INFORMATION CONTENT AND ERROR ANALYSIS

Clive D Rodgers

Atmospheric, Oceanic and Planetary Physics

University of Oxford

ESA Advanced Atmospheric Training Course

September 15th – 20th, 2008

INFORMATION CONTENT OF A MEASUREMENT

Information in a general qualitative sense:

$$\textit{Conceptually, what does } \mathbf{y} \textit{ tell you about } \mathbf{x}?$$

We need to answer this

- to determine if a conceptual instrument design actually works
- to optimise designs

Use the linear problem for simplicity to illustrate the ideas.

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

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:

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

  $M(\mathbf{x})$ is a measure function. We will take it to be constant.

- Compare this with the statistical mechanics definition of entropy:

$$S = -k \sum_i p_i \ln p_i$$

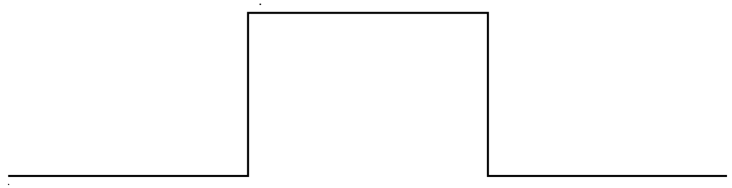
  $P(\mathbf{x})d\mathbf{x}$ corresponds to p_i. $1/M(\mathbf{x})$ is a kind of scale for $d\mathbf{x}$.

- 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:

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

What does this all mean?

ENTROPY OF A BOXCAR PDF



Consider a uniform p.d.f in one dimension, constant in (0,a):

$$P(x) = 1/a \quad 0 < x < a$$

and zero outside. The entropy is given by

$$S = -\int_0^a \frac{1}{a} \ln\left(\frac{1}{a}\right) \mathrm{d}x = \ln a$$

Similarly, the entropy of any constant pdf in a finite volume V of arbitrary shape is:

$$S = -\int_V \frac{1}{V} \ln\left(\frac{1}{V}\right) \mathrm{d}v = \ln V$$

i.e the entropy is the log of the volume of state space occupied by the p.d.f.

ENTROPY OF A GAUSSIAN PDF

Consider the Gaussian distribution:

$$P(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{n}{2}}|\mathbf{S}|^{\frac{1}{2}}} \exp[-\frac{1}{2}(\mathbf{x}-\bar{\mathbf{x}})^T \mathbf{S}^{-1}(\mathbf{x}-\bar{\mathbf{x}})]$$

If you evaluate the entropy of a Gaussian distribution you will find it is proportional to $\log|\mathbf{S}|^{\frac{1}{2}}$.

- The contours of $P(\mathbf{x})$ in n-space are ellipsoidal, described by

$$(\mathbf{x}-\bar{\mathbf{x}})^T \mathbf{S}^{-1}(\mathbf{x}-\bar{\mathbf{x}}) = \text{constant}$$

- The principal axes of the ellipsoid are the eigenvectors of $\mathbf{S}$, and their lengths are proportional to the square roots of the corresponding eigenvalues.
- The volume of the ellipsoid is proportional to the root of the product of the eigenvalues, which is proportional to $|\mathbf{S}|^{\frac{1}{2}}$.
- Therefore **entropy** is the log of the volume enclosed by some particular contour of $P(\mathbf{x})$. A 'volume of uncertainty'.

ENTROPY AND INFORMATION

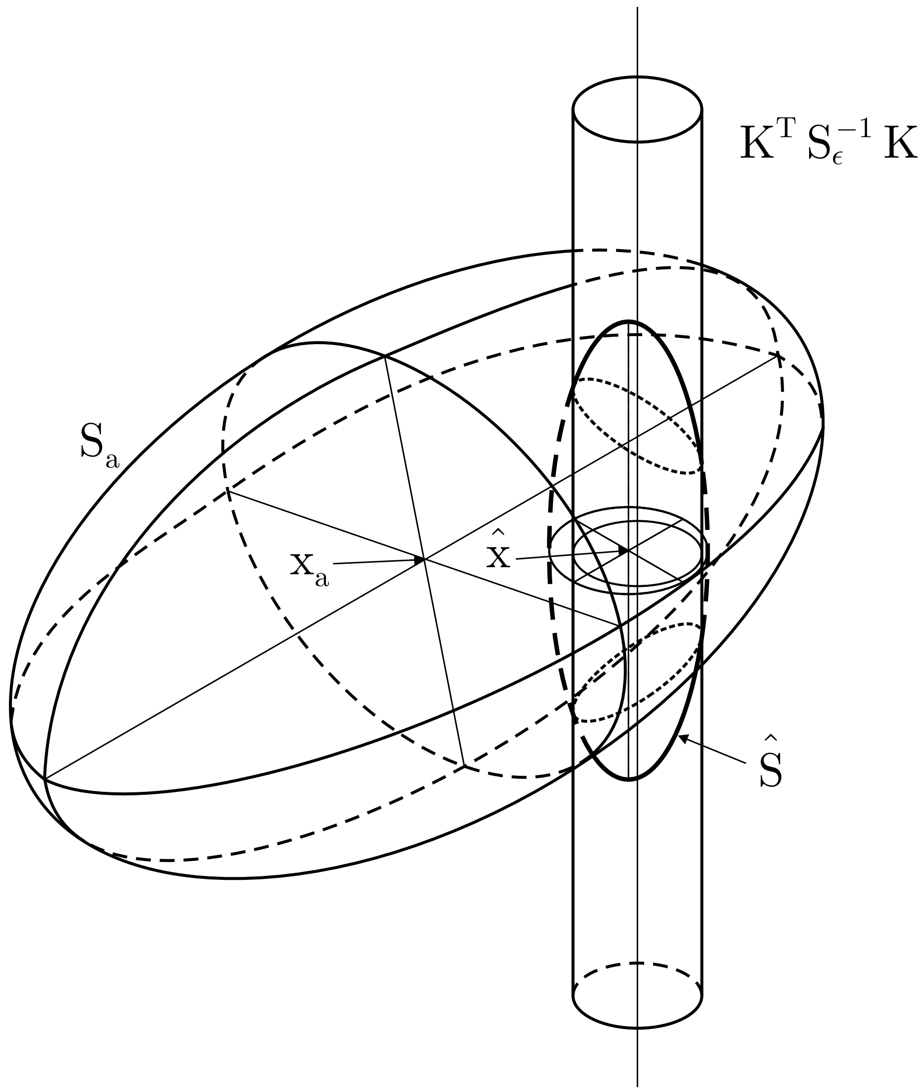
Information content is the change in entropy,

- i.e. the log of the ratio of the volumes of uncertainty before and after making a measurement.
- A generalisation of 'signal to noise'.
- In the Boxcar case, the log of the ratio of the volumes before and after
- In the Gaussian case:

$$H = \log|\mathbf{S}_a| - \log|\hat{\mathbf{S}}| = -\log|(\mathbf{K}^T\mathbf{S}_\epsilon^{-1}\mathbf{K} + \mathbf{S}_a^{-1})^{-1}\mathbf{S}_a^{-1}|$$

- minus the log of the determinant of the weight of $\mathbf{x}_a$ in the Bayesian expectation.

