Information Content of Advanced Sounders

Clive D. Rodgers

AOPP, Clarendon Laboratory, Oxford, OX1 3PU, United Kingdom
c.rodgers@physics.ox.ac.uk

ABSTRACT

The latest generation of sounders has far more channels per profile measurement than we are likely to be able to use, so that we need to think carefully to make use of the information content of the data with least computation. The ideas of information theory provide a useful basis on which to understand the information content of measurements, and to develop methods for making best use of it. Shannon information can be used as a tool for optimising instruments and data analysis systems. An illustration is presented for channel selection for high spectral resolution instruments. For developing methods of efficiently assimilating such data, the preservation of information can be used as a constraint on transformations applied to the data before assimilation. An illustration of such a transformation is discussed.

1 Introduction

The latest generation of sounders has far more channels per profile measurement than we are likely to be able to use. For example an AIRS spectrum contains about 2400 spectral points, compared with a few tens of points for a radiometer. IASI will have more, around 8500 spectral points, and the really high resolution instruments, MIPAS and TES include about 16 spectra per profile, with a total number of spectral elements closer to a million.

Consequently we need to think carefully to make best use of this kind of data with the computational power available. We need techniques to make use of the information content of the data with least computation. This will entail a rethink of the current techniques of assimilation of retrievals and of radiances.

2 Information

We start with the concept of information content itself. There are many definitions, of which three are particularly useful. The first is Shannon's Information Content, which is a a scalar quantity which relates prior knowledge to posterior knowledge, rather like a signal/noise ratio. Consequently it can be used as a target for optimisation. The second is the Fisher Information Matrix, which is a matrix measure of size and shape of the region of state space containing the uncertainty of our knowledge of the state. It refers only to posterior knowledge, and not at all to prior knowledge. The third is another scalar, the degrees of freedom for signal, which is an effective number of independent quantities whose uncertainty has been improved by the measurement, and so also related prior to posterior knowledge.

These concepts can all be developed from the Bayesian perspective, in which knowledge is represented in terms of a set of probability density functions (*pdf*s): $P(\mathbf{x})$, the *a priori pdf* of the state – describing what we know about the state before we make the measurement; $P(\mathbf{y})$, the *a priori pdf* of the measurement; $P(\mathbf{x},\mathbf{y})$, is the joint *a priori pdf* of $\mathbf{x}$ and $\mathbf{y}$; $P(\mathbf{y}|\mathbf{x})$, is the *pdf* of the measurement given the state – this depends on experimental

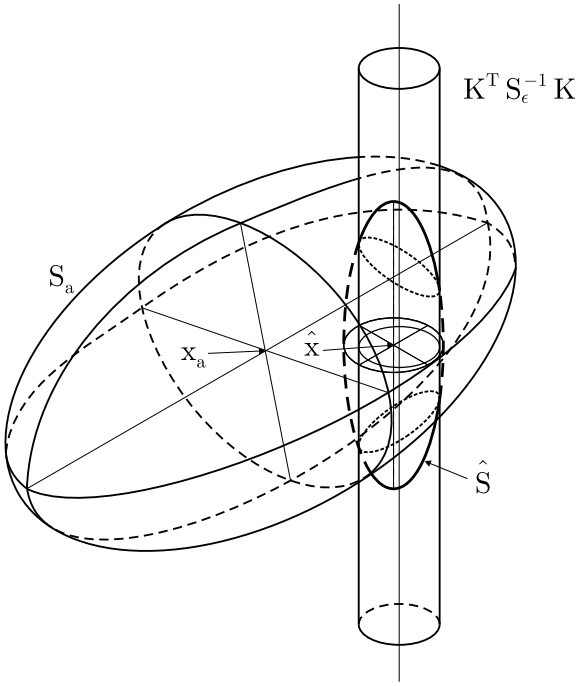

Figure 1: A geometric interpretation of the relationship between the prior state estimate, the measurement mapped into state space, and the posterior estimate, for a 3D state space and a 2D measurement space. The large ellipsoid is a contour of the prior pdf*, the cylinder is a contour of the* pdf *of the state given only the measurement, and the small ellipsoid is a contour of the posterior* pdf.

error and the forward function and $P(\mathbf{x}|\mathbf{y})$, the *pdf* of the state given the measurement – describing what we know about the state after we make the measurement, as illustrated in Figure 1. Information is encapsulated in the relevant *pdf*s, and most useful measures of 'information content' are functions of these *pdf*s.

