DEPARTMENT OF STATISTICS
University of Wisconsin
1210 West Dayton St.
Madison, WI 53706

TECHNICAL REPORT NO. 920

April 27, 1994

Adaptive tuning of numerical weather prediction models: Part I: randomized GCV and related methods in three and four dimensional data assimilation [1]

by

Grace Wahba

Donald R. Johnson

Feng Gao

and

Jianjian Gong

[1]Dedicated to the memory of Henry Fleming and his contributions to Meteorology and Mathematics.

Abstract

We consider the problem of ingesting scattered, noisy, heterogeneous data into an imperfect, evolving mathematical model for the purpose of prediction of a dynamical system. In particular we consider models which perform a three or four dimensional data assimilation using variational/statistical methods. This paper is written with global scale medium range numerical weather prediction models in mind, but the results are applicable more generally to computer models of dynamical systems which ingest scattered, noisy data.

In variational data assimilation, optimal ingestion of the observational data, and optimal use of prior physical and statistical information involves the choice of numerous weighting, smoothing and tuning parameters which control the filtering and merging of divers sources of information. Generally these weights must be obtained from a partial and imperfect understanding of various sources of errors, and are frequently chosen by a combination of historical information, physical reasoning and trial and error.

It is our purpose to show how the methods of generalized cross validation (GCV) and unbiased risk(UBR), and two extensions of GCV and UBR, namely Partial GCV (PGCV) and Partial UBR (PUBR) can be used to estimate certain of the most important tuning, weighting and smoothing parameters, in the context of extremely large operational models (10^6 degrees of freedom). For comparison we review how maximum likelihood (ML) and generalized maximum likelihood(GML) can be used for some of the same purposes when the model is correctly specified. A theoretical proof of the properties of PUBR is given.

In order to be useful in operational numerical weather prediction, the estimates must have the potential for operational implementation in the context of the iterative solution of extremely large variational problems. We carry out two 'toy' simulations which suggest how this may be done, via the technique of randomized trace estimation. In many examples of approximate solution of extremely large variational problems via iterative methods, the number of iterations can also play a role as a 'tuning' parameter. In the second 'toy' simulation it is demonstrated how UBR and GCV can also be used to decide when to stop the iteration.

Following these demonstrations, which are carried out in variational problems without time dependency, we describe how the methods may be implemented in several four dimensional contexts that appear in the current literature.

In this paper, we explicitly only describe the choice of parameters which play the role of signal-to-noise ratios, and number of iterations. Other estimable parameters are specifically described only in passing. In the second paper in this series, we develop a general theory of parameter estimability, which will provide a general yardstick for selecting parameters which can be estimated by these methods. In the third paper we give a new GCV-like method for the estimation of relative accuracy of heterogeneous sources of data.

Contents

1 Introduction

In modern global scale numerical weather prediction models, it is common to update the state vector (x), which will serve as initial conditions to integrate forward the primitive equations, by combining observations, forecast and possibly other physical constraints in a manner that is (some approximation to) the solution to a minimization problem of the following general form: Find x to minimize

$$J(x) = (y - K(x))'S^{-1}(y - K(x)) + (x - x_*)'\Sigma^{-1}(x - x_*). \tag{1.1}$$

Here y is a vector of data which are related to the state vector x by the possibly partly nonlinear operator K, x_* is a known vector and the matrices S and Σ embody information concerning how close y is expected to be to $K(x)$ and how close x is expected to be to x_*. At this point we are being deliberately vague about y because it may include forecast 'data' as well as observational data, alternatively, forecast 'data' may be incorporated into x_*. Formally, the matrices S and Σ may be derived as covariance matrices under certain statistical assumptions. See Parrish and Derber(1992), Lorenc, Bell and McPherson(1991), Lorenc(1986), Wahba(1990, 1985, 1982), Wahba and Wendelberger(1980), Kimeldorf and Wahba(1971). Four dimensional assimilation with the model as a strong constraint can be put in this framework by, for example, letting x be the state at an initial time and including model integrations in K. Physically based penalties, for example, energy in gravity waves, may be incorporated in the penalty term $(x-x_*)\Sigma^{-1}(x-x_*)$. See Lewis and Derber(1985), Courtier and Talagrand(1987,1990), Zou, Navon and LeDimet(1992), Zou, Navon and Sela(1993), Rabier *et al* (1992), Li, Navon, Courtier and Gautier(1993) and references cited there. The model strong constraint can be weakened, while remaining within the realm of (1.1) by allowing a bias or forcing term, see Derber(1989), Wergen(1992), Zupanski(1993), and the model itself may also be put in as a weak constraint.

The problem is to choose certain unknown parameters in Σ, K and S. This fairly simple statement of the problem conceals many choices that must be made in practice. The entries in Σ and S contain numerous smoothing, tuning and weighting parameters, that in practice are obtained from putative error covariances which aim to take into account measurement error, forecast error, errors of representativeness, and model error, from prior physical and statistical information about the atmosphere, from physical intuition, and from trial and error. K may contain instrument calibration constants, physical parameters, and so forth. Some of the discrepancies between y and $K(x)$ are fairly well understood (e.g. radiosonde measurements) while others, particularly satellite radiances are not, due in part to the difficulty of modeling the forward problem accurately. This problem is exacerbated in four dimensional assimilation where K includes model integrations. Forecast error covariances can to a certain extent be studied empirically from historical data by comparing forecast and observation over a period of time. See Hollingsworth and Lonnberg(1986), Lonnberg and Hollingsworth(1986), Bartello and Mitchell(1993), Goerss and Phoebus(1993).

It is the purpose of this paper to initiate the development of a general theory of adaptive estimation of smoothing, weighting and tuning parameters based on generalized cross validation and unbiased risk methods, for parameters which are hidden in Σ and in K , and also to some extent in S, which is applicable to the tuning of three and four dimensional numerical weather prediction models, and to other data assimilation problems solved via variational problems which can be put in the general form (1.1).

There is, of course much interest in developing objective methods of obtaining these parameters. Aside from work in the spirit of Hollingsworth and Lonnberg, which involves directly fitting parametric or semiparametric models of covariances to large historical data sets of forecasts minus observations, there are several other trains of research with similar goals, based on Kalman filter theory. The Kalman filter theory shows how the forecast error covariance evolves, given past data patterns and past observational error and 'plant noise' (usually considered to be model errors in this context). Cohn(1993) looks at stochastic dynamic equations for the growth and propagation of forecast errors. Dee, Cohn, Dalcher and Ghil(1985) and later Dee(1990), use a simplified model for the evolution of the forecast error covariance, and fit the simplified model using historical data. Recently Daley(1992a,b,c,d) develops some ingenious methods for estimating the stationary isotropic part of certain required covariances in the context of Kalman filtering, after having parameterized them with a small number of unknowns.

The estimates that we will discuss here are all "adaptive" or "dynamic", in the sense that they are carried out *simultaneously* with the estimation of x , unlike the methods described in the references given above which use historical "after the fact" data. By this we mean that a historical sequence of estimates of x's are obtained from a model using whatever parameters exist in the model. Then new parameters are estimated given this series of x's along with their associated series of forecasts and data. Once these new parameters are obtained they are then substituted in the model. After this substitution, the model error properties may also change, so that this procedure needs to be iterated, see for example, Daley(1982d), Section 3b. The proposed adaptive methods can be used to monitor or fine tune certain parameters dynamically which have been obtained from historical data, from a Kalman filter method, or from other methods. We will be discussing four methods of estimation of parameters in Σ, S and K above, namely, maximum likelihood (ML), generalized maximum likelihood (GML)[2], generalized cross validation(GCV), and unbiased risk (UBR) estimators. The GCV and UBR are so-called 'performance-oriented' methods. They are appropriate for estimating certain of the smoothing, weighting and tuning parameters within a framework or structure which may be misspecified - these methods attempt to find the parameters which optimize some performance criteria within whatever structure is given, *correct or not.* The performance criteria are in terms of some distance to the *true* atmosphere, which, of course is unknown. The GCV and UBR are computable *estimates* or *proxies* for this distance. In contrast, ML or GML is in theory optimum if the overall model is correctly specified in a certain statistical sense, (except for the parameters being estimated); but may behave badly from a 'performance-oriented' criteria if it is not. See Cramer (1954), Wahba(1985a). GCV is the most general performance oriented method, but UBR can be used to advantage when certain error variances are known well (for example, error variances in 500hpa heights from radiosondes). Dee(1993, 1994) discusses ML estimation in a form similar to that which we will consider here. When the second term in (1.1) is used to impose a physically based penalty, the ML estimate of, say, a multiplier on this penalty, is not properly defined, but a sound argument can be made for the use of GCV or UBR to choose it. Other examples of penalty terms where an argument can be made for the use of GCV can be found in Hoffman (1984, 1985), Hoffman and Louis (1990), Bennett and Budgell (1987).

In Section 2 we first give a general background on (three dimensional) smoothing, weighting, and tuning parameters, in the context of variational objective analysis.

In Section 3 we describe the standard ML, GML, GCV and UBR estimates, and new generalizations of GCV and UBR, namely partial GCV (PGCV) and partial UBR (PUBR)), and give a short discussion of the properties of PUBR and PGCV. The PUBR and PGCV estimates may be particularly useful in optimizing parameters from the point of view of a forecast criteria in four dimensional assimilation.

All of the theory would be, to some extent, moot in operating NWP systems without feasible methods to compute the estimates. The GCV and UBR estimates require the trace of the so-

[2]This is actually a special case of ML in which a noise variance parameter is factored out.

called influence matrix (known as the degrees of freedom for signal, equivalently the trace of the inverse Hessian of the variational problem), and the PUBR and PGCV require the trace of a certain submatrix. This might appear to be an insurmountable task for large problems. Recently Girard(1987,1989,1991) and Hutchinson (1989) have demonstrated that randomized GCV estimates can be efficiently computed by estimating these traces by Monte Carlo methods. In Section 4 we describe how randomized GCV and UBR estimation can be carried out in the context of a three dimensional operational system which computes the solution of a monster variational problem by a conjugate gradient or other iterative method, for example as in Parrish and Derber(1992). Future four dimensional systems are future candidates, for example, Li, Navon, Yang, Zou, Bates, Moorthi and Higgins(1994). We carry out some modest experiments on a 'toy' problem solved by a conjugate gradient algorithm. The results demonstrate that the randomized GCV and UBR estimates can be used to simultaneously choose both the number of iterations in the algorithm and a multiplier on a penalty term.

In Section 5 we show how GCV, UBR, PGCV and PUBR may be implemented in several different four dimensional variational assimilation methods that appear in the recent literature. Various authors have discussed the relationship of Kalman filtering to four dimensional variational problems, see Jazwinski(1970), Ghil and Malanotte-Rizzoli(1991), Thepaut and Courtier(1991), Thepaut, Hoffman and Courtier(1993). In this section we also (re)derive an exact result which relates four dimensional variational assimilation with the model as a weak constraint, and Kalman filtering. This result is then used to suggest how GML, GCV and UBR estimates may be applied in this context. In Section 6 we make a few remarks concerning the estimation of forecast accuracy based on GCV and UBR estimates, and Section 7 is a Summary. In an Appendix we propose a rationale for using the model as a strong constraint even when there is model error.

2 Background, three dimensional variational data assimilation

We review the well known statistical assumptions which lead to (1.1) and related procedures. See also Kimeldorf and Wahba(1970,1971), Lorenc(1986,1988), Wahba(1982,1985,1990).

Let x be the state vector of a NWP system. We suppose that x has some climatological mean that has already been subtracted out, and we will suppose that the mean values of the components of x can be treated as though they are zero.

If u is a vector, we will use the notation $u \sim \mathcal{N}(\mu, M)$ to mean, we will treat u as though it has a multivariate Gaussian distribution with mean μ and covariance matrix M. For now we will suppose $x \sim \mathcal{N}(0, \Sigma)$. We next suppose that there are (without loss of generality) two sources of information, y_1 and y_2 concerning x which satisfy the model

$$y_1 = K_1(x) + \epsilon_1 \tag{2.1}$$

$$y_2 = K_2(x) + \epsilon_2 \tag{2.2}$$

where $\epsilon_1 \sim \mathcal{N}(0, S_1)$ and $\epsilon_2 \sim \mathcal{N}(0, S_2)$, and it is assumed that ϵ_1 and ϵ_2 are independent of each other. The components of $K_1(x)$ and $K_2(x)$ represent functionals of x, if they are linear functionals then we may represent the K's as matrices. See O'Sullivan and Wahba(1985), Wahba(1985,1990), Lorenc(1986). In the case of satellite radiances the functionals in K are mildly nonlinear integrals. See O'Sullivan and Wahba(1985) for a direct treatment of this nonlinearity using B-splines for basis functions. Until noted otherwise we will assume for ease of exposition that the K's are linear or have been linearized. It is generally desirable, to linearize as late as is practicable.

Given the statistical assumptions on x, ϵ_1 and ϵ_2 above, the conditional expectation $\hat{x}$ of x, , given the data y_1 and y_2 is given by the minimizer of

$$(y_1 - K_1(x))' S_1^{-1} (y_1 - K_1(x)) + (y_2 - K_2(x))' S_2^{-1} (y_2 - K_2(x)) + x' \Sigma^{-1} x. \tag{2.3}$$

Proceeding under the assumption that the K's are linear, then the minimizer $\hat{x}$ of (2.3) is

$$\hat{x} = \left(K_1' S_1^{-1} K_1 + K_2' S_2^{-1} K_2 + \Sigma^{-1} \right)^{-1} \left(K_1' S_1^{-1} y_1 + K_2' S_2^{-1} y_2 \right). \tag{2.4}$$

It is well known that the posterior covariance matrix for $\hat{x}$ given y_1, y_2 is

$$(K_1' S_1^{-1} K + K_2' S_2^{-1} K_2 + \Sigma^{-1})^{-1}, \tag{2.5}$$

which is the inverse Hessian for the problem (2.3). This covariance may be derived by using the formulas for the conditional covariance of one random normal vector given another given in *e. g.* Anderson(1958, p 26) and the useful identity

$$\Sigma - \Sigma K' (K \Sigma K' + S)^{-1} K \Sigma = (K' S^{-1} K + \Sigma^{-1})^{-1}. \tag{2.6}$$

In three dimensional objective analysis typically the analysis takes place not on the state vector x, but on the so called analysis increment $\delta = x - x*$ where $x*$ is the forecast. That setup also fits into the theory we are discussing here, with a slightly different interpretation on the penalty, we now describe how that happens. Let y_2 be the forecast, that is, let $K_2 = I$, and let S_2 be the forecast error covariance. Conceptually we consider the (possibly fictional) approach that a forecast with covariance S_2 is then smoothed based on the prior covariance Σ. Then the model is

$$y_1 = K_1 x + \epsilon_1 \tag{2.7}$$

$$y_2 = x + \epsilon_2 \tag{2.8}$$

where x and ϵ_2 have the prior covariances Σ and S_2 respectively. Then $x*$, the conditional expectation of x given y_2 is given by

$$x* = (S_2^{-1} + \Sigma^{-1})^{-1} S_2^{-1} y_2 \tag{2.9}$$

and $x*$ has the posterior covariance

$$(S_2^{-1} + \Sigma^{-1})^{-1} = S*, \tag{2.10}$$

say. Now, the conditional expectation of x given $x*$ and the data y_1 is the minimizer $\hat{x}$ of

$$(y_1 - K_1 x)' S_1^{-1} (y_1 - K_1 x) + (x * - x)' S *^{-1} (x * - x). \tag{2.11}$$

Letting $\delta = x - x*$, which we will call the analysis increment, and

$$y_1 * = y_1 - K_1 x* \tag{2.12}$$

which we will call the observation increment, we may then rewrite (2.11) as

$$(y_1 * - K_1 \delta)' S_1^{-1} (y_1 * - K_1 \delta) + \delta' S *^{-1} \delta, \tag{2.13}$$

which is then solved for the analysis increment δ in terms of the forecast increment $y*$. The minimizer $\hat{\delta}$ is given by

$$\hat{\delta} = (K_1' S_1^{-1} K_1 + S*^{-1})^{-1} K_1' S_1^{-1} y_1 * . \tag{2.14}$$

It can be checked that $\hat{x} = \hat{\delta} + x*$ with $\hat{\delta}$ given by (2.14) is the same as $\hat{x}$ given by (2.4). Furthermore, the posterior covariance of $\hat{\delta}$ given $x*$ and y_1 is

$$(K_1' S_1^{-1} K_1 + S*^{-1})^{-1} \tag{2.15}$$

which agrees with (2.5) with $K_2 = I$. The matrix $S*$ given by (2.10) is in some sense doing double duty - containing information concerning the forecast error covariance as well as prior information concerning the true but unknown state vector x.