The log of the ratio of the volumes of the ellipsoids



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

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

Using tr(**CD**)=tr(**DC**), we can see that

$$\begin{aligned} d_s &= E\{\mathrm{tr}([\hat{\mathbf{x}} - \mathbf{x}_a][\hat{\mathbf{x}} - \mathbf{x}_a]^T \mathbf{S}_a^{-1})\} \\ &= \mathrm{tr}(E\{[\hat{\mathbf{x}} - \mathbf{x}_a][\hat{\mathbf{x}} - \mathbf{x}_a]^T\} \mathbf{S}_a^{-1}) \end{aligned} \tag{6}$$

DEGREES OF FREEDOM FOR SIGNAL AND NOISE II

With some manipulation we can find

$$\begin{aligned} d_s &= \mathrm{tr}((\mathbf{K}^T\mathbf{S}_\epsilon^{-1}\mathbf{K} + \mathbf{S}_a^{-1})^{-1}\mathbf{K}^T\mathbf{S}_\epsilon^{-1}\mathbf{K}) \\ &= \mathrm{tr}(\mathbf{K}\mathbf{S}_a\mathbf{K}^T(\mathbf{K}\mathbf{S}_a\mathbf{K}^T + \mathbf{S}_\epsilon)^{-1}) \end{aligned} \tag{7}$$

and

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

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.

Therefore 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. [>>]

SQUARE ROOTS OF MATRICES

The square root of an arbitrary matrix is defined as $\mathbf{A}^{\frac{1}{2}}$ where

$$\mathbf{A}^{\frac{1}{2}}\mathbf{A}^{\frac{1}{2}} = \mathbf{A}$$

Using $\mathbf{A}^n = \mathbf{R}\mathbf{\Lambda}^n\mathbf{L}^T$ for $n = 1/2$:

$$\mathbf{A}^{\frac{1}{2}} = \mathbf{R}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{L}^T$$

This square root of a matrix is not unique, because the diagonal elements of $\mathbf{\Lambda}^{\frac{1}{2}}$ in $\mathbf{R}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{L}^T$ can have either sign, leading to 2^n possibilities.

We only use square roots of symmetric covariance matrices. In this case $\mathbf{S}^{\frac{1}{2}} = \mathbf{L}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{L}^T$ is symmetric.

SQUARE ROOTS OF SYMMETRIC MATRICES

Symmetric matrices can also have non-symmetric roots satisfying $\mathbf{S} = (\mathbf{S}^{\frac{1}{2}})^T \mathbf{S}^{\frac{1}{2}}$, of which the Cholesky decomposition:

$$\mathbf{S} = \mathbf{T}^T \mathbf{T}$$

where $\mathbf{T}$ is upper triangular, is the most useful.

There are an infinite number of non-symmetric square roots: if $\mathbf{S}^{\frac{1}{2}}$ is a square root, then clearly so is $\mathbf{X}\mathbf{S}^{\frac{1}{2}}$ where $\mathbf{X}$ is any orthonormal matrix.

The inverse symmetric square root is $\mathbf{S}^{-\frac{1}{2}} = \mathbf{L}\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{L}^T$.

The inverse Cholesky decomposition is $\mathbf{S}^{-1} = \mathbf{T}^{-1}\mathbf{T}^{-T}$.

The inverse square root $\mathbf{T}^{-1}$ is triangular, and its numerical effect is implemented efficiently by back substitution.

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.

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

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.

Therefore 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:

$$\begin{aligned} H &= S\{\mathbf{S}'_a\} - S\{\hat{\mathbf{S}}'\} \\ &= -\frac{1}{2}\log(|\mathbf{I}_n|) + \frac{1}{2}\log(|(\mathbf{\Lambda}^2 + \mathbf{I})^{-1}|) \\ &= \sum_i \frac{1}{2}\log(1 + \lambda_i^2) \end{aligned} \tag{9}$$

DEGREES OF FREEDOM

Degres of Freedom in the Transformed Basis

The number of independent quantities measured can be thought of as the number of singular values for which $\lambda_i \gg 1$

The degrees of freedom for signal is

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

It is also the sum of the eigenvalues of $\mathbf{I}_n - \hat{\hat{\mathbf{S}}}$.

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

THE OBSERVABLE AND NULL SPACES

- The part of measurement space that can be seen is that spanned by the weighting functions. Anything outside that is in the null space of $\mathbf{K}$.

- Any orthogonal linear combination of the weighting functions will form a basis (coordinate system) for the observable space. An example is those singular vectors of $\mathbf{K}$ which have non-zero singular values.

- The vectors which have zero singular values form a basis for the null space.

- Any component of the state in the null space maps onto the origin in measurement space.

- This implies that there are distinct states, in fact whole subspaces, which map onto the same point in measurement space, and cannot be distinguished by the measurement.

THE NEAR NULL SPACE

However

- the solution can have components in the null space - from the *a priori*.
- components observable in principle can have near zero contributions from the measurement, the '*near null space*'

In the $\mathbf{x}'$, $\mathbf{y}'$ system:

- vectors with $\lambda = 0$ are in the null space
- vectors with $\lambda \ll 1$ are in the near null space
- vectors with $\lambda \gtrsim 1$ are in the non-null space, and are observable.

Diagnostics for the Standard Case

Table 1: Singular values of $\tilde{\mathbf{K}}$, together with contributions of each vector to the degrees of freedom d_s and information content H for both covariance matrices. Measurement noise variance is 0.25 K^2.

	Diagonal covariance			Full covariance		
i	λ_i	$d_{\mathrm{s}i}$	H_i (bits)	λ_i	$d_{\mathrm{s}i}$	H_i (bits)
1	6.51929	0.97701	2.72149	27.81364	0.99871	4.79865
2	4.79231	0.95827	2.29147	18.07567	0.99695	4.17818
3	3.09445	0.90544	1.70134	9.94379	0.98999	3.32105
4	1.84370	0.77269	1.06862	5.00738	0.96165	2.35227
5	1.03787	0.51858	0.52731	2.39204	0.85123	1.37443
6	0.55497	0.23547	0.19368	1.09086	0.54337	0.56546
7	0.27941	0.07242	0.05423	0.46770	0.17948	0.14270
8	0.13011	0.01665	0.01211	0.17989	0.03135	0.02297
totals		4.45653	8.57024		5.55272	16.75571

$$\mathbf{S}_a = 100\mathbf{I}\ \mathrm{K}^2 \qquad\qquad S^a_{ij} = 100[1 - \exp(-\frac{|z_i - z_j|}{h})]$$

where $h = 1$ in $\ln(p)$ coordinates.

FTIR Measurements of CO

30 levels; 894 measurements:

Apparently heavily overconstrained.

