially varying fields.

The simplest choice for the form of the kernel of the spatial smoother $L$ is a Gaussian function, or a suitable smooth generalization of it on the sphere. Good approximations to such smoothers can be realized through the applications of carefully designed recursive filters. Such filters have already been developed at the U.K. Meteorological Office (e.g., Purser and McQuigg 1982, Lorenc 1992) and are being refined at NCEP primarily for use in the generation of the quasi-Gaussian components $\mathbf{G}_k$ of the operators $\mathbf{C}$, as described by (3.5). The appropriate horizontal scale for the present use of such filters is presumably related directly to the coherence scale implied by a more complete statistical model for the prior probability density of a geographically varying $\boldsymbol{\lambda}$.

## 6. Idealized experimental illustration of the method

A highly idealized one-dimensional simulation will serve to illustrate some of the methods proposed here. We consider a gridded one-dimensional domain, coordinate $z$, on which a

variational analysis of a single variable is carried out. The error in the background estimate, $\mathbf{x}^b$, is unbiased and normally distributed with a covariance $\mathbf{B}$ of the form,

$$B_{ij} = \sqrt{v(z_i)v(z_j)}\exp\left[-(s(z_i)-s(z_j))^2/2\right], \tag{6.1}$$

where $v(z)$ is the variance of background error at $z$ and $s(z)$ is an "effective distance" in units proportional to the covariance's coherence scale, obtained by integrating a local metric term, $ds/dz$. Two statistical parameters, $\boldsymbol{\lambda} \equiv (\lambda_1, \lambda_2)$, control the magnitude of the variance and the local coherence scale through the parameterizations,

$$v(z) = v^{(0)}\exp(\lambda_1(z)/\sigma_1), \tag{6.2}$$

$$\frac{ds(z)}{dz} = \frac{1}{z^{(0)}}\exp(\lambda_2(z)/\sigma_2). \tag{6.3}$$

The measurements are distributed along a central portion of the domain (to allow a simplification in the numerical treatment of boundaries) in a quasi-random fashion; the measurement locations all fall on a unit-spaced grid in $z$ with random spacing of consecutive locations equally likely to be any integer between extremes, $\text{sep}^-$ and $\text{sep}^+$. The variance of the errors in the observations, $R_{ii}$, is constant and these errors are uncorrelated.

In the first instance, we shall restrict the parameters $\boldsymbol{\lambda}$ to being global constants, so that all the available measurements provide information about the same pair of parameters. The following experimental constants are used:

$$\begin{aligned} \text{sep}^- &= 2, \\ \text{sep}^+ &= 12, \\ R_{ii} &= 1.0, \\ v^{(0)} &= 25.0, \\ \sigma_1 &= 0.45, \\ z^{(0)} &= 5.0, \\ \sigma_2 &= 0.25. \end{aligned} \tag{6.4}$$

We perform a simulation using the above constants on a domain large enough to contain $m = 283$ (simulated) measurements. We choose the "true" parameters of the simulation to take on five pairs of values, the first of which sets both components to zero to serve as a test of the stability of the method. In each case, the retrieved estimates are obtained using only one iteration of the Newton-Raphson procedure. Since, in a simulation of this simplicity we are able to compute the components of the matrix $\mathbf{Q}$ explicitly, we can calculate the "exact" terms of the gradient and Hessian of $l_1$ *without* recourse to the stochastic method (alternatively, we could have run an experiment with an extremely large number of stochastic vectors $\boldsymbol{\epsilon}$ and allowed the law of large numbers to give us essentially the same result). We shall refer to this "exact" estimate as $\boldsymbol{\lambda}^{(E)}$, but we should bear in mind that strictly, there remain errors of truncation due to the parameter-space finite differencing (a very small error, as the stencil size is quite small), and the fact that we still only take one Newton-Raphson step. Also, although in $\boldsymbol{\lambda}^{(E)}$ we have formally eliminated the artificial component of stochastic noise, we still retain the

effective weighting multiplier of one half, appropriate to $p = 1$, in order to make the numerical estimates more directly comparable with those that *do* involve artificial stochastic terms. The next estimator, $\boldsymbol{\lambda}^{(S)}$, is the one using single-$\boldsymbol{\epsilon}$ stochastic trace estimates for the gradient and Hessian of the log-likelihood, and includes all four terms of the Hessian as given by (3.19). Finally, $\boldsymbol{\lambda}^{(P)}$ is the estimator which, like $\boldsymbol{\lambda}^{(S)}$, employs the stochastic trace method, but now with the further simplifying approximation of the "partial" Hessian suggested by (3.25). Table 2 lists the estimators obtained in the five experiments, all of which used the same pseudo-random number sequences throughout.