Thus the general structure

$$y = Kx + \epsilon, \quad \epsilon \sim \mathcal{N}(0, S) \tag{2.16}$$

$$x \sim \mathcal{N}(0, \Sigma) \tag{2.17}$$

can be made to cover both the case where the state vector x is estimated, and, by replacing x and Σ by δ and $S*$, the case where the analysis increment δ is estimated.

3 Estimation of smoothing, tuning, and weighting parameters

In this Section we first describe the well known maximum likelihood (ML), Generalized Maximum likelihood (GML), Generalized Cross Validation (GCV) and Unbiased Risk (UBR) estimators in the context of the variational problems discussed previously. We then describe two extensions of GCV and UBR, namely, partial GCV (PGCV) and partial UBR (PUBR), which will be particularly useful in four dimensional variational methods.

3.1 Maximum likelihood (ML) and generalized maximum likelihood (GML) estimation

With the general model

$$y = Kx + \epsilon, \quad \epsilon \sim \mathcal{N}(0, S), \quad x \sim \mathcal{N}(0, \Sigma) \tag{3.1}$$

we have that the conditional expectation of x given y is the minimizer of

$$(y - Kx)'S^{-1}(y - Kx) + x'\Sigma^{-1}x. \tag{3.2}$$

The inverse (unconditional) covariance matrix of y is

$$\begin{aligned}(K\Sigma K' + S)^{-1} &= S^{-1/2}(I - S^{-1/2}K(K'S^{-1}K + \Sigma^{-1})^{-1}K'S^{-1/2})S^{-1/2} & (3.3)\\ &= S^{-1/2}(I - A)S^{-1/2}, & (3.4)\end{aligned}$$

say, where

$$A = S^{-1/2}K(K'S^{-1}K + \Sigma^{-1})^{-1}K'S^{-1/2}. \tag{3.5}$$

Note that A is a 'smoother matrix', that is, it is a symmetric, non-negative definite matrix with all its eigenvalues in the interval [0,1]. We suppose that Σ, K and possibly S depend on some unknown parameter vector θ, that is, $\Sigma = \Sigma(\theta), K = K(\theta), S = S(\theta)$. Then the maximum likelihood (ML) estimate of these parameters is obtained by maximizing with respect to θ, the likelihood $(2\pi)^{n/2}f_\theta(y)$ given by

$$f_\theta(y) = \frac{1}{detS^{1/2}det(I - A)^{-1/2}}e^{-\frac{1}{2}y'S^{-1/2}(I-A)S^{-1/2}y}, \tag{3.6}$$

or, equivalently, the log likelihood $logf_\theta(y)$. ML in the context of Kalman filtering has recently been studied by Dee (1993, 1994).

To emphasize the role of certain key smoothing or penalty parameters, we next reparameterize the problem as: $\epsilon \sim \mathcal{N}(0, \sigma^2 S), x \sim \mathcal{N}(0, b\Sigma)$, where S and Σ are either known or depend on some other unknowns (θ) satisfying identifiability conditions, and σ^2 and b are unknown. Letting $\lambda = \sigma^2/b$, then the estimate $\hat{x}_\lambda$ of x is given by the minimizer of

$$(y - Kx)'S^{-1}(y - Kx) + \lambda x'\Sigma^{-1}x, \tag{3.7}$$

here λ is called the smoothing or regularization parameter. If we replace S by $\sigma^2 S$ and Σ by $b\Sigma$ in (3.6), let $\lambda = \sigma^2/b$ and maximize (3.6) with respect to σ^2 we get that the maximum likelihood estimate σ^2_{ML} for σ^2 is

$$\sigma^2_{ML} = \frac{1}{n}y'S^{-1/2}(I - A(\lambda))S^{-1/2}y, \tag{3.8}$$

where $A(\lambda)$ is given by

$$A(\lambda) = S^{-1/2}K(K'S^{-1}K + \lambda\Sigma^{-1})^{-1}K'S^{-1/2}. \tag{3.9}$$

By substituting (3.8) into (3.6) we then obtain that the maximum likelihood estimate of λ and any other parameters θ in $\Sigma = \Sigma(\theta), S = S(\theta)$, and $K = K(\theta)$ are obtained by minimizing the GML (Generalized Maximum Likelihood) function $M(\lambda, \theta)$

$$M(\lambda, \theta) = \frac{y'S^{-1/2}(I - A(\lambda, \theta))S^{-1/2}y}{[detS^{-1}]^{1/n}[det(I - A(\lambda, \theta))]^{1/n}}. \tag{3.10}$$

We are writing $A(\lambda, \theta)$ to indicate the possible presence of other parameters in A. See Anderssen and Bloomfield(1974), Wahba(1975,1985a), Wahba, Johnson and Reames(1991). Note that the form of the likelihood function to be minimized depends on the set of unknown parameters which are included in the likelihood function. If, for example, σ^2 is known *a priori*, then the equation to be minimized for λ is different from the equation to be minimized if it is not, compare Dee(1993). For further examples of this little known but possibly important phenomenon, see O'Hagan(1976).

It will in many applications be desirable to allow multiple penalty functionals with separate smoothing parameters. This can be done by replacing $(b\Sigma)^{-1}$ in the model by $\sum_\alpha b_\alpha^{-1} J_\alpha$ (where the J_α are, of course sufficiently different), then replacing σ^2/b_α by λ_α. With some abuse of notation, letting $\lambda = (\lambda_1, ..., \lambda_p)$, then the estimate $\hat{x}_\lambda$ is the minimizer of

$$(y - Kx)'S^{-1}(y - Kx) + \sum_\alpha \lambda_\alpha x' J_\alpha x. \tag{3.11}$$

The GML estimates of the components of λ and any other parameters θ are obtained by maximizing $M(\lambda, \theta)$ with $\lambda\Sigma^{-1}$ replaced by $\sum_\alpha \lambda_\alpha J_\alpha$ in $A(\lambda, \theta)$ of (3.9).

It is clear that there is a limit on the number and nature of the parameters θ that can be estimated with one or a limited number of cycles of data. The most important parameter(s) for the variational problem are usually the λ's, which play the role of signal to noise ratios. If confidence intervals on x are to be obtained, then σ^2 is also important, and usually, these parameters are the easiest to estimate. In Wahba and Johnson, in preparation, we discuss in some detail identifiability and error properties of GML estimates of $\sigma^2, \lambda, \theta$. It is clear, however, that it is desirable that the Hessian of M with respect to λ, θ be well conditioned.

If the statistical model is correct, then the maximum likelihood estimates are generally the best, from the point of view of minimizing the covariances of the errors in the estimates of the unknown $\sigma^2, \lambda, \theta$, and hence, usually, the resulting estimate of x. The drawback of maximum likelihood estimates in NWP and elsewhere occurs if the model is misspecified, that is, there are (sufficiently bad) errors in the specification of the fixed parts of K, S, and Σ. Then, depending on the nature and severity of the misspecification, the maximum likelihood estimates can lead to poor estimates of x, at least relative to performance oriented or predictive mean square error criteria like GCV and UBR, and poor estimates of accuracy. (See Wahba (1985a), Stein(1990)).

3.2 Generalized cross validation (GCV) estimation

GCV estimates need (some) data with at least roughly uncorrelated error structure, or data that can be transformed so that it is at least roughly uncorrelated, to 'cross-validate' against. See Wahba(1990), Section 4.9.

In this section we will assume that the observational error covariance matrix is known up to a constant, that is

$$y = Kx + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2 S) \tag{3.12}$$

where S is known and σ^2 is unknown. In the first part of this discussion we assume K is completely known. The estimate of x is taken as the minimizer of

$$(y - Kx)'S^{-1}(y - Kx) + \lambda x' \Sigma^{-1} x. \tag{3.13}$$

However, the origin of the quadratic form $x'\Sigma^{-1}x$ is irrelevant to the use of GCV, provided that it is penalizing something that is *a priori* known to be relatively small, for example, unbalanced mass or the energy in gravity waves. As before, we may let $\lambda = (\lambda_1, .., \lambda_p)$ and replace $\lambda x' \Sigma^{-1} x$ by $\sum_\alpha \lambda_\alpha x' J_\alpha x$.

The GCV estimate $\hat{\lambda}$ of λ and other (identifiable) parameters in Σ is the minimizer of $V(\lambda)$ (or $V(\lambda,\theta)$) given by

$$V(\lambda) = \frac{y'S^{-1/2}(I-A(\lambda))^2S^{-1/2}y}{(tr(I-A(\lambda)))^2} \tag{3.14}$$

$$= \frac{\|(I-A(\lambda))S^{-1/2}y\|^2}{(tr(I-A(\lambda)))^2}, \tag{3.15}$$

where "*tr*" is trace, and A is given by (3.9). Setting $S = I$ wherever it occurs results in the familiar form of V in the literature, where it is assumed that the data vector y has already been rescaled by $S^{-1/2}$ to provide the required 'white noise'.

We note the important fact here, to be used later, that A is the influence matrix for this problem. The influence matrix $A(\lambda)$ (or $A(\lambda,\theta)$) is the matrix which relates the (rescaled) data vector $S^{-1/2}y$ to the fitted data vector $S^{-1/2}\hat{y}$. That is, (suppressing θ), if $\hat{y} = K\hat{x}_\lambda$, then

$$(S^{-1/2}\hat{y}) \equiv S^{-1/2}K\hat{x}_\lambda \equiv A(\lambda)(S^{-1/2}y). \tag{3.16}$$

The GCV estimate of λ,θ is based on a predictive mean square error criteria which attempts to obtain a 'best' estimate of x within the family of possible minimizers of (3.13), parameterized by λ,θ. It will do this under rather general conditions on Σ. It differs from the maximum likelihood estimate, which will provide the optimum estimate of λ with respect to many reasonable criteria provided that the (sampled) true solution behaves like a sample data vector with covariance a multiple of Σ, but which may provide a relatively poor estimate of λ if this assumption is not reasonably correct.

The 'predictive mean square error' is with respect to data with 'white noise' errors, or data normalized by a covariance matrix so that the errors are at least roughly 'white'. To be specific, let the predictive mean square error $R(\lambda)$ be defined by

$$R(\lambda) = \frac{1}{n}\|S^{-1/2}(Kx_{true} - K\hat{x}_\lambda)\|^2. \tag{3.17}$$

Here x_{true} is the true but unknown state vector and we consider it fixed (not a random vector). Then the GCV estimate $\hat{\lambda}$ of λ is a good estimate of the λ *which minimizes* $R(\lambda)$, under fairly general conditions on x_{true}, irrespective of whether x_{true} is considered to be a fixed vector satisfying certain conditions, or whether it is considered as a random vector with covariance a multiple of Σ, see Craven and Wahba(1979), Wahba and Wendelberger(1980), Speckman(1985), Li(1986). Generally $\hat{\lambda}$ is also a good estimate of the minimizer of $D(\lambda) = \|\hat{x}_\lambda - x_{true}\|^2$ under some fairly but not completely general conditions, see Wahba and Wang(1990). A cross-validation based estimate for σ^2 that has been shown to work well in examples is

$$\sigma^2_{GCV} = \frac{RSS(\hat{\lambda})}{tr(I-A(\hat{\lambda}))}, \tag{3.18}$$

here $RSS(\hat{\lambda}) = \|(I-A(\hat{\lambda}))S^{-1/2}y\|^2$ is the (scaled) residual sum of squares when $\hat{\lambda}$ is used, and $\hat{\lambda}$ is the GCV estimate of λ, that is, the minimizer of $V(\lambda)$. Viewing $trA(\hat{\lambda})$ as the degrees of freedom for signal, this estimate is the analogue of the usual estimate of the variance after linear regression. See Wahba(1983). Note that, due to the presence of S in the theoretical loss function R of (3.17) GCV in this form is not in general appropriate to estimate unknown parameters in S . Partial GCV (see below) may be used when a sufficiently large submatrix of S is known, and other GCV related methods are discussed in Gao(1992,1993) and Gao, Wahba and Johnson, in preparation.

Various parameters θ in $\Sigma = \Sigma(\theta)$ are known to be amenable to estimation by GCV, by replacing Σ by $\Sigma(\theta)$ in (3.13), and minimizing V of (3.14) with respect to both λ and θ. See for example, Wahba and Wendelberger(1980), Hutchinson, Kalma and Johnson(1984), and Chapter 3 of Wahba(1990).

Certain parameters θ in $K = K(\theta)$ may also be estimated this way, see Wahba(1990a) and Wahba, Gong and Johnson (in preparation), by setting $K = K(\theta)$ in the ingredients of V. Then the minimizer of V should be a good estimate of the minimizer of R given by

$$R(\lambda, \theta) = \frac{1}{n}\|S^{-1/2}(K(\theta_{true})x_{true} - K(\theta)\hat{x}_{\lambda,\theta})\|^2, \tag{3.19}$$

where θ_{true} contains the true (but unknown) components of θ in K and x_{true} is the 'true' (but unknown) state vector.

The minimization of $V(\lambda_1, ..., \lambda_p)$ as well as of $M(\lambda_1, ..., \lambda_p)$ can be carried out for medium sized data sets for both GML and GCV via the algorithm of Gu and Wahba(1991), which uses truncated matrix decomposition methods. The code RKPACK which implements this algorithm is available over the internet through netlib. Write netlib@ornl.gov in the body and the robot mailserver will respond with instructions for using the system. Matrix decomposition methods are not at present suitable for general systems of the size that occur in global scale numerical weather prediction. In Section 4 we will provide numerical methods for computing V that can be implemented in conjunction with iterative methods for minimizing (1.1) with extremely large state vectors.