Table 2: Singular Values of K

5.345	3.498	0.033	0.0046	7.15e-4	2.56e-4
7.76e-5	2.13e-3	1.71e-5	1.38e-5	9.01e-6	6.73e-6
5.82e-6	4.79e-6	2,87e-6	3.52e-6	3.75e-6	1.91e-6
9.83e-7	2.37e-7	7.71e-7	1.18e-7	1.48e-6	1.95e-7
1.37e-7	6.67e-8	3.50e-8	3.37e-8	5.83e-9	6.29e-9

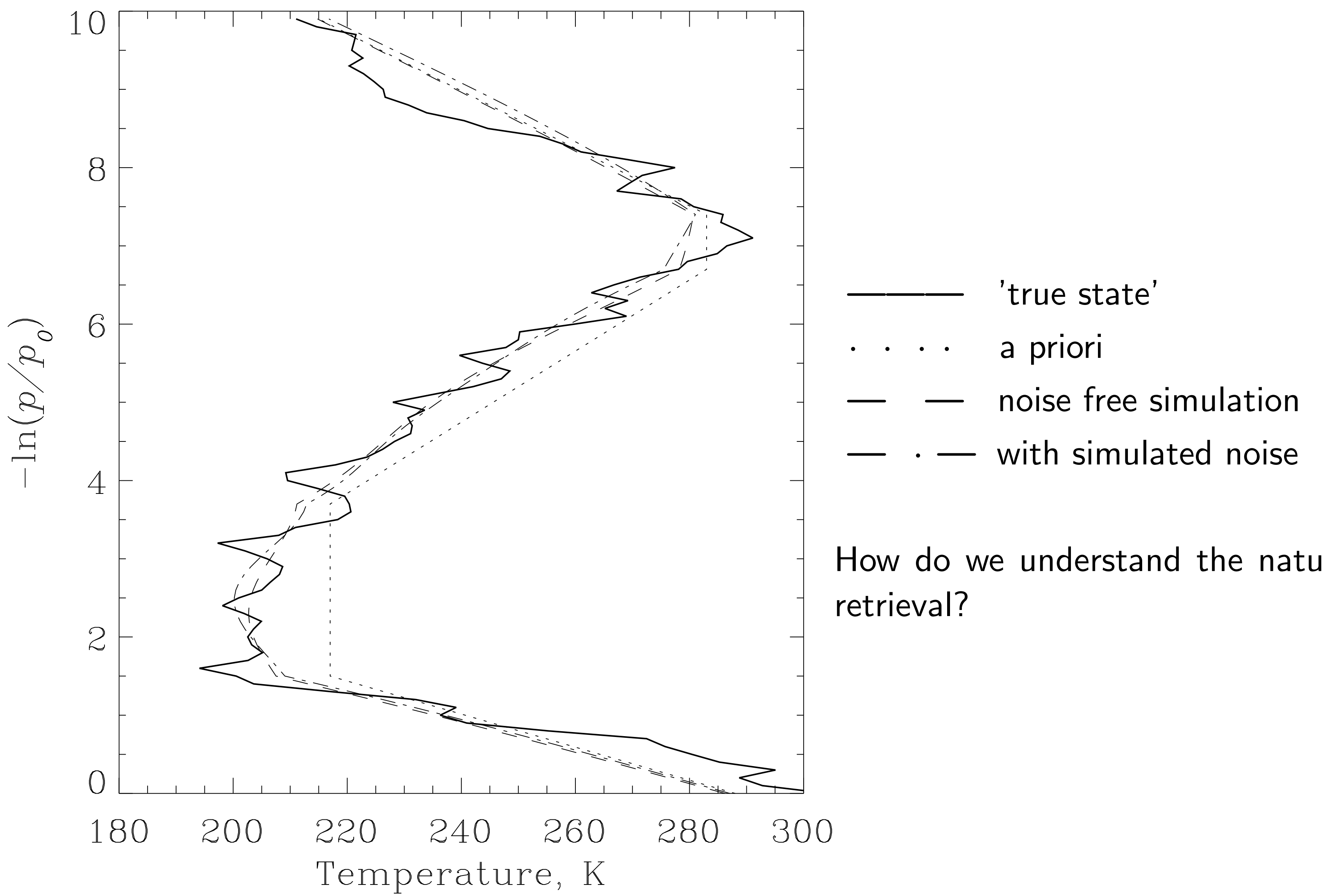
Noise is 0.03 in these units;

Prior variance is about 1.

There are 2.5 degrees of freedom.

Lots of near null space.

A SIMULATED RETRIEVAL



How do we understand the nature of a retrieval?

ERROR ANALYSIS AND CHARACTERISATION

Conceptually

- The measurement $\mathbf{y}$ is a function of some unknown state $\mathbf{x}$:

$$\mathbf{y} = \mathbf{f}(\mathbf{x}, \mathbf{b}) + \boldsymbol{\epsilon}$$

  where:
  - $\mathbf{y}$ is the measurement vector, length m
  - $\mathbf{x}$ is the state vector, length n
  - $\mathbf{f}$ is a function describing the physics of the measurement,
  - $\mathbf{b}$ is a set of 'known' parameters of this function,
  - $\boldsymbol{\epsilon}$ is measurement error, with covariance $\mathbf{S}_\epsilon$.

ERROR ANALYSIS & CHARACTERISATION II

- The retrieval $\hat{\mathbf{x}}$ is a function of the form:

$$\hat{\mathbf{x}} = \mathbf{R}(\mathbf{y}, \hat{\mathbf{b}}, \mathbf{c})$$

where:

- $\mathbf{R}$ represents the retrieval method
- $\hat{\mathbf{b}}$ is our estimate of the forward function parameters $\mathbf{b}$
- $\mathbf{c}$ represents any parameters used in the inverse method that do not affect the measurement, e.g. *a priori*.

The Transfer function

Thus the retrieval is related to the 'truth' $\mathbf{x}$ formally by:

$$\hat{\mathbf{x}} = \mathbf{R}(\mathbf{f}(\mathbf{x}, \mathbf{b}) + \boldsymbol{\epsilon}, \hat{\mathbf{b}}, \mathbf{c})$$

which may be regarded as the transfer function of the measurement and retrieval system as a whole:

$$\hat{\mathbf{x}} = \mathbf{T}(\mathbf{x}, \mathbf{b}, \boldsymbol{\epsilon}, \hat{\mathbf{b}}, \mathbf{c})$$

ERROR ANALYSIS & CHARACTERISATION III

Characterisation means evaluating:

- $\partial\hat{\mathbf{x}}/\partial\mathbf{x} = \mathbf{A}$, sensitivity to actual state: *Averaging Kernel Matrix*

Error analysis involves evaluating:

- $\partial\hat{\mathbf{x}}/\partial\epsilon = \mathbf{G}_y$, sensitivity to noise (or to measurement!)
- $\partial\hat{\mathbf{x}}/\partial\mathbf{b} = \mathbf{G}_b$, sensitivity to non-retrieved parameters
- $\partial\hat{\mathbf{x}}/\partial\mathbf{c} = \mathbf{G}_c$, sensitivity to retrieval method parameters

and understanding the effect of replacing $\mathbf{f}$ by a numerical *Forward Model* $\mathbf{F}$.

THE FORWARD MODEL

We often need to approximate the forward function by a *Forward Model*:

$$\mathbf{F}(\mathbf{x}, \mathbf{b}) \simeq \mathbf{f}(\mathbf{x}, \mathbf{b})$$

Where $\mathbf{F}$ *models* the physics of the measurement, including the instrument, as well as we can.

