

**Estimation of Entropy Reduction  
and Degrees of Freedom for  
Signal for Large Variational  
Analysis Systems**

Michael Fisher

Research Department

January 2003

**For additional copies please contact**

The Library  
ECMWF  
Shinfield Park  
Reading, Berks RG2 9AX

library@ecmwf.int

**Series: ECMWF Technical Memoranda**

A full list of ECMWF Publications can be found on our web site under:

<http://www.ecmwf.int/publications.html>

**© Copyright 2003**

European Centre for Medium Range Weather Forecasts  
Shinfield Park, Reading, Berkshire RG2 9AX, England

Literary and scientific copyrights belong to ECMWF and are reserved in all countries. This publication is not to be reprinted or translated in whole or in part without the written permission of the Director. Appropriate non-commercial use will normally be granted under the condition that reference is made to ECMWF.

The information within this publication is given in good faith and considered to be true, but ECMWF accepts no liability for error, omission and for loss or damage arising from its use.



## Summary

An algorithm due to Bai *et al.* (1996) allows accurate estimates to be made of the traces of certain functions of large matrices. We show that, by applying the algorithm in the context of variational data assimilation, it is possible to calculate two important measures of the information provided by observations. Specifically, we calculate the entropy reduction and degrees of freedom for signal produced by assimilating many thousands of observations in a version of the ECMWF four-dimensional variational analysis system. We verify the method by comparing it with explicit calculation for a simplified case with few observations. We describe an alternative (more expensive) method for calculating degrees of freedom for signal, based on an algorithm by Wahba (1995). We show that the estimate produced by this method converges to that produced using Bai *et al.*'s algorithm as the number of iterations of its minimization algorithm is increased.

## 1. Introduction

It is useful to be able to quantify the amount of information provided by an observation or by an observing system. In the development of remote-sounding instruments, two popular measures of information content are entropy-reduction and degrees of freedom for signal (see, for example, Eyre, 1990; Rodgers, 1996; Rabier *et al.*, 2002; Fourrié and Thépaut, 2003). These quantities are typically calculated for small-scale, one-dimensional analysis or retrieval systems, for which the covariance matrices of background and observation error are small enough to allow explicit matrix manipulations.

Wahba *et al.* (1995) presented an algorithm, in the context of generalized cross-validation (Golub *et al.*, 1979; Wahba and Wendleberger, 1980; Wahba, 1990), for calculating the degrees of freedom for signal for a large variational assimilation system. Another algorithm due to Cardinali (personal communication) is discussed below. We discuss the relative merits of these algorithms in section 6.

We know of no calculations of entropy reduction for large-scale analysis systems, although we note that algorithms exist (e.g. Liu, 2000; Barry and Pace, 1999) for estimating the log of the determinant of large matrices. It is likely that these algorithms could be used to estimate entropy reduction. A particular advantage of the method presented here is that estimates of degrees of freedom for signal and entropy reduction are produced simultaneously.

The structure of this paper is as follows. In sections 2 and 3, we briefly review the concepts of entropy-reduction and degrees of freedom for signal. The reader is referred to Rodgers (2000) for a more complete discussion. In section 4, we show that in variational data assimilation, both quantities may be expressed in terms of the trace of a function of the Hessian matrix of the analysis cost function, and may consequently be estimated using the algorithm of Bai *et al.* (1996). We apply the method to the ECMWF four-dimensional data assimilation system (Rabier *et al.*, 2000; Mahfouf and Rabier, 2000; Klinker *et al.*, 2000) in section 5. In section 6, we verify the method by applying it to a case for which explicit calculation of entropy reduction and degrees of freedom for signal is possible. We show that the estimates produced by the method are correct to within the random uncertainty inherent in the algorithm. We also compare our estimate of degrees of freedom for signal with that produced by a variant of Wahba *et al.*'s (1995) algorithm. The final section of the paper discusses the utility of the method, and suggests some possibilities for its wider application.

## 2. Entropy reduction

In information theory, entropy is a real-valued functional that characterizes probability density functions (Shannon and Weaver, 1949):

$$H(P) = -\int P(\mathbf{x}) \log_2(P(\mathbf{x})) d\mathbf{x} \quad (1)$$

(The logarithm may be taken to any convenient base. Base 2 is conventional, and will be used in this paper. For this choice of base, the units of entropy are called “bits”.)

For a Gaussian distribution, with covariance matrix  $\mathbf{C}$ , the entropy can be shown to be  $H(P) = \frac{1}{2} \log_2 |\mathbf{C}|$  (see, for example, Rodgers 2000). In this case, the entropy may be interpreted as a measure of the volume in phase space enclosed by a surface of constant probability.

The entropy-reduction  $S$  due to the use by the analysis of one or more observations (represented by the vector  $\mathbf{y}$ ) is simply the difference in entropy between the prior and posterior densities:

$$S = H(P(\mathbf{x})) - H(P(\mathbf{x}|\mathbf{y})) \quad (2)$$

For Gaussian distributions it is easy to show that entropy reduction is invariant under a linear transformation of the state vector. To see this, consider the transformation  $\mathbf{z} = \mathbf{L}\mathbf{x}$  (where  $\mathbf{L}$  is a non-singular square matrix) for a system with prior covariance matrix  $\mathbf{B}$  and posterior covariance matrix  $\mathbf{P}^a$ . In terms of  $\mathbf{x}$ , the entropy reduction is:

$$S_{\mathbf{x}} = \frac{1}{2} \log_2 |\mathbf{B}| - \frac{1}{2} \log_2 |\mathbf{P}^a| \quad (3)$$

The prior and posterior covariance matrices for  $\mathbf{z}$  are  $\mathbf{L}\mathbf{B}\mathbf{L}^T$  and  $\mathbf{L}\mathbf{P}^a\mathbf{L}^T$ , giving an entropy reduction for  $\mathbf{z}$  of:

$$S_{\mathbf{z}} = \frac{1}{2} \log_2 |\mathbf{L}\mathbf{B}\mathbf{L}^T| - \frac{1}{2} \log_2 |\mathbf{L}\mathbf{P}^a\mathbf{L}^T| \quad (4)$$

We now use the fact that the determinant of a product of square matrices is equal to the product of their determinants to write:

$$\begin{aligned} S_{\mathbf{z}} &= \frac{1}{2} \log_2 (|\mathbf{L}||\mathbf{B}||\mathbf{L}^T|) - \frac{1}{2} \log_2 (|\mathbf{L}||\mathbf{P}^a||\mathbf{L}^T|) \\ &= \frac{1}{2} (\log_2 |\mathbf{B}| + \log_2 |\mathbf{L}| + \log_2 |\mathbf{L}^T|) - \frac{1}{2} (\log_2 |\mathbf{P}^a| + \log_2 |\mathbf{L}| + \log_2 |\mathbf{L}^T|) \\ &= \frac{1}{2} \log_2 |\mathbf{B}| - \frac{1}{2} \log_2 |\mathbf{P}^a| \\ &= S_{\mathbf{x}} . \end{aligned} \quad (5)$$

Of particular interest are transformations for which  $\mathbf{L}\mathbf{B}\mathbf{L}^T$  is the identity matrix. Transformations of this form are used in most variational data assimilation systems to define the control variable for the minimization algorithm (Courtier *et al.*, 1998; Gauthier *et al.*, 1999; Gustafsson *et al.*, 2001; Lorenc *et al.*, 2000). In this case, the entropy reduction is  $S = -\frac{1}{2} \log_2 |\mathbf{P}_z^a|$ , where  $\mathbf{P}_z^a = \mathbf{L}\mathbf{P}^a\mathbf{L}^T$  is the covariance matrix for the posterior



distribution of  $\mathbf{z}$ . Noting that the logarithm of the determinant of a symmetric matrix is equivalent to the trace of its logarithm, we have:

$$S = -\frac{1}{2} \text{trace}(\log_2(\mathbf{P}_z^a)) \quad (6)$$

### 3. Degrees of freedom for signal

Degrees of freedom for signal may also be defined by means of a transformation  $\mathbf{L}$  that reduces the prior covariance matrix to the identity. For the prior, the components of the transformed state vector  $\mathbf{z}$  are statistically independent, and have unit variance. Each component corresponds to an independent “degree of freedom”.