TABLE 2. Simulated retrievals of global statistical parameters, $\boldsymbol{\lambda}$

| Realization | $\boldsymbol{\lambda}$ | $\boldsymbol{\lambda}^{(E)}$ | $\boldsymbol{\lambda}^{(S)}$ | $\boldsymbol{\lambda}^{(P)}$ |
|---|---|---|---|---|
| 1 | ( 0.0, 0.0) | ( 0.06, -0.18) | (-0.07, -0.29) | (-0.08, -0.41) |
| 2 | ( 1.0, 0.0) | ( 0.84, -0.19) | ( 0.77, -0.26) | ( 1.05, -0.70) |
| 3 | (-1.0, 0.0) | (-1.00, -0.24) | (-1.22, -0.41) | (-0.80, -0.23) |
| 4 | ( 0.0, 1.0) | (-0.01, 1.63) | (-0.22, 1.09) | (-0.17, 0.62) |
| 5 | ( 0.0, -1.0) | ( 0.22, -0.70) | ( 0.13, -0.75) | ( 0.01, -2.07) |

In a second class of experiments, we permit the parameters $\boldsymbol{\lambda}$ some freedom to vary by making them random variables, normally distributed with a mean, $\langle \boldsymbol{\lambda} \rangle = \mathbf{0}$ and a covariance,

$$\langle \boldsymbol{\lambda}(z_i) \boldsymbol{\lambda}(z_j) \rangle \equiv \mathbf{I} \exp\left( -\frac{(z_i - z_j)}{2\sigma_0^2} \right) \tag{6.5}$$

The scale of coherence, $\sigma_0$, of this covariance must be assumed to be large compared to the coherence scale of the covariance function in $\mathbf{B}$ that the parameters $\boldsymbol{\lambda}$ describe, otherwise the inadequate sampling of background variability will make it impossible, in principle, for the data to provide useful information about the $\boldsymbol{\lambda}$. Here we have selected a value,

$$\sigma_0 = 500 \tag{6.6}$$

We use the methods of Section 5, with the Gaussian function of (6.5) acting as the kernel of the smoothing operator. A Monte Carlo method was used to generate the parameter fields $\boldsymbol{\lambda}(z)$ consistent with the covariance model (6.5). We show the results of the retrieval of these parameters using one stochastic vector, $\boldsymbol{\epsilon}$, in Figure 2. The solid curve shows the true parameter in each case, the dotted curve shows the retrieval using the full four-term stochastic approximation to the Hessian. The dashed curve shows the retrieval using the single-term approximation corresponding to (3.25). Over most of the domain, the retrieved profiles of both parameters track the true parameters quite closely. Evidently, there is a problem at the left edge of the domain, especially when the full four-term approximation to the Hessian is used. The evidence suggests that this approximation to the Hessian is giving a matrix which is no longer, or only barely, positive definite. Figure 3 shows the result of employing four independent stochastic vectors $\boldsymbol{\epsilon}$. The fitting is significantly improved and the instability of the retrieved parameter, $\lambda_2$, associated with the four-term Hessian approximation has been removed. The fact that this retrieved parameter still undergoes a large excursion suggests the possibility that the true Hessian might actually be almost singular here. To investigate this possibility, Figures 4 and 5

display the results of modifying the local estimate on $l_{\alpha\beta}$ through a substitution based on the smooth function:

$$x'(x) = \frac{1}{2}(x + \sqrt{1 + x^2}) \tag{6.7}$$

applied to the Hessian matrix in the same manner as is described (for the logarithmic function) in Appendix A. The result is a smooth modification of the Hessian, having little impact where it is already strongly positive, but dramatically changing its character where the original becomes indefinite or negative, so that the modified hessian is everywhere positive definite. In the case of the single stochastic vector approximation, shown in Figure 4, we see that this empirical device is as effective a remedy as inflating the "ensemble" of $\boldsymbol{\epsilon}$ to four members. However, when the four-member $\boldsymbol{\epsilon}$ ensemble is treated using this same regularizing substitution (6.7), the resulting change is very small, suggesting either that, through sampling effects, the data themselves are misleading the estimation method, or that the single application of the Newton-Raphson technique that we have restricted ourselves to is inadequate in this particular case.

These idealized examples give an indication of the potential usefulness of the method, suggesting that it is indeed a valuable tool for applications of 3D-VAR where the covariances are required to vary in space. The examples also illustrate some of the potential pitfalls, particularly in the problematic estimation of the Hessian components. Our examples suggest that a greater robustness is achieved through the use of the one-term simplified approximation to the Hessian, even though this approximation may be quite poor when the true and default parameters are widely different.