3.3 Unbiased risk (UBR) estimation

Continuing with the model (3.12), suppose σ^2 is known, i. e. the covariance matrix of the observational errors is completely known. Then smoothing parameters (and certain other parameters) in the penalty functional can be estimated by an unbiased risk estimate. The unbiased risk estimate is discussed and demonstrated in the context of the smoothing spline in Craven and Wahba(1979), based on Mallows(1973) C_L. More complete theoretical properties of the unbiased risk estimate are given in Li(1986). Let

$$U(\lambda) = \|(I - A(\lambda))S^{-1/2}y\|^2 + 2\sigma^2 trA(\lambda), \tag{3.20}$$

where A is as in (3.9) and possible dependence on θ in $K = K(\theta)$ and $\Sigma = \Sigma(\theta)$ has been suppressed. The unbiased risk estimate of λ (including $\lambda = (\lambda_1, ..., \lambda_p)$) is the minimizer of $U(\lambda)$. For *any* fixed unknown x_{true}, $\frac{1}{n}U(\lambda)$ plus a constant not depending on λ is an unbiased estimate of $ER(\lambda) = E\frac{1}{n}\|S^{-1/2}(Kx_{true} - K\hat{x}_\lambda)\|^2$ of (3.17), or, more generally of $R(\lambda, \theta)$ of (3.19). Thus the UBR estimate of λ, θ can be expected to choose a value of λ, θ which minimizes the predictive mean square error $R(\lambda, \theta)$.

More generally, let W be a given weight matrix and let $U_W(\lambda, \theta)$ be defined by

$$U_W(\lambda, \theta) = \|W(I - A(\lambda, \theta))S^{-1/2}y\|^2 + 2\sigma^2 trWA(\lambda, \theta)W'. \tag{3.21}$$

Then $U_W(\lambda, \theta) + \sigma^2 trWW'$ is an unbiased estimate of $ER_W(\lambda, \theta)$ where $R_W(\lambda, \theta)$ is defined by

$$R_W(\lambda, \theta) = \|WS^{-1/2}(K(\theta_{true})x_{true} - K(\theta)\hat{x}_{\lambda,\theta})\|^2, \tag{3.22}$$

that is,

$$EU_W(\lambda, \theta) + 2\sigma^2 trWW' = ER_W(\lambda, \theta), \tag{3.23}$$

where the expectation (E) is taken over the random vector ϵ. Thus we have the weighted unbiased risk estimate of λ, θ as the minimizer of $U_W(\lambda, \theta)$ of (3.21). The proof of (3.23) goes as follows: Let $g = S^{-1/2}K(\theta_{true})x_{true}$, then

$$\begin{aligned} S^{-1/2}y &= g + S^{-1/2}\epsilon && (3.24)\\ S^{-1/2}\hat{y}_{\lambda,\theta} &= S^{-1/2}K(\theta)\hat{x}_{\lambda,\theta} \equiv A(\lambda,\theta)(g + S^{-1/2}\epsilon), && (3.25)\end{aligned}$$

say. Then

$$\begin{aligned} ER_W(\lambda) &= E\|WS^{-1/2}(K(\theta_{true})x_{true} - K(\theta)\hat{x}_{\lambda,\theta})\|^2 && (3.26)\\ &= E\|W(g - A(\lambda,\theta)(g + S^{-1/2}\epsilon))\|^2 && (3.27)\\ &= \|W(I - A(\lambda,\theta))g\|^2 + \sigma^2 trWA(\lambda)^2W' && (3.28)\end{aligned}$$

whereas

$$\begin{aligned} EU_W(\lambda,\theta) &= \|W(I - A(\lambda,\theta))g\|^2 + \sigma^2 trW(I - A(\lambda,\theta))^2W' + 2\sigma^2 trWA(\lambda,\theta)W'\\ &= ER_W(\lambda) + \sigma^2 trWW'. && (3.29)\end{aligned}$$

If one wished to choose λ, θ to minimize $E\|(x_{true} - \hat{x}_{\lambda,\theta})\|^2$ it is tempting to try to set W so that $K'(\theta_{true})S^{-1/2}W'WS^{-1/2}K(\theta_{true})$ (say) is close to I. Care must be taken that this does not result in a numerically unstable problem, and/or a U_W with an unacceptably large variance.

3.4 Partial unbiased risk (PUBR) and partial generalized cross validation (PGCV) estimates

A special case of the weighted UBR estimate has the potential for obtaining (performance-based) estimates of some parameters involved in the error structure of some of the data. To do this, it is necessary that a sufficiently large subvector y_1 of y has completely known error structure, independent of the other data errors. Suppose $y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$, where

$$\begin{aligned} y_1 = &\ K_1x + \epsilon_1, \quad \epsilon_1 \sim \mathcal{N}(0, \sigma^2 S_1)\\ y_2 = &\ K_2x + \epsilon_2, \quad \epsilon_2 \sim \mathcal{N}(0, \sigma^2 S_2(\theta)).\end{aligned}$$

Then $K = \begin{pmatrix} K_1 \\ K_2 \end{pmatrix}$ and $S = \begin{pmatrix} S_1 & 0 \\ 0 & S_2(\theta) \end{pmatrix}$. Let $\hat{x}_{\lambda,\theta}$ be the minimizer of

$$(y - Kx)'S^{-1}(y - Kx) + \lambda x'\Sigma^{-1}x \qquad (3.30)$$

as before, except that now $S = S(\theta)$. Let $W = \begin{pmatrix} W_1 & 0 \\ 0 & 0 \end{pmatrix}$ where the dimensions of W_1 conform to y_1, then the predictive mean square error $R_W(\lambda, \theta)$ of (3.22) becomes

$$R_{W_1}(\lambda,\theta) = \|W_1 S_1^{-1/2}(K_1 x_{true} - K_1\hat{x}_{\lambda,\theta})\|^2, \qquad (3.31)$$

or, more generally, if $K = K(\theta)$

$$R_{W_1}(\lambda,\theta) = \|W_1 S_1^{-1/2}(K_1(\theta_{true})x_{true} - K_1(\theta)\hat{x}_{\lambda,\theta})\|^2. \qquad (3.32)$$

Let $A_{11}(\lambda, \theta)$ be the upper left block of $A(\lambda, \theta)$, with dimensions conforming to y_1. Following the same argument as before

$$U_{W_1}(\lambda, \theta) = \|W_1(y_1 - K_1(\theta)\hat{x}_{\lambda,\theta})\|^2 + 2\sigma^2 trW_1A_{11}(\lambda, \theta)W_1' \tag{3.33}$$

is an unbiased estimate of $ER_{W_1}(\lambda, \theta) + \sigma^2 trW_1'W_1$ for each λ, θ, and its minimizer may be used to obtain values of λ and θ for those parameters which are sufficiently estimable. We will see how the PUBR estimate may be used in the context of four dimensional variational analysis in Section 5.

A partial GCV estimate may be used when σ^2 is unknown, and is defined as the minimizer of V_{W_1} given by

$$V_{W_1}(\lambda, \theta) = \frac{\|W_1S_1^{-1/2}(y_1 - K_1(\theta)\hat{x}_{\lambda,\theta})\|^2}{(tr(W_1W_1' - W_1A_{11}(\lambda, \theta)W_1'))^2}. \tag{3.34}$$

It can be shown, using the same theoretical method in Craven and Wahba(1979), that under certain circumstances the minimizer of V_{W_1} is also an estimate of the minimizer of R_{W_1}.

3.5 GCV and UBR for nonlinear problems

So far we have been assuming that y is related to x linearly, but in general we have $y = K(\theta_{true}, x) + \epsilon$, where the dependence on x is nonlinear. We will assume that K is reasonably linearizable, and, to simplify the notation, will suppress any dependence on any components of an unknown θ_{true} in K that is to be estimated. The (possibly weighted) residual sum of squares term $\|W(I - A(\lambda, \theta))S^{-1/2}y\|^2$ in the GCV and UBR functions simply becomes $\|WS^{-1/2}(y - \hat{y})\|^2 = \|WS^{-1/2}(y - K(\hat{x}_{\lambda,\theta}))\|^2$, compare (3.16). The trace of the influence matrix $A(\lambda, \theta)$ now becomes the trace of the linearized version of the influence operator, $A_y(\lambda, \theta)$, linearized in the neighborhood of the minimizer. That is, let $\hat{x}_{\lambda,\theta}(y)$ be the estimate of x for a given λ, θ and y, and let $\hat{y} = K(\hat{x}_{\lambda,\theta}(y)) \approx A_y(\lambda, \theta)(y)$ plus quantities independent of y. See O'Sullivan and Wahba(1985) and Wahba(1990), Section 8.3. Then $A_y(\lambda, \theta)$ replaces A in Equations (3.14, 3.20, 3.33, 3.34). Practical computation of the traces involving $A_y(\lambda, \theta)$ will be described in Sections 4 and 5.

4 Fast randomized computation of the GCV and UBR functions and certain GML functions, for extremely large data sets

4.1 Exact and randomized GCV

To calculate the GCV and UBR estimates requires the calculation of $trA(\lambda, \theta)$ or $trWA(\lambda, \theta)W'$. For square matrices of dimension less than, say, 1,000, these traces can be obtained on large workstations with algorithms using matrix decompositions, in particular the QR decomposition and the singular value decomposition. See Wahba(1990), and the documentation for GCVPACK (Bates, Lindstrom, Wahba and Yandell(1986)), and RKPACK (Gu (1989)). However, matrix decomposition methods are out of the question for matrices of the size occurring in global scale NWP. Fast methods involving special structure of the matrices involved are available, however the matrices involved in NWP models generally do not have the special structure required. In this section we describe a method which may be used to calculate GCV, UBR and certain GML estimates, in the context of operational global scale NWP, whenever the means is available to solve the variational problem for the state vector with a (single) random perturbation of the data, along with the original data. The idea is to estimate the required trace by Monte Carlo, or randomized, methods, and was proposed in connection with the calculation of GCV functions like those of (3.14) by

Girard(1987, 1989, 1991) and Hutchinson(1989). Girard (1991) provided a theorem that showed that the error due to the randomization part was generally negligible in the context of estimation of certain parameters by GCV. Let $B(\theta)$ be any $n \times n$ matrix depending on some parameter vector θ, with i,jth entry $b_{ij}(\theta)$, and let ξ be an n dimensional random vector with components $\{\xi_i\}$ satisfying $E\xi_i = 0, E\xi_i\xi_j = 1, i = j, = 0, i \neq j$. Then $\frac{1}{n}\xi' B(\theta)\xi = \frac{1}{n}\sum_{i,j=1}^{n} \xi_i\xi_j b_{ij}(\theta)$ and $E\frac{1}{n}\xi' B(\theta)\xi = \frac{1}{n}\sum_{i=1}^{n} b_{ii}(\theta) = \frac{1}{n} trB(\theta)$. The randomized trace estimate of $\frac{1}{n}trB(\theta)$ is then given by $\frac{1}{n}\xi' B(\theta)\xi$ where ξ comes from a random number generator. If $\xi \sim \mathcal{N}(0, I)$, then the standard deviation of this randomized trace estimate is $\sqrt{\frac{2}{n}}[\frac{1}{n}trB^2(\theta)]^{1/2}$ (see Girard 1989). If $B(\theta)$ is a smoother matrix, that is, it is symmetric, non-negative definite, with all its eigenvalues between 0 and 1, as is the case for any matrix A of the form of (3.5), then $0 \leq \frac{1}{n}trB \leq 1$, and the standard deviation of $\frac{1}{n}\xi' B(\theta)\xi$ is no greater than $\sqrt{\frac{2}{n}}[\frac{1}{n}trB(\theta)]^{1/2}$.

We have run a small 'toy' problem which demonstrates that an estimate of a smoothing parameter λ in this problem, calculated via Randomized GCV gives, for all practical purposes, just as good a value for λ as one calculated more exactly using matrix decompositions. We generated data from the model

$$y_i = f(x(i)) + \epsilon_i, \quad i = 1, ..., n, \tag{4.1}$$

where $x(i) = (x_1(i), x_2(i))$ is a point in the unit square and $\epsilon = (\epsilon_1, ..., \epsilon_n)' \sim \mathcal{N}(0, \sigma^2 I)$. We took $f(x)$ as Franke's principal test function. A formula and plot of f appear in Wahba(1983), where f was used to test Bayesian 'confidence intervals' along with GCV, see Section 6 below. f is a smooth function with two round peaks and a rounded valley with minimum height near 0 and maximum height approximately 1.2. The standard deviation σ of the noise ϵ was 0.1, and $676 = 26^2$ values of $x(i) = (x_1(i), x_2(i))$ were chosen on a regular 26×26 grid on the unit square. GCVPACK was used to estimate f given y by a thin plate smoothing spline (Wahba and Wendelberger(1980)). The thin plate spline used here is the solution to the minimization problem: Find f_λ (in an appropriate function space) to minimize

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x(i)))^2 + \lambda \int_{\infty}^{\infty}\int_{\infty}^{\infty}(f_{x_1,x_1}^2 + 2f_{x_1,x_2}^2 + f_{x_2,x_2}^2)dx_1 dx_2, \tag{4.2}$$

where $x = (x_1, x_2)$. An analytical representation for f_λ and for the influence matrix $A(\lambda)$ which satisfies $(f_\lambda(x(1)), f_\lambda(x(2)), ..., f_\lambda(x(n)))' = A(\lambda)y$ is part of GCVPACK. GCVPACK uses matrix decomposition methods to compute

$$V(\lambda) = \frac{y'(I - A(\lambda))^2 y}{(tr(I - A(\lambda)))^2} \tag{4.3}$$

to a large number of significant figures. The solid line in Figure 4.1 is a plot of $V(\lambda)$ as a function of λ as computed by GCVPACK. The dotted line is $R(\lambda)$, the predictive mean square error (PMSE) function, defined by

$$R(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(f(x(i)) - f_\lambda(x(i)))^2. \tag{4.4}$$

$R(\lambda)$ can only be plotted in a synthetic experiment when the 'truth' is known, and is used to check the performance of the GCV in synthetic experiments. Note that the minimizer of $V(\lambda)$, indicated by a diamond in the figure, is very close to the minimizer of $R(\lambda)$, indicated by the circle. This is as predicted by the theory. (See Wahba(1990) or Wahba and Wendelberger(1980) for further discussion and references). The dashed line in Figure 4.1 is a plot of $RanV(\lambda)$, given by

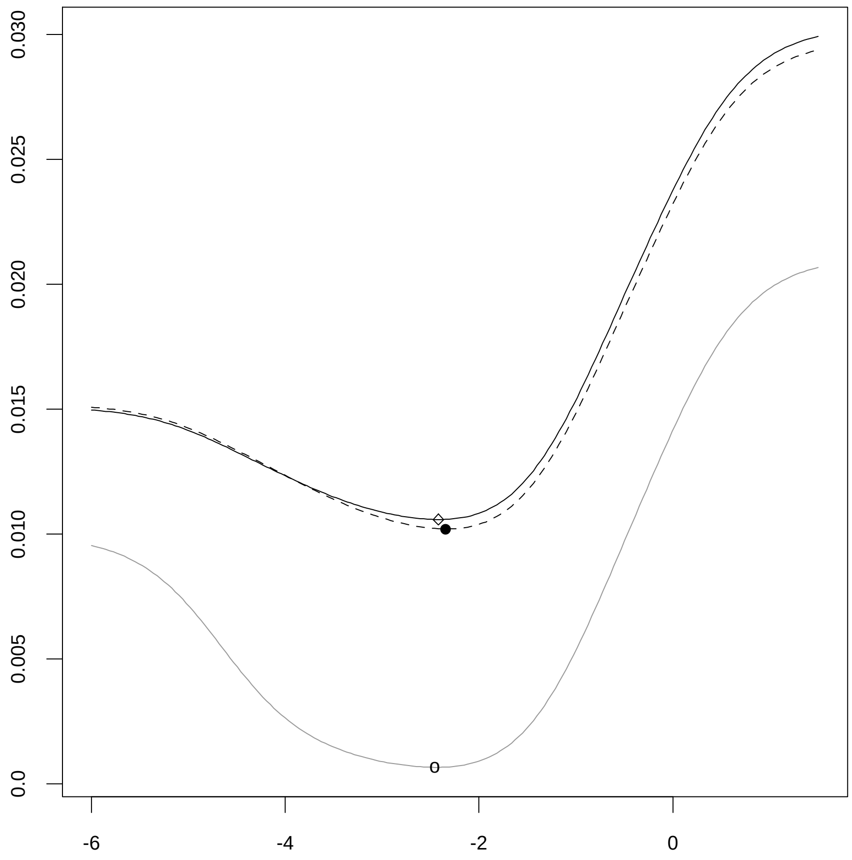

Figure 4.1: The exact GCV function and one replicate of the randomized GCV function with the PMSE function.

$$RanV(\lambda) = \frac{y'(I - A(\lambda))^2 y}{(\xi'(I - A(\lambda))\xi)^2}, \tag{4.5}$$

where ξ came from a random number generator $\xi \sim \mathcal{N}(0, I)$, and $tr(I - A(\lambda))$ in $V(\lambda)$ has been replaced by a randomized estimate of it. It is important to note that the *same* ξ is used for all values of λ. Figure 4.2 contains the exact GCV function and PMSE function of Figure 4.1 on an expanded scale as well as ten replicates of $RanV$. Each replicate represents a different ξ obtained from a random number generator. It can be seen that the minimizers of all ten replicates do an excellent job of estimating the minimizer of $R(\lambda)$, even though the heights of the replicates vary.