The transformation  $\mathbf{L}$  is not uniquely determined, since we may replace  $\mathbf{L}$  by  $\mathbf{X}^T\mathbf{L}$ , where  $\mathbf{X}$  is an orthogonal matrix. In this case, we have  $\mathbf{X}^T\mathbf{L}\mathbf{B}\mathbf{L}^T\mathbf{X} = \mathbf{X}^T\mathbf{X} = \mathbf{I}$ . By choosing  $\mathbf{X}$  to be the matrix of eigenvectors of  $\mathbf{P}_z^a$ , we simultaneously reduce  $\mathbf{B}$  to the identity matrix, and  $\mathbf{P}_z^a$  to the diagonal matrix of its eigenvalues. In this case, we may interpret the eigenvalues  $\lambda_i$  as giving the fractional reduction in variance in each of the  $N$  statistically independent directions corresponding to the  $N$  components of the transformed state vector. Where an eigenvalue is close to zero, the direction is well observed. The direction is said to be a “degree of freedom for signal”. By contrast, where an eigenvalue is close to one, the direction is unconstrained by the observations, and is said to be a “degree of freedom for noise”. An indication of the effective number of degrees of freedom for signal is given by:

$$d = N - \sum_i \lambda_i \quad (7)$$

That is:

$$d = N - \text{trace}(\mathbf{P}_z^a) \quad (8)$$

Note that Wahba *et al.* (1995) define degrees of freedom for signal as the trace of the so-called influence matrix,  $\mathbf{A} = \mathbf{R}^{-1/2}\mathbf{H}\mathbf{P}^a\mathbf{H}^T\mathbf{R}^{-1/2}$  (where  $\mathbf{R}$  is the covariance matrix of observation error, and  $\mathbf{H}$  is the matrix that maps the model state vector to the observed quantities). We show in appendix A that the two definitions of degrees of freedom for signal are equivalent.

### 4. Estimating $S$ and $d$ in a variational analysis system

It is well known (Gauthier, 1992; Rabier and Courtier, 1992) that, for a perfect model and correctly specified covariance matrices of background and observation error, the covariance matrix of analysis error is equal to the inverse of the Hessian matrix of the analysis cost function. Specifically, if the control variable for the minimization is defined by the transformation  $\mathbf{z} = \mathbf{L}\mathbf{x}$  (as noted above, this is usually the case), then  $\mathbf{P}_z^a = (\mathbf{J}_z'')^{-1}$ , where  $\mathbf{J}_z''$  is the Hessian matrix of the cost function with respect to the control variable.

In this case we have:

$$S = -\frac{1}{2} \text{trace}(\log_2(\mathbf{J}_z'')^{-1}) \quad (9)$$

and:

$$d = N - \text{trace}\left(\left(J_z^r\right)^{-1}\right) \quad (10)$$

Both quantities require the calculation of the trace of a function of the Hessian matrix.

An elegant algorithm for estimating the trace of a function of a large matrix is described by Bai *et al.* (1996). The algorithm may be split into two parts. (Details are given in appendix B.) First, it is noted that the trace of a matrix  $\mathbf{C}$  may be estimated using the method of Hutchinson (1989). This method calculates:

$$\text{trace}(\mathbf{C}) \approx \mathbf{u}^T \mathbf{C} \mathbf{u} \quad (11)$$

where  $\mathbf{u}$  is a vector whose elements take the values  $\pm 1$  randomly and independently with probability  $\frac{1}{2}$ .

Use of this randomized trace estimate reduces the problem of calculating the trace of a function  $f$  of a matrix  $\mathbf{C}$  to that of calculating  $\mathbf{u}^T f(\mathbf{C}) \mathbf{u}$ . Bai *et al.* (1996) use the algorithm of Golub and Meurant (1993) and Golub and Strakos (1993) for this. The algorithm is based on manipulations of the tri-diagonal matrix that is generated by applying the Lanczos algorithm to  $\mathbf{C}$  with initial vector  $\mathbf{u}$ . For functions  $f(\lambda)$  whose even derivatives have consistent sign, and whose odd derivatives have the opposite sign for all  $\lambda$  lying between the smallest and largest eigenvalue of  $\mathbf{C}$ , it is possible to calculate strict upper and lower bounds for  $\mathbf{u}^T f(\mathbf{C}) \mathbf{u}$ . In particular, this property holds for the logarithm and for the inverse.

Application of Bai *et al.*'s algorithm to the problem of calculating  $S$  and  $d$  in a variational analysis system is straightforward, and requires only the calculation of Hessian-vector products. These may be calculated using the fact that for a strictly quadratic cost function the product of the Hessian with a vector may be calculated exactly as a finite difference of gradients. (In practice, the requirement that the cost function should be strictly quadratic is not very restrictive. Most variational assimilation systems use a cost function that is quadratic, or that may easily be converted to quadratic by modifying a few non-quadratic terms.)

The computational cost of applying the algorithm in a variational analysis system is overwhelmingly dominated by the cost of the gradient calculations required to determine the tri-diagonal matrix. Once this matrix has been determined, the cost of the manipulations required to calculate upper and lower bounds for degrees of freedom for signal and for entropy reduction is negligible. Bounds on both quantities may be determined simultaneously.

## 5. Estimates of $S$ and $d$ for the ECMWF 4dVar analysis system

Figure 1 shows ten estimates of the information content of the ECMWF 4dVar system for an analysis for 1<sup>st</sup> October 2002. (Note that the vertical axis corresponds to a relatively small range of values of entropy.) For these calculations, the spectral resolution of the analysis system was T159 with T95 increments, and the duration of the 4dVar analysis window was 6 hours. In all other aspects, the analysis system was equivalent to the version of the analysis system that became operational on 14<sup>th</sup> January 2003. The dimension of the analysis control vector was 2842383, and the rank of the observation error covariance matrix was 604981. The cost function included a  $J_c$  term to penalize rapid oscillations in the time evolution of the analysis increment for divergence. This term provides a significant amount of information to the analysis, and its contribution is included in the estimated entropy reduction. The estimates shown in Figure 1 correspond to different sequences of random numbers. Each sequence results in a different estimate of the trace of



$\log_2 (J_z^n)^{-1}$ , as a result of the random errors introduced by the trace estimator (equation 11). For each sequence, Bai *et al.*'s (1996) algorithm generates upper and lower bounds on the estimated entropy reduction. These bounds are shown as a function of the number of Lanczos iterations. One cost-function gradient calculation is required for each iteration. Note that the upper and lower bounds converge to within the error introduced by the randomization after about 25 iterations. This is less than the number of iterations of minimization performed in a typical analysis.

