Information Content of Advanced Sounders

Clive D Rodgers

Atmospheric, Oceanic and Planetary Physics

University of Oxford

ECMWF Workshop on Assimilation of high spectral resolution sounders in NWP

June 28 - July 1 2004

OUTLINE

- Advanced Sounders
- What is Information?
- Selecting Information-rich Data for Retrieval/Assimilation
- Systematic Errors and Non-optimal Retrievals
- Data Transformation for Assimilation
- Preserving Information

ADVANCED SOUNDERS

The latest generation of sounders has far more channels per data point than we are likely to be able to use:

- AIRS: $\sim$2400 spectral points per observation.
- IASI: $\sim$8500 spectral points per observation.
- MIPAS: $\sim$60,000 spectral points per spectrum, $\sim 10^6$ points per profile
- TES: $\sim$20,000 spectral points per nadir spectrum,
  $\sim$80,000 spectral points per limb spectrum, $\sim 3 \times 10^6$ per observation,
- GOME: $\sim$2000 spectral points per observation.
- GOMOS
- SCIAMACHY
- etc. . .

All high spectral resolution compared to the previous generation exemplified by ATOVS and TOMS.

We need to think carefully to make best use of this kind of data.

How do we make use of the information content with least computation?

What do we assimilate?

INFORMATION

- What is Information?
- And how do we know when we have captured enough?

There are at least three relevant quantities:

- **Shannon Information**

  A scalar quantity which relates prior knowledge to posterior knowledge, rather like signal/noise ratio.

- **Fisher Information**

  A matrix quantity related to the region of state space containing the uncertainty of our knowledge of the state. Refers only to posterior knowledge.

- **Degrees of freedom for signal**

  An effective number of independent quantities whose uncertainty has been improved by the measurement.

THE BAYESIAN PERSPECTIVE

This is the most general view to the problem (that I know of).

Knowledge is represented in terms of $probability\ density\ functions$:

- $P(\mathbf{x})$ is the *a priori* p.d.f. of the state – describing what we know about the state before we make the measurement.
- $P(\mathbf{y})$ is the *a priori* p.d.f. of the measurement.
- $P(\mathbf{x}, \mathbf{y})$ is the joint *a priori* p.d.f. of $\mathbf{x}$ and $\mathbf{y}$.
- $P(\mathbf{y}|\mathbf{x})$ is the p.d.f. of the measurement given the state – this depends on experimental error and the forward function.
- $P(\mathbf{x}|\mathbf{y})$ is the p.d.f. of the state given the measurement – describing what we know about the state after we make the measurement.

Information is encapsulated in the relevant pdf's.

Most useful 'information content' parameters are some function of these pdf's.

SHANNON INFORMATION

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

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

  $M(\mathbf{x})$ is a measure function. We will take it to be constant.
  Qualitatively, entropy is the log of the volume of state space occupied by the pdf.

- The Shannon information content of a measurement is the change in entropy between the p.d.f. before, $P(\mathbf{x})$, and the p.d.f. after, $P(\mathbf{x}|\mathbf{y})$, the measurement:

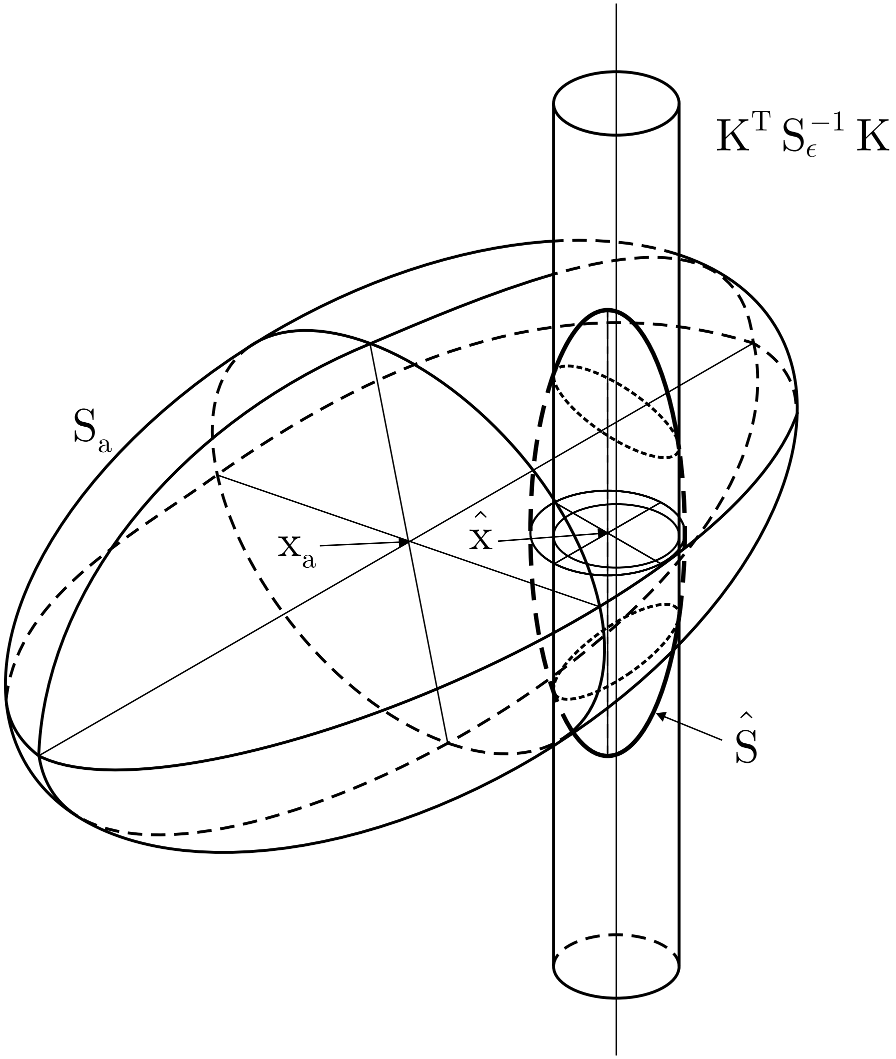
$$H = S\{P(\mathbf{x})\} - S\{P(\mathbf{x}|\mathbf{y})\}$$

  It can be thought of as the log of the ratio of the posterior to prior volumes of state space.
  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:

$$H = \frac{1}{2}\log|\mathbf{S}_{\rm prior}| - \frac{1}{2}\log|\mathbf{S}_{\rm posterior}|$$

A GEOMETRIC INTERPRETATION



1 ≫ 2 ≫ 3 ≫

DEGREES OF FREEDOM FOR SIGNAL AND NOISE

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

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

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:

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

and

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

With some manipulation we can find

$$\begin{aligned} d_s &= \mathrm{tr}((\mathbf{K}^T \mathbf{S}_\epsilon^{-1} \mathbf{K} + \mathbf{S}_a^{-1})^{-1} \mathbf{K}^T \mathbf{S}_\epsilon^{-1} \mathbf{K}) \\ &= \mathrm{tr}(\mathbf{K}\mathbf{S}_a\mathbf{K}^T (\mathbf{K}\mathbf{S}_a\mathbf{K}^T + \mathbf{S}_\epsilon)^{-1}) \qquad (1) \end{aligned}$$

$$\begin{aligned} d_n &= \mathrm{tr}((\mathbf{K}^T \mathbf{S}_\epsilon^{-1} \mathbf{K} + \mathbf{S}_a^{-1})^{-1} \mathbf{S}_a^{-1}) + m - n \\ &= \mathrm{tr}(\mathbf{S}_\epsilon (\mathbf{K}\mathbf{S}_a\mathbf{K}^T + \mathbf{S}_\epsilon)^{-1}) \qquad (2) \end{aligned}$$

INDEPENDENT MEASUREMENTS

If the measurement error covariance is not diagonal, the elements of the $\mathbf{y}$ vector will not be statistically independent. Likewise for the *a priori*.

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.

First, statistical independence. Define:

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

The transformed covariances $\tilde{\mathbf{S}}_a$ and $\tilde{\mathbf{S}}_\epsilon$ both become unit matrices.

The forward model becomes:

$$\tilde{\mathbf{y}} = \tilde{\mathbf{K}}\tilde{\mathbf{x}} + \tilde{\boldsymbol{\epsilon}}$$

where $\tilde{\mathbf{K}} = \mathbf{S}_\epsilon^{-\frac{1}{2}}\mathbf{K}\mathbf{S}_a^{\frac{1}{2}}$.