4.2 Randomized GCV and UBR in the iterative solution of extremely large variational problems

In modern operational global scale numerical weather prediction models, some iterative method, with k iterations, is used to obtain an approximate solution $\hat{x}^k$ to the minimizer of

$$(y - K(x))' S^{-1} (y - K(x)) + x' \Sigma^{-1} x, \tag{4.6}$$

given K, S^{-1}, and Σ^{-1}. See for example Parrish and Derber(1992), who use a conjugate gradient method, and Lorenc, Bell and McPherson(1991), Lorenc(1992) who discuss a successive correction method. The L-BFGS algorithm (Liu and Nocedal 1989) is also popular. The particular method used is not important in what follows. We only assume that there is an operational code, which we shall call the 'black box', which, given K, S^{-1}, Σ^{-1} and y, returns (after k iterations), $\hat{x}^k = \hat{x}^k(y)$, which is an approximation to the minimizer $\hat{x}$ of (4.6); and this 'black box' also returns $K(\hat{x}^k)$. In what follows, we assume that S is known, and incorporated into $\tilde{y} = S^{-1/2} y$ and $\tilde{K} = S^{-1/2} K$ and we drop the ˜ on y and K in the rest of this Section. The aforementioned 'black box' may now be used to obtain a randomized estimate of the trace of the matrix which plays the role of $A(\theta)$, where now θ represents the unknown parameters (including λ). Even if K is linear in x, the relationship between y and $K(\hat{x}^k(y))$ is not necessarily linear in y if the iteration is stopped before the exact minimizer $\hat{x}$ has been found. This happens, in, for example, the conjugate gradient algorithm if $k < n$. Thus we no longer have an influence matrix $A(\theta)$ which satisfies $K\hat{x}_\theta = A(\theta) y$, but an influence operator $A^k(\theta, y)$ defined by $K(\hat{x}^k_\theta(y)) = A^k(\theta, y)$. Define the matrix $A^k_y(\theta)$ by

$$K(\hat{x}^k_\theta) \approx A^k_y(\theta) y, \tag{4.7}$$

that is, $A^k_y(\theta)$ is the linearized version of the influence operator implicitly defined by the 'black box', evaluated at y. If the 'black box' is always used to find the approximate minimizer of (4.6), then it is actually $tr\ A^k_y(\theta)$ and not $tr\ A(\theta)$ that should be used in computing the UBR function $U(\theta)$ of (3.20) and the GCV function $V(\theta)$ of (3.14), and similarly for their partial versions.

In any case, let ξ come from a random number generator with $\xi \sim \mathcal{N}(0, \sigma_\xi^2 I)$ and let $\hat{x}^k_\theta(y + \xi)$ be the estimate for x obtained by substituting $(y + \xi)$ for y in (4.6), and using the 'black box' to obtain the approximate minimizer. Then

$$K(\hat{x}^k_\theta(y + \xi)) \approx A^k_{y+\xi}(\theta)(y + \xi). \tag{4.8}$$

A randomized estimate of $tr\ A^k_y(\theta)$ is given by

$$\frac{1}{\sigma_\xi^2}[\xi' A^k_{y+\xi}(\theta)(y + \xi) - \xi' A^k_y(\theta) y] \tag{4.9}$$

$$\approx \frac{1}{\sigma_\xi^2}\xi'[K(\hat{x}^k_\theta(y + \xi)) - K(\hat{x}^k_\theta(y))], \tag{4.10}$$

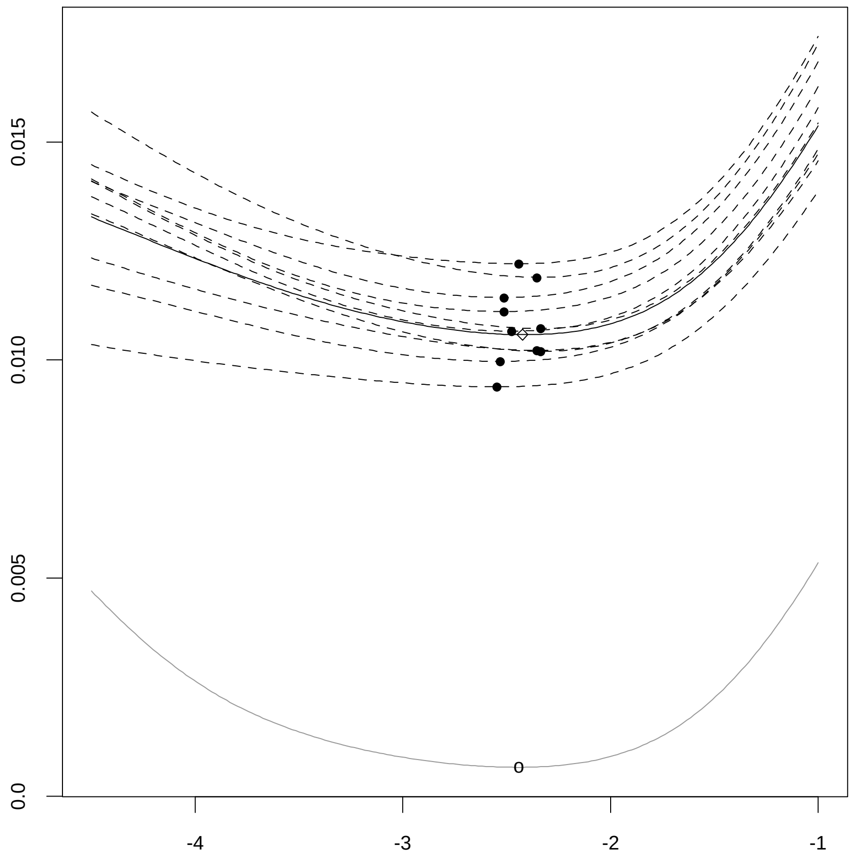

Figure 4.2: Ten replicates of the randomized GCV function along with the exact GCV function and the PMSE function.

where we have written $\hat{x}_\theta^k(y+\xi)$ for $\hat{x}_\theta^k$ based on the data $y+\xi$. A randomized estimate for $trWA_y^k(\theta)W'$ is obtained by replacing ξ by $W'\xi$ here.

In the case of linear iterative methods, such as the Richardson/Landweber/Fridman/Pickard/Cimino (RLFPC) iteration (with K linear), simple exact formulas can be obtained for $A_y^k(\theta)$, not depending on y, see Wahba(1987). It is shown there, and in Fleming(1990) that truncated iteration with such a linear method, is a form of regularization, or low-pass filtering. Roughly speaking, a truncated iteration tends to project the exact solution towards eigenvectors corresponding to large eigenvalues, and these eigenvectors tend to be 'smooth'. Thus both k, the number of iterations, and λ, a multiplier on Σ^{-1} can be thought of as 'smoothing' or regularization parameters. It was suggested that GCV could be used to choose both k and λ simultaneously in Wahba(1987), however no procedure for carrying out the calculations with large data sets was provided there.

We have constructed a semi-realistic 'toy' problem to test and demonstrate the feasibility and efficiency of choosing both k and λ via GCV or UBR, in conjunction with the randomized trace estimation of (4.10). Rather than use an RLFPC algorithm for this demonstration, we have chosen to use a preconditioned conjugate gradient algorithm, since conjugate gradients are used operationally, and are well known to have favorable properties. The experimental setup we use here is a part of the experimental setup in Gao(1993) and Gao, Wahba and Johnson, in preparation. ECMWF Gridded Level IIIB FGGE data for the 500mb height for January 2, 1979, was used to obtain a spherical harmonic representation for this 500mb height field of the form

$$f(P) = \sum_{\ell=0}^{30} \sum_{s=-\ell}^{\ell} x_{\ell s} Y_{\ell s}(P), \tag{4.11}$$

where P is a point on the sphere, and the $Y_{\ell s}$ are spherical harmonics. This representation was obtained by solving a variational problem given the gridded data. The amount of smoothing was chosen to make the resulting contour plots match the ECMWF plots visually. See Gao(1993) for details. Simulated observational data at $n = 600$ North American radiosonde stations was generated by

$$y_i = f(P_i) + \epsilon_i \tag{4.12}$$

where $\epsilon = (\epsilon_1, ..., \epsilon_{600})' \sim \mathcal{N}(0, \sigma^2 I)$, and the P_i are station locations. We chose $\sigma = 9m.$ to represent observational error. A spline on the sphere analysis analogous to Wahba and Wendelberger(1980) is obtained by letting $\hat{x}_\lambda = (\hat{x}_{00,\lambda}, \hat{x}_{10,\lambda}, ...)$ be the minimizer of

$$\sum_{i=1}^{n}(y_i - \sum_{\ell=0}^{30}\sum_{s=-\ell}^{\ell} x_{\ell s} Y_{\ell s}(P_i))^2 + \lambda \sum_{\ell=0}^{30}\sum_{s=-\ell}^{\ell} [(\ell)(\ell+1)]^2 x_{\ell s}^2. \tag{4.13}$$

The penalty functional $J(f) = \sum_{\ell s}[(\ell)(\ell+1)]^2 x_{\ell s}^2$ is a multiple of $J(f) = \int_{\mathcal{S}}(\triangle f)^2$ where $\triangle$ is the Laplacian on the sphere (see Wahba(1981,1982a)). Letting K be the 600×960 matrix with entries $Y_{\ell s}(P_i)$ and D be the diagonal matrix with $\ell s, \ell s$ entries $[\ell(\ell+1)]^2$, then the minimizer $\hat{x}_\lambda$ satisfies

$$(K'K + \lambda D)\hat{x}_\lambda = K'y. \tag{4.14}$$

A preconditioned conjugate gradient algorithm with (symmetric, invertible) preconditioner C replaces $\hat{x}_\lambda$ by $C^{-1}w$ in (4.14) and solves for w in

$$C^{-1}(K'K + \lambda D)C^{-1}w = C^{-1}K'y. \tag{4.15}$$

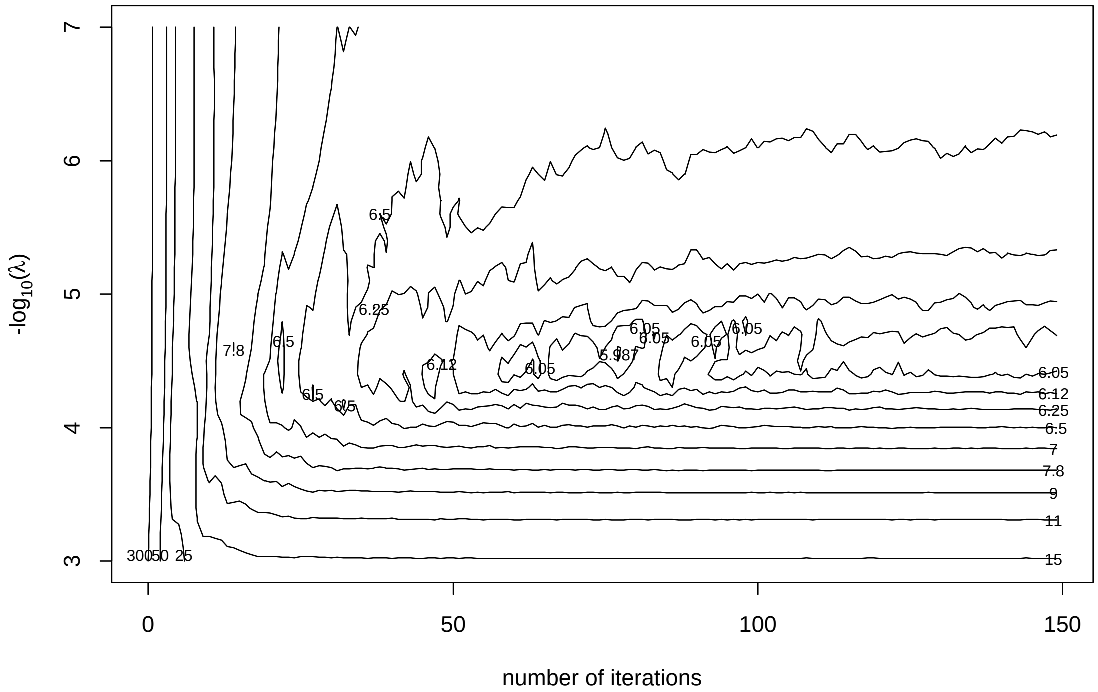

Figure 4.3: The root predictive mean square error $R^{1/2}$ as a function of $\log_{10}(\lambda)$ and k.

See Golub and van Loan (1989), Section 10.3. In the experiment below C was taken as $[diag(K'K+\lambda D)]^{1/2}$. The starting point x^0 of the iteration was taken as $x^0 = (\widetilde{K'K+\lambda D})^{-1}K'y$ where $(\widetilde{K'K+\lambda D})^{-1}$ is a diagonal approximation to $(K'K+\lambda D)^{-1}$ obtained by letting

$$\widetilde{K'K+\lambda D} = \begin{pmatrix} b_{11} & b'_{12} \\ b_{12} & diag(B_{22}+\lambda D_1) \end{pmatrix},$$

and inverting analytically.