## 7. Discussion

This note provides an outline of a proposed procedure to identify the spatially varying parameters of a variational assimilation scheme from the data themselves. The discussion has focused on 3D-VAR but these techniques should be equally applicable within the framework of a 4D-VAR scheme where, as discussed by Fisher and Courtier (1995), there is still a need to estimate the forecast error covariances. It is necessary to experiment with the empirical aspects of this proposed method, particularly the spatial scale of the smoother and the construction of the measurement projection operators that carry the inner-product contributions into the grid domain. We made an estimate for the extent to which the artificial stochastic vector inputs degrade the statistical efficiency of the estimation procedure, but our assessment relied on the assumption that the neglect of the stochastic contributions in the Hessian is not an important omission. Clearly, it would be desirable to carry out a more thorough analysis of this situation in order to determine whether our assumption is justifiable and, if not, precisely how one should set about deriving a more correct adjustment to the effective estimation efficiency.

It is hoped that this method will supply a tool for the objective modification of statistical parameters required by the new generation of spatially adaptive covariance models for variational meteorological data analysis. The emphasis of the present note has been the implementation of these methods within the measurement space. However, it is recognized that, in the variational analysis context, there exists a formal duality between 'model space' and 'measurement space' formulations, as recently discussed by Courtier (1997); it is therefore desirable to determine the extent to which a corresponding duality extends to the case of parameter estimation. This

becomes particularly relevant when, as in the case of the operational NCEP analyses (Parrish and Derber 1992), there are numerical efficiencies related to the treatment of weakly nonlinear observation operators (generalizing our $\mathbf{H}$) which are only easily exploited by the analysis formulation implemented in the model-space. Another question worthy of investigation is whether the somewhat *ad hoc* empirical treatment of the spatial variability of the statistical parameter fields $\boldsymbol{\lambda}$ could be dealt with in a practical implementation of the more rigorous Bayesian treatment of these quantities' spatial covariances. The problem of handling the strong nonlinearity is fairly daunting in this case, but might not be totally beyond the bounds of practicality if the appropriate numerical methods can be developed, or perhaps evolved from the present tentative start on this problem.

Finally, we note that, while the emphasis of the present note has been entirely focused on the modes of estimation involving the likelihood function, it might alternatively prove to be feasible and valuable to carry through the corresponding exercise of making a spatially adaptive "GCV-density" regularized by an appropriately modelled input of 'prior' information; as with the likelihood calculations, the component terms are all expressible as trace estimates which can be rendered 'local', through appropriate spatial smoothing, in exactly the same fashion we suggested in the case of the likelihood calculations.


## Acknowledgments

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.


## Appendix A

*Derivation of the identity of Equation (3.1)*

Since $\mathbf{Q}$ is a symmetric positive definite matrix, there exists a diagonal matrix, $\hat{\mathbf{Q}}$, comprising the positive eigenvalues, related to $\mathbf{Q}$ through a rotation:

$$\mathbf{V}^T\mathbf{Q}\mathbf{V} = \hat{\mathbf{Q}}, \tag{A.1}$$

where the square matrix $\mathbf{V}$ is orthogonal:

$$\mathbf{V}^T\mathbf{V} = \mathbf{I}, \tag{A.2}$$

and hence,

$$\mathbf{Q} = \mathbf{V}\hat{\mathbf{Q}}\mathbf{V}^T. \tag{A.3}$$

Any analytic function, say $f(z)$, valid for real and positive arguments, $z$, has a natural matrix-valued generalization for arguments of the class of matrices exemplified by $\mathbf{Q}$, namely:

$$f(\mathbf{Q}) = \mathbf{V}f(\hat{\mathbf{Q}})\mathbf{V}^T, \tag{A.4}$$

where $f(\hat{\mathbf{Q}})$ is diagonal with components,

$$\left[f(\hat{\mathbf{Q}})\right]_{i,i} = f(\hat{Q}_{i,i}). \tag{A.5}$$

In particular,

$$\log(\mathbf{Q}) = \mathbf{V}(\log \hat{\mathbf{Q}})\mathbf{V}^T. \tag{A.6}$$

Now, since, by the orthogonality of $\mathbf{V}$,

$$|\mathbf{Q}| = |\hat{\mathbf{Q}}|, \tag{A.7}$$

and,

$$\text{trace}\left[\mathbf{V}(\log \hat{\mathbf{Q}})\mathbf{V}^T\right] = \text{trace}(\log \hat{\mathbf{Q}}), \tag{A.8}$$

and since,

$$\log|\hat{\mathbf{Q}}| = \log\left(\prod_i \hat{Q}_{i,i}\right) = \sum_i \log \hat{Q}_{i,i} = \text{trace}(\log \hat{\mathbf{Q}}) \tag{A.9}$$

we obtain the desired result:

$$\log|\mathbf{Q}| = \text{trace}(\log \mathbf{Q}). \tag{A.10}$$

## Appendix B

*Finite difference stencils for gradient and Hessian evaluations*

In constructing a differencing stencil for the gradient and Hessian calculations based on simplex arrangements of the data points it is probably desirable that the dispersion of these points as measured by their averaged second moments is comparable with or smaller than the assumed prior covariance, which is the identity matrix in this case.