The solution covariance becomes:

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

TRANSFORM AGAIN

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

$$\tilde{\mathbf{y}} = \tilde{\mathbf{K}}\tilde{\mathbf{x}} + \tilde{\boldsymbol{\epsilon}} \qquad \rightarrow \qquad \tilde{\mathbf{y}} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^T\tilde{\mathbf{x}} + \tilde{\boldsymbol{\epsilon}}$$

Define:

$$\mathbf{x}' = \mathbf{V}^T\tilde{\mathbf{x}} \quad \mathbf{y}' = \mathbf{U}^T\tilde{\mathbf{y}} \quad \boldsymbol{\epsilon}' = \mathbf{U}^T\tilde{\boldsymbol{\epsilon}}$$

The forward model becomes:

$$\mathbf{y}' = \boldsymbol{\Lambda}\mathbf{x}' + \boldsymbol{\epsilon}' \tag{1}$$

The Jacobian is now diagonal, $\boldsymbol{\Lambda}$, and the *a priori* and noise covariances are still unit matrices, hence the solution covariance becomes:

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

which is diagonal, and the solution itself is

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

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

- Elements for which $\lambda_i \gg 1$ or $(1 + \lambda_i^2)^{-1} \ll 1$ are well measured
- Elements for which $\lambda_i \ll 1$ or $(1 + \lambda_i^2)^{-1} \gg 1$ are poorly measured.

≪

INFORMATION

Shannon Information in the Transformed Basis

Because it is a ratio of volumes, the linear transformation does not change the information content. So consider information in the $\mathbf{x}'$, $\mathbf{y}'$ system:

$$\begin{aligned} H &= S\{\mathbf{S}'_a\} - S\{\hat{\mathbf{S}}'\} \\ &= -\frac{1}{2}\log(|\mathbf{I}_n|) + \frac{1}{2}\log(|(\mathbf{\Lambda}^2 + \mathbf{I})^{-1}|) \\ &= \sum_i \frac{1}{2}\log(1+\lambda_i^2) \end{aligned} \tag{3}$$

DEGREES OF FREEDOM

Degres 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

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

Summary: for each independent component x_i'

- The information content is $\frac{1}{2} \log(1 + \lambda_i^2)$
- The degrees of freedom for signal is $\lambda_i^2 (1 + \lambda_i^2)^{-1}$

TWO APPLICATIONS OF INFORMATION CONCEPTS

- **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.

- **Data transformation for assimilation**

  For efficiency of data assimilation, we can extract a representation of the information content of a measurement in a number of pseudo-channels corresponding to the degrees of freedom for signal.

DATA SUBSETTING

- **Channel selection.**

  For an instrument like AIRS or IASI, where the FM computes radiances separately for each channel.

  Basic strategy is to select channels independently and sequentially:

  - Recompute information content of each channel not yet selected
  - Select the channel providing the most information.
  - Repeat until enough information has been gathered.