Following (3.17) we define the predictive mean square error as

$$R(\lambda, k) = \frac{1}{n}\sum_{i=1}^{n}(f_\lambda^k(P_i) - f(P_i))^2 \tag{4.16}$$

where now

$$f_\lambda^k(P) = \sum_{\ell s} \hat{x}_{\ell s,\lambda}^k Y_{\ell s}(P), \tag{4.17}$$

$$\hat{x}_\lambda^k = \{\hat{x}_{\ell s,\lambda}^k\}, \tag{4.18}$$

and $\hat{x}_\lambda^k$ is the approximate solution after k iterations. Figure 4.3 gives a plot of $R^{1/2}(\lambda, k)$ as a function of λ and k where k is the number of iterations in the conjugate gradient iterative solution of (4.15).

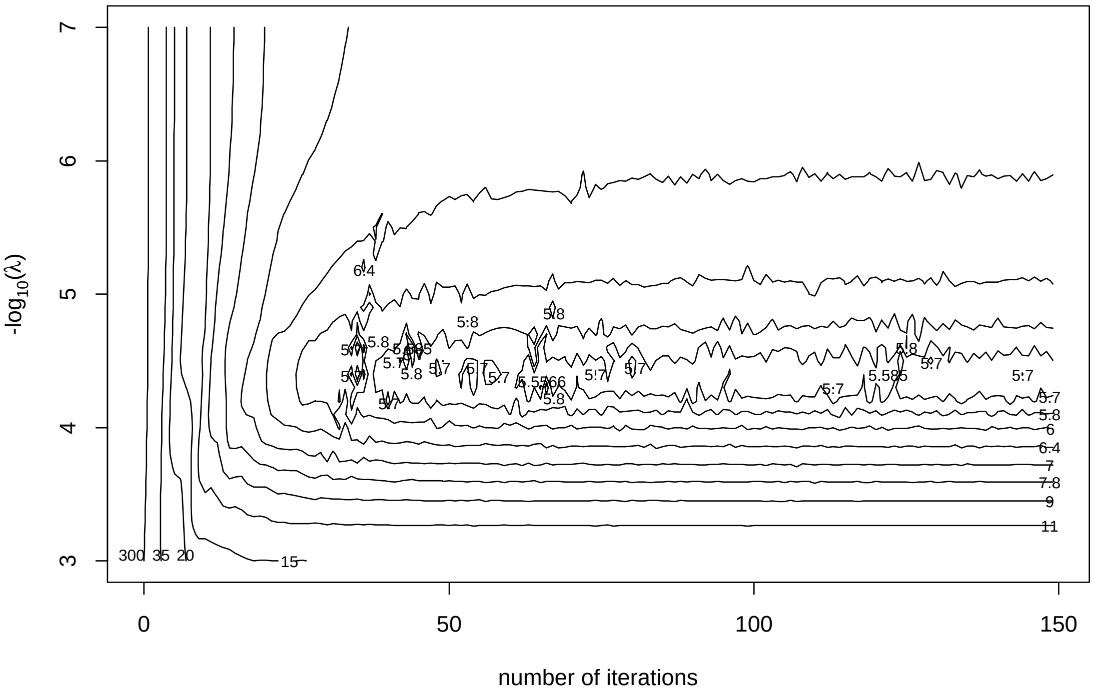

Figure 4.4: $RanU^{1/2}$ as a function of $log_{10}(\lambda)$ and k.

This kind of plot is, of course, only available in a simulation study where the 'ground truth' is known. $R(\lambda, k)$ is minimized at around $-log_{10}(\lambda) = 4.5$, and $k = 75$. Note that the value of $R^{1/2}(\lambda, k)$ at the minimum is about $6m$. Assuming that good approximation to these optimal values of λ and k can be found, then the smoothing procedure has resulted in a smoothed minus true standard deviation which is about 1/3 less than the observational standard deviation. (This would be reduced further if unbiased forecast data were also available).

Figure 4.4 gives a plot of the square root of the randomized version of the UBR function U of (3.20), as a function of k and λ. The randomized UBR function is computed as

$$RanU(\lambda, k) = \frac{1}{n}\|y - K\hat{x}_\lambda^k\|^2 + \frac{2\sigma^2}{n}\left\{\frac{1}{\sigma_\xi^2}\xi'[K\hat{x}_\lambda^k(y+\xi) - K\hat{x}_\lambda^k(y)]\right\}, \tag{4.19}$$

where ξ came from a random number generator, $\xi \sim \mathcal{N}(0, \sigma_\xi^2 I)$. The first term in (4.19) is the mean residual sum of squares and the expression in large brackets is the randomized trace estimate, compare (3.20). Here the true value $9m$ of σ is assumed known, and used in (4.19). Otherwise, $RanU$ is based on the data and the model alone. The same ξ was used for the entire plot. The standard deviation σ_ξ for the random vector ξ should be chosen carefully if $A_y^k(\theta)y$ is not linear in y. If σ_ξ is too small, then the calculation of the difference in (4.10) may be unstable, while if σ_ξ is too large the behavior at $A_y^k(\theta)$ may not be captured. After a little trial and error we found that a σ_ξ of the order of, but smaller than the presumed σ of the noise in y worked well. In these experiments we took $\sigma_\xi = 3m = \frac{1}{3}\sigma$. The striking thing to observe about $RanU^{1/2}(\lambda, k)$, is how

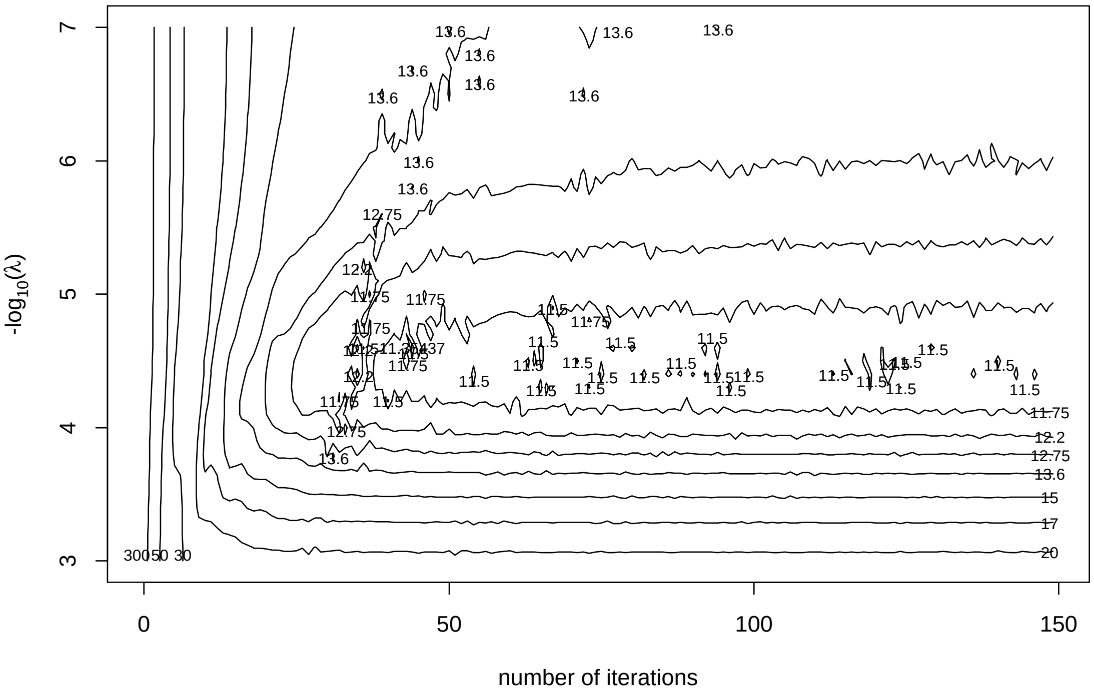

Figure 4.5: $RanV^{1/2}$ as function of $\log_{10}(\lambda)$ and k.

well it actually estimates $R^{1/2}(\lambda, k)$. The smallest value of $R^{1/2}(\lambda, k)$ is 5.987. The minimum of $RanU^{1/2}(\lambda, k)$ is located in a region for which the value of $R^{1/2}$ is less than or equal to 6.12 in the R plot, so that if a value of λ and a stopping rule k based on minimizing $RanU$ were used, then the ratio of the resulting predictive mean square error to the minimum possible predictive mean square error (known as the inefficiency), would be no larger than $6.12/5.987 = 1.022$.

Figure 4.5 gives a plot of the randomized version of the GCV function of (3.14) as a function of k and λ. The randomized GCV function is computed as

$$RanV(\lambda, k) = \frac{\frac{1}{n}\|y - K\hat{x}_\lambda^k\|^2}{\left(\frac{1}{n}\left\{\frac{1}{\sigma_\xi^2}\xi'[\xi - (K\hat{x}_\lambda^k(y+\xi) - K\hat{x}_\lambda^k(y))]\right\}\right)^2}. \tag{4.20}$$

Although this plot is a bit harder to read, it provides essentially the same information as the previous two plots. The value of $RanV^{1/2}$ at the minimum (11.354) is roughly an estimate of $\sqrt{\min_{\lambda,k} R(\lambda, k) + \sigma^2} = 10.8$, as predicted by the theory, see Wahba(1990). The minimum GCV score is located in a region for which the PMSE score $R^{1/2}$ is less than or equal to 6.25 so that the inefficiency is no bigger than $6.25/5.987 = 1.044$.

Before this experiment was conducted we had conjectured that there would be a minimum in $R(\lambda, k)$ and consequently, in $RanU(\lambda, k)$ and $RanV(\lambda, k)$ at some value of $k << n$, but here the minimum is fairly shallow, and the (λ, k) surfaces flatten out as a function of k as k gets larger. In practice, of course, one could stop as soon as the surface has flattened out. In another experiment, where the noise standard deviation σ was inadvertently set to the unrealistically large value of $120m$. there was a very distinct minimum in k, suggesting that k was an important regularization parameter. We also conjectured that the best λ might depend on k (that is, smaller λ might want a smaller k, due to a regularizing effect of the smaller number of iterations), but that is not evident in Figure 4.3, since the minimizing λ appears not to depend on k. However we do not rule out this phenomena in other experiments or in practice.

Note that randomized partial UBR and partial GCV estimates may be computed in the same way, since for any symmetric non-negative definite matrix B an unbiased estimate of $traceW'BW$ is given by $\frac{1}{\sigma_\xi^2}\xi_W' B\xi_W$ where $\xi_W = W\xi$, with $\xi \sim \mathcal{N}(0, \sigma_\xi^2 I)$.

4.3 Randomized ML and GML

Let $G(\theta) = S^{-1/2}(I - A)S^{-1/2}$, with A given by (3.5) and where at first we are not indicating how G depends on θ. To estimate θ by ML or GML it is desired to minimize, respectively,

$$-2log\ f_\theta = y'G(\theta)y - log\ detG(\theta) \tag{4.21}$$

$$-logM(\theta) = log\ y'G(\theta)y - \frac{1}{n}log\ detG(\theta). \tag{4.22}$$

We can obtain a minimum using GML by solving

$$-\frac{d}{d\theta}\log M(\theta) = \frac{y'\frac{d}{d\theta}G(\theta)y}{y'G(\theta)y} - \frac{1}{n}\frac{d}{d\lambda}\log\det G(\theta) = 0 \tag{4.23}$$

and similarly for ML. It can be shown that $\frac{d}{d\theta}log\ detG(\theta) = trG(\theta)^{-1}\frac{d}{d\theta}G(\theta)$, see for example Dee(1993). Dee shows that the relevant calculations can be done using matrix decompositions in $O(n^3)$ flops. In the special case that $\theta = \lambda$ and $A(\lambda)$ is of the form (3.9) we can write $A(\lambda) = \Gamma D(\lambda)\Gamma'$ where Γ is an orthogonal matrix independent of λ, and $D(\lambda)$ a diagonal matrix with diagonal entries of the form $\frac{\lambda_\nu}{\lambda_\nu+\lambda}$ if Σ^{-1} is of full rank, or $\frac{\lambda_\nu}{\lambda_\nu+\lambda}$ or 1 if Σ^{-1} is not of full rank. Since

$$\frac{d}{d\lambda}(I - D(\lambda)) = \frac{1}{\lambda}D(\lambda)(I - D(\lambda)) \tag{4.24}$$

$$\frac{d}{d\lambda}\log\det(I - D(\lambda)) = \frac{1}{\lambda}trD(\lambda) \tag{4.25}$$

we have

$$\lambda\frac{d}{d\lambda}\log M(\lambda) = \frac{y'S^{-1/2}A(\lambda)(I - A(\lambda))S^{-1/2}y}{y'S^{-1/2}(I - A(\lambda))S^{-1/2}y} - \frac{1}{n}trA(\lambda). \tag{4.26}$$

We assume we have a 'black box' which finds $\hat{x}_\lambda$ to approximately minimize $(y-Kx)'S^{-1}(y-Kx) + \lambda x'\Sigma^{-1}x$. We have suppressed any indication of the number of iterations k, since k apparently does not admit of estimation by GML. Then we have a randomized version of (4.26), namely

$$Ran\lambda\frac{d}{d\lambda}\log M(\lambda) = \frac{(K\hat{x}_\lambda)'S^{-1}(y - K\hat{x}_\lambda)}{y'S^{-1}(y - K\hat{x}_\lambda)} - \frac{1}{n}\left\{\frac{1}{\sigma_\xi^2}\xi'[S^{-1/2}(K\hat{x}_\lambda(y + S^{1/2}\xi) - K\hat{x}_\lambda^k(y))]\right\}, \tag{4.27}$$

where $\hat{x}_\lambda(y + S^{1/2}\xi)$ is the estimate of x based on the data vector y perturbed by the random vector $S^{1/2}\xi$. The analogous calculation can be made for ML.

5 Four dimensional variational data assimilation

In this Section we will describe how ML, GML, GCV and UBR may be applied in some assimilation setups involving time, which have been considered in recent literature.

To carry out this description we will move back and forth between Kalman filter and four-dimensional variational arguments. A number of authors have discussed the relationship between Kalman filtering and variational problems, see, for example Jazwinski(1970), Ghil and Malanotte-Rizzoli(1991). See also Daley(1991), Ghil *et al* (1981), Gauthier, Courtier and Moll(1993), Rabier *et al*(1993). We will review some of these results in the process of providing the ML, GML, GCV and UBR estimates.

5.1 ML, GML, GCV and UBR based on the Kalman Filter at a single time step

Dee(1993) has described a ML estimate in this context. We first review the assumptions behind the usual Kalman filter with data at discrete times. In this section, integer subscripts will refer to time. x_t refers to the true but unknown state vector at time t, and y_t refers to the observation vector at time t.