2.1 Shannon Information

The Shannon information content of a measurement of $\mathbf{x}$ is the change, as a result of making the measurement, in the *entropy* of the probability density function describing our knowledge of $\mathbf{x}$. Entropy is defined by:

$$S\{P\} = -\int P(\mathbf{x}) \log_2\{P(\mathbf{x})/M(\mathbf{x})\} d\mathbf{x} \tag{1}$$

where $M(\mathbf{x})$ is a measure function which we will take it to be constant. Qualitatively, entropy can be thought of as the log of the volume of state space occupied by the *pdf*. The Shannon information content of a measurement is the reduction in entropy between the *pdf* before, $P(\mathbf{x})$, and the *pdf* after, $P(\mathbf{x}|\mathbf{y})$, the measurement:

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

It can be thought of as the log of the ratio of the posterior to prior volumes of state space (the small and large ellipsoids in Figure 1), i.e. the log of a generalisation of signal/noise ratio, measured in bits. For Gaussian *pdf*s, the entropy can be obtained from the covariance $\mathbf{S}$, and the information content becomes:

$$H = \frac{1}{2}\log_2|\mathbf{S}_{\text{prior}}| - \frac{1}{2}\log_2|\mathbf{S}_{\text{posterior}}| \tag{3}$$

2.2 Degrees of freedom for signal and for noise

The state estimate that maximises $P(\mathbf{x}|\mathbf{y})$ in the linear Gaussian case is the one which minimises

$$\chi^2 = [\mathbf{y}-\mathbf{K}\mathbf{x}]^T\mathbf{S}_\varepsilon^{-1}[\mathbf{y}-\mathbf{K}\mathbf{x}] + [\mathbf{x}-\mathbf{x}_a]^T\mathbf{S}_a^{-1}[\mathbf{x}-\mathbf{x}_a] \tag{4}$$

The r.h.s. has initially $m+n$ degrees of freedom, of which n are fixed by choosing $\mathbf{x}$ to be $\hat{\mathbf{x}}$, so the expected value of χ^2 is m. These m degrees of freedom can be assigned to degrees of freedom for noise d_n and degrees of freedom for signal d_s according to:

$$d_n = E\{[\mathbf{y}-\mathbf{K}\hat{\mathbf{x}}]^T\mathbf{S}_\varepsilon^{-1}[\mathbf{y}-\mathbf{K}\hat{\mathbf{x}}]\} \tag{5}$$

and

$$d_s = E\{[\hat{\mathbf{x}}-\mathbf{x}_a]^T\mathbf{S}_a^{-1}[\hat{\mathbf{x}}-\mathbf{x}_a]\} \tag{6}$$

Degrees of freedom for signal is a measure of the number of independent quantities for which information is greater than noise. With some manipulation we can find

$$\begin{aligned} d_s &= \mathrm{tr}((\mathbf{K}^T\mathbf{S}_\varepsilon^{-1}\mathbf{K}+\mathbf{S}_a^{-1})^{-1}\mathbf{K}^T\mathbf{S}_\varepsilon^{-1}\mathbf{K}) \\ &= \mathrm{tr}(\mathbf{K}\mathbf{S}_a\mathbf{K}^T(\mathbf{K}\mathbf{S}_a\mathbf{K}^T+\mathbf{S}_\varepsilon)^{-1}) \qquad (7) \\ d_n &= \mathrm{tr}((\mathbf{K}^T\mathbf{S}_\varepsilon^{-1}\mathbf{K}+\mathbf{S}_a^{-1})^{-1}\mathbf{S}_a^{-1}) + m - n \\ &= \mathrm{tr}(\mathbf{S}_\varepsilon(\mathbf{K}\mathbf{S}_a\mathbf{K}^T+\mathbf{S}_\varepsilon)^{-1}) \qquad (8) \end{aligned}$$