In $\kappa$ Euclidean dimensions we define the $\kappa+1$ vectors $\mathbf{e}^{(\beta)}, \beta = 0, \ldots, \kappa$:

$$e_\alpha^{(0)} = \frac{1}{2}\left(\begin{array}{c}\alpha+1\\2\end{array}\right)^{-\frac{1}{2}} \tag{B.1}$$

$$e_\alpha^{(\beta)} = \begin{cases} e_\alpha^{(0)} & : \quad \alpha > \beta \\ -\alpha e_\alpha^{(0)} & : \quad \alpha = \beta \\ 0 & : \quad \alpha < \beta \end{cases} \tag{B.2}$$

These points define a regular unit-sided simplex centered at the origin, as we can verify from the identities,

$$(\mathbf{e}^{(\beta)} - \mathbf{e}^{(0)})\cdot(\mathbf{e}^{(\gamma)} - \mathbf{e}^{(0)}) = \begin{cases} 1 & : \quad \gamma = \beta \\ 1/2 & : \quad \gamma \neq \beta \end{cases} \tag{B.3}$$

for $\beta, \gamma \neq 0$, which shows that each pair of edges radiating from $e^{(0)}$ forms an equilateral triangle. The matrix of averaged second moments of these points is

$$\overline{\mathbf{e}\mathbf{e}^T} = \frac{1}{2(\kappa+1)}\mathbf{I}. \tag{B.4}$$

Writing $\psi_j \equiv \psi(\mathbf{e}^{(j)})$, we find that the partial derivatives evaluated using these points as a finite differencing stencil are given by:

$$\frac{\partial \psi}{\partial \lambda_\beta} = \left( \begin{array}{c} \beta+1 \\ 2 \end{array} \right)^{-\frac{1}{2}} \left( \sum_{\alpha=0}^{\beta-1} (\psi_\alpha - \psi_\beta) \right). \tag{B.5}$$

If, in addition to these $\kappa+1$ points, we include the $\kappa(\kappa+1)/2$ points at the edge midpoints, we have a stencil suitable also for the calculation of the Hessian components (with the bonus that the gradient calculations can now be carried out at a higher order of accuracy). The condition of collocation of the stencil values with the interpolating second-degree polynomial is probably the simplest method for supplying the differencing coefficients in this case.

## Appendix C

*Using model space to invert the innovation covariance*

The most expensive part of the algorithm for parameter estimation is the solution of the linear inversion problems, such as (3.7) and (3.9). These inversions are defined in observation space, while the machinery for some existing operational 3D-VAR and 4D-VAR systems is defined in model space. However, because of the duality of the two forms, there is a simple relationship that can be exploited so that the required inversions can be obtained without altering existing codes.

The vector $\mathbf{x}^a$ which minimizes (2.9) results in a vanishing gradient of (2.9),

$$\mathbf{B}^{-1}(\mathbf{x}^a - \mathbf{x}^b) - \mathbf{H}^T\mathbf{R}^{-1}(\mathbf{y}^o - \mathbf{H}\mathbf{x}^a) = 0 \tag{C.1}$$

or

$$\mathbf{x}^a = \mathbf{x}^b + \mathbf{B}\mathbf{H}^T\mathbf{R}^{-1}(\mathbf{y}^o - \mathbf{H}\mathbf{x}^a). \tag{C.2}$$

Comparing (C.2) with (2.10), it is apparent that the vector $\mathbf{f}$ also satisfies

$$\mathbf{f} = \mathbf{R}^{-1}(\mathbf{y}^o - \mathbf{H}\mathbf{x}^a). \tag{C.3}$$

The model space algorithms generate $\mathbf{x}^a$, given $\mathbf{y}^o$, $\mathbf{x}^b$, $\mathbf{R}$, and $\mathbf{B}$. Using (C.3), it is relatively simple to obtain $\mathbf{f}$.

To obtain a random vector $\mathbf{r}$ with covariance $\mathbf{Q}^{-1}$, as for example in (3.9), a simple procedure that still makes use of the unmodified model space algorithm is to perturb the observation vector $\mathbf{y}^o$ with the random vector $\mathbf{q}$,

$$\hat{\mathbf{y}}^o = \mathbf{y}^o + \epsilon\mathbf{q}, \tag{C.4}$$

for an appropriate scalar, $\epsilon$. Using $\hat{\mathbf{y}}^o$, a vector $\hat{\mathbf{x}}^a$ is generated, and finally $\hat{\mathbf{f}}$ is obtained from (C.3). It is then easy to show that

$$\mathbf{r} = \epsilon^{-1}(\hat{\mathbf{f}} - \mathbf{f}). \tag{C.5}$$

It is possible that these computations are still valid even when the forward operator $\mathbf{H}$ is non-linear, but this needs careful examination to be sure that the statistical arguments are still valid.

