Application of Principal Component Analysis to TES data

Clive D Rodgers
Clarendon Laboratory
University of Oxford

Madison, Wisconsin, 27th April 2006

My take on the PCA business

What is the best estimate of a spectrum given many measured spectra?

Simple approach using optimal estimation:

- Forward model is

$$\mathbf{y} = \mathbf{x} + \varepsilon$$

$\mathbf{x}$ is true spectrum, $\mathbf{y}$ is measurement, ε is noise

- Maximum a posteriori estimate of $\mathbf{x}$ is

$$\mathbf{x}_r = \mathbf{x}_a + \mathbf{S}_a(\mathbf{S}_a + \mathbf{S}_\varepsilon)^{-1}(\mathbf{y} - \mathbf{x}_a)$$

Continued...

- If we have a large sample of spectra:
  - Expect that $\mathbf{x}_a = \langle \mathbf{x} + \varepsilon \rangle = \langle \mathbf{y} \rangle$
  - Can estimate $\mathbf{S}_a + \mathbf{S}_\varepsilon$ from statistics of $\mathbf{y}$
  - Should have a good idea of $\mathbf{S}_\varepsilon$ a priori

- But:
  - $\mathbf{S}_a(\mathbf{S}_a + \mathbf{S}_\varepsilon)^{-1}$ will be a large matrix
  - $\mathbf{S}_a$ found from $\mathbf{S}_a + \mathbf{S}_\varepsilon$ and $\mathbf{S}_\varepsilon$ is likely to be ill-conditioned

- $\mathbf{S}_a$ is likely to have a 'small' number of eigenvalues greater than noise

Singular Vectors
(or Principal Components)

- Let the ensemble of ($\mathbf{y}$ - <$\mathbf{y}$>) be N columns of a matrix $\mathbf{Y}$
- Represent $\mathbf{Y}$ as its singular vector decomposition:

$$\mathbf{Y} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^{T}$$

where $\boldsymbol{\Lambda}$ is diagonal, $\mathbf{U}^{T}\mathbf{U}=\mathbf{I}$ and $\mathbf{V}^{T}\mathbf{V}=\mathbf{I}$

- The j'th individual spectrum $\mathbf{y}_j$ is then

$$\mathbf{y}_j = \langle\mathbf{y}\rangle + \Sigma_i\, \boldsymbol{u}_i \lambda_i v_{ij}^{T}$$

- The spectrum is represented as a sum of columns $\boldsymbol{u}_i$ of $\mathbf{U}$, with coefficients $\lambda_i v_{ij}^{T}$.
- Because $\mathbf{U}^{T}\mathbf{U}=\mathbf{I}$, we can compute $\lambda_i v_{ij}^{T}$ for any spectrum as $\mathbf{U}^{T}(\mathbf{y}_j - \langle\mathbf{y}\rangle)$.

What do we expect?

- $^1/_N\mathbf{YY}^T$ is the covariance matrix $\mathbf{S}_y$ of the spectra
- Left singular vectors $\mathbf{U}$ are the same as eigenvectors of $\mathbf{YY}^T$, singular values are the square roots of its eigenvalues
- In the linear case with independent constant noise, $\mathbf{S}_y$ would be

$$\mathbf{S}_y = \mathbf{S}_a + \sigma_\varepsilon^2\ \mathbf{I}$$

- $\mathbf{S}_a$ has rank $\leq n$, $\mathbf{I}$ is of dimension $m \gg n$, where n is the degrees of freedom of the atmosphere
- The eigenvalues of $\mathbf{S}_y$ are λ_i^2/N
- The eigenvalues of $\mathbf{S}_a$ should be $\lambda_i^2/N - \sigma_\varepsilon^2$

Reconstructing Spectra

- We can drop terms with $\lambda_i^2/N \sim \sigma_\varepsilon^2$, i.e. $i > n$, without significant loss
  - they correspond to noise only
  - Better, multiply retained terms by something like $(\lambda_i^2 - N\sigma_\varepsilon^2)/\lambda_i^2$
- So spectra can be reconstructed from the first few coefficients.
- The noise can be reconstructed from the rest…
- Reconstructed spectra have much reduced noise

Information Content

- The Shannon information content of a single spectrum relative to the ensemble is

$$H = {}^{1}/_{2}\sum_i \ln(\lambda_i^{2}/N\sigma_\varepsilon^{2})$$

- The degrees of freedom for signal is

$$d_s = \sum_i (1 - N\sigma_\varepsilon^{2}/\lambda_i^{2})$$

- But reality isn't quite like that. . .