The assumptions are: The data at time $t = 0, 1, ..., T$ consists of

$$\begin{array}{lclcll} x_* & & & x* & \sim & \mathcal{N}(x_0, S*) \\ y_0 & = & K_0(x_0) + \epsilon_0 & \epsilon_0 & \sim & \mathcal{N}(0, S_o) \\ y_1 & = & K_1(x_1) + \epsilon_1 & \epsilon_0 & \sim & \mathcal{N}(0, S_1) \\ & & \cdots & \cdots & & \\ y_T & = & K_T(x_T) + \epsilon_T & \epsilon_T & \sim & \mathcal{N}(0, S_T). \end{array} \tag{5.1}$$

Here $x*$ is a starting guess for x_0, which could be climatology or a previous forecast [3], and the ϵ_t are independent from time to time. The true, but unknown system $x_0, x_1, ...$ evolves according to $x_t = \mathcal{M}_{t-1}(x_{t-1}, z)$, z represents 'true' state information not contained in the finite dimensional state vector x_t, and $\mathcal{M}_t$, for $t = 0, 1, ...$ represents the 'true' evolution. However, only a computer model $M_t(x_t)$ is known, and it is assumed that

$$\begin{array}{lclcll} x_1 & = & M_0(x_0) + \eta_0 & \eta_0 & \sim & \mathcal{N}(0, Q_0) \\ x_2 & = & M_1(x_1) + \eta_1 & \eta_1 & \sim & \mathcal{N}(0, Q_1) \\ & & \cdots & \cdots & & \\ x_T & = & M_{T-1}(x_{T-1}) + \eta_{T-1} & \eta_{T-1} & \sim & \mathcal{N}(0, Q_{T-1}) \end{array} \tag{5.2}$$

where the η_t are proxies for the model error and are assumed to be independent from time to time and independent also of x_t. These latter assumptions are, of course fairly suspect for numerical weather prediction models since model errors are likely to have a component systematic in time, and may depend as well on the true state. However, these assumptions are necessary for the usual recursive formulae for the Kalman filter to hold. [4]

We review the well known Kalman filter recursion relations, for the case that $K_t(x_t) = K_t x_t$ and $M_t(x_t) = M_t x_t$, that is, K_t and M_t are matrices. Below $\hat{x_t}$ is the estimate (analysis) at time t

[3] If $x*$ is a forecast, it is assumed to differ by the 'truth' by a random vector with covariance $S*$, if $x*$ is taken as climatology, then it is assumed that the 'truth' x_0 differs from climatology by a random vector with covariance $S*$. The mathematics is indifferent to which interpretation is intended, but $S*$ should be chosen with the appropriate interpretation in mind.

[4] The lack of independence can sometimes be gotten around by enlarging the state vector, we omit details here.

and x_t* is the forecast at time t. Then

$$\begin{array}{rcll} \hat{x}_t & = & x_t * + G_t(y_t - K_t x_t *), & t = 0, 1, ... \\ x_t * & = & M_{t-1}\hat{x}_{t-1}, & t = 1, 2, ... \\ G_t & = & S_t * K_t'(K_t S_t * K_t' + S_t)^{-1}. & t = 1, 2, ... \end{array} \tag{5.3}$$

For $t = 1, 2, ..., S_t*$ is the forecast error covariance, which satisfies the recursion formula

$$S_t * = M_{t-1} P_{t-1}^a M_{t-1}' + Q_{t-1}, \tag{5.4}$$

where P_t^a is the analysis error covariance, which, in turn satisfies

$$P_t^a = (I - G_t K_t) S_t * = S_{t*} - S_{t*} K_t' (K_t S_t * K_t' + S_t)^{-1} K_t S_{t*}. \tag{5.5}$$

Letting $\delta_t = \hat{x}_t - x_t*$ and $y_t * = y_t - K_t x_t*$ and recalling the discussion around (2.13) we have that at time t, δ_t is the minimizer of

$$(y_t * - K_t \delta_t)' S_t^{-1} (y_t * - K_t \delta_t) + \delta_t' S_t *^{-1} \delta_t. \tag{5.6}$$

Recent work has focused on Kalman filter methods for obtaining an approximation to S_t*, including methods for nonlinear M_t, see Cohn(1993), and also Todling and Cohn(1993), and references cited there.

Dee(1993,1994) observes that the Kalman filter model entails that

$$y_t * \sim \mathcal{N}(0, K_t S_t * K_t' + S_t) \tag{5.7}$$

and he supposes that at time t,

$$S_t * = S_t * (\theta) = M_t P_{t-1}^a M_t' + Q_{t-1}(\theta) \tag{5.8}$$

where $S_t * (\theta)$ is completely known except for θ. Compare (5.4). Then he notes that the maximum likelihood estimate of θ may be obtained by maximizing the likelihood function corresponding to (5.7), with respect to θ. This done in the notation of the present paper by making the following identifications in (3.2) through (3.5):

$$\begin{array}{rcl} y_t * & \leftrightarrow & y \\ K_t & \leftrightarrow & K \\ \delta_t & \leftrightarrow & x \\ S_t & \leftrightarrow & S \\ S_t * (\theta) & \leftrightarrow & \Sigma. \end{array} \tag{5.9}$$

Bouttier(1993) has used a low resolution Kalman filter model with $Q_t(\theta)$ depending on a single parameter determined by trial and error by matching predicted error statistics with observational error statistics. Unfortunately we have not been able to find a randomized version of ML or GML to estimate θ. If, however, instead of estimating θ in (5.8), $K_t S_t * K_t'$ is replaced by $b K_t S_t * K_t'$ and S_t is replaced by $\sigma^2 S_t$ with σ^2 and $\lambda = \sigma^2 / b$ the unknowns, then the randomized version of GML of (4.26) can be used to estimate λ and σ^2, so that these parameters can be obtained for n too large for matrix decomposition methods. If σ^2 is known, then λ can be estimated using the randomized version of ML.

If S_t is completely known then (sufficiently identifiable!) parameters θ in $S_t*(\theta)$ may be estimated by minimizing $RanU(\theta)$. Let $\xi \sim N(0, \sigma_\xi^2 I)$ from a random number generator as before, and let $\hat{\delta}_{t,\theta}$ be the minimizer of

$$(y_t * - K_t\delta_t)' S_t^{-1}(y_t * - K_t\delta_t) + \delta_t' S_t * (\theta)^{-1}\delta_t. \tag{5.10}$$

Note that $\hat{\delta}_{t,\theta} = \hat{\delta}_{t,\theta}(y*) = \hat{x}_t - x*$ is produced by the model. Let $\hat{\delta}_{t,\theta}(y * + S^{1/2}\xi)$ be the corresponding quantity produced by perturbing y_t* by an amount $S^{1/2}\xi$, this is effected by adding the perturbation to y_t and retrieving the new model estimate for δ_t. Then [5]

$$RanU(\theta) = \frac{1}{n}(y_t* - K_t\hat{\delta}_{t,\theta})' S_t^{-1}(y_t* - K_t\hat{\delta}_{t,\theta}) + \frac{2}{n}\left\{\frac{1}{\sigma_\xi^2}\xi'[S_t^{-1/2}(K_t\hat{\delta}_{t,\theta}(y_t * + S^{1/2}\xi) - K_t\hat{\delta}_{t,\theta}(y_t*))]\right\}. \tag{5.11}$$

Randomized GCV is appropriate when S_t is replaced by $\sigma^2 S_t$ with σ^2 unknown. Replace $S_t*(\theta)$ of (5.4) by

$$S_t * (\theta) = b[M_{t-1}P_{t-1}^a M_{t-1}' + Q_{t-1}(\theta)], \tag{5.12}$$

where b may be known or unknown, and let $\lambda = \sigma^2/b$. Note that G_t then depends on λ, which will appear as a multiplier on S_t in (5.3), but G_t does not depend on b. Then $\hat{\delta}_{t,\theta,\lambda}$ will be the minimizer of

$$(y_t * - K_t\delta_t)' S_t^{-1}(y_t * - K_t\delta_t) + \lambda\delta_t' S_t *^{-1}\delta_t. \tag{5.13}$$

$RanV(\theta, \lambda)$ can then be minimized to estimate θ and λ, it is given by

$$RanV(\theta,\lambda) = \frac{\frac{1}{n}(y_t * - K_t\hat{\delta}_{t,\theta,\lambda})' S_t^{-1}(y_t * - K_t\hat{\delta}_{t,\theta,\lambda})}{\left(\frac{1}{n}\left\{\frac{1}{\sigma_\xi^2}\xi'[\xi - (S_t^{-1/2}(K_t\hat{\delta}_{t,\theta}(y_t * + S_t^{1/2}\xi) - K_t\hat{\delta}_{t,\theta}(y_t*)))]\right\}\right)^2}. \tag{5.14}$$

Again, some parameters in K may be estimable by including them in the minimization of $RanU$ or $RanV$.

5.2 The weak four dimensional variational problem, and the Kalman Filter

Four-Dimensional Variational Lemma 1 *Let $x*, y_t, t_0, ..., T$ and $x_t, t = 1, ..., T$ satisfy (5.1) and (5.2), and consider x_0 as an unknown constant.*

Then, the joint density of $(x, x_1, ...x_T, y_1, ...y_T)$ is given by a constant times*

$$\frac{1}{\left[(\prod_{t=0}^T \det S_t)(\prod_{t=0}^{T-1}\det Q_t)\det S*\right]^{1/2}}$$

$$\times\, e^{-\frac{1}{2}\left[\sum_{t=0}^T (y_t - K_t(x_t))' S_t^{-1}(y_t - K_t(x_t)) + \sum_{t=1}^T (x_t - M_{t-1}(x_{t-1}))' Q_{t-1}^{-1}(x_t - M_{t-1}(x_{t-1})) + (x* - x_0)' S*^{-1}(x* - x_0)\right]}. \tag{5.15}$$

Proof:

$$f(x*, x_1, ..., x_T, y_1, ..., y_T) = f(y_1, ..., y_T | x*, x_1, ..., x_T) f(x_1, ..., x_T) f(x*) \tag{5.16}$$

[5]Note that σ^2 of (3.20) has been absorbed into S_t. Otherwise it would have appeared after the 2.

where $f(u|v)$ means the conditional density of u given v. We have

$$f(y_1, ..., y_T | x*, x_1, ..., x_T) = const \times \frac{1}{(\prod_{t=0}^{T} \det S_t)^{1/2}} \times e^{-\frac{1}{2}\left[\sum_{t=0}^{T}(y_t - K_t(x_t))' S_t^{-1}(y_t - K_t(x_t))\right]}. \quad (5.17)$$

Now,

$$f(x_1, ..., x_T) = f(x_1)f(x_2|x_1)f(x_3|x_1, x_2)...f(x_T|x_1, ..., x_{T-1}) \quad (5.18)$$

Since

$$f(x_t|x_{t-1}, ..., x_1) = f(x_t|x_{t-1}), t = 1, 2, ..., T, \quad (5.19)$$

$$f(x_1, ..., x_T) = const \times \frac{1}{(\prod_{t=0}^{T-1} \det Q_t)^{1/2}} e^{-\frac{1}{2}\sum_{t=1}^{T}(x_t - M_{t-1}(x_{t-1}))' Q_{t-1}^{-1}(x_t - M_{t-1}(x_{t-1}))} \quad (5.20)$$

and

$$f(x*) = const \times \frac{1}{(det S*)^{1/2}} e^{\frac{1}{2}(x* - x_0)' S*^{-1}(x* - x_0)} \quad (5.21)$$

the result follows.

Therefore, the maximum likelihood estimates of $x_0, x_1, ..., x_T$ given $x*, y_1, ..., y_T$ are obtained by minimizing

$$\sum_{t=0}^{T}(y_t - K_t(x_t))' S_t^{-1}(y_t - K_t(x_t)) + \sum_{t=1}^{T}(x_t - M_{t-1}(x_{t-1}))' Q_{t-1}^{-1}(x_t - M_{t-1}(x_{t-1})) + (x_0 - x*)' S*^{-1}(x_0 - x*). \quad (5.22)$$

In the linear case (K_t, M_t matrices) where all the random variables are Gaussian, the Kalman filter estimates, which are conditional expectations (see Jazwinski(1970)) are also maximum likelihood estimates. It follows, that in the linear case, we have the

Kalman Filter Variational Lemma 1 *$\hat{x}_T$ obtained from (5.3, 5.4, 5.5) will be the same as $\hat{x}_T$ obtained from the minimization of (5.22).*

Note, however that $\hat{x}_t$ for $t < T$ does not necessarily coincide with $\hat{x}_t$ produced from the Kalman filter, since (5.22) uses y_T to estimate x_t for $t < T$, while the Kalman filter does not.

Let $y = (y_o, ..., y_T)'$ and $x = (x_o, ..., x_T)'$. Rewrite (5.22) as

$$(y - K(x))' S^{-1}(y - K(x)) + J_\theta(x, x*), \quad (5.23)$$

in an obvious notation, θ are unknown parameters in the ingredients $\{M_t, Q_t, \}$ and $S*$ of J_θ. Let $\hat{x}_\theta(y)$ and $\hat{x}_\theta(y + S^{1/2}\xi)$ be, respectively, the solutions to the variational problem of (5.22) given the data y and the perturbed data $y + S^{1/2}\xi$. Then the randomized UBR estimate of θ is the minimizer of

$$RanU(\theta) = \frac{1}{n}(y - K(\hat{x}_\theta(y)))' S^{-1}(y - K(\hat{x}_\theta(y))) + \frac{2}{n}\left\{\frac{1}{\sigma_\xi^2}\xi'[S^{-1/2}(K\hat{x}_\theta(y + S^{1/2}\xi) - K\hat{x}_\theta(y))]\right\}. \quad (5.24)$$

$RanV(\theta)$ would be used if, for example S in (5.23) were replaced by $\sigma^2 S$ with σ^2 unknown and then (5.23) was multiplied through by σ^2.

5.3 Using the model as a strong constraint

5.3.1 Perfect model case

As before, let M_t be the computer implementation of a model which is integrated forward in time to obtain x_t, given x_{t-1}, that is,

$$x_t \approx M_{t-1}(x_{t-1}). \tag{5.25}$$

Assuming that there is no model error, then the state x_t is related to the state x_o by

$$x_t = M_{t-1}(M_{t-2}(...M_o(x_o))) = \tilde{M}_t(x_o), \quad \text{say.} \tag{5.26}$$

Let y_t be a vector of observations at time t related to x_t by

$$y_t = K_t(x_t) + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, S_t). \tag{5.27}$$

Four dimensional variational data assimilation with the model as a strong constraint based on the statistical model of (5.12) with the $Q_t \equiv 0$ would find the initial state x_o to minimize [6]

$$\sum_{t=0}^{T}(y_t - K_t(\tilde{M}_t(x_o)))'S_t^{-1}(y_t - K_t(\tilde{M}_t(x_o))) + (x_o - x*)'S*^{-1}(x_o - x*). \tag{5.28}$$

Sophisticated methods for finding x_o to minimize (5.28) are a subject of great interest. See, for example, Zou *et al* (1992). Letting $K = (K_o(\tilde{M}_o(\cdot))', K_1(\tilde{M}_1(\cdot))', ..., K_T(\tilde{M}_T(\cdot))')'$ we may write (5.28) as

$$(y - K(x_o))'S^{-1}(y - K(x_o)) + (x_o - x_*)'S*^{-1}(x_o - x_*). \tag{5.29}$$

Letting θ be unknown parameters in K and $S*$, and letting $\hat{x}_{o,\theta}(y)$ and $\hat{x}_{o,\theta}(y + S^{1/2}\xi)$ be the minimizers of (5.28) based on data y and $y + S^{1/2}\xi$ respectively, the randomized UBR estimate of θ is found by minimizing

$$RanU(\theta) = \frac{1}{n}(y-K(\hat{x}_{o,\theta}(y)))'S^{-1}(y-K(\hat{x}_{o,\theta}(y)))+\frac{2}{n}\left\{\frac{1}{\sigma_\xi^2}\xi'[S^{-1/2}(K\hat{x}_{o,\theta}(y + S^{1/2}\xi) - K\hat{x}_{o,\theta}(y))]\right\} \tag{5.30}$$

and, similarly for GCV.