## References


Courtier, P. 1997 Dual formulations of four-dimensional variational assimilation. *Quart. J. Roy. Meteor. Soc.*, **123,** 2449–2461.

Craven, P., and G. Wahba 1979 Smoothing noisy data with spline functions. *Numer. Math.*, **31,** 377–403.

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

Dee, D. P., 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.

Desroziers, G. 1997 A coordinate change for data assimilation in spherical geometry of frontal structure. *Mon. Wea. Rev.*, **125,** 3030–3038.

Evensen, G. 1994 Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res.*,**99,**(C5), 10143–10162

Fisher, M. 1998 Development of a simplified Kalman filter. *ECMWF Technical Memorandum* 260, 16pp.

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

Gear, C. W. 1971 *Numerical Initial Value Problems in Ordinary Differential Equations* Prentice-Hall, Englewood Cliffs, NJ

Gill, P. E., W. Murray, and M. H. Wright 1981 *Practical Optimization.* Academic Press, London, 401pp.

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

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

Gong, J., G. Wahba, D. R. Johnson, and J. Tribbia 1998 Adaptive tuning of numerical weather prediction models: simultaneous estimation of weighting, smoothing, and physical parameters. *Mon. Wea. Rev.*, **126,** 210–231.

Hollingsworth, A., and P. Lönnberg 1986 The statistical structure of short-range forecast errors as determined from radiosonde data. Part I: The wind field. *Tellus*, **38A,** 111–136.

Huang, H.-L., and R. J. Purser 1996 Objective measures of the information density of satellite data. *Meteor. Atmos. Phys.*, **60,** 105–117.

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.

Lorenc, A. C. 1992 Iterative analysis using covariance functions and filters. *Quart. J. Roy. Meteor. Soc.*, **118,** 569–591.

Mitchell, H. L., and P. L. Houtekamer 2000 An adaptive ensemble Kalman filter. *Mon. Wea. Rev.*, **128,** 416–433.

Parrish, D. F., and J. C. Derber 1992 The National Meteorological Center's Spectral Statistical Interpolation Analysis System. *Mon. Wea. Rev.*, **120,** 1747–1763.

Purser, R. J., and R. McQuigg 1982 A successive correction analysis scheme using recursive numerical filters. Met. O 11 Tech. Note, No. 154, British Meteorological Office. 17pp.

Purser, R. J., and H.-L. Huang 1993 Estimating the effective data density in a satellite retrieval or an objective analysis. *J. Appl. Meteor.*, **32,** 1092–1107.

Riishøjgaard, L.-P. 1998 A direct way of specifying flow-dependent background error correlations for meteorological analysis systems. *Tellus*, **50A,** 42–57.

Swinbank, R., R. Menard, and L. P. Riishøjgaard 2000 Anisotropic error correlation modelling in a simple 2-D assimilation system. (Abstract) Third WMO International Symposium on Assimilation of Observations in Meteorology and Oceanography, Montreal, Canada, 1999.

Thiébaux, H. J. 1976 Anisotropic correlation functions for objective analysis. *Mon. Wea. Rev.*, **104,** 994–1002

Thiébaux, H. J., H. L. Mitchell, and D. W. Shantz 1986 Horizontal structure of hemispheric forecast error correlations for geopotential and temperature. *Mon. Wea. Rev.*, **114,** 1048–1066

Wahba, G. 1990 *Spline Models for Observational Data*, SIAM. CBMS-NSF Regional Conference Series in Applied Mathematics, Vol. 59, 165 pp.

Wahba, G., and J. Wendelberger 1980 Some new mathematical methods for variational objective analysis using splines and cross-validation. *Mon. Wea. Rev.*, **108,** 1122–1145.

Wahba, G., D. R. Johnson, F. Gao, and J. Gong 1995 Adaptive tuning of numerical weather prediction models: randomized GCV in three- and four-dimensional data assimilation. *Mon. Wea. Rev.*, **123,** 3358–3369.