- It usually has parameters $\mathbf{b}$ which have experimental error
- The vector $\mathbf{b}$ is not a target for retrieval
- There may be parameters $\mathbf{b}'$ of the forward function that are not included in the forward model:

$$\mathbf{F}(\mathbf{x}, \mathbf{b}) \simeq \mathbf{f}(\mathbf{x}, \mathbf{b}, \mathbf{b}')$$

LINEARISE THE TRANSFER FUNCTION

The retrieved quantity is expressed as:

$$\hat{\mathbf{x}} = \mathbf{R}(\mathbf{f}(\mathbf{x}, \mathbf{b}, \mathbf{b}') + \boldsymbol{\epsilon}, \hat{\mathbf{b}}, \mathbf{x}_a, \mathbf{c})$$

where we have also separated $\mathbf{x}_a$, the *a priori*, from other components of $\mathbf{c}$.

Replace $\mathbf{f}$ by $\mathbf{F} + \Delta\mathbf{f}$:

$$\hat{\mathbf{x}} = \mathbf{R}(\mathbf{F}(\mathbf{x}, \mathbf{b}) + \Delta\mathbf{f}(\mathbf{x}, \mathbf{b}, \mathbf{b}') + \boldsymbol{\epsilon}, \hat{\mathbf{b}}, \mathbf{x}_a, \mathbf{c})$$

where $\Delta\mathbf{f}$ is the error in the forward model relative to the real physics:

$$\Delta\mathbf{f} = \mathbf{f}(\mathbf{x}, \mathbf{b}, \mathbf{b}') - \mathbf{F}(\mathbf{x}, \mathbf{b})$$

Linearise $\mathbf{F}$ about $\mathbf{x} = \mathbf{x}_a$, $\mathbf{b} = \hat{\mathbf{b}}$:

$$\hat{\mathbf{x}} = \mathbf{R}(\mathbf{F}(\mathbf{x}_a, \hat{\mathbf{b}}) + \mathbf{K}_x(\mathbf{x} - \mathbf{x}_a) + \mathbf{K}_b(\mathbf{b} - \hat{\mathbf{b}}) + \Delta\mathbf{f}(\mathbf{x}, \mathbf{b}, \mathbf{b}') + \boldsymbol{\epsilon}, \hat{\mathbf{b}}, \mathbf{x}_a, \mathbf{c})$$

where $\mathbf{K}_x = \partial\mathbf{F}/\partial\mathbf{x}$ (the *weighting function*) and $\mathbf{K}_b = \partial\mathbf{F}/\partial\mathbf{b}$.

LINEARISE THE TRANSFER FUNCTION II

$$\hat{\mathbf{x}} = \mathbf{R}(\mathbf{F}(\mathbf{x}_a, \hat{\mathbf{b}}) + \mathbf{K}_x(\mathbf{x} - \mathbf{x}_a) + \mathbf{K}_b(\mathbf{b} - \hat{\mathbf{b}}) + \Delta\mathbf{f}(\mathbf{x}, \mathbf{b}, \mathbf{b}') + \boldsymbol{\epsilon}, \hat{\mathbf{b}}, \mathbf{x}_a, \mathbf{c})$$

Linearise $\mathbf{R}$ with respect to its first argument, $\mathbf{y}$:

$$\hat{\mathbf{x}} = \mathbf{R}[\mathbf{F}(\mathbf{x}_a, \hat{\mathbf{b}}), \hat{\mathbf{b}}, \mathbf{x}_a, \mathbf{c}] + \mathbf{G}_y[\mathbf{K}_x(\mathbf{x} - \mathbf{x}_a) + \mathbf{K}_b(\mathbf{b} - \hat{\mathbf{b}}) + \Delta\mathbf{f}(\mathbf{x}, \mathbf{b}, \mathbf{b}') + \boldsymbol{\epsilon}]$$

where $\mathbf{G}_y = \partial\mathbf{R}/\partial\mathbf{y}$ (the *contribution function*)

CHARACTERISATION

Some rearrangement gives:

$$\begin{aligned} \hat{\mathbf{x}} - \mathbf{x}_a &= \mathbf{R}(\mathbf{F}(\mathbf{x}_a, \hat{\mathbf{b}}), \hat{\mathbf{b}}, \mathbf{x}_a, \mathbf{c}) - \mathbf{x}_a & \textit{bias} \\ & \quad + \mathbf{A}(\mathbf{x} - \mathbf{x}_a) & \textit{smoothing} \\ & \quad + \mathbf{G}_y \epsilon_y & \textit{error} \end{aligned} \tag{10}$$

where

$$\mathbf{A} = \mathbf{G}_y \mathbf{K}_x = \partial \hat{\mathbf{x}} / \partial \mathbf{x}$$

the *averaging kernel*, and

$$\epsilon_y = \mathbf{K}_b(\mathbf{b} - \hat{\mathbf{b}}) + \Delta \mathbf{f}(\mathbf{x}, \mathbf{b}, \mathbf{b}') + \epsilon$$

is the total error in the measurement relative to the forward model.

BIAS

This is the error you would get on doing a simulated error-free retrieval of the *a priori*.

$A\ priori$ is what you know about the state before you make the measurement. Any reasonable retrieval method should return the *a priori* given a measurement vector that corresponds to it, so the bias should be zero.

But check your system to be sure it is. . .

THE AVERAGING KERNEL

The retrieved state is a smoothed version of the true state with smoothing functions given by the rows of $\mathbf{A}$, plus error terms:

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

You can either:

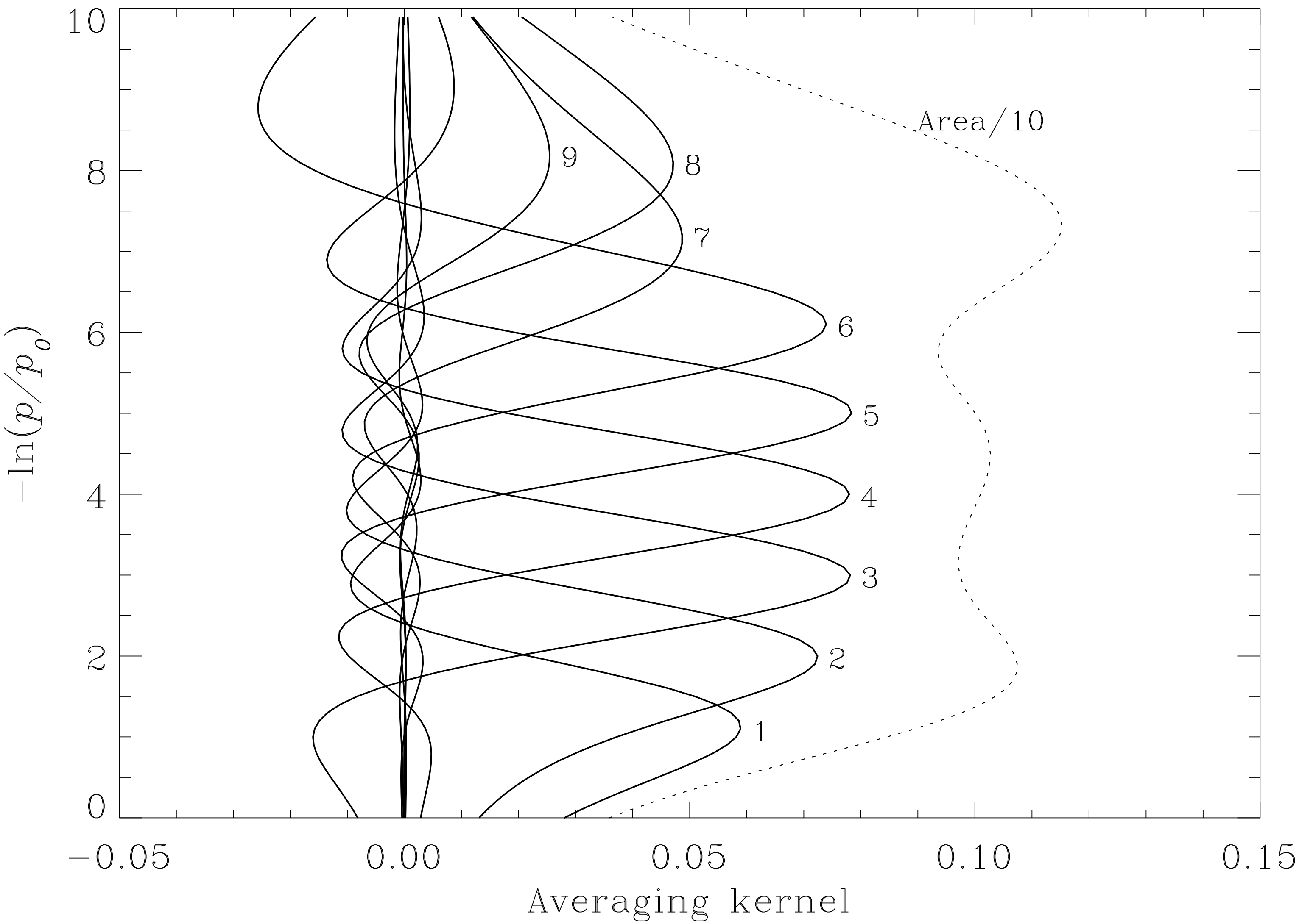
- accept that the retrieval is an estimate of a smoothed state, not the true state

  or
- consider the retrieval as an estimate of the true state, with an error contribution due to smoothing.

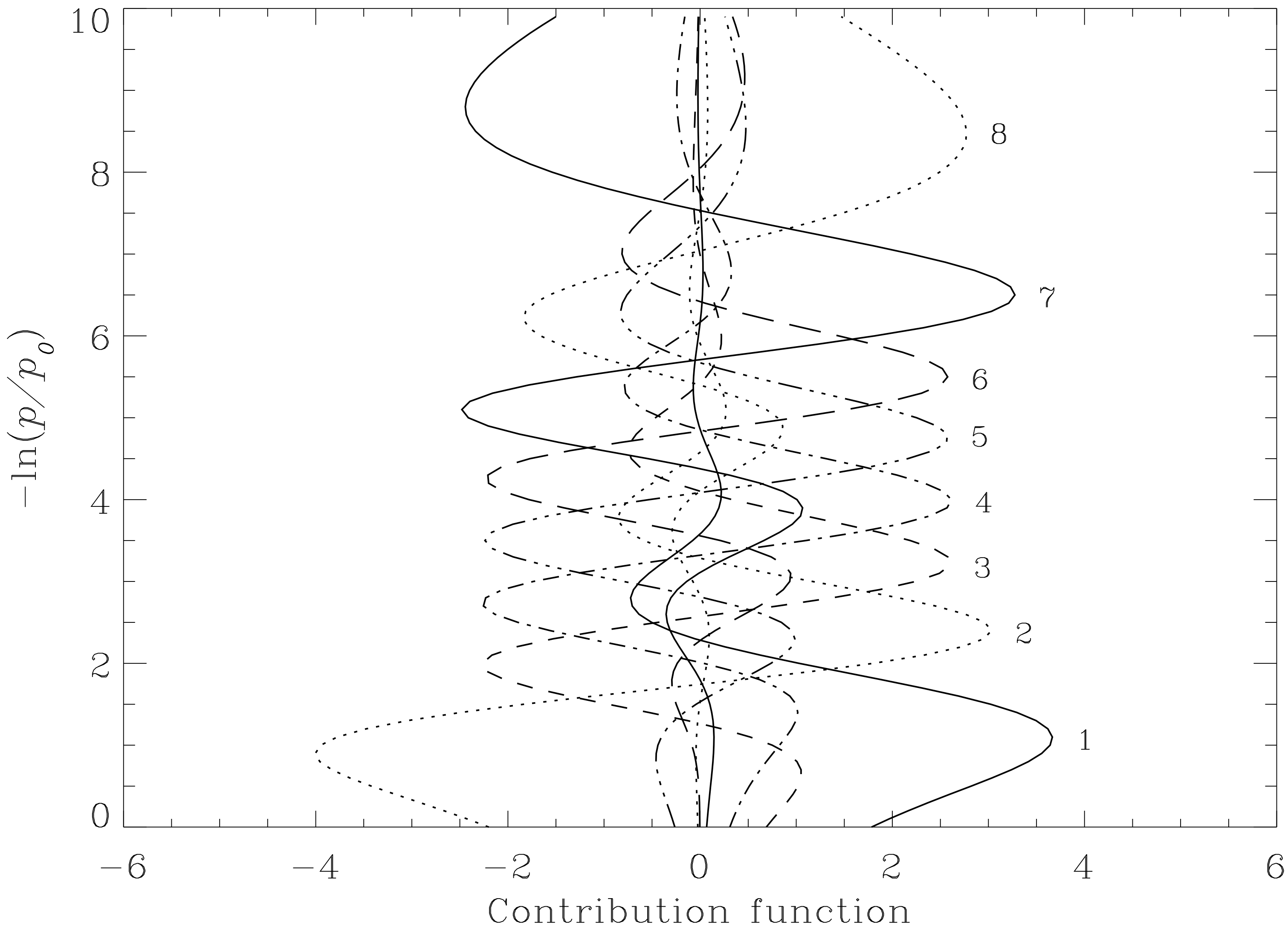
The error analysis is different in the two cases, because in the second case there is an extra term.

- If the state represents a profile, then the averaging kernel is a smoothing function with a *width* and an *area*.
  - The *width* is a measure of the resolution of the observing system
  - The *area* (generally between zero and unity) is a simple measure of the amount of real information that appears in the retrieval.

STANDARD EXAMPLE: AVERAGING KERNEL



STANDARD EXAMPLE: RETRIEVAL GAIN



THE OTHER TWO PARAMETERS

"The retrieved quantity is expressed as:

$$\hat{\mathbf{x}} = \mathbf{R}(\mathbf{f}(\mathbf{x}, \mathbf{b}, \mathbf{b}') + \boldsymbol{\epsilon}, \hat{\mathbf{b}}, \mathbf{x}_a, \mathbf{c})$$

where we have also separated $\mathbf{x}_a$, the *a priori*, from other components

of $\mathbf{c}$."