2.3 Independent measurements

If the measurement error covariance is not diagonal, the elements of the $\mathbf{y}$ vector will not be statistically independent, and similarly for any *a priori*. Further, the measurements will not be independent functions of the state if $\mathbf{K}$ is not diagonal. It helps to understand where the information comes from if we transform to a different basis. In the context of Figure refball, we scale state space so that the large ellipsoid is spherical, then rotate to the principal axes of the scaled small ellipsoid. At the same time, we scale measurement space so that the measurement error ellipsiod is spherical. First, statistical independence. Scale the state and measurement spaces to define:

$$\tilde{\mathbf{y}} = \mathbf{S}_\varepsilon^{-\frac{1}{2}}\mathbf{y} \qquad \tilde{\mathbf{x}} = \mathbf{S}_a^{-\frac{1}{2}}\mathbf{x} \tag{9}$$

The transformed covariances $\tilde{\mathbf{S}}_a$ and $\tilde{\mathbf{S}}_\varepsilon$ both become unit matrices, and the forward model becomes:

$$\tilde{\mathbf{y}} = \tilde{\mathbf{K}}\tilde{\mathbf{x}} + \tilde{\varepsilon} \tag{10}$$

where $\tilde{\mathbf{K}} = \mathbf{S}_\varepsilon^{-\frac{1}{2}}\mathbf{K}\mathbf{S}_a^{\frac{1}{2}}$. The solution covariance becomes:

$$\hat{\tilde{\mathbf{S}}} = (\mathbf{I}_n + \tilde{\mathbf{K}}^T\tilde{\mathbf{K}})^{-1} \tag{11}$$

Now make $\tilde{\mathbf{K}}$ diagonal. Rotate both $\mathbf{x}$ and $\mathbf{y}$ to yet another basis, defined by the singular vectors of $\tilde{\mathbf{K}}$:

$$\tilde{\mathbf{y}} = \tilde{\mathbf{K}}\tilde{\mathbf{x}} + \tilde{\varepsilon} \qquad \rightarrow \qquad \tilde{\mathbf{y}} = \mathbf{U}\Lambda\mathbf{V}^T\tilde{\mathbf{x}} + \tilde{\varepsilon} \tag{12}$$

Define:

$$\mathbf{x}' = \mathbf{V}^T\tilde{\mathbf{x}} \quad \mathbf{y}' = \mathbf{U}^T\tilde{\mathbf{y}} \quad \varepsilon' = \mathbf{U}^T\tilde{\varepsilon} \tag{13}$$

The forward model becomes:

$$\mathbf{y}' = \Lambda\mathbf{x}' + \varepsilon' \tag{14}$$

The Jacobian is now diagonal, Λ, and the *a priori* and noise covariances are still unit matrices, hence the solution covariance becomes:

$$\hat{\tilde{\mathbf{S}}} = (\mathbf{I}_n + \tilde{\mathbf{K}}^T\tilde{\mathbf{K}})^{-1} \quad \rightarrow \quad \hat{\mathbf{S}}' = (\mathbf{I}_n + \Lambda^2)^{-1} \tag{15}$$

which is diagonal, and the solution itself is

$$\hat{\mathbf{x}}' = (\mathbf{I}_n + \Lambda^2)^{-1}(\Lambda\mathbf{y}' + \mathbf{x}'_a) \tag{16}$$

not $\hat{\mathbf{x}}' = \Lambda^{-1}\mathbf{y}'$ as you might expect from (1).

We can summarise by noting that elements for which $\lambda_i \gg 1$ or $(1+\lambda_i^2)^{-1} \ll 1$ are well measured, and elements for which $\lambda_i \ll 1$ or $(1+\lambda_i^2)^{-1} \gg 1$ are poorly measured.

2.3.1 Shannon Information in the Transformed Basis

Because it is a ratio of volumes, the linear transformations do not change the Shannon information content. Thus information in the $\mathbf{x}'$, $\mathbf{y}'$ system is given by

$$\begin{aligned} H &= S\{\mathbf{S}'_a\} - S\{\hat{\mathbf{S}}'\} \\ &= -\frac{1}{2}\log(|\mathbf{I}_n|) + \frac{1}{2}\log(|(\Lambda^2+\mathbf{I})^{-1}|) \\ &= \sum_i \frac{1}{2}\log(1+\lambda_i^2) \end{aligned} \tag{17}$$

2.3.2 Degrees of Freedom in the Transformed Basis

The number of independent quantities measured is qualitatively the number of singular vectors for which $\lambda_i \gg 1$. The degrees of freedom for signal is

$$d_s = \sum_i \lambda_i^2(1+\lambda_i^2)^{-1} \tag{18}$$

In the example of Figure 1, it is clear that there would be two small λ's, and one large one (close to unity), so that the number of degrees of freedom for signal would be near two.