DATA SUBSETTING

- **Microwindow selection**

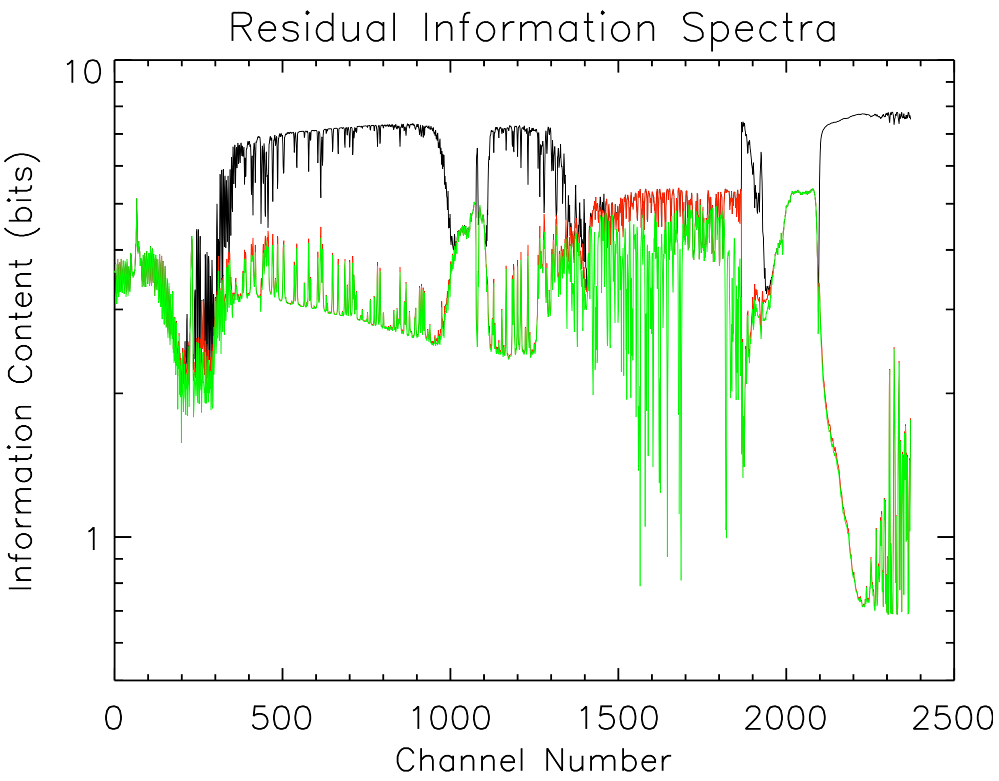
  For an instrument like MIPAS, where the FM computes the spectrum monochromatically on a fine grid, and then convolves with a spectral response function, there is a computational advantage in computing several adjacent spectral points.

  The same applies to adjacent vertical points for a limb-sounder with a finite field of view function.

  Basic strategy is:

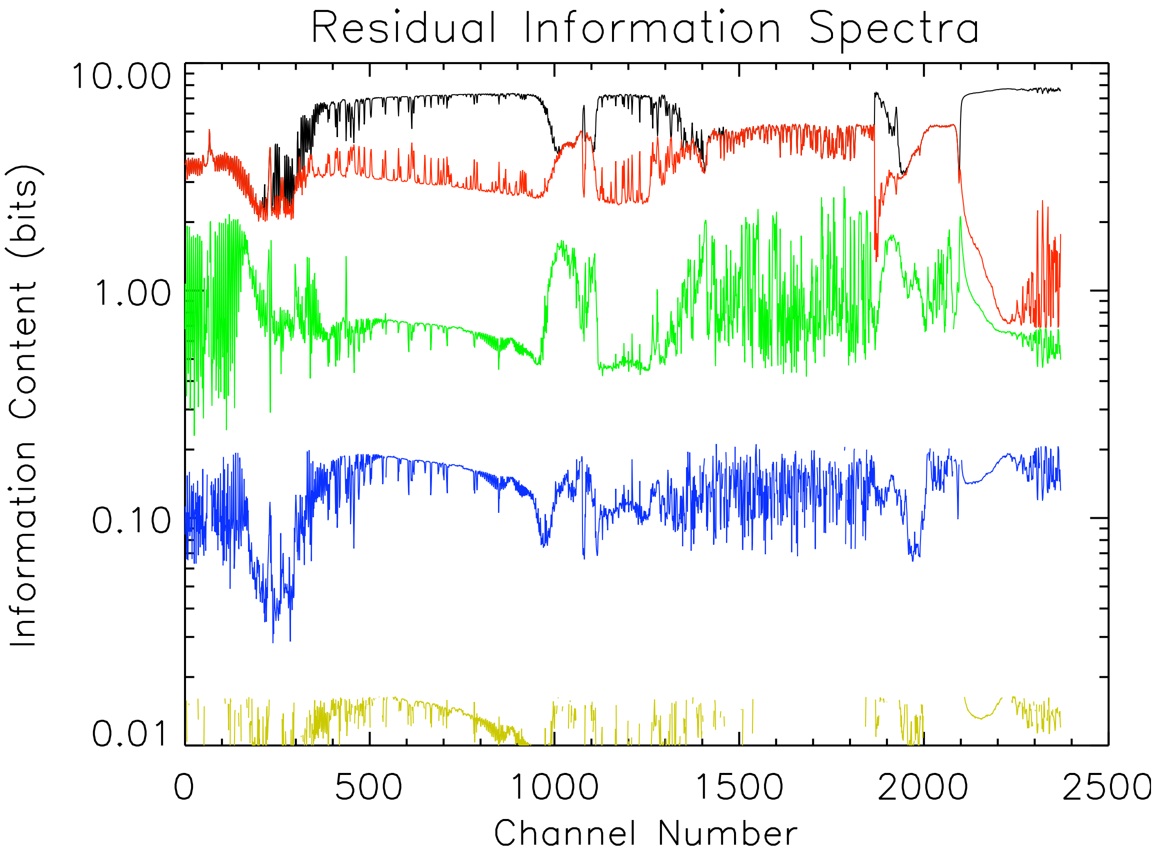
  - Select a 'seed' channel from the whole spectrum providing most information
  - Select adjacent channels (spectrally or in tangent altitude) providing the most information
  - When no increase in information is obtained, or the microwindow reaches a predetermined size, select a new seed for a new window