We should also look at the sensitivity of the retrieval to the inverse model parameters, $\mathbf{x}_a$ and $\mathbf{c}$.

The linear expansion in $\mathbf{x}$, $\mathbf{b}$ and $\boldsymbol{\epsilon}$ gave:

$$\hat{\mathbf{x}} = [\mathbf{R}(\mathbf{F}(\mathbf{x}_a, \hat{\mathbf{b}}), \hat{\mathbf{b}}, \mathbf{x}_a, \mathbf{c})] + \mathbf{A}(\mathbf{x} - \mathbf{x}_a) + \mathbf{G}_y\boldsymbol{\epsilon}_y$$

We argued that the term in square brackets should be equal to $\mathbf{x}_a$, for any reasonable retrieval.

This has the consequence that, at least to first order, $\partial\mathbf{R}/\partial\mathbf{c} = 0$ for any reasonable retrieval.

THE OTHER TWO PARAMETERS II

$$\hat{\mathbf{x}} = [\mathbf{R}(\mathbf{F}(\mathbf{x}_a, \hat{\mathbf{b}}), \hat{\mathbf{b}}, \mathbf{x}_a, \mathbf{c})] + \mathbf{A}(\mathbf{x} - \mathbf{x}_a) + \mathbf{G}_y \boldsymbol{\epsilon}_y$$

It also follows that:

$$\frac{\partial \mathbf{R}}{\partial \mathbf{x}_a} = \frac{\partial \hat{\mathbf{x}}}{\partial \mathbf{x}_a} = \mathbf{I}_n - \mathbf{A}$$

For any reasonable retrieval.

Thus the reasonableness criterion has the consequence that there can be no inverse model parameters $\mathbf{c}$ that matter, other than $\mathbf{x}_a$.

Incidentally:

The term in square brackets should not depend on $\hat{\mathbf{b}}$ either.

This implies that $\mathbf{G}_b + \mathbf{G}_y \mathbf{K}_b = 0$, which is no more than:

$$\frac{\partial \hat{\mathbf{x}}}{\partial \mathbf{b}} + \frac{\partial \hat{\mathbf{x}}}{\partial \mathbf{y}} \frac{\partial \mathbf{y}}{\partial \mathbf{b}} = 0$$

ERROR ANALYSIS

Some further rearrangement gives for the error in $\hat{\mathbf{x}}$:

$$\begin{aligned} \hat{\mathbf{x}} - \mathbf{x} = \quad & (\mathbf{A} - \mathbf{I})(\mathbf{x} - \mathbf{x}_a) && \textit{smoothing} \\ & +\mathbf{G}_y \epsilon_y && \textit{measurement error} \end{aligned}$$

where the bias term has been dropped, and:

$$\epsilon_y = \mathbf{K}_b(\mathbf{b} - \hat{\mathbf{b}}) + \Delta\mathbf{f}(\mathbf{x}, \mathbf{b}, \mathbf{b}') + \epsilon$$

Thus the error sources can be split up as:

$$\begin{aligned} \hat{\mathbf{x}} - \mathbf{x} = \quad & (\mathbf{A} - \mathbf{I})(\mathbf{x} - \mathbf{x}_a) && \textit{smoothing} \\ & +\mathbf{G}_y \mathbf{K}_b(\mathbf{b} - \hat{\mathbf{b}}) && \textit{model parameters} \\ & +\mathbf{G}_y \Delta\mathbf{f}(\mathbf{x}, \mathbf{b}, \mathbf{b}') && \textit{modelling error} \\ & +\mathbf{G}_y \epsilon && \textit{measurement noise} \end{aligned}$$

Some of these are easy to estimate, some are not.

MEASUREMENT NOISE

$$\textit{measurement noise } = \mathbf{G}_y \boldsymbol{\epsilon}$$

This is the easiest component to evaluate.

$\boldsymbol{\epsilon}$ is usually random noise, and is often unbiassed and uncorrelated between channels, and has a known covariance matrix.

The covariance of the measurement noise is:

$$\mathbf{S}_n = \mathbf{G}_y \mathbf{S}_\epsilon \mathbf{G}_y^T$$

Note that $\mathbf{S}_n$ is not necessarily diagonal, so there will be errors which are correlated between different elements of the state vector.

This is true of all of the error sources.

SMOOTHING ERROR

To estimate the *actual* smoothing error, you need to know the true state:

$$\textit{smoothing error} = (\mathbf{A} - \mathbf{I})(\mathbf{x} - \mathbf{x}_a)$$

To characterise the statistics of this error, you need its mean and covariance over some ensemble.

The mean should be zero.

The covariance is:

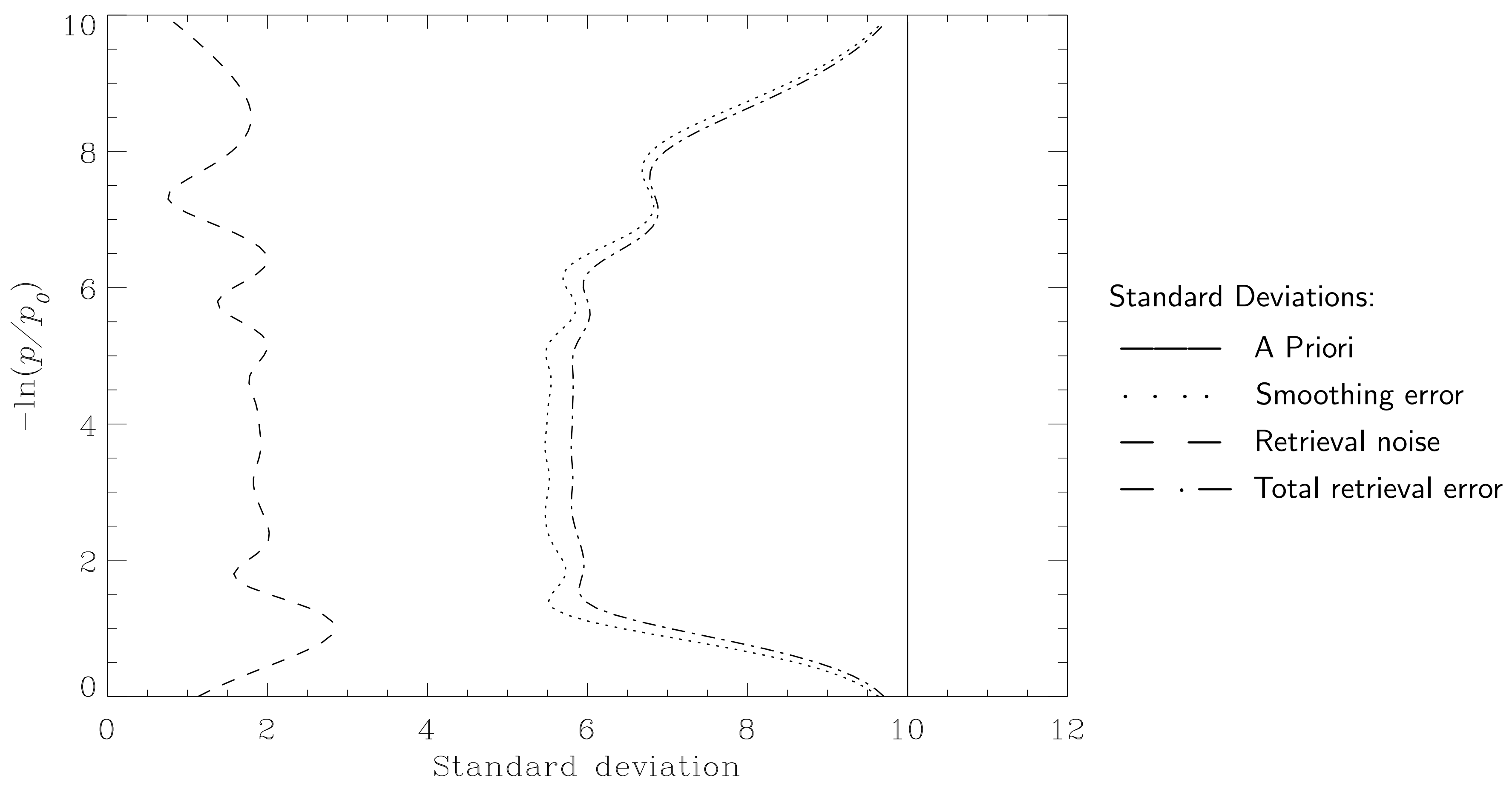
$$\begin{aligned}
\mathbf{S}_s &= \mathcal{E}\{(\mathbf{A} - \mathbf{I})(\mathbf{x} - \mathbf{x}_a)(\mathbf{x} - \mathbf{x}_a)^T(\mathbf{A} - \mathbf{I})^T\} \\
&= (\mathbf{A} - \mathbf{I})\mathcal{E}\{(\mathbf{x} - \mathbf{x}_a)(\mathbf{x} - \mathbf{x}_a)^T\}(\mathbf{A} - \mathbf{I})^T \\
&= (\mathbf{A} - \mathbf{I})\mathbf{S}_a(\mathbf{A} - \mathbf{I})^T
\end{aligned}$$