Thus for each independent component x'_i, the information content is $\frac{1}{2}\log(1+\lambda_i^2)$, and the degrees of freedom for signal is $\lambda_i^2(1+\lambda_i^2)^{-1}$

3 Data Subsetting for Retrieval/assimilation Efficiency

High spectral resolution instruments provide more channels than can be used, or are needed. There is duplication of information. We need a means of selecting the channels or microwindows which contain most of the information.

For an instrument like AIRS or IASI, where the forward model computes radiances separately for each channel, the basic strategy is to select channels independently and sequentially. Starting with no channels selected: (1) Compute the information content of each channel not yet selected, relative to those selected; (2) Select the channel providing the most information; (3) Repeat until enough information has been gathered. The effect is

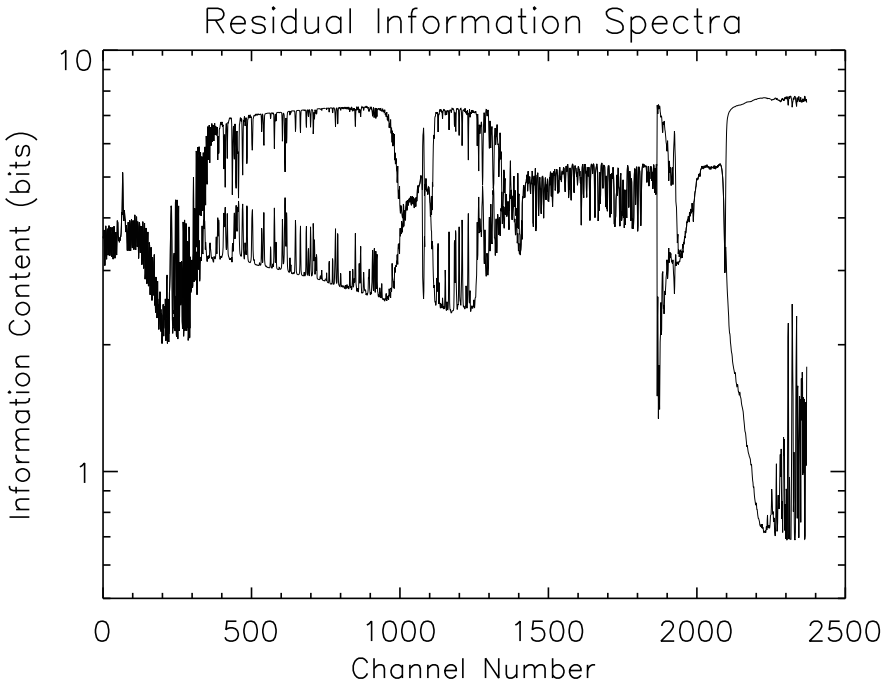

Figure 2: A residual information spectrum, before (upper) and after (lower) selecting the channel with the most information. This example for AIRS uses an early set of early weighting functions supplied by Allen Huang. The detailed algebra and numerical method for this process can be found in Rodgers (1996).

illustrated in figures 2, 3 and 4. In figure 2, the top curve is the information content of each channel considered individually. The channel with most information is a window channel at number 2316. The other curve is the information content of the remaining channels after 2316 has been selected. It can be seen, for example, that nearby channels no longer have significant information and that other window channels elsewhere in the spectrum also have reduced information. This is because they are very similar to the one selected, and convey more or less the same information. Figure 3 shows the same two curves, plus the residual information after 10, 100 and 1000 channels have been selected. A cumulative information spectrum is shown in figure 4. This indicates that almost all of the information can be obtained with a relatively small number of channels. It is generally found that the cumulative information content of the channels selected increases approximately logarithmically with their number.

For an instrument like MIPAS, where the forward model computes the spectrum monochromatically on a fine grid, and then convolves with a spectral response function, there is a computational advantage in computing several adjacent spectral points, or 'microwindows'. The same applies to adjacent vertical points for a limb-sounder with a finite field of view function. The basic strategy is then: (1) Select a 'seed' channel from the whole spectrum, providing most information; (2) Select adjacent channels (spectrally or in tangent altitude) providing the most information; (3) When no significant increase in information is obtained, or the microwindow reaches a predetermined size, select a new seed for a new window.