EXAMPLE: CHANNEL SELECTION



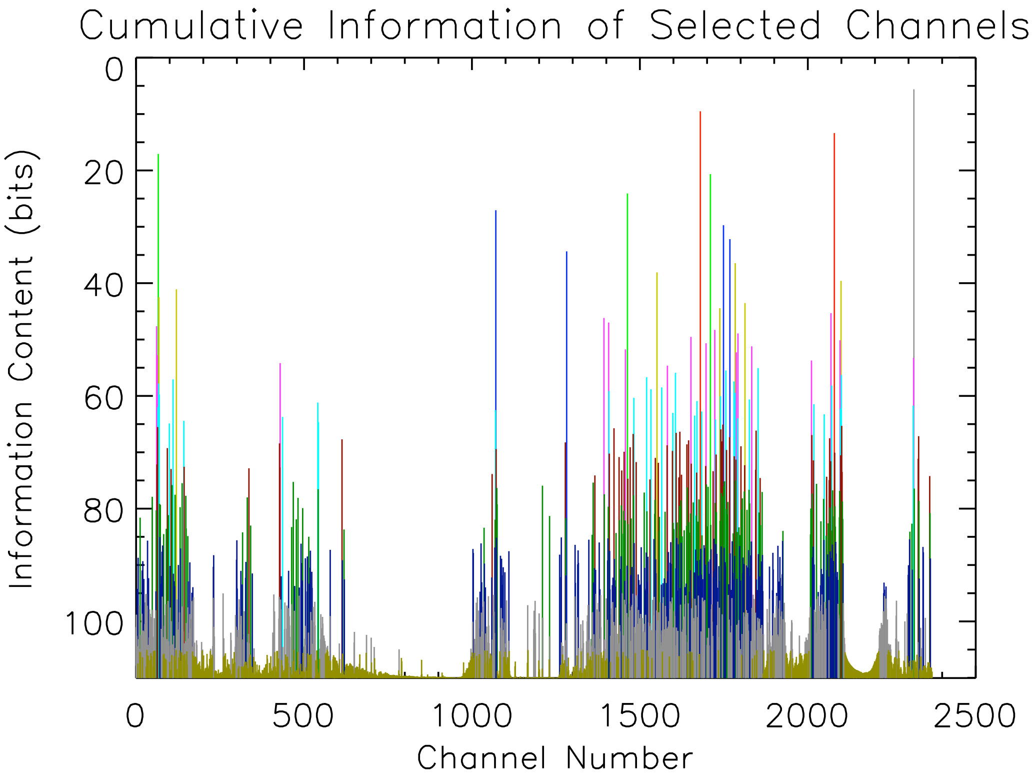
Using early weighting functions for AIRS supplied by Allen Huang. The algebra and numerical method for this process can be found in Rodgers, (1996).

EXAMPLE: CHANNEL SELECTION



Residual information after selecting 1, 10, 100 and 1000 channels.

EXAMPLE: CHANNEL SELECTION



PRACTICAL CONSIDERATIONS: NON-OPTIMAL RETRIEVAL

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

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

It is possible for the information content to be negative in these circumstances.

So we can 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.

ASSIMILATION OF HIGH SPECTRAL RESOLUTION DATA

How do we represent the information content of the measurement for efficient assimilation?

ASSIMILATING RETRIEVALS

To assimilate any observation we need:

- A forward model for the observation
- Jacobian of the forward model
- Error characteristics of the observation

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
- The error covariance is taken to be constant

BUT

- The retrieval contains a priori, and often has poor vertical resolution
- If it is to be regarded as an estimate of the true profile, its error depends on the profile
- Its error covariance neither diagonal nor constant

ASSIMILATING RADIANCES

To assimilate radiances we need:

- A forward model for the observed radiances
- Jacobian of the forward model
- Error characteristics of the radiances

This looks conceptually more straightforward, but. . .