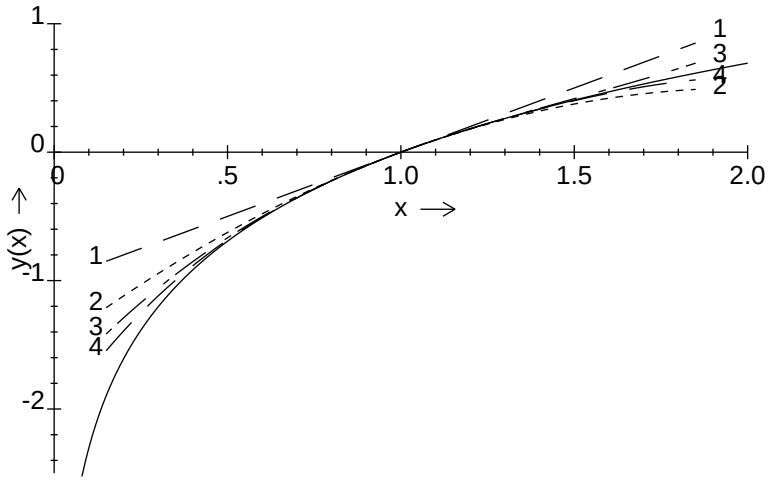


Figure 1. Fit to the logarithmic function by polynomials of successive degrees. The solid curve is the exact logarithm of $x$; the degrees of the approximating polynomials are indicated.

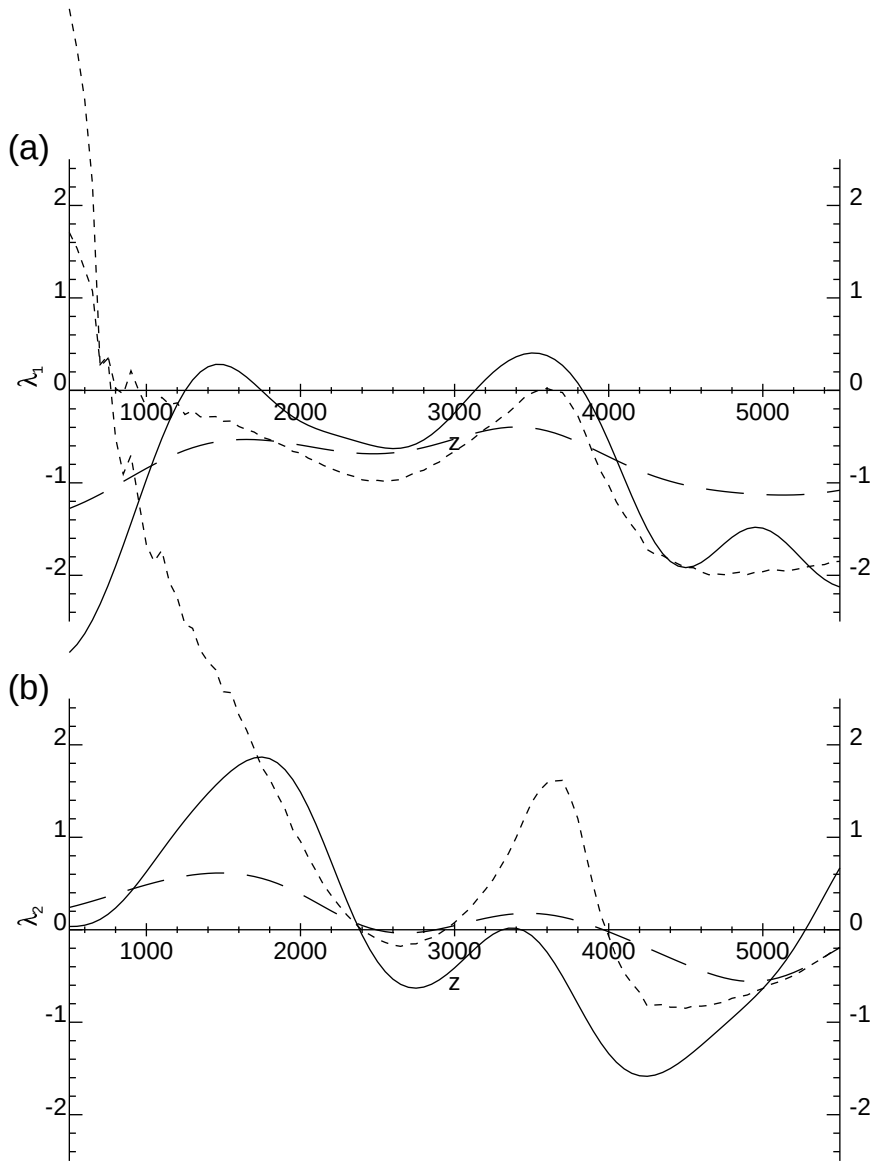


Figure 2. Retrieval of amplitude parameter, $\lambda_1$, and scale parameter, $\lambda_2$, using the Bayesian method with stochastically approximated trace terms based on a single stochastic vector, $\epsilon$. The solid curves are the true paremeter profiles. The curves formed by short dashes show the retrievals using the complete four-term approximation to the Hessian of the likelihood function; the long-dashed curves show the corresponding retrievals when the one-term approximation to the Hessian is used. Note that the retrievals using the four-term Hessian approximation, especially of $\lambda_2$, become unstable near the left of the domain in this example, owing to the presence of small eigenvalues in the estimated Hessian.