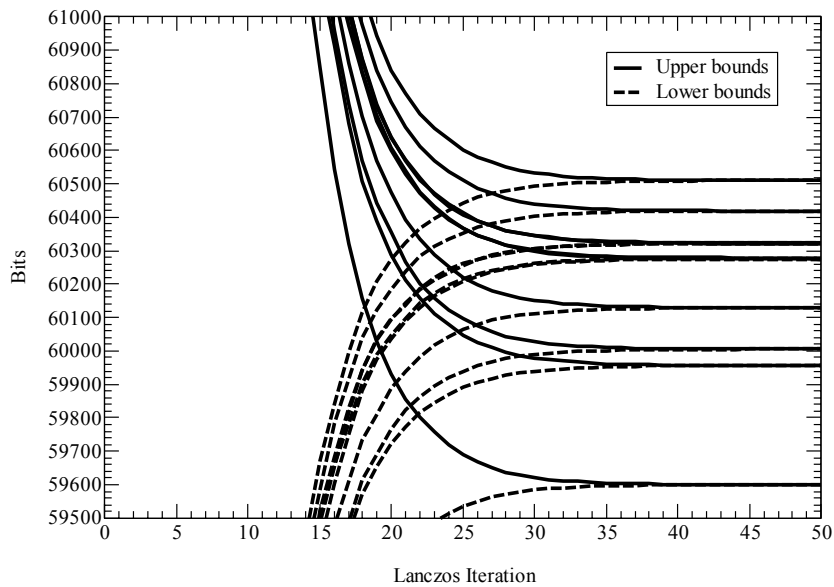


Figure 1: Estimates of Entropy Reduction for the ECMWF 4dVar System.

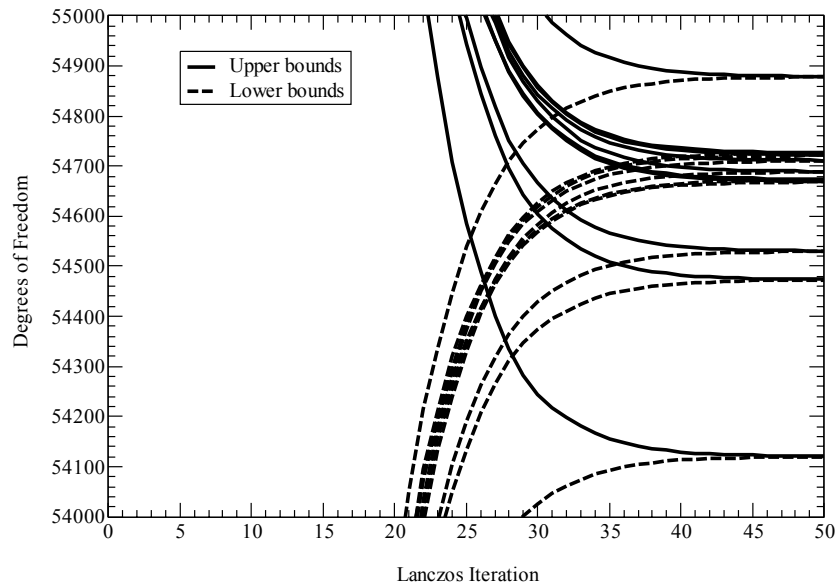


Figure 2: Estimates of Degrees of Freedom for Signal for the ECMWF 4dVar System.

If the random error in the estimates of information content is assumed to be Gaussian, we may assign confidence intervals for the information content using Student's  $t$  distribution (see, for example, Barlow 1989). The 95% confidence interval is  $60181.7 \pm 181.6$ .

Figure 2 shows corresponding estimates of the degrees of freedom for signal. Again, the bounds converge rapidly and the randomization error is small. The 95% confidence interval is  $54619.0 \pm 140.4$ .

The similar values for  $S$  and  $d$  may be understood from the spectral properties of  $\mathbf{P}_z^a$ . As a result of the transformation  $\mathbf{L}$ , all the eigenvalues of  $\mathbf{P}_z^a$  lie between zero and one, and the vast majority are close to one. Writing  $\varepsilon_i = 1 - \lambda_i$ , we have:

$$d = \sum_i \varepsilon_i$$

and

$$S = -\frac{1}{2} \sum_i \log_2(1 - \varepsilon_i)$$

For  $\varepsilon_i$  close to zero,  $-\frac{1}{2} \log_2(1 - \varepsilon_i) \approx \frac{\varepsilon_i}{2 \ln 2} \approx 0.72\varepsilon_i$ . Relatively few terms have  $\varepsilon_i$  close to one. These terms contribute much more to  $S$  than to  $d$ , giving a value of  $S$  somewhat larger than  $0.72d$ .

## 5.1 Interpretation

At first sight, the values of entropy and of degrees of freedom for signal calculated for the full analysis system are surprisingly small. On average, each observation contributes only 0.1 bits of information, and roughly 10 observations are required for each well-observed degree of freedom. Moreover, only about 2% of the degrees of freedom represented by the dimension of the analysis control vector are "for signal".

The degrees of freedom of the control vector are mostly accounted for by the 5 primary analysis variables: vorticity, divergence, temperature, specific humidity and ozone. However, some of these degrees of freedom correspond to gravity-wave oscillations. These are unlikely to be well observed, since gravity waves generally have small amplitude in the atmosphere. Discounting gravity waves reduces the observable degrees of freedom to 3 variables: a balanced dynamical variable, together with ozone and specific humidity.

It is likely that the dynamical fields are significantly better observed than the ozone or specific humidity. To confirm this, entropy and degrees of freedom for signal were estimated for an analysis in which the assumed standard deviations of background error for specific humidity and ozone were multiplied by  $10^{-3}$ . In this case, essentially no information was provided to the analysis by observations of humidity or ozone. As a consequence, degrees of freedom for signal and entropy were both reduced. However, the reduction was small: roughly 6000 degrees of freedom and 9000 bits respectively. This leaves approximately 48000 degrees of freedom to describe the balanced dynamical flow. This number of degrees of freedom for signal could be achieved, for example, by observing the balanced flow to good accuracy everywhere with a vertical resolution roughly one third that of the model, and to a horizontal resolution of T48 (roughly 400km). A value of 48000 for  $S$  corresponds to halving the standard deviation of error in these 48000 degrees of freedom (Eyre, 1990).





## 5.2 Example

As an example of the possible utility of the algorithm, Figure 3 and Figure 4 show the change in entropy reduction and degrees of freedom for signal that results from the removal of certain satellite data. The thick solid and dotted curves show the same upper and lower bounds as are plotted in Figure 1 and Figure 2. This analysis included HIRS data from the NOAA 16 satellite and AMSU-A data from both NOAA 15 and NOAA 16. The thin lines show upper and lower bounds for three randomized trace estimates of entropy reduction (Figure 3) and degrees of freedom for signal (Figure 4) for an analysis in which data from the AMSU-A and HIRS instruments was removed. Clearly, data from these instruments contribute significantly to the information content of the analysis.

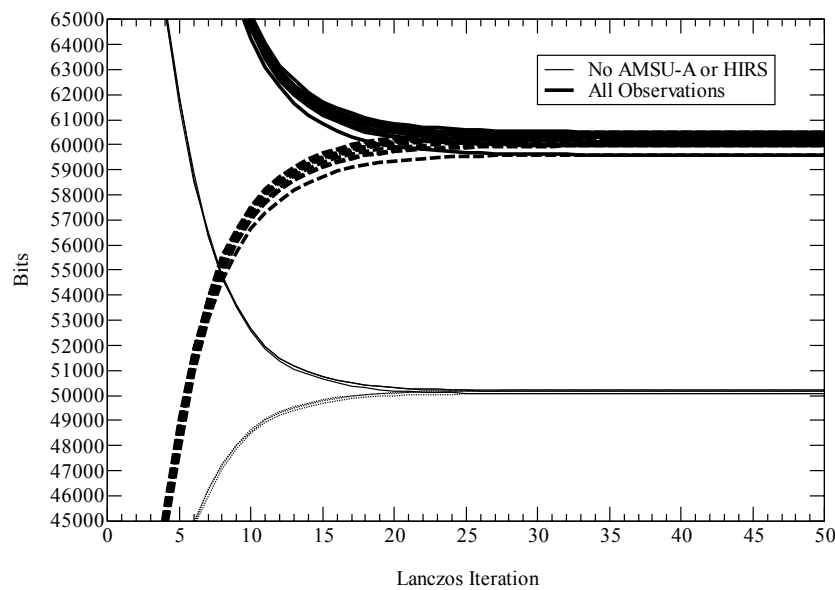


Figure 3: Estimates of entropy reduction for analyses with all observations (thick lines) and analyses with AMSU-A and HIRS data excluded (thin lines).