The simple approach outlined applies to instruments where the error analysis is complete and the retrieval is optimal. The sub-optimal case includes for example situations where there are systematic errors not included in the error covariance used in the estimator. In these cases, the effective information content of the retrieval

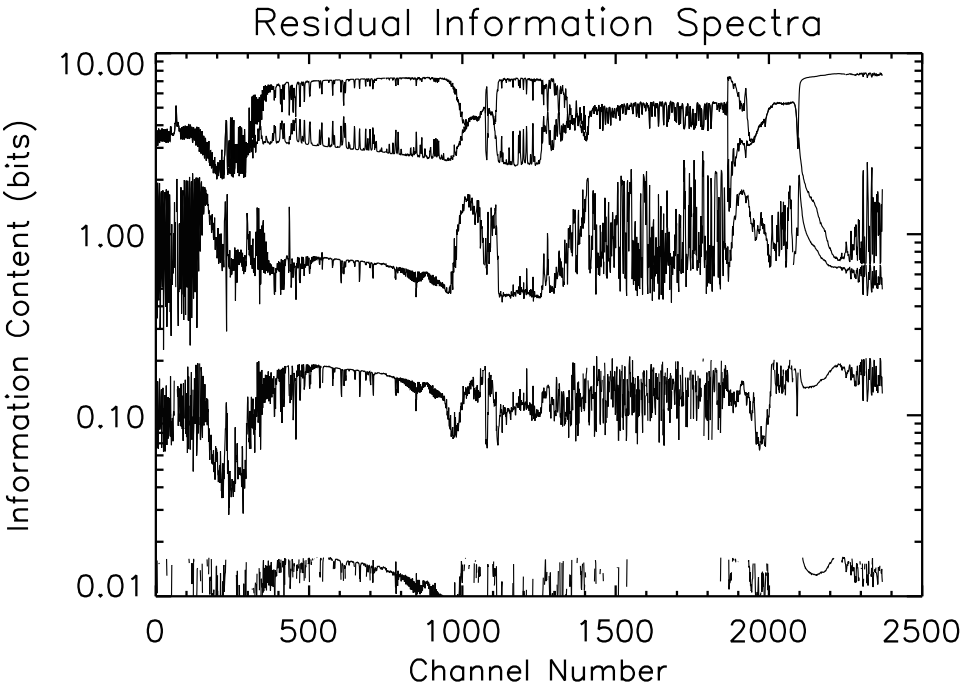

Figure 3: Residual information after selecting 0, 1, 10, 100 and 1000 channels.

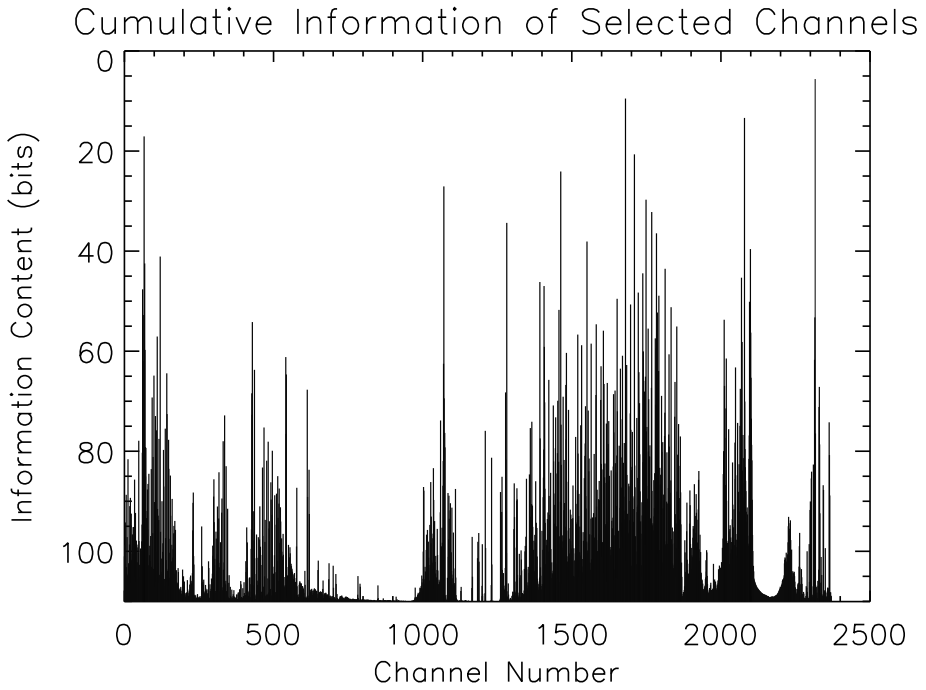

Figure 4: Cumulative information content as each channel is selected. The tallest spikes correspond to the channels first selected.