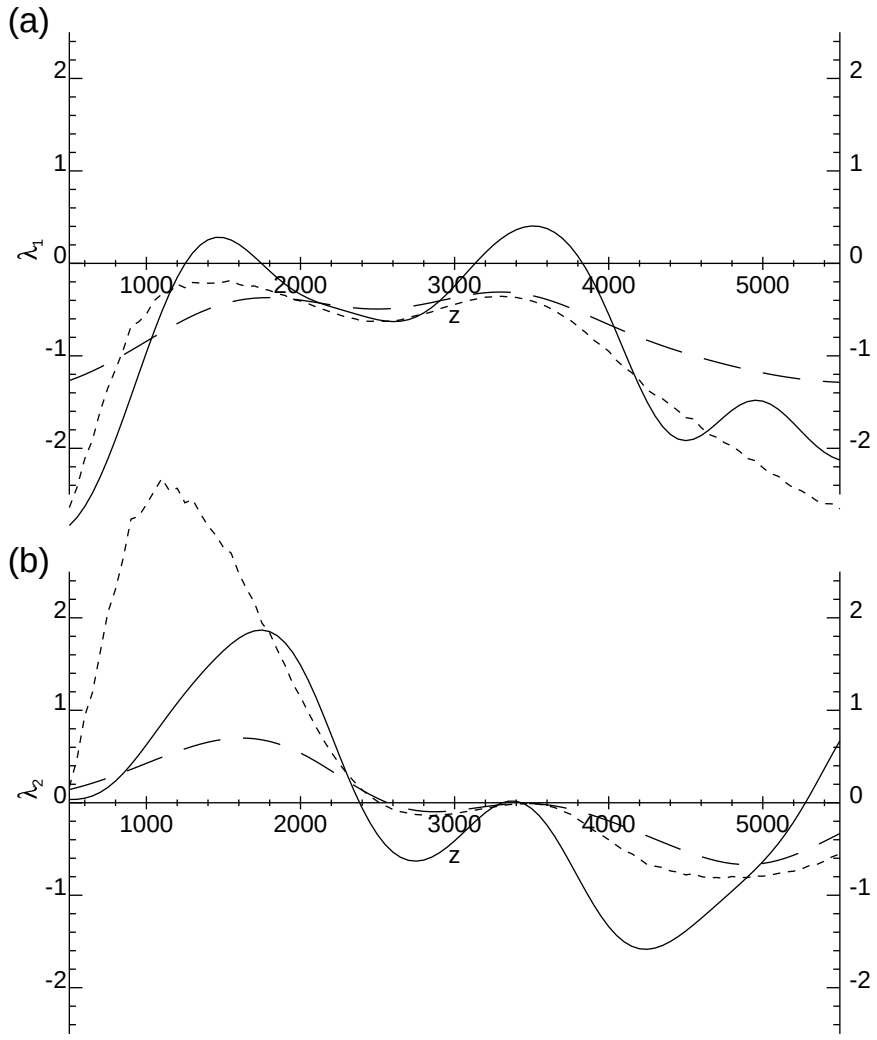


Figure 3. As in Figure 2, but using an ensemble of four stochastic vectors $\epsilon$. The instability evident in Figure 2 has been removed but the remaining erroneous excursion of the second parameter near the left boundary suggests that the full Hessian actually possesses small eigenvalues in this region.