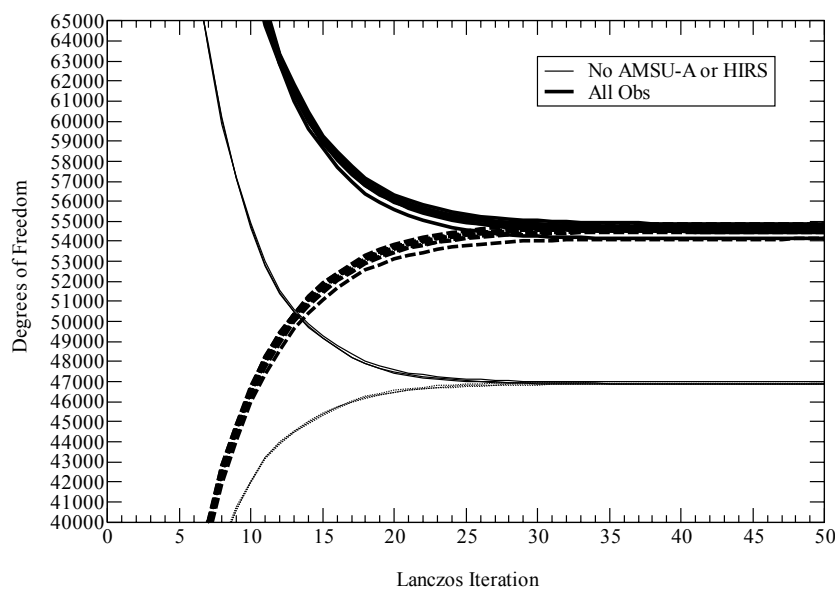


Figure 4: Estimates of degrees of freedom for signal for analyses with all observations (thick lines) and analyses with AMSU-A and HIRS data excluded (thin lines).

## 6. Verification of the Method

### 6.1 Comparison with Direct Calculation for Few Observations

To check the correctness of the algorithm, an analysis was performed for which the number of observations was drastically reduced, by eliminating all observations outside a  $2^\circ \times 2^\circ$  area centred on  $45^\circ\text{N}$ ,  $45^\circ\text{W}$ . Only 40 observations remained, all of them radiances from 5 profiles of the AMSU-A instrument. The  $J_c$  term of the analysis cost function was also removed. As a consequence, only 40 eigenvalues of the analysis Hessian differed from one, and these could be determined to good accuracy using the Lanczos algorithm. Knowledge of the full eigenspectrum of the Hessian allows direct calculation of  $S$  and  $d$ , and in this case gives  $S=2.2219$  and  $d=2.3418$ .

Estimates of  $S$  and  $d$  were calculated for ten different random vectors  $\mathbf{u}$ . The estimates are shown in Figure 5 and Figure 6. Thick dashed horizontal lines mark the directly calculated values. Note that the upper and lower bounds converge within one or two iterations. Note also that the variance introduced by the randomized trace estimate is relatively large for both  $d$  and  $S$ . For  $d$ , this is because the inverse of the analysis Hessian in this case is very close to the identity matrix, so that its trace is close to the dimension of the control vector,  $N$ . The randomized trace calculation estimates the trace of the inverse of the Hessian extremely accurately, with a random error of a few parts per million, but this corresponds to a large relative error in  $d = N - \text{trace}\left(\left(J_z''\right)^{-1}\right)$ . In the case of  $S$ , the diagonal elements of the log of the Hessian are close to zero, so that the conditions (decay away from the diagonal) required for an accurate estimate of the trace are not satisfied.

The means of the estimated values are  $\bar{S} = 2.3723$  and  $\bar{d} = 2.6162$ . Assuming that the randomized values of  $S$  and  $d$  have Gaussian distributions with means equal to the directly calculated values, but with unknown variances, we may use Student's  $t$  distribution to evaluate the deviations of the sample means from the directly-calculated values. The probabilities that deviations at least as large as those obtained could arise by chance are 0.642 for  $S$  and 0.453 for  $d$ . The observed deviations are therefore consistent with the hypothesis that  $\bar{S}$  and  $\bar{d}$  are drawn from distributions with means equal to the directly-calculated values of  $S$  and  $d$ .

The value of entropy-reduction for this experiment compares well with that given by Eyre (1990) for a single AMSU-A profile. (Eyre gives a value of around 4.36 over sea (*op. cit.* table 7), but uses a definition of entropy that is a factor of two larger than that used in this paper.) The fact that our 5 profiles do not provide significantly more information than his single profile is probably due to a combination of factors. First, the profiles are in close proximity. The assumed horizontal correlation of background error introduces a degree of redundancy into the information provided by the observations within the scale length of the correlations (about 300km for temperature). Second, the variances of background error for temperature used in the current ECMWF assimilation system are significantly smaller than those used by Eyre. Third, Eyre (*op. cit.*) uses all 15 AMSU-A channels for his retrieval, whereas an average of 8 channels per profile were used by our analysis system for the 5 selected profiles.

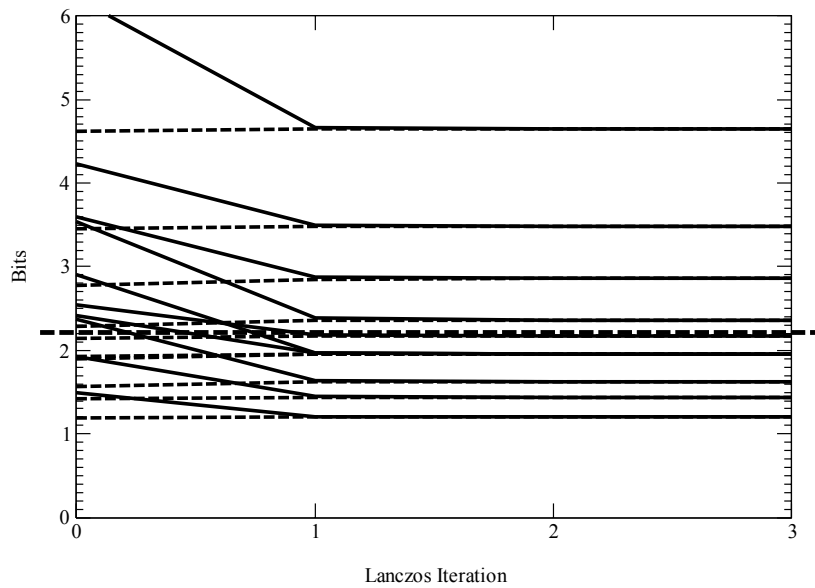


Figure 5: Estimates of Entropy Reduction for 40 AMSU-A radiances.