must be calculated with the prior *pdf*, and a posterior *pdf* that includes all known errors, notionally:

$$H = \frac{1}{2}\log_2 |\mathbf{S}_{\text{prior}}| - \frac{1}{2}\log_2 |\mathbf{S}_{\text{posterior}} + \mathbf{S}_{\text{systematic}}| \tag{19}$$

It is possible for the information content to be negative in these circumstances, so we add another stopping criterion for selecting microwindows: when the incremental information is not positive. Details of the calculations required can be found in Dudhia et al. (2002).

4 Data Transformation for Assimilation

We can also use the ideas of information content to develop an efficient way of assimilating data from instruments with large numbers of channels. We can construct a representation of the information content of a measurement using a number of pseudo-channels corresponding approximately to the number of degrees of freedom for signal.

To assimilate *any* observation we need a forward model for the observation, a Jacobian of the forward model and the error characteristics of the observation. The usual approach to assimilating retrievals as observations involves approximations: the retrieved profile is taken to be an estimate of the true profile, the forward model is a unit matrix (at its simplest) or, more generally, an interpolation operator, the error covariance is taken to be diagonal and the error covariance is taken to be constant. However the retrieval contains a priori, and often has poor vertical resolution. If it is regarded as an estimate of the true profile, then its error depends on the profile, and its error covariance neither diagonal nor constant.

4.1 Assimilating radiances

A better approach is to assimilate radiances, for which we need: a forward model for the observed radiances; a Jacobian of the forward model and error characteristics of the radiances. This looks conceptually more straightforward, but the forward model is much more complex than for assimilating retrievals. It involves modelling such things as: radiative transfer equation with several absorbers; instrument spectral response; instrument field of view response and instrument scan strategy. It is particularly time consuming for an instrument such as AIRS with around 2300 channels or MIPAS with $\sim 10^6$ channels per observation. especially as the Jacobian must be computed at the same time.

The error characteristics are generally simpler than for retrievals, the covariance is usually taken to be diagonal and constant, but systematic errors should really be taken into account, These are not diagonal, and are correlated between successive observations.

4.2 What else could we assimilate?

For simplicity and efficiency of assimilation, we would like a data representation to: have a trivial forward model, preferably linear; have no more observation elements than the number of degrees of freedom; have a diagonal error covariance; have no *a priori* component; represent all of the information contained in the measurement and have the bulk of the calculation done offline, before the assimilation, and preferably by the data supplier

Any information-preserving transformation of the data can be used, provided it has a forward model, a Jacobian and an error analysis. A critical insight is that any linearisation need be valid only over an appropriate range of

the parameters, and this range is, in effect, the error bounds of a retrieval.

Possibilities include a linearised and prewhitened radiance model, evaluated at an offline retrieval, and an averaging kernel representation of a retrieval, prewhitened. Either of these can be compressed by the use of singular vectors of the linear model, and both contain all of the information of the measurement, and are algebraically simple for the assimilator.

We note that if a transformation preserves the relative sizes and shapes of the prior and posterior probability density functions, then the information in the measurement is preserved. Any full-rank linear transformation will do this, e.g. rotations and scale changes in state space. Any complete description of the posterior *pdf* contains all of the information in the measurement relative to the prior.

We should also note that for assimilation, we only want to preserve the information content (i.e. the *pdf*) of the *measurement* – the prior doesn't matter. But the only part of the *pdf* that matters is that part of the prior *pdf* that lies in the (smaller) region of the posterior *pdf*.