5.3.2 Imperfect model case

In this section we consider the case where the model will be used as a strong constraint, even though it is known that there are model errors. The computer model, given x_{t-1}, produces x_t as $x_t = M_{t-1}(x_{t-1})$, and, given x_o, will produce

$$x_t = M_{t-1}(M_{t-2}(...M_o(x_o))) = \tilde{M}_t(x_o), \quad \text{say.} \tag{5.31}$$

[6] Most (but not all) authors experimenting with four dimensional variational data assimilation have found that some penalty term of the form $(x_0 - x*)'S*^{-1}(x_0 - x*)$ ($x*$ may be 0) based on balance or related physical considerations is necessary, or, at least improves the analysis. From a mathematical point of view, in order for a variational problem of the form $(y - Kx)'S^{-1}(y - Kx) + x'\Sigma^{-1}x$, to have a unique minimizer, it is necessary the intersection of the null spaces of K and Σ^{-1} are empty, that is, $Kx = 0$ and $\Sigma^{-1}x = 0$ imply that $x = 0$. If this penalty term is omitted, and there are a large number of degrees of freedom in the model, this may not be true. See also Bennett and Miller(1991). Of course the larger T is the less important this penalty term may become. Balance or other penalty terms may also be imposed at some time other than t_o.

Let y_t be a vector of observations at time t related to x_t by

$$y_t = K_t(x_t) + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, S_t). \tag{5.32}$$

We now define a control variable x_o^c, to be found, which, if the model were perfect, would have the meaning of an estimate of the state of the system at time o. However, using the model as a strong constraint while knowing that the model is not perfect, we actually want to choose x_o^c in such a way that the analysis $\hat{x}_T$ at some future time is the best obtainable estimate of the state at time T. Alternatively, for a historical reanalysis, we may want to choose x_o^c so that some measure of the sequence of analysis errors from time o to T is minimized. In the presence of model errors, this may or may not be achieved by an x_o^c best estimating the state at time o. A typical four dimensional variational problem with the model as a strong constraint might find x_0^c to minimize

$$\begin{aligned}(y_T - K_T(\tilde{M}_T(x_0^c)))'S_T^{-1}(y_T - K_T(\tilde{M}_T(x_0^c))) \quad &+ \quad \sum_{t=0}^{T-1}(y_t - K_t(\tilde{M}_t(x_0^c)))'(V_t(\theta) + S_t)^{-1}(y_t - K_t(\tilde{M}_t(x_0^c))) \\ &+ \quad (x_0^c - x*)'(V_*(\theta) + S*)^{-1}(x_0^c - x*)\end{aligned} \tag{5.33}$$

where the V_t are included to compensate for model error. Note that this formulation is consistent with the desire to estimate x_T as well as possible, and then it would be expected that the V_t would increase going backwards in time, consistent with the desire to rely more strongly on recent data when there is model error. A rationale for this argument is given in the Appendix.

A partial UBR estimate for θ based on y_T (whose error structure is assumed known) [7] , may be based on the criteria for estimating θ which minimizes a proxy for $R_{W_T}(\theta)$ given by

$$R_{W_T}(\theta) = \|W_T S_T^{-1/2}(K_T(x_T) - K_T(\tilde{M}_T(\hat{x}_0^c(\theta))))\|^2, \tag{5.34}$$

where x_T is the true, but unknown state at time T. The partial UBR estimate of θ is the minimizer of

$$U_{W_T}(\theta) = \|W_T S_T^{-1/2}(y_T - K_T(\tilde{M}_T(\hat{x}_0^c(\theta))))\|^2 + 2tr\ W_T A_{TT}(\theta) W_T', \tag{5.35}$$

where $A_{TT}(\theta)$ satisfies

$$S_T^{-1/2} K_T(\tilde{M}_T(\hat{x}_0^c(\theta))) \equiv S_T^{-1/2}\hat{y}_T \approx A_{TT}(\theta) S_T^{-1/2} y_T + terms\ in\ x_*, x_0, y_1, \cdots, y_{T-1}. \tag{5.36}$$

To estimate $trW_T A_{TT}(\theta)W_T'$, write $\hat{x}_o^c(\theta, y_T)$ to indicate the dependence of $S_T^{-1/2}\hat{y}_T = S_t^{-1/2} K_T(\tilde{M}_T(x_o^c(\theta, y_T)))$ on the data y_T. (Of course $\hat{y_T}$ depends on $x*, x_o, y_1, ..., y_{T-1}$ also.) Then the randomized estimate of $trW_T A_{TT}(\theta)W_T'$ is given by

$$Ran\, trW_T A_{TT}(\theta) W_T' = \frac{1}{\sigma_\xi^2} W_T \xi' [S_T^{-1/2}(K_T(\tilde{M}_T(x_c^o(\theta, y_T + S_T^{1/2}\xi))) - K_T(\tilde{M}_T(x_c^o(\theta, y_T + S_T^{1/2} W_T'\xi))))]. \tag{5.37}$$

Note that just choosing θ to minimize the first term on the right of (5.35) may lead to overfitting.

If S_T is replaced by $\sigma^2 S_T$, with σ^2 unknown, then GCV can be used.

[7] A hidden assumption that there exists a control vector x_o^c that will lead to the true x_T is implicitly being made here.

5.4 The VCA method

Derber(1989) in the so-called variational continuous assimilation (VCA) method, suggested relaxing the requirement that the model be satisfied exactly, by supposing

$$x_t \approx M_{t-1}(x_{t-1}) + c_t\phi, \tag{5.38}$$

where the c_t are given constants depending on t and ϕ is a bias or forcing term (generally taken independent of t). ϕ may be known or may have a few unknown parameters. See also Wergen(1992), Zupanski(1993). The model then produces x_t from the $c_t\phi$ and a control variable x_o^c having the same meaning as before, via

$$x_t = \tilde{M}_t(x_o, c_t\phi) \tag{5.39}$$

where $\tilde{M}_t$ is defined by the recursion (5.38). x_o^c is then chosen to minimize (5.33) with $\tilde{M}_t(x_o^c)$ replaced by $\tilde{M}_t(x_o^c, c_t\phi)$. Then identifiable parameters θ in c_t or in ϕ may also be estimated by GCV or UBR as in the preceding Section.

6 Accuracy estimation

The statistical theory, including Kalman Filter theory, provides (at least in the linear or linearized case) posterior covariances for the estimated state vector. These covariances with possibly a few parameters estimated by ML or GML, will provide good accuracy estimates if the other ingredients of the model are specified correctly. For these covariances to be realistic however, the various input covariances which generated them must be reasonably correctly specified. Referring to the general case of Section 2, if $y = Kx + \epsilon$ with $\epsilon \sim \mathcal{N}(0, \sigma^2 S)$, $x \sim \mathcal{N}(0, b\Sigma)$, then the posterior covariance of x given y is given by

$$\begin{aligned} cov\,[x|y] &= (K'(\sigma^2 S)^{-1}K + (b\Sigma)^{-1})^{-1} \\ &= \sigma^2(K'S^{-1}K + \lambda\Sigma^{-1})^{-1}. \end{aligned} \tag{6.1}$$

See (2.6) and (5.5). Substituting in σ^2_{GCV} from (3.18)and $\hat{\lambda}$ and other parameters chosen by UBR or GCV into (6.1) may provide very good estimates of the average variance of the components of x even if K, S and Σ are misspecified to some extent. Evidence for this remark in some simple smoothing spline cases has appeared, but further work is warranted to establish this in more general contexts, see Wahba(1983), Nychka (1988), Gao(1993), Gu and Wahba(1993). In the cases studied in those papers, where the 'truth' is known, the so-called Bayesian 'confidence intervals', which are 95% confidence intervals based on the variances in the analogue of (6.1) have, approximately, the average coverage property. That is, about 95% of the components of $\hat{x}$, the estimate of x, have their confidence intervals covering the corresponding components of x_{true}. This happens even if Σ is not actually the covariance of x, but represents an energy penalty.

7 Summary

We have reviewed the use of ML and GML; and described the use of GCV and UBR in the tuning of numerical weather prediction models and other dynamical systems when data is assimilated via the solution of a variational problem or an equivalent method. We have generalized the GCV and UBR to PGCV and PUBR. We have shown how the technique of randomized trace estimation makes the implementation of GCV and UBR feasible in very large models, and shown how it may be used to choose the number of iterations when the variational problem is solved by an iterative method. Finally we have described how the methods may be used in several variations of four dimensional data assimilation that appear in the literature. In future papers in this series we will develop a theory of parameter estimability applicable to the setup considered here and we will develop a new GCV-like statistic for estimating the relative accuracy of various sources of data.

Acknowledgments. The first author wishes to thank a number of people for helpful discussions and comments, and early versions of their manuscripts: Phillipe Courtier, Roger Daley, John Derber, Michael Ghil, Andrew Lorenc, Mike Navon, Dave Parrish, Jim Pfaendtner, Jim Purser, Florence Rabier and Olivier Talagrand. Thanks also are due to Ken Bergman of NASA Headquarters for his encouragement. This research was supported by NASA Grant NAGW-2961 and NSF Grant DMS 9121003. This paper is dedicated to the memory of Henry Fleming and his many contributions to Meteorology and Mathematics.

A Appendix

The model as a strong constraint when there are model errors

In this appendix we provide one rationale for the four dimensional variational problem where the model is used as a strong constraint, even though it is known that there are model errors, as in (5.33). The idea of this Appendix was a result of several private communications with Olivier Talagrand during the LaGarde Summer School.

Letting

$$x_t = M_{t-1}x_{t-1} + \eta_{t-1}, \quad t = 1, 2, ..., T, \tag{A.1}$$

where the η_t are proxies for model error, we solve (symbolically) for each x_t in terms of x_T, giving

$$\begin{aligned} x_{T-1} &= M_{T-1}^{-1}x_T + \psi_{T-1} \\ x_{T-2} &= M_{T-2}^{-1}M_{T-1}^{-1}x_T + \psi_{T-1} \\ \cdots & \quad \cdots \\ x_0 &= M_0^{-1}M_1^{-1}\cdots M_{T-1}^{-1}x_T + \psi_0, \end{aligned} \tag{A.2}$$

where

$$\begin{aligned} \psi_{T-1} &= -M_{T-1}^{-1}\eta_{T-1} \\ \psi_{T-2} &= -M_{T-2}^{-1}M_{T-1}^{-1}\eta_{T-1} - M_{T-2}^{-1}\eta_{T-2} \\ \cdots & \quad \cdots \\ \psi_0 &= -M_0^{-1}M_1^{-1}\cdots M_{T-1}^{-1}\eta_{T-1} \\ & \quad -M_0^{-1}M_1^{-1}\cdots M_{T-2}^{-1}\eta_{T-2}\cdots - M_0^{-1}M_1^{-1}\eta_1 - M_0^{-1}\eta_0. \end{aligned} \tag{A.3}$$

Now *define* the control variable x_0^c by

$$x_0^c = M_0^{-1}\cdots M_{T-1}^{-1}x_T \equiv \tilde{M}_T^{-1}x_T \equiv x_0 - \psi_o. \tag{A.4}$$

We will estimate x_0^c by $\hat{x}_0^c$ and determine the desired estimate $\hat{x}_T$ by $\tilde{M}_T(\hat{x}_0^c)$. Letting $x* = x_0 + \epsilon*$, where $\epsilon* \sim \mathcal{N}(0, S*)$, and eliminating $x_1, \cdots, x_T$ from the model (5.1,5.2) gives

$$x_* = x_o^c + \psi_0 + \epsilon *$$

$$
\begin{aligned}
y_0 &= K_0(\tilde{M}_0(x_0^c)) + K_0\psi_0 + \epsilon_0 \\
y_1 &= K_1(\tilde{M}_1(x_0^c)) + K_1\psi_1 + \epsilon_1 \\
\cdots & \quad \cdots \\
y_{T-1} &= K_{T-1}(\tilde{M}_{T-1}(x_0^c)) + K_{T-1}\psi_{T-1} + \epsilon_{T-1} \\
y_T &= K_T(\tilde{M}_T(x_0^c)) + \epsilon_T.
\end{aligned} \tag{A.5}
$$

(A.6)

We may find the variational problem that goes with this model upon supposing that the ψ's and ϵ's are zero mean Gaussian random variables. We need the covariance matrix V, say, for the random vector

$$
\begin{pmatrix}
\psi_0 & + & \epsilon* \\
K_0\psi_0 & + & \epsilon_0 \\
\cdots & & \cdots \\
K_{T-1}\psi_{T-1} & + & \epsilon_{T-1} \\
 & & \epsilon_T
\end{pmatrix}. \tag{A.7}
$$

Letting Q_{st} be the covariance of $\psi_s, \psi_t, s, t = 0, \cdots, T-1$, this covariance matrix is given by

$$
V = \begin{pmatrix}
Q_{00} + S* & Q_{01}K_0' & \cdots & Q_{0,T-1}K_{T-1}' & 0 \\
K_0Q_{00} & K_0Q_{00}K_0' + S_0 & \cdots & K_{T-1}Q_{T-1,1}K_0' & 0 \\
\cdots & \cdots & \cdots & \cdots & 0 \\
K_{T-1}Q_{T-1,0} & K_{T-1}Q_{T-1,T-1}K_0' & \cdots & K_{T-1}, Q_{T-1,T-1}K_{T-1}' + S_{T-1} & 0 \\
0 & 0 & 0 & 0 & S_T
\end{pmatrix}. \tag{A.8}
$$

Given the covariances of the η_t's, the covariances $Q_{st} = E\psi_s'\psi_t$ are obtained by substituting in the right hand side of (A.3) and taking expectations. Note that Q_{st} for $s \neq t$ is not zero in general even though the η_t's are independent, and that the Q_{tt} matrices tend to increase as t decreases, due to the way the original model errors (η_t's) accumulate to form the ψ_t's. Let $V^{st}, s, t = *, 0, \cdots, (T-1)$ be the s, tth block of the inverse of the covariance matrix V in (A.8). Then the variational problem to be minimized to get x_0^c becomes: find x_0^c to minimize

$$
\begin{aligned}
(y_T - K_T(\tilde{M}_T(x_0^c)))'S_T^{-1}(y_T - K_T(\tilde{M}_T(x_0^c))) &+ \sum_{s,t=0}^{T-1}(y_s - K_s(\tilde{M}_s(x_0^c)))'V^{st}(y_t - K_t(\tilde{M}_t(x_0^c))) \\
&+ (x_0^c - x*)'V^{**}(x_0^c - x*) \\
&+ 2\sum_{s=0}^{T-1}(y_s - K_s(\tilde{M}_s(x_0^c)))'V^{s*}(x_0^c - x*).
\end{aligned} \tag{A.9}
$$

Note that in principle the covariances Q_{st} depend on the model matrices M_t^{-1}. In practice, it is likely that it will be necessary to approximate the covariance matrix V of (A.8) by an easily computable block diagonal matrix which depends only on a few unknown parameters, and whose diagonal blocks reflect the growth of the model error backwards in time. For example, ignoring the off diagonal blocks in (A.9) and replacing $K_tQ_{tt}K_t'$ by $V_t(\theta)$ say, would give: Find x_0^c to minimize

$$
\begin{aligned}
(y_T - K_T(\tilde{M}_T(x_0^c)))'S_T^{-1}(y_T - K_T(\tilde{M}_T(x_0^c))) &+ \sum_{t=0}^{T-1}(y_t - K_t(\tilde{M}_t(x_0^c)))'(V_t(\theta) + S_t)^{-1}(y_t - K_t(\tilde{M}_t(x_0^c))) \\
&+ (x_0^c - x*)'(V_*(\theta) + S*)^{-1}(x_0^c - x*).
\end{aligned} \tag{A.10}
$$

References

Anderson, T. (1958), *An Introduction to Multivariate Analysis*, Wiley, New York.