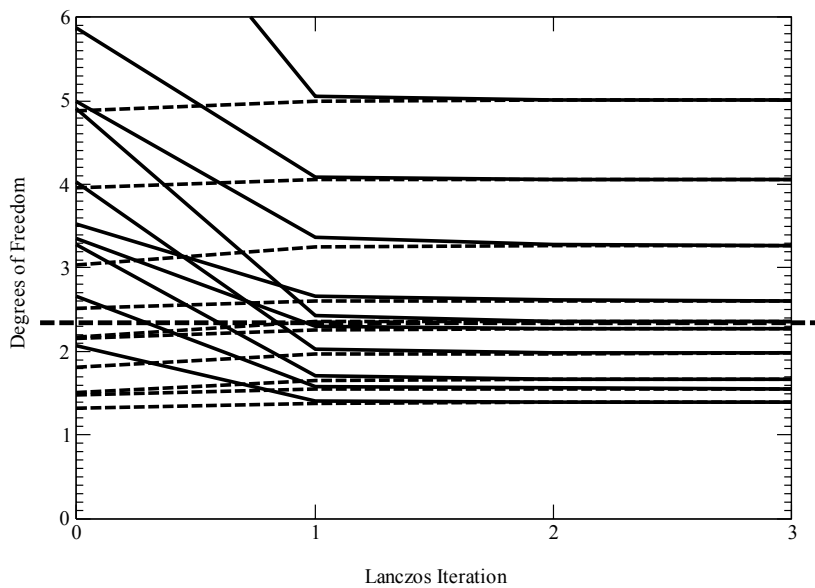


Figure 6: Estimates of Degrees of Freedom for Signal for 40 AMSU-A radiances.

## 6.2 Comparison with an Alternative Method

Wahba *et al.* (1995) give an alternative algorithm for estimating the degrees of freedom for signal for a large variational analysis system. Their algorithm requires observations to be perturbed by amounts characteristic of observation error. In the ECMWF analysis system, it is more convenient to perturb the analysis control vector, so we have devised the following variant of Wahba *et al.*'s method.

Let us write the analysis as a function of the background and observations:

$$\mathbf{x}_a(\mathbf{x}_b, \mathbf{y}) = \mathbf{P}^a (\mathbf{B}^{-1} \mathbf{x}_b + \mathbf{H}^T \mathbf{R}^{-1} \mathbf{y}) \tag{12}$$

Now consider the difference between two analyses with the same observations, but with different backgrounds,  $\mathbf{x}_b$  and  $\mathbf{x}_b + \mathbf{x}_\varepsilon$  :

$$\mathbf{x}_a(\mathbf{x}_b + \mathbf{x}_\varepsilon, \mathbf{y}) - \mathbf{x}_a(\mathbf{x}_b, \mathbf{y}) = \mathbf{P}^a \mathbf{B}^{-1} \mathbf{x}_\varepsilon \quad (13)$$

Defining  $\mathbf{z}_a(\cdot, \cdot) = \mathbf{L} \mathbf{x}_a(\cdot, \cdot)$ , and  $\mathbf{z}_\varepsilon = \mathbf{L} \mathbf{x}_\varepsilon$ , we have:

$$\mathbf{z}_a(\mathbf{x}_b + \mathbf{x}_\varepsilon, \mathbf{y}) - \mathbf{z}_a(\mathbf{x}_b, \mathbf{y}) = \mathbf{P}_z^a \mathbf{z}_\varepsilon \quad (14)$$

In particular, we may choose  $\mathbf{z}_\varepsilon$  to be a vector whose elements take the values  $\pm 1$  randomly and independently with probability  $\frac{1}{2}$ . In this case, Hutchinson's (1989) estimate for the trace of  $\mathbf{P}_z^a$  is given by:

$$\mathbf{z}_\varepsilon^T \mathbf{P}_z^a \mathbf{z}_\varepsilon = \mathbf{z}_\varepsilon^T (\mathbf{z}_a(\mathbf{x}_b + \mathbf{x}_\varepsilon, \mathbf{y}) - \mathbf{z}_a(\mathbf{x}_b, \mathbf{y})) \quad (15)$$

The corresponding estimate of degrees of freedom for signal is:

$$d \approx N - \mathbf{z}_\varepsilon^T (\mathbf{z}_a(\mathbf{x}_b + \mathbf{x}_\varepsilon, \mathbf{y}) - \mathbf{z}_a(\mathbf{x}_b, \mathbf{y})) \quad (16)$$

The computational cost of this estimate is equivalent to that of two analyses. This is also the case for Wahba *et al.*'s (1995) method. In comparison, the computational cost of the method proposed in section 4 is somewhat less than the cost of one analysis. This is because the upper and lower bounds on the randomized trace estimate tend to converge to sufficient accuracy in fewer iterations than are required to converge the analysis to acceptable accuracy. Moreover, the latter method provides an estimate of entropy reduction for negligible additional computational effort.

Although computationally more expensive, the algorithms represented by equation 16 and described by Wahba *et al.* (1995) have some advantage over the method proposed in section 4. Variational data assimilation uses an iterative procedure to calculate the analysis. In the case that the iterative procedure is stopped before full convergence has been achieved, the analysis will not be given by equation 12, but will in fact be a complicated nonlinear function of the background and observations. In addition, the analysis cost function may itself contain non-quadratic terms that lead to a nonlinear relationship between the analysis and the background and observations. Despite these nonlinearities, it is still possible to calculate the expression on the right hand side of equation 16 (or the equivalent expression involving  $\mathbf{A}$ , given by Wahba *et al.*, 1995, equation 3.9), and define an effective number of degrees of freedom for signal as:

$$\tilde{d} = N - E \left[ \mathbf{z}_\varepsilon^T (\mathbf{z}_a(\mathbf{x}_b + \mathbf{x}_\varepsilon, \mathbf{y}) - \mathbf{z}_a(\mathbf{x}_b, \mathbf{y})) \right] \quad (17)$$

where  $E$  denotes expectation.

It is interesting to consider the evolution of  $\tilde{d}$  as the minimization proceeds. If the starting point for the minimization is the background, then initially we will have  $\mathbf{z}_a(\mathbf{x}_b + \mathbf{x}_\varepsilon, \mathbf{y}) - \mathbf{z}_a(\mathbf{x}_b, \mathbf{y}) = \mathbf{z}_\varepsilon$ , so that  $\tilde{d} = 0$ . During subsequent iterations, both the perturbed analysis and the unperturbed analysis will move towards the observations, and away from their respective backgrounds. This will decrease the norm of the difference between them. The tendency will be for  $E \left[ \mathbf{z}_\varepsilon^T (\mathbf{z}_a(\mathbf{x}_b + \mathbf{x}_\varepsilon, \mathbf{y}) - \mathbf{z}_a(\mathbf{x}_b, \mathbf{y})) \right]$  to decrease and for  $\tilde{d}$  to increase. Effectively, the iterations of the minimization algorithm act to transfer information from the observations to the analysis.



The estimate of  $d$  calculated using the algorithm described in section 4 corresponds to the value of  $d$  that would be obtained for a fully converged minimization. We suggest that the difference between this estimate and  $\tilde{d}$  might provide a useful diagnostic to determine whether an analysis system performs enough iterations to extract most of the information from the observations. (Wahba (1987) has suggested that generalized cross-validation might also be used for this purpose.)

Figure 7 shows estimates of  $d$  for the case shown in Figure 1 and Figure 2, calculated using equation 15 with a particular choice of random vector. The estimates are shown as a function of the number of iterations of the conjugate gradient algorithm that was used to determine the perturbed and unperturbed analyses. (For this experiment, the minimization algorithm's objective stopping criterion was suppressed, so that a prescribed number of iterations would be performed.) The estimated value of  $d$  increases with the number of iterations used for the minimizations, eventually saturating after 50 to 100 iterations. The dashed horizontal line shows the estimate of  $d$  (after the upper and lower bounds have converged) calculated using the algorithm presented in section 4 for the same choice of random vector. This estimate of  $d$  is roughly 2% larger than the value estimated using equation 15 with either 100 or nearly 200 iterations. It is likely that this discrepancy results from the different effects of rounding errors in the two algorithms, rather from a lack of convergence of the perturbed and unperturbed analyses. After nearly 200 iterations the norm of the gradient of the cost function for both analyses had been reduced by more than 12 orders of magnitude, indicating that the analyses themselves were very accurately determined.