where $\mathbf{S}_a$ is the covariance of an ensemble of states about the *a priori* state. This is best regarded as the covariance of a climatology.

To estimate the smoothing error, you need to know the climatological covariance matrix.

To do the job properly, you need the real climatology, not just some *ad hoc* matrix that has been used as a constraint in the retrieval. The real climatology is often not available. Much of the smoothing error can be in fine spatial scales that may never have been measured.

STANDARD EXAMPLE: ERROR COMPONENTS



FORWARD MODEL ERRORS

Forward model parameters

$$\textit{forward model parameter error} = \mathbf{G}_y\mathbf{K}_b(\mathbf{b} - \hat{\mathbf{b}})$$

This one is easy. (In principle)

If you have estimated the forward model parameters properly, their individual errors will be unbiassed, so the mean error will be zero.

The covariance is:

$$\mathbf{S}_f = \mathbf{G}_y\mathbf{K}_b\mathbf{S}_b\mathbf{K}_b^T\mathbf{G}_y^T$$

where $\mathbf{S}_b$ is the error covariance matrix of $\mathbf{b}$, namely $\mathcal{E}\{(\mathbf{b} - \hat{\mathbf{b}})(\mathbf{b} - \hat{\mathbf{b}})^T\}$

However remember that this is most probably a systematic, not a random error.

FORWARD MODEL ERRORS II

Modelling error

$$\textit{modelling error } = \mathbf{G}_y \Delta \mathbf{f} = \mathbf{G}_y(\mathbf{f}(\mathbf{x}, \mathbf{b}, \mathbf{b}') - \mathbf{F}(\mathbf{x}, \mathbf{b}))$$

Note that this is evaluated at the true state, and with the true value of $\mathbf{b}$, but hopefully its sensitivity to these quantities is not large.

This can be hard to evaluate, because it requires a model $\mathbf{f}$ which includes the correct physics. If $\mathbf{F}$ is simply a numerical approximation for efficiency's sake, it may not be too difficult, but if $\mathbf{f}$ is not known in detail, or so horrendously complex that no proper model is feasible, then modelling error can be tricky to estimate.

This is also usually a systematic error.

ERROR COVARIANCE MATRIX

An Error Covariance Matrix $\mathbf{S}$ is defined as

$$S_{ij} = \mathcal{E}\{\epsilon_i \epsilon_j\}$$

Diagonal elements are the familiar error variances.

The corresponding probability density function (PDF), if Gaussian, is:

$$P(\mathbf{y}) \propto \exp\left(-\frac{1}{2}(\mathbf{y} - \bar{\mathbf{y}})^T \mathbf{S}^{-1}(\mathbf{y} - \bar{\mathbf{y}})\right)$$

Contours of the PDF are

$$(\mathbf{y} - \bar{\mathbf{y}})^T \mathbf{S}^{-1}(\mathbf{y} - \bar{\mathbf{y}}) = const$$

i.e. ellipsoids, with arbitrary axes.

CORRELATED ERRORS

- How do we understand an error covariance matrix?
- What corresponds to *error bars* for a profile?

We would really like a PDF to be of the form

$$P(\mathbf{z}) \propto \prod_i \exp(-z_i^2/2\sigma_i^2)$$

i.e. each z_i to have independent errors.

This can be done by diagonalising $\mathbf{S}$, and substituting $\mathbf{S} = \mathbf{L}\Lambda\mathbf{L}^T$, where $\mathbf{L}$ is matrix of eigenvectors $\mathbf{l}_i$. Then

$$P(\mathbf{x}) \propto \exp\left(-\frac{1}{2}(\mathbf{x}-\bar{\mathbf{x}})^T\mathbf{L}\Lambda^{-1}\mathbf{L}^T(\mathbf{x}-\bar{\mathbf{x}})\right)$$

So if we put $\mathbf{z} = \mathbf{L}^T(\mathbf{x}-\bar{\mathbf{x}})$ then the z_i are independent with variance λ_i.