ASSIMILATING RADIANCES

- **The forward model**
  - is much more complex than for assimilating retrievals
  - It involves modelling:
    - radiative transfer equation with several absorbers
    - instrument spectral response
    - instrument field of view response
    - instrument scan strategy
  - Especially tedious for an instrument with e.g.
    - 2371 channels per observation (AIRS)
    - $\sim 10^6$ channels per observation (MIPAS)
- **The Jacobian**
  - must be computed at the same time as the radiances
- **The error characteristics**
  - Are generally simpler than for retrievals.
  - Usually taken to be diagonal and constant, but. . .
  - Systematic errors should really be taken into account, and they are not diagonal, and are correlated between successive observations.

WHAT ELSE COULD WE ASSIMILATE?

We would like it 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
- have the bulk of the calculation done offline, before the assimilation
- and preferably by the data supplier. . .

AN APPROACH

- Any information-preserving transformation of the data can be used, provided it has:
  - A forward model
  - A Jacobian
  - An error analysis
- Critical insight:
  - Any linearisation need be valid only over an appropriate range of the parameters
  - This range is, in effect, the error bounds of a retrieval
- Possibilities are:
  - A linearised and prewhitened radiance model, evaluated at an offline retrieval.
  - Averaging kernel representation of a retrieval, prewhitened

  Either of these can be compressed by the use of singular vectors of the linear model

  Both contain all of the information of the measurement, and are algebraically simple, for the assimilator.

DATA TRANSFORMATION FOR ASSIMILATION

- If any transformation we make 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.

- 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.

≫

A TRANSFORMED RETRIEVAL

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

$$\hat{\mathbf{x}} = \mathbf{x}_a + \mathbf{A}(\mathbf{x} - \mathbf{x}_a) + \mathbf{G}\boldsymbol{\epsilon}_y$$

where:

- $\mathbf{x}_a$ is a priori
- $\mathbf{A}$ is Averaging kernel
- $\mathbf{G}$ is the Kalman gain
- $\boldsymbol{\epsilon}_y$ is the measurement error

It 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 $\mathbf{G}\boldsymbol{\epsilon}_y$.

A TRANSFORMED RETRIEVAL II

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 $\mathbf{G}\boldsymbol{\epsilon}_y$.

This has a forward model

$$\mathbf{z} = \mathbf{S}_{\hat{\mathbf{x}}}^{-\frac{1}{2}}\mathbf{A}\mathbf{x} + \boldsymbol{\epsilon}_z$$

where $\boldsymbol{\epsilon}_z$ has covariance $\mathbf{I}$.

- $\mathbf{z}$ contains no a priori contribution
- It 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.

A TRANSFORMED RETRIEVAL III

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

$$\mathbf{z} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T\mathbf{x} + \boldsymbol{\epsilon}_z$$

define $\mathbf{z}'$:

$$\mathbf{z}' = \mathbf{U}^T\mathbf{z} = \mathbf{\Lambda}\mathbf{V}^T\mathbf{x} + \mathbf{U}^T\boldsymbol{\epsilon}_z$$

The covariance of $\mathbf{U}^T\boldsymbol{\epsilon}_z$ is still unity.

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.

[We can be more rigorous by prewhitening $\mathbf{x}$ with $\mathbf{S}_a^{-\frac{1}{2}}$.]

Procedure:

- Retrieve a profile
- Carry out the above transformations
- Provide the assimilation with the truncated $\mathbf{z}'$ and $\mathbf{\Lambda}\mathbf{V}^T$

COMMENTS

- The complex part of the retrieval, radiative transfer, is done only once. It is not needed for every iteration of the assimilation. It 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 forward model

  $$\mathbf{y} = \mathbf{f}(\mathbf{x}_0) + \mathbf{K}(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0) + \boldsymbol{\epsilon}_y$$

  where $\mathbf{x}_0$ is any linearisation point, e.g. a retrieval.

- This will not be optimal for grossly nonlinear retrieval problems.

SUMMARY

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

References

Rodgers, C. D. 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, 1996.

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

End