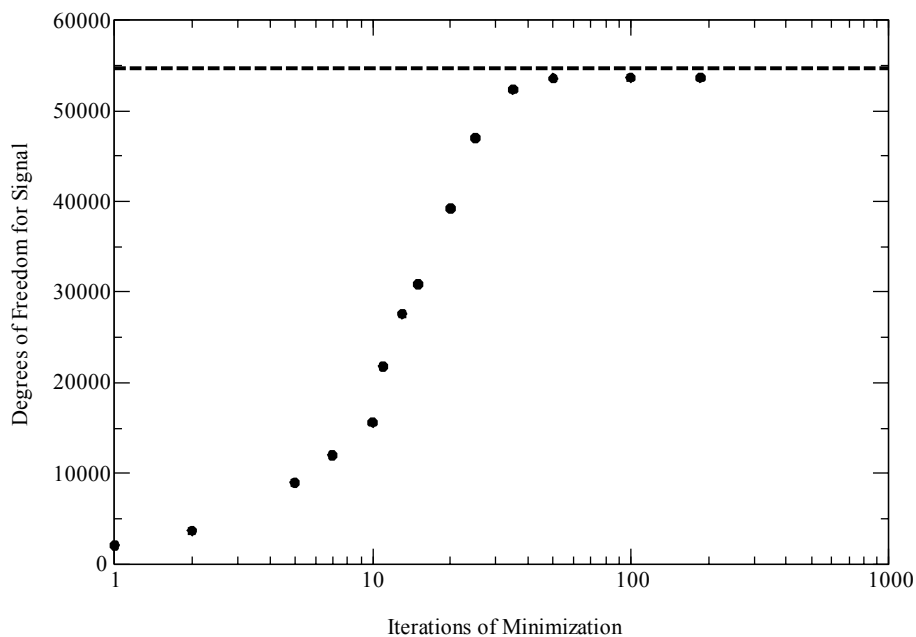


Figure 7: Estimates of degrees of freedom for signal using equation 16, as a function of the number of iterations. The dashed line shows the converged estimate given by the algorithm described in section 4

### 6.3 Another approach

Another approach to calculating degrees of freedom for signal has been developed by Cardinali (personal communication). She approximates the covariance matrix of analysis error as:

$$\mathbf{P}^a \approx \sum_{i=1}^K \mathbf{u}_i \mathbf{u}_i^T - \sum_{i=1}^L \mathbf{w}_i \mathbf{w}_i^T \quad (18)$$

using the method described by Fisher and Courtier (1995). Here,  $K$  and  $L$  are in the range of a few tens to a few hundreds.

Given this approximation, we may write the influence matrix as:

$$\mathbf{A} \approx \sum_{i=1}^K \tilde{\mathbf{u}}_i \tilde{\mathbf{u}}_i^T - \sum_{i=1}^L \tilde{\mathbf{w}}_i \tilde{\mathbf{w}}_i^T \quad (19)$$

where  $\tilde{\mathbf{u}}_i = \mathbf{R}^{-1/2} \mathbf{H} \mathbf{u}_i$  and  $\tilde{\mathbf{w}}_i = \mathbf{R}^{-1/2} \mathbf{H} \mathbf{w}_i$ . It is straightforward and fast to calculate the individual elements of this approximation. In particular, the individual elements of the diagonal may be determined. Cardinali (personal communication) interprets these elements as indicating the importance of individual observations to the analysis. The sum of these diagonal elements is an approximation to the degrees of freedom for signal. The disadvantage of this approach is that it is difficult to determine the accuracy to which the degrees of freedom for signal are determined. We suggest that the methods presented in this paper might be used to determine the mean error of the diagonal elements estimated using Cardinali's method.

## 7. Summary and concluding remarks

An algorithm has been presented that allows accurate estimates to be obtained for the entropy reduction and degrees of freedom for signal in a large-scale variational analysis system. There are several obvious applications for the method. An objective calculation of the information provided by different observing systems would provide valuable information to designers of new observing systems or special observing campaigns. Many satellite data require thinning before they can be assimilated. This is done partly to reduce the spatial correlation of observation errors, but also has a large impact on the computational cost of the analysis. Knowledge of the information content provided by the observations should help in choosing appropriate thinning densities. For some instruments, such as the space-borne Doppler wind lidar of the European Space Agency's ADM-Aeolus mission (Stoffelen *et al.* 2003), there is a trade-off between the accuracy of individual measurements and their spatial resolution. The information content could be a useful tool in determining the optimum trade-off.

It is worth noting that the diagnostics presented in this paper depend strongly on the equivalence between the inverse of the analysis Hessian and the covariance matrix of analysis error. This equivalence holds only if the covariance matrices of background and observation error are correctly specified. (In the case of 4dVar, model error is also neglected.) Incorrectly specified covariance matrices are likely to give misleading estimates of  $S$  and  $d$ . Moreover, it should be noted that the introduction of new observations into the analysis system is likely to change the characteristics of background error, necessitating a re-tuning of the background error covariance matrix. This re-tuning may significantly change the estimated information content of the new observations. In fact, the need for a re-tuning of the background error covariance matrix is an indication of temporal redundancy in the information provided by the observations: assimilation of the new observations during the preceding analysis cycle has reduced the background errors for the current cycle, and as a consequence, has reduced the impact of the new observations on the current cycle.



In section 5.1, we attempted to interpret degrees of freedom for signal and entropy for the full analysis system in terms of an effective resolution for the well-observed component of the analysis. However, these global numbers are difficult to interpret, and are likely be of limited practical use. Of much greater interest is the comparative information content of different observation types, and the degree to which there is redundancy between the information provided by different types of observation. We calculated in section 5.2 the change in entropy and degrees of freedom for signal caused by the removal of data from the AMSU-A and HIRS instruments. In a future paper we intend to conduct a systematic investigation of the information content of all the observation types currently used in the ECMWF analysis system.

There are several other uses for Bai *et al.*'s (1996) algorithm. Golub and von Matt (1997) describe how it may be used to calculate the generalised cross-validation function (Golub, Heath and Wahba, 1979). A second possible application of the algorithm is in maximum-likelihood parameter estimation. Dee (1995) considers a covariance matrix  $\mathbf{S}$  that is a function of a few parameters  $\boldsymbol{\alpha}$ . He shows that the maximum likelihood estimate for  $\mathbf{S}(\boldsymbol{\alpha})$ , given a sample  $\mathbf{v}$  drawn from a zero-mean normal distribution with covariance matrix  $\mathbf{S}$  is given by the parameters that minimize the function:

$$f(\boldsymbol{\alpha}) = \log|\mathbf{S}(\boldsymbol{\alpha})| + \mathbf{v}^T \mathbf{S}^{-1}(\boldsymbol{\alpha}) \mathbf{v}$$

Both terms on the right hand side of this expression may be evaluated using Bai *et al.*'s (1996) algorithm.

## Acknowledgements

The author would like to thank Tony Hollingsworth, Erik Andersson, Lars Isaksen and Jean Noël Thépaut for their helpful comments.