ERROR PATTERNS

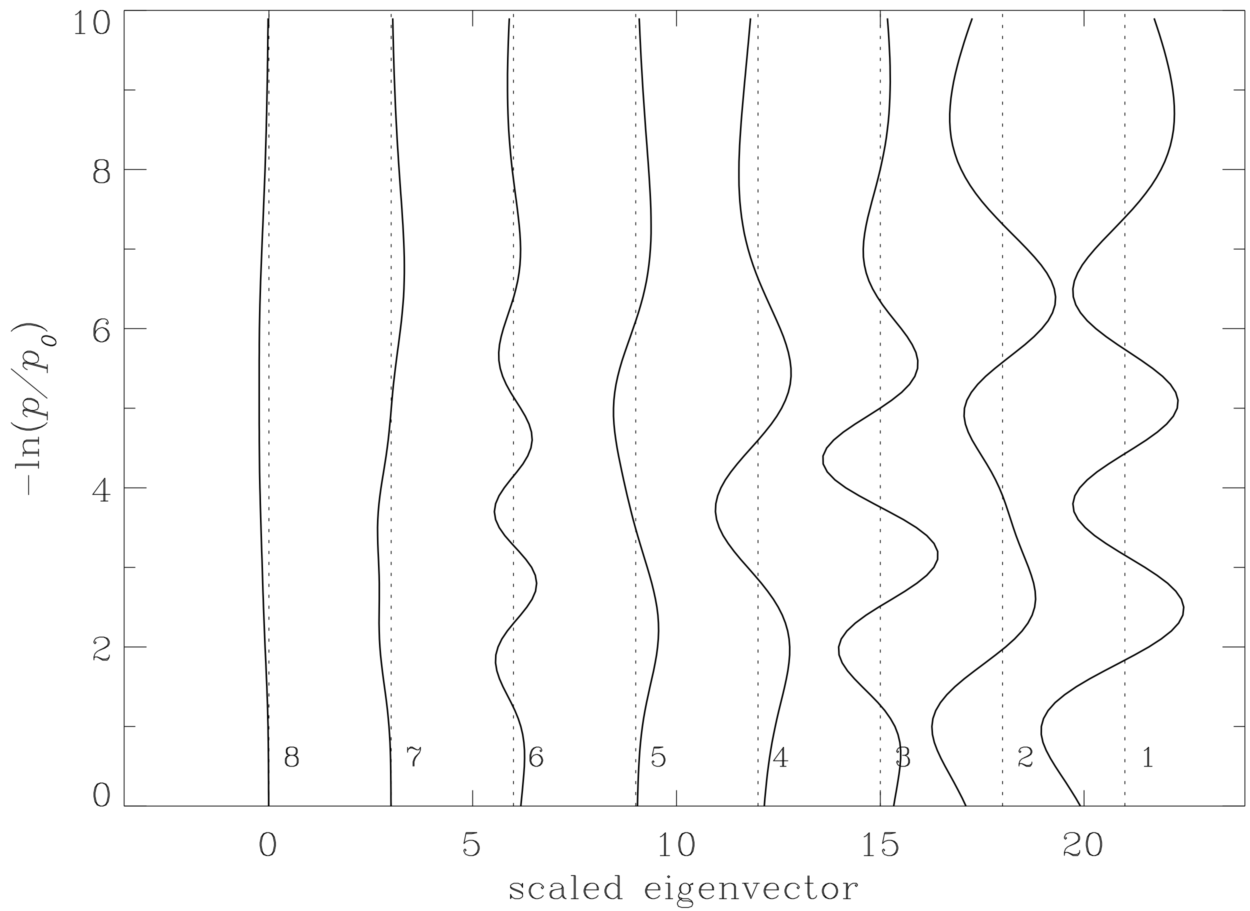
- Error covariance is described in state space by an ellipsoid
- We have, in effect, transformed to the principal axes of the ellipsoid
- We can express the error in e.g. a state vector estimate $\hat{\mathbf{x}}$, due to an error covariance $\hat{\mathbf{S}}$, in terms of *Error Patterns* $\mathbf{e}_i = \lambda_i^{\frac{1}{2}} \mathbf{l}_i$ such that the total error is of the form

$$\hat{\mathbf{x}} - \mathbf{x} = \sum_i \beta_i \mathbf{e}_i$$

  where the error patterns are orthogonal, and the coefficients β are independent random variables with unit variance.

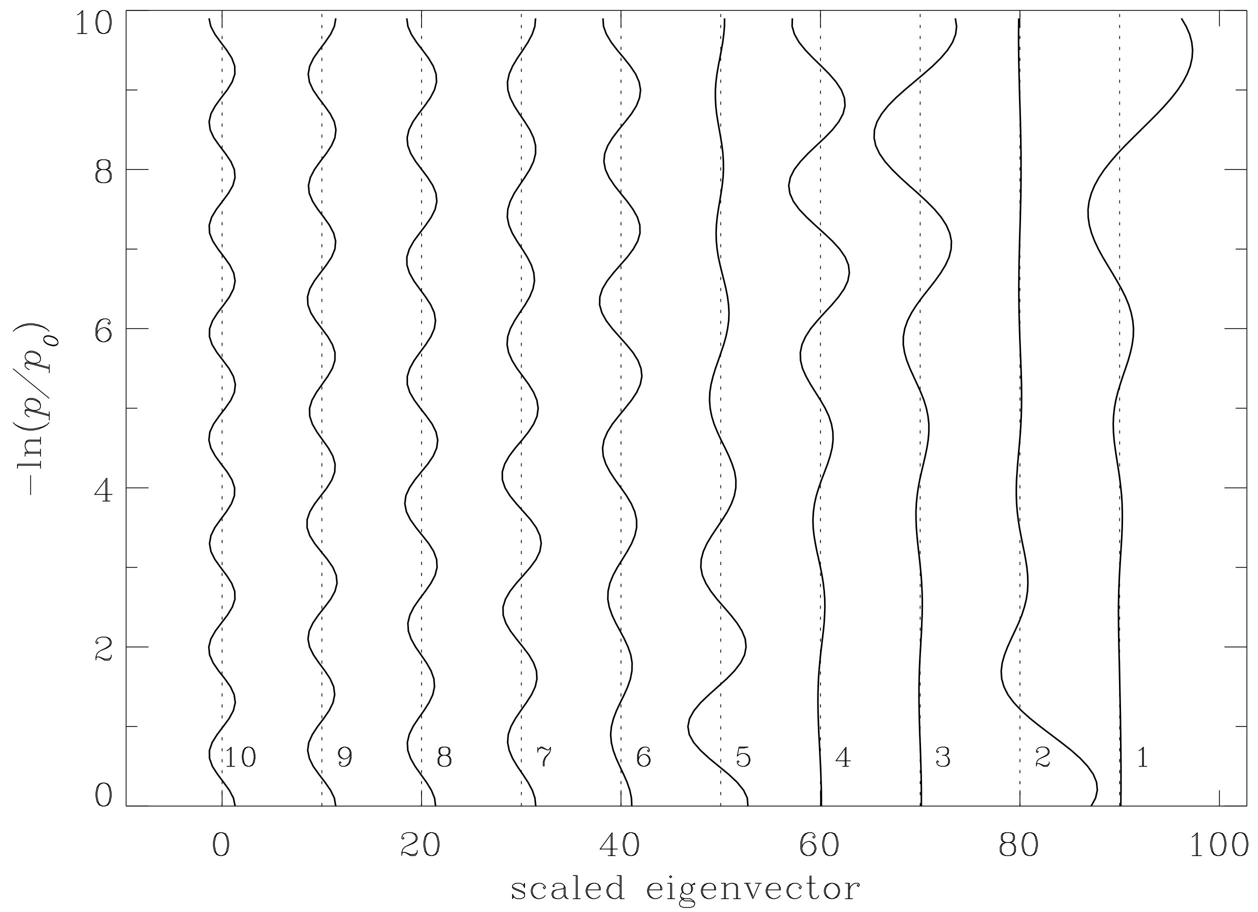
STANDARD EXAMPLE: NOISE ERROR PATTERNS

There are only eight, same as the rank of the problem.



STANDARD EXAMPLE: SMOOTHING ERROR PATTERNS

The largest ten out of one hundred, the dimension of the state vector.



Unsolved Problem

How do you describe an error covariance to the naive data user?

End of Section