4.3 A transformed retrieval

The retrieval characterisation contains all of the information of the measurement:

$$\hat{\mathbf{x}} = \mathbf{x}_a + \mathbf{A}(\mathbf{x} - \mathbf{x}_a) + \mathbf{G}\varepsilon_y \tag{20}$$

where $\mathbf{x}_a$ is *a priori*, $\mathbf{A}$ is Averaging kernel, $\mathbf{G}$ is the Kalman gain and ε_y is the measurement error. Thus the retrieval can be interpreted as providing a measurement $\hat{\mathbf{x}}$ of $\mathbf{x}$ with a linear forward model $\mathbf{x}_a + \mathbf{A}(\mathbf{x} - \mathbf{x}_a)$ and errors $\mathbf{G}\varepsilon_y$.

Now define $\mathbf{z} = \mathbf{S}_{\hat{\mathbf{x}}}^{-\frac{1}{2}}[\hat{\mathbf{x}} + (\mathbf{A} - \mathbf{I})\mathbf{x}_a]$, where $\mathbf{S}_{\hat{\mathbf{x}}}$ is the covariance of $\mathbf{G}\varepsilon_y$. This has a forward model

$$\mathbf{z} = \mathbf{S}_{\hat{\mathbf{x}}}^{-\frac{1}{2}}\mathbf{A}\mathbf{x} + \varepsilon_z \tag{21}$$

where ε_z has covariance $\mathbf{I}$. We note that $\mathbf{z}$ contains no *a priori* contribution and contains all of the information content of the original measurement. The representation will be valid as long as the radiance forward model is linear within the error bounds of the retrieval.

It may be possible to simplify further by using a singular vector expansion, $\mathbf{S}_{\hat{\mathbf{x}}}^{-\frac{1}{2}}\mathbf{A} = \mathbf{U}\Lambda\mathbf{V}^T$. Then:

$$\mathbf{z} = \mathbf{U}\Lambda\mathbf{V}^T\mathbf{x} + \varepsilon_z \tag{22}$$

We can now define $\mathbf{z}'$:

$$\mathbf{z}' = \mathbf{U}^T\mathbf{z} = \Lambda\mathbf{V}^T\mathbf{x} + \mathbf{U}^T\varepsilon_z \tag{23}$$

The covariance of $\mathbf{U}^T\varepsilon_z$ is still unity. The elements of $\mathbf{z}'$ corresponding to small singular values can be ignored. The number kept should be approximately equal to the degrees of freedom for signal. (This derivation has been slightly simplified. we can be more rigorous by prewhitening $\mathbf{x}$ with $\mathbf{S}_a^{-\frac{1}{2}}$.)

4.4 Comments

The procedure to be carried out on the raw data before assimilation is to: retrieve a profile by any appropriate method in order to find a linearisation point, carry out the above transformations, and provide the assimilation with the truncated $\mathbf{z}'$ as a measurement with $\Lambda\mathbf{V}^T$ providing the forward model.

The complex part of the retrieval, radiative transfer, is done only once. It is not needed for every iteration of the assimilation, and can be done offline, before the assimilation. The retrieval method need not be optimal, as long as it is within linear reach of the true state, and has a proper characterisation and error analysis, and preserves information. Retrieval *a priori* does not pollute the information given to the assimilation. It is only used to provide a linearisation point. A similar process can be carried out for radiances, based on the linearised radiance forward model

$$\mathbf{y} = \mathbf{f}(\mathbf{x}_0) + \mathbf{K}(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0) + \varepsilon_y \tag{24}$$

where $\mathbf{x}_0$ is any linearisation point, e.g. a retrieval. Finally, the approach will not be optimal for grossly nonlinear retrieval problems, but in that case, a direct assimilation also is not optimal.

5 Summary

Information Content is a useful conceptual tool for optimisation, it can be applied effectively to selection of raw data for retrieval and assimilation, and can also be applied to the preparation of data for optimal assimiliation

References

Dudhia A, Jay V. L., and Rodgers C. D. (2002), Microwindow selection for high-spectral-resolution sounders *Appl Optics* **41** (18): pp. 3665-3673.

Fisher, R. A. (1921), "On the mathematical foundation of theoretical statistics", *Phil. Trans. R. Soc. Lond.*, **A222**, 309.

Rodgers, C. D. (1996), Information content and optimisation of high spectral resolution measurements, SPIE, Vol **2830**, *Optical Spectroscopic Techniques and Instrumentation for Atmospheric and Space Research II*, Paul B. Hays and Jinxue Wang, eds., p 136-147.

Shannon, C. E. and Weaver, W. (1949), *The Mathematical Theory of Communication*, Paperback edition, University of Illinois Press, Urbana, 1962.