## References

- Bai Z., Fahey M. and Golub G.H., 1996: Some Large Scale Matrix Computation Problems. *J. Comput. Appl. Math.*, **74**, 21-89.
- Barlow, R.J., 1989, Statistics: A Guide to the Use of Statistical Methods in the Physical Sciences. *The Manchester Physics Series*, Wiley, 204pp.
- Barry R. and Pace R.K., 1999: A Monte Carlo Estimator of the Log Determinant of Large Sparse Matrices. *Linear Algebra and its Applications*. **289 No.1-3**, 41-54.
- Courtier P., Andersson E., Heckley W., Pailleux J., Vasiljević D., Hamrud M., Hollingsworth A., Rabier F. and Fisher M., 1998: The ECMWF implementation of three dimensional variational assimilation (3DVar). Part 1: Formulation. *Quart. J. Roy. Meteor. Soc.*, **124**, 1783-1808.
- Dee D.P., 1995: On-line Estimation of Error Covariance Parameters for Atmospheric Data Assimilation. *Mon. Wea. Rev.*, **123**, 1128-1144.
- Eyre, J. R., 1990, The Information Content of Data From Satellite Sounding Systems: A Simulation Study. *Quart. J. Roy. Meteor. Soc.*, **116**, 401-434.



- Fisher M. and Courtier P., 1995: Estimating the Covariance Matrices of Analysis and Forecast Error in Variational Data Assimilation. *ECMWF Research Department Technical Memorandum 220*.
- Fourrié, N. and Thépaut, J.-N., 2003, Validation of the NESDIS Near Real Time AIRS Channel Selection. *Submitted to Quart. J. Roy. Meteor. Soc.*
- Gauthier, P., 1992: Chaos and quadri-dimensional data assimilation: a study based on the Lorenz model. *Tellus*, **44A**, 2-17.
- Gauthier P., Charette C., Fillion L., Koclas P. and Laroche S., 1999: Implementation of a 3D Variational Data Assimilation System at the Canadian Meteorological Centre. Part 1: The Global Analysis. *Atmosphere Ocean*. **XXXVII No. 2**, 103-156.
- Golub G.H., Heath M. and Wahba G., 1979: Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter. *Technometrics*, **21**, 215-223.
- Golub G.H. and Meurant G., 1993: Matrices, Moments and Quadrature. in *Numerical Analysis*, eds. D.F. Griffiths and G.A. Watson, Longman, Essex, England, 1994, 105-156.
- Golub G.H. and von Matt U., 1997: Generalized cross-validation for large-scale problems. *J. Comput. Graph. Stat.*, **6**, 1-34.
- Golub G.H. and Strakos Z., 1994: Estimates in Quadratic Formulas. *Numer. Algor.*, **8**, 241-268.
- Gustafsson N., Berre L., Hörnquist S., Huang X.-Y., Lindskog M., Navascués B., Mogensen S. and Thorsteinsson S., 2001: Three-dimensional variational data assimilation for a limited area model. *Tellus*, **53A**, 425-446.
- Hutchinson M., 1989: A Stochastic Estimator of the Trace of the Influence Matrix for Laplacian Smoothing Splines. *Commun. Statist. Simula.*, **18**, 1059-1076.
- Klinker E., Rabier F., Kelly G. and Mahfouf J.-F., 2000: The ECMWF operational implementation of four dimensional variational assimilation. Part III: Experimental results and diagnostics with operational configuration. *Quart. J. Roy. Meteor. Soc.*, **126**, 1191-1215.
- Liu K.-F., 2000: A Noisy Monte Carlo Algorithm with Fermion Determinant. *Chinese Journal of Physics*. **38 3-II**, 605-614.
- Lorenc A.C., Ballard S.P., Bell R.S., Ingleby B., Andrews P.L.F., Barker D.M., Bray J.R., Clayton A.M., Dalby T., Li D., Payne T.J. and Saunders F.W., 2000: The Met. Office Global three-dimensional variational data assimilation scheme. *Quart. J. Roy. Meteorol. Soc.*, **126**, pp 2991-3012.
- Mahfouf J.-F. and Rabier F., 2000: The ECMWF operational implementation of four dimensional variational assimilation. Part II: Experimental results with improved physics. *Quart. J. Roy. Meteor. Soc.*, **126**, 1171-1190.
- Rabier, F. and Courtier, P., 1992: Four-dimensional assimilation in the presence of baroclinic instability. *Quart. J. Roy. Meteor. Soc.*, **118**, 649-672.





Rabier, F., Fourrié, N., Chafaï, D., and Prunet, P., 2002: Channel selection methods for infrared atmospheric sounding interferometer radiances. *Quart. J. Roy. Meteor. Soc.*, **128**, 1011-1027.

Rabier F., Järvinen H., Klinker E., Mahfouf J.-F., Simmons A., 2000: The ECMWF operational implementation of four dimensional variational assimilation. Part I: Experimental results with simplified physics. *Quart. J. Roy. Meteor. Soc.*, **126**, 1143-1170.

Rodgers, C.D., 1996: Information Content and Optimisation of High Spectral Resolution Measurements. *Optical Spectroscopic Techniques and Instrumentation for Atmospheric and Space Research*, **SPIE Volume 2830**, 136-147.

Rodgers, C.D., 2000: Inverse Methods for Atmospheres: Theory and Practice. *Series on Atmospheric, Oceanic and Planetary Physics*, World Scientific Publ., Singapore, 238pp.

Shannon, C.E. and Weaver, W., 1949: The Mathematical Theory of Communication. *University of Illinois Press, Urbana*.

Stoffelen A., Pailleux J., Källén E., Vaughan J.M., Isaksen L., Flamant P., Wergen W., Andersson E., Schyberg H., Culoma A., Meynard R., Endemann M. and Ingmann P., 2003: The Atmospheric Dynamics Mission for Global Wind Field Measurement. *Submitted to Bull. Amer. Meteorol. Soc.*

Wahba G., 1987: Three topics in ill posed problems. *Proc. of the Alpine-U.S. Seminar on Inverse and Ill Posed Problems*. Academic Press, H.Engl and C. Groetsch, eds., 37-51.

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

Wahba G., Johnson D.R., Gao F., Gong J., 1995: Adaptive Tuning of Numerical Weather Prediction Models: Randomized GCV in Three- and Four-Dimensional Data Assimilation. *Monthly Weather Review*, **123**, 3358-3369.

Wahba G. and Wendleberger J., 1980: Some New Mathematical Methods for Variational Objective Analysis using Splines and Cross-Validation. *Mon. Wea. Rev.*, **108**, 1122-1145.

## Appendix A: Equivalence of Two Definitions for Degrees of freedom for signal

The covariance matrix of analysis error for the control variable may be written as:

$$\begin{aligned}\mathbf{P}_z^a &= \mathbf{L}(\mathbf{H}^T \mathbf{R}^{-1} \mathbf{H} + \mathbf{L}^T \mathbf{L})^{-1} \mathbf{L}^T \\ &= (\mathbf{L}^{-T} \mathbf{H}^T \mathbf{R}^{-1} \mathbf{H} \mathbf{L}^{-1} + \mathbf{I})^{-1}\end{aligned}$$

where  $\mathbf{H}$  and  $\mathbf{R}$  have their conventional meanings, and where  $\mathbf{L}^T \mathbf{L} = \mathbf{B}^{-1}$ , the covariance matrix of background error. ( $\mathbf{B}$  is assumed to be non-singular.)