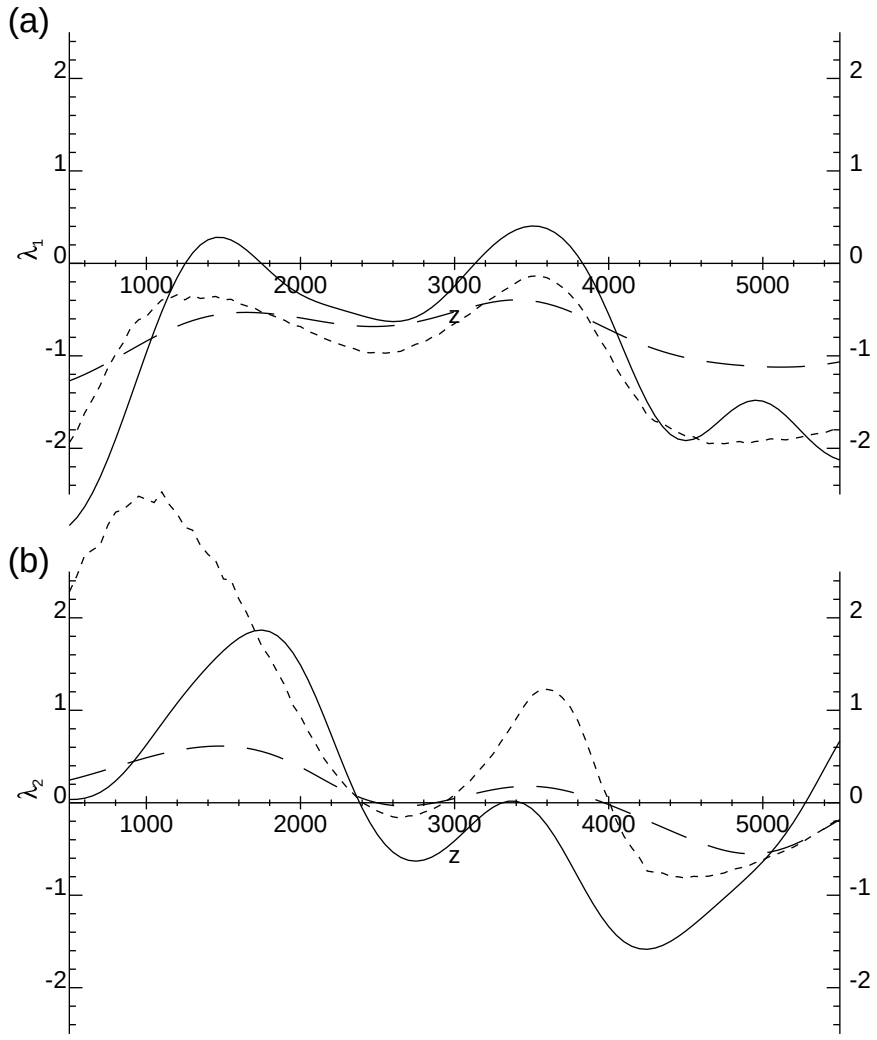


Figure 4. As in Figure 2, and with a single stochastic vector, but with the Hessian estimates modified to smoothly avoid negative or small eigenvalues

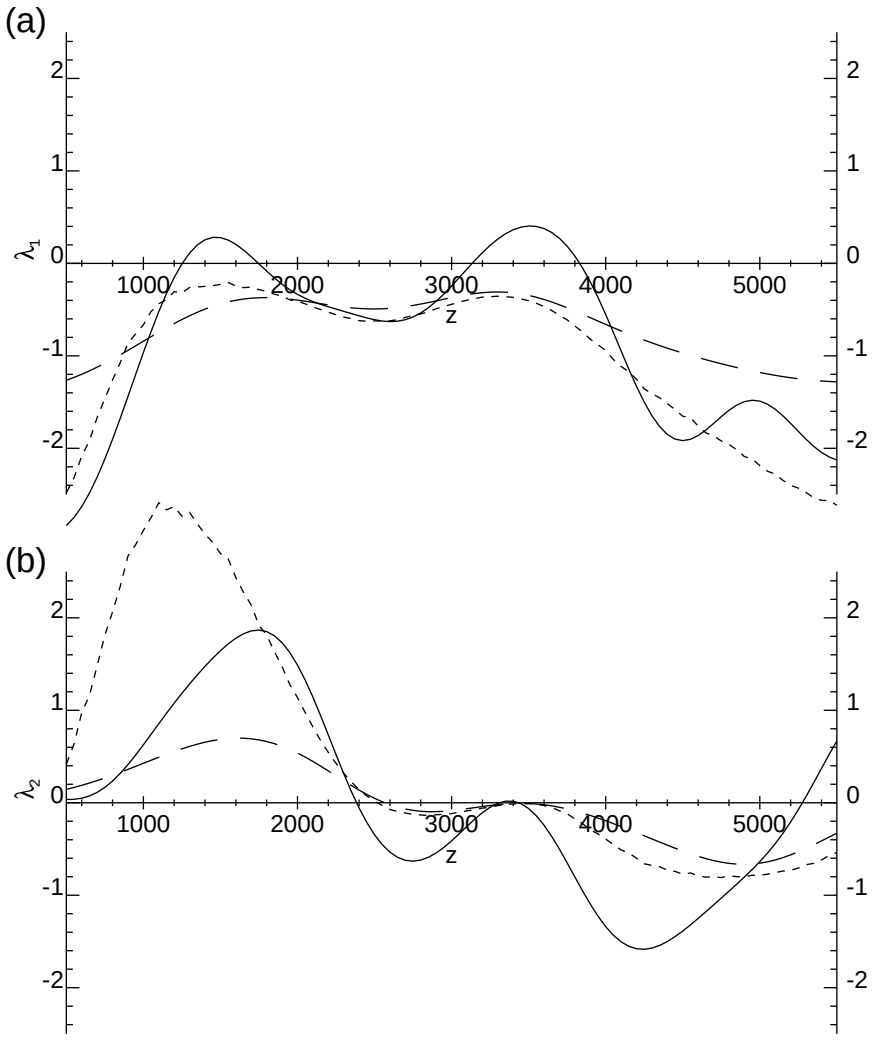


Figure 5. As in Figure 4, but with an ensemble of four stochastic vectors