Anderssen, R. & Bloomfield, P. (1974), 'A time series approach to numerical differentiation', *Technometrics* **16**, 69–75.

Bartello, P. & Mitchell, H. (1993), 'A continuous three-dimensional model of short-range forecast error covariances', *Tellus* **44A**, 217–235.

Bates, D., Lindstrom, M., Wahba, G. & Yandell, B. (1986), GCVPACK-Routines for generalized cross validation, Technical Report 775(rev.), Dept. of Statistics, University of Wisconsin. An earlier version has appeared in Commun. Statist. Simul. Comput., 1987, v. 16, pp263-297.

Bennett, A. & Budgell, W. (1987), 'Ocean data assimilation and the Kalman filter: spatial reglarity', *J. Physical Oceanography* **17**, 1583–1601.

Bennett, A. F. & Miller, R. N. (1991), 'Weighting initial conditions in variational assimilation schemes', *Mon. Wea. Rev.* **119**, 1098–1102.

Bouttier, F. (1993), A dynamical estimation of error covariances in an assimilation system, manuscript, Meteo-France/CNRM.

Cohn, S. (1993), 'Dynamics of short-term univariate forecast error covariances', *Monthly Weather Review* **121**, 3123–3149.

Courtier, P. & Talagrand, O. (1987), 'Variational assimilation of meteorological observations with the adjoint vorticity equation. ii: Numerical results', *Quart. J. R. Meteorol. Soc.* **113**, 1329–1347.

Cramer, H. (Sixth Printing, 1954), *Mathematical Methods of Statistics*, Princeton University Press.

Craven, P. & Wahba, G. (1979), 'Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation', *Numer. Math.* **31**, 377–403.

Daley, R. (1991), *Atmospheric Data Analysis*, Cambridge University Press.

Daley, R. (1992*a*), 'The effect of serially correlated observation and model error on atmospheric data assimilation.', *Mon. Wea. Rev.* **120**, 164–177.

Daley, R. (1992*b*), 'Estimating model-error covariances for application to atmospheric data assimilation', *Mon. Wea. Rev.* **120**, 1735–1746.

Daley, R. (1992*c*), 'Forecast error statistics for homogeneous and inhomogeneous observation networks.', *Mon. Wea. Rev.* **120**, 627–643.

Daley, R. (1992*d*), 'The lagged innovation covariance: a performance diagnostic for atmospheric data assimilation', *Mon. Wea. Rev.* **120**, 178–196.

Dee, D. (1990), Simplified adaptive Kalman filtering for large-scale geophysical models, *in* M. Kaashoek, J. van Schuppen & A. Ram, eds, 'Realization and modelling in system theory: Proceedings of the international symposium MTNS-89, v1', Birkhauser, pp. 567–574.

Dee, D. (1993), A simple scheme for tuning forecast error covariance parameters, *in* ECMWF, ed., 'Variational assimilation, with special emphasis on three-diensional aspects', European Center for Medium-Range Weather Forecasts, Reading, England, pp. 191–202.

Dee, D. (1994), On-line estimation of error covariance parameters for atmospheric data assimilation, manuscript.

Dee, D., Cohn, S., Dalcher, A. & Ghil, M. (1985), 'An efficient algorithm for estmating noise covariances in distributed systems', *IEEE Trans. Autom. Control* **AC-30**, 1057–1065.

Derber, J. (1989), 'A variational continuous assimilation technique', *Mon. Wea. Rev.* **117**, 2437–2466.

Derber, J., Parrish, D. & Lord, S. (1991), 'The new global operational analysis system at the National Meteorological Center', *Weather Forecasting* **6**, 538–547.

Fleming, H. (1990), 'Equivalence of regularization and truncated iteration in the solution of ill-posed image reconstruction problems', *Lin. Alg. Appl.* **130**, 133–150.

Gao, F. (1992), On combining data from multiple sources with unknown relative weights, Technical Report 894, Dept. of Statistics, University of Wisconsin, Madison, WI. To appear *Commun. Statist.*.

Gao, F. (1993), On combining data from multiple sources with unknown relative weights (thesis), Technical Report 902, Dept. of Statistics, University of Wisconsin, Madison, WI.

Gauthier, P., Courtier, P. & Moll, P. (1993), 'Assimilation of simulated wind lidar data with a Kalman filter', *Mon. Wea. Rev.* **121**, 1803–1820.

Ghil, M. & Malanotte-Rizzoli, P. (1991), 'Data assimilation in meteorology and oceanography', *Advances in Geophysics* **33**, 141–183.

Ghil, M., Cohn, S., Tavantzis, J., Bube, K. & Isaacson, E. (1981), Applications of estimation theory to numerical weather prediction, *in* L. Bengtsson, M. Ghil & E. Kallen, eds, 'Dynamic Meteorology: data assimilation methods.', Springer-Verlag, pp. 139–284.

Girard, D. (1987), A fast 'Monte Carlo cross validation' procedure for large least squares problems with noisy data, Technical Report RR 687-M, IMAG, Grenoble, France.

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. Statist.* **19**, 1950–1963.

Goerss, J. & Phoebus, P. (1993), The mulivariate optimum interpolation analysis of meteorological data at the fleet numerical oceanography center, Technical Report NRL/FR/7531-92-9413, Naval Research Laboratory, Monterey, CA.

Golub, G. & Loan, C. V. (1989), *Matrix Computations, Second Edition*, Johns Hopkins University Press.

Gu, C. (1989), RKPACK and its applications: fitting smoothing spline models, *in* 'Proceedings of the Statistical Computing Section', American Statistical Association, pp. 42–51. Code available thru `netlib`.

Gu, C. & Wahba, G. (1991), 'Minimizing GCV/GML scores with multiple smoothing parameters via the Newton method', *SIAM J. Sci. Statist. Comput.* **12**, 383–398.

Gu, C. & Wahba, G. (1993), 'Smoothing spline ANOVA with component-wise Bayesian "confidence intervals"', *J. Computational and Graphical Statistics* **2**, 97–117.

Hoffman, R. (1984), 'SASS wind ambiguity removal by direct minimization II. Effect of smoothing and dynamical constraints', *Monthly Weather Review* **112**, 1829–1852.

Hoffman, R. (1985), Using smoothness constraints in retrievals, *in* A. Deepak, H. Fleming & M. Chahine, eds, 'Advances in Remote Sensing Retrieval Methods', A. Deepak Publishing, Hampton, VA, pp. 411–436.

Hoffman, R. & Louis, J.-F. (1990), 'The influence of atmospheric stratification on scatterometer winds', *J. Geophys. Res.* **95**, 9723–9730.

Hollingsworth, A. & Lonnberg, P. (1986), 'The statististical structure of short-range forecast errors as determined from radiosonde data. part i: The wind field', *Tellus* **38A**, 111–136.

Hutchinson, M. (1989), 'A stochastic estimator for the trace of the influence matrix for Laplacian smoothing splines', *Commun. Statist.-Simula.* **18**, 1059–1076.

Hutchinson, M., Kalma, J. & Johnson, M. (1984), 'Monthly estimates of windspeed and wind run for australia', *J. Climatology* **4**, 311–324.

Jazwinski, A. (1970), *Stochastic processes and filtering theory*, Academic Press.

Kimeldorf, G. & Wahba, G. (1970), 'A correspondence between Bayesian estimation of stochastic processes and smoothing by splines', *Ann. Math. Statist.* **41**, 495–502.

Kimeldorf, G. & Wahba, G. (1971), 'Some results on Tchebycheffian spline functions', *J. Math. Anal. Applic.* **33**, 82–95.

Lewis, J. & Derber, J. (1985), 'The use of adjoint equations to solve a variational adjustment problem with advective constraints', *Tellus* **37A**, 309–322.

Li, K. C. (1986), 'Asymptotic optimality of C_L and generalized cross validation in ridge regression with application to spline smoothing', *Ann. Statist.* **14**, 1101–1112.

Li, Y., Navon, I., Yang, W., Zou, X., Bates, J., Moorthi, S. & Higgins, R. (1994), 'Four-dimensional variational data assimilation experiments with a multilevel semi-lagrangian semi-implicit general circulation model', *Monthly Weather Review* **122**, xx–xx.

Li, Y., Navon, M., Courtier, P. & Gauthier, P. (1993), 'variational data assimilation with a semi-Lagrangian semi-implicit global shallow-water equation model and its adjoint', *Mon. Wea. Rev.* **121**, 1759–1769.

Liu, D. & Nocedal, J. (1989), 'On the limited memory BFGS method for large scale optimization', *Mathematical Programming* **45**, 503–528.

Lonnberg, P. & Hollingsworth, A. (1986), 'The statistical structure of short-range forecast errors as determined from radiosonde data, Part II: The covariance of height and wind errors', *Tellus* **38A**, 137–161.

Lorenc, A. (1986), 'Analysis methods for numerical weather prediction', *Quart. J. R. Met. Soc.* **112**, 1177–1194.

Lorenc, A. (1988), 'Optimal nonlinear objective analysis', *Quart. J. R. Met. Soc.* **114**, 205–240.

Lorenc, A. (1992), 'Iterative analysis using covariance functions and filters', *Q. J. R. Meterol. Soc.* **118**, 569–591.

Lorenc, A. C., Bell, R. S. & MacPherson, B. (1991), 'The Meteorological Office analysis correction data assimilation scheme', *Q. J. R. Meteorol. Soc.* **117**, 59–89.

Mallows, C. (1973), 'Some comments on C_p', *Technometrics* **15**, 661–675.

Nychka, D. (1988), 'Bayesian confidence intervals for smoothing splines', *J. A. S. A.* **83**, 1134–1143.

O'Hagan, A. (1976), 'On posterior joint and marginal modes', *Biometrika* **63**, 329–333.

O'Sullivan, F. & Wahba, G. (1985), 'A cross validated Bayesian retrieval algorithm for non-linear remote sensing', *J. Comput. Physics* **59**, 441–455.

Parrish, D. & Derber, J. (1992), 'The National Meteorological Center's spectral statistical interpolation analysis system', *Monthly Weather Review* pp. 1747–1763.

Rabier, F. & Courtier, P. (1992), 'Four-dimensional assimilation in the presence of baroclinic instability', *J. R. Meteorol. Soc.* **118**, 649–672.

Rabier, F., Courtier, P., Pailleux, J., Talagrand, O., Thepaut, J. & Vasiljevic, D. (1993), 'A comparison between four dimensional variational assimilation and simplified sequential assimilation relying on three dimensional analysis', *Quart. J. Roy. Meteor.Soc.* **119**, 845–880.

Speckman, P. (1985), 'Spline smoothing and optimal rates of convergence in nonparametric regression', *Ann. Statist.* **13**, 970–983.

Stein, M. (1990), 'A comparison of generalized cross validation and modified maximum likelihood for estimating the parameters of a stochastic process', *Ann. Statist.* **18**, 1139–1157.

Thepaut, J. & Courtier, P. (1991), 'Four-dimensional variational data assimilation using the adjoint of a multilevel primitive-equation model', *Q. J. R. Meteorol. Soc.* **117**, 1225–1254.

Thepaut, J., Hoffman, R. & Courtier, P. (1993), 'Interactions of dynamics and observations in a four-dimensional variational assimilation', *Monthly Weather Review* **12**, 3393–3414.

Todling, R. & Cohn, S. (1993), Suboptimal schemes for atmospheric data assimilation based on the Kalman filter, submitted.

Wahba, G. (1975), A canonical form for the problem of estimating smooth surfaces, Technical Report 420, Dept. of Statistics, University of Wisconsin-Madison.

Wahba, G. (1981), 'Spline interpolation and smoothing on the sphere', *SIAM J. Sci. Stat. Comput.* **2**, 5–16.

Wahba, G. (1982), Variational methods in simultaneous optimum interpolation and initialization, *in* D. Williamson, ed., 'The Interaction Between Objective Analysis and Initialization', Atmospheric analysis and Prediction Division, National Center for Atmospheric Research, Boulder, CO, pp. 178–185.

Wahba, G. (1982a), 'Erratum: Spline interpolation and smoothing on the sphere', *SIAM J. Sci. Stat. Comput.* **3**, 385–386.

Wahba, G. (1983), 'Bayesian "confidence intervals" for the cross-validated smoothing spline', *J. Roy. Stat. Soc. Ser. B* **45**, 133–150.

Wahba, G. (1985), Variational methods for multidimensional inverse problems, *in* A. Deepak, H. Fleming & M. Chahine, eds, 'Remote Sensing Retrieval Methods', A. Deepak Publishing, pp. 385–408.

Wahba, G. (1985a), 'A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem', *Ann. Statist.* **13**, 1378–1402.

Wahba, G. (1987), Three topics in ill posed problems, *in* H. Engl & C. Groetsch, eds, 'Proceedings of the Alpine-U.S. Seminar on Inverse and Ill Posed Problems', Academic Press, pp. 37–51.

Wahba, G. (1990), *Spline Models for Observational Data*, SIAM. CBMS-NSF Regional Conference Series in Applied Mathematics, vol. 59.

Wahba, G. (1990*a*), Regularization and cross validation methods for nonlinear, implicit, ill-posed inverse problems, *in* A. Vogel, C. Ofoegbu, R. Gorenflo & B. Ursin, eds, 'Geophysical data Inversion Methods and Applications', Vieweg, Weisbaden-Braunschweig, pp. 3–13.

Wahba, G. & Wang, Y. (1990), 'When is the optimal regularization parameter insensitive to the choice of the loss function?', *Commun. Statist.-Theory Meth.* **19**, 1685–1700.

Wahba, G. & Wendelberger, J. (1980), 'Some new mathematical methods for variational objective analysis using splines and cross-validation', *Monthly Weather Review* **108**, 1122–1145.

Wahba, G., Johnson, D. R. & Reames, F. (1991), Multiple smoothing and weighting parameters in direct variational methods for objective analysis of meteorological information, *in* F.-X. LeDimet & O. Talagrand, eds, 'Assimilation of Observations in Meteorology and Oceanography', World Meteorologican Organization, pp. 448–453.

Wergen, W. (1992), 'The effect of model errors in variational assimilation', *Tellus* **44A**, 297–313.

Zou, X., Navon, I. & LeDimet, F.-X. (1992), 'Incomplete observations and control of gravity waves in variational data assimilation. part ii, applications and numerical results', *Tellus* **44A**, 297–313.

Zou, X., Navon, I. & Sela, J. (1993), 'Control of gravitational oscillations in variational data assimilation', *Monthly Weather Review* pp. 272–289.

Zupanski, M. (1993), 'Regional four-dimensional variational data assimilation in a quasi-operational forecasting environment', *Monthly Weather Review* **121**, 2396–2408.