Now consider the eigendecomposition  $\mathbf{L}^{-T} \mathbf{H}^T \mathbf{R}^{-1} \mathbf{H} \mathbf{L}^{-1} = \mathbf{V} \mathbf{\Lambda} \mathbf{V}^T$ . We have:

$$\mathbf{P}_z^a = \mathbf{V}(\mathbf{\Lambda} + \mathbf{I})^{-1} \mathbf{V}^T.$$

That is, the eigenvalues  $\mu$  of  $\mathbf{P}_z^a$  are related to the eigenvalues  $\lambda$  of  $\mathbf{L}^{-T} \mathbf{H}^T \mathbf{R}^{-1} \mathbf{H} \mathbf{L}^{-1}$  by:

$$\mu_i = \frac{1}{1 + \lambda_i}.$$

Next, consider the eigenvalues  $\nu$  of the influence matrix  $\mathbf{A} = \mathbf{R}^{-1/2} \mathbf{H} \mathbf{P}_z^a \mathbf{H}^T \mathbf{R}^{-1/2}$ . If  $\mathbf{y}$  is the eigenvector corresponding to  $\nu$ , then:

$$\mathbf{R}^{-1/2} \mathbf{H} \mathbf{P}_z^a \mathbf{H}^T \mathbf{R}^{-1/2} \mathbf{y} = \nu \mathbf{y}$$

Now let  $\mathbf{x} = \mathbf{L}^{-T} \mathbf{H}^T \mathbf{R}^{-1/2} \mathbf{y}$ . Rewriting the eigenvector equation in terms of  $\mathbf{x}$  gives:

$$\mathbf{L}^{-T} \mathbf{H}^T \mathbf{R}^{-1} \mathbf{H} \mathbf{L}^{-1} \mathbf{P}_z^a \mathbf{x} = \nu \mathbf{x}.$$

In terms of the eigendecomposition of  $\mathbf{L}^{-T} \mathbf{H}^T \mathbf{R}^{-1} \mathbf{H} \mathbf{L}^{-1}$ , this is:

$$\mathbf{V} \mathbf{\Lambda} (\mathbf{\Lambda} + \mathbf{I})^{-1} \mathbf{V}^T \mathbf{x} = \nu \mathbf{x}.$$

There are two possible ways in which this equation may be satisfied. The first possibility is that  $\mathbf{x}$  is identically zero. In this case, the corresponding vector  $\mathbf{y}$  lies in the nullspace of  $\mathbf{A}$ , so that  $\nu = 0$ . Alternatively, we must have  $\nu = \lambda_i / (1 + \lambda_i)$  for some eigenvalue  $\lambda_i$  of  $\mathbf{L}^{-T} \mathbf{H}^T \mathbf{R}^{-1} \mathbf{H} \mathbf{L}^{-1}$ .

If the dimension  $M$  of  $\mathbf{A}$  is smaller than the dimension  $N$  of  $\mathbf{L}^{-T} \mathbf{H}^T \mathbf{R}^{-1} \mathbf{H} \mathbf{L}^{-1}$ , then some eigenvectors of the latter matrix cannot be written in the form  $\mathbf{x} = \mathbf{L}^{-T} \mathbf{H}^T \mathbf{R}^{-1/2} \mathbf{y}$ . Let us consider one such eigenvector.

We have:

$$\mathbf{L}^{-T} \mathbf{H}^T \mathbf{R}^{-1} \mathbf{H} \mathbf{L}^{-1} \mathbf{x} = \lambda \mathbf{x}$$

Now let  $\mathbf{y} = \mathbf{R}^{-1/2} \mathbf{H}^T \mathbf{L}^{-1} \mathbf{x}$ . Then  $\mathbf{L}^{-T} \mathbf{H}^T \mathbf{R}^{-1/2} \mathbf{y} = \lambda \mathbf{x}$ . Clearly, we must have  $\lambda = 0$  to avoid contradicting the assumption that  $\mathbf{x}$  cannot be written in the form  $\mathbf{x} = \mathbf{L}^{-T} \mathbf{H}^T \mathbf{R}^{-1/2} \mathbf{y}$ .



We have shown that for  $i=1\dots N$ , either  $\lambda_i = 0$  (in which case  $\mu_i = 1$ ), or there exists an eigenvalue  $v_j$  of  $\mathbf{A}$  such that  $v_j = 1 - \mu_i$ . Likewise, for  $j=1\dots M$ , either  $v_j = 0$ , or  $v_j = \lambda_i / (1 + \lambda_i) = 1 - \mu_i$  for some  $i$ .

In particular, we may conclude that  $\sum_{j=1}^M v_j = \sum_{i=1}^N (1 - \mu_i)$ . That is,  $\text{trace}(\mathbf{A}) = N - \text{trace}(\mathbf{P}_z^d) = d$ .

## Appendix B: A Brief Description of Bai *et al.*'s (1996) Algorithm

Hutchinson's (1989) randomized trace estimate is defined as  $t = \mathbf{u}^T \mathbf{A} \mathbf{u}$ , where  $\mathbf{u}$  is a vector whose elements take the values  $\pm 1$  randomly and independently with probability  $1/2$ . Writing this in terms of elements of  $\mathbf{A}$  and  $\mathbf{u}$ , we have:

$$\begin{aligned} t &= \sum_{i,j} A_{ij} u_i u_j \\ &= \sum_i A_{ii} + \sum_{i \neq j} A_{ij} u_i u_j . \end{aligned}$$

The first sum on the right hand side is the trace of  $\mathbf{A}$ . The second sum is a random variable with zero mean and variance  $2 \sum_{i \neq j} A_{ij}^2$ . For a large matrix whose entries decay with distance from the diagonal, this variance is small compared with the trace, and  $t$  is an accurate and unbiased estimate of the trace.

To evaluate  $\mathbf{u}^T f(\mathbf{A}) \mathbf{u}$ , we write  $\mathbf{A}$  in terms of its eigen-decomposition,  $\mathbf{A} = \mathbf{Q} \mathbf{\Lambda} \mathbf{Q}^T$ :

$$\begin{aligned} \mathbf{u}^T f(\mathbf{A}) \mathbf{u} &= \mathbf{u}^T \mathbf{Q} f(\mathbf{\Lambda}) \mathbf{Q}^T \mathbf{u} \\ &= \tilde{\mathbf{u}}^T f(\mathbf{\Lambda}) \tilde{\mathbf{u}} \\ &= \sum_i f(\lambda_i) \tilde{u}_i^2 , \end{aligned}$$

where  $\tilde{\mathbf{u}} = \mathbf{Q}^T \mathbf{u}$ .

The sum may be regarded as a Riemann-Stieltjes integral:

$$\sum_i f(\lambda_i) \tilde{u}_i^2 = \int_{\lambda_{\min}}^{\lambda_{\max}} f(\lambda) d\mu$$

where  $\mu(\lambda)$  is a staircase function with steps of height  $\tilde{u}_i^2$  at each of the eigenvalues  $\lambda_i$ .

Expressing  $\mathbf{u}^T f(\mathbf{A}) \mathbf{u}$  as an integral allows it to be evaluated using Gauss-type quadrature rules. Bai *et al.* (1996) show that the weights and nodes for the quadrature rules may be calculated from the tri-diagonal matrix  $\mathbf{T}_j$  of coefficients that is generated during the Lanczos algorithm. Specifically, for Gauss quadrature:

$$\int_{\lambda_{\min}}^{\lambda_{\max}} f(\lambda) d\mu = \sum_i w_i^2 f(\theta_i) + R[f]$$

where  $\theta_i$  are the eigenvalues of  $\mathbf{T}_j$ , and  $w_i$  are the first elements of the corresponding eigenvectors.  $R[f]$  is the residual error. The sign of the residual error is known for functions whose even derivatives are all of one sign, and whose odd derivatives are all of the opposite sign. Related quadrature rules, with different residual errors, are given by similar formulae involving the eigenvalues and the first elements of the eigenvectors of slightly modified tri-diagonal matrices. The calculations performed for this paper used Gauss quadrature to provide upper bounds on  $S$  and  $d$ , and Gauss-Radau quadrature to provide lower bounds.