Tropospheric Emission Spectrometer

- Connes-type four-port Fourier transform spectrometer
- Nadir view: 0.06 cm^{-1}
- Limb view: 0.015 cm^{-1}
- 16 element detector arrays in 4 focal planes:
  - 1A (1900-3050), 1B (820-1150), 2A (1100-1950), 2B (650-900)
- 0.5 X 5 km nadir; 2.3 X 23 km limb
- Global Survey mode: Along track nadir, 216 views/orbit.

SVD Applications

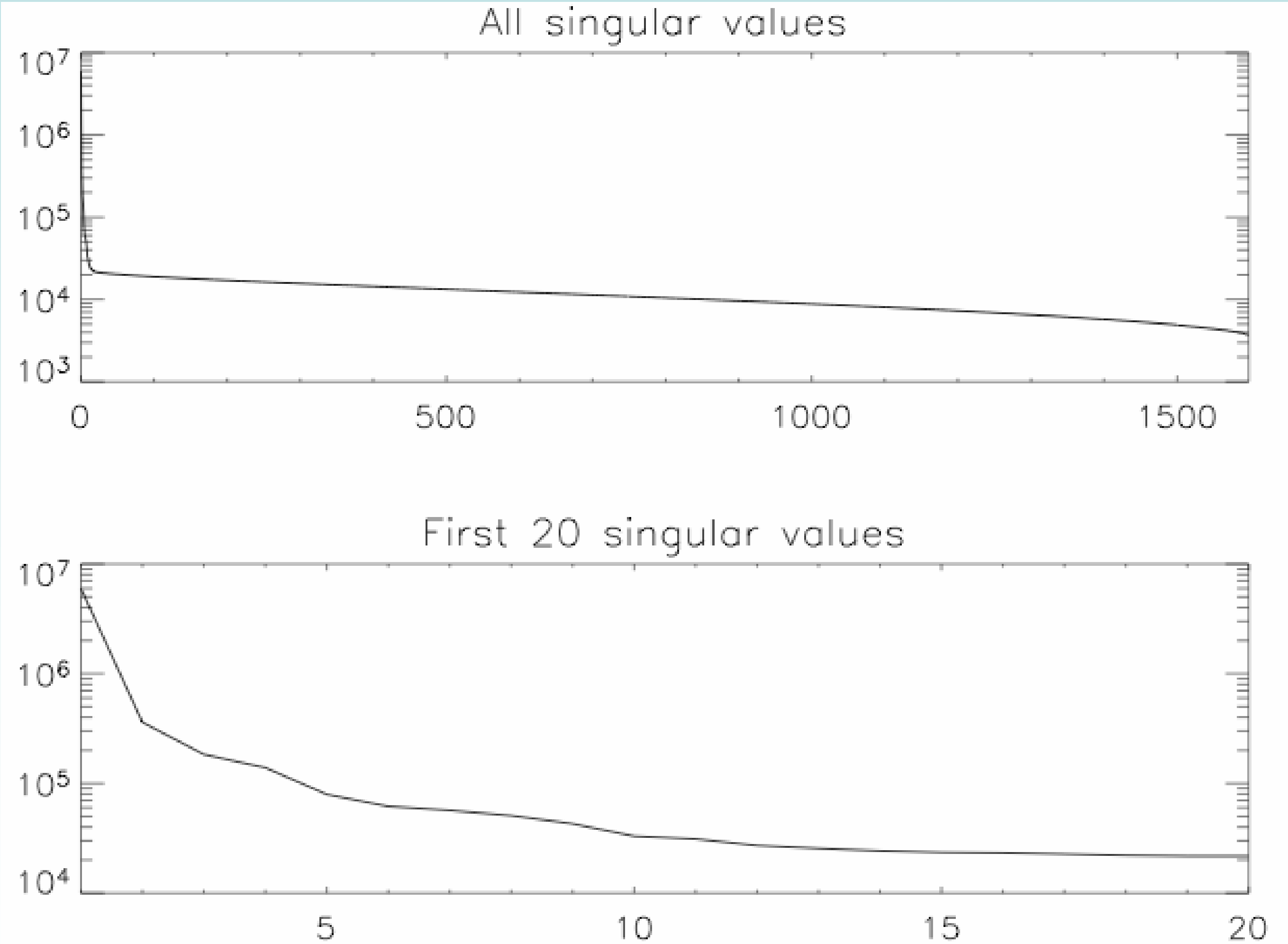
- Information content analysis
- Validation
  - Instrument performance
  - Forward model
- De-noised spectra
- Retrieval?

A typical example

Run 2147
21 Sept 2004?

Channel 1B2
923-1160 cm^{-1}

100 views * 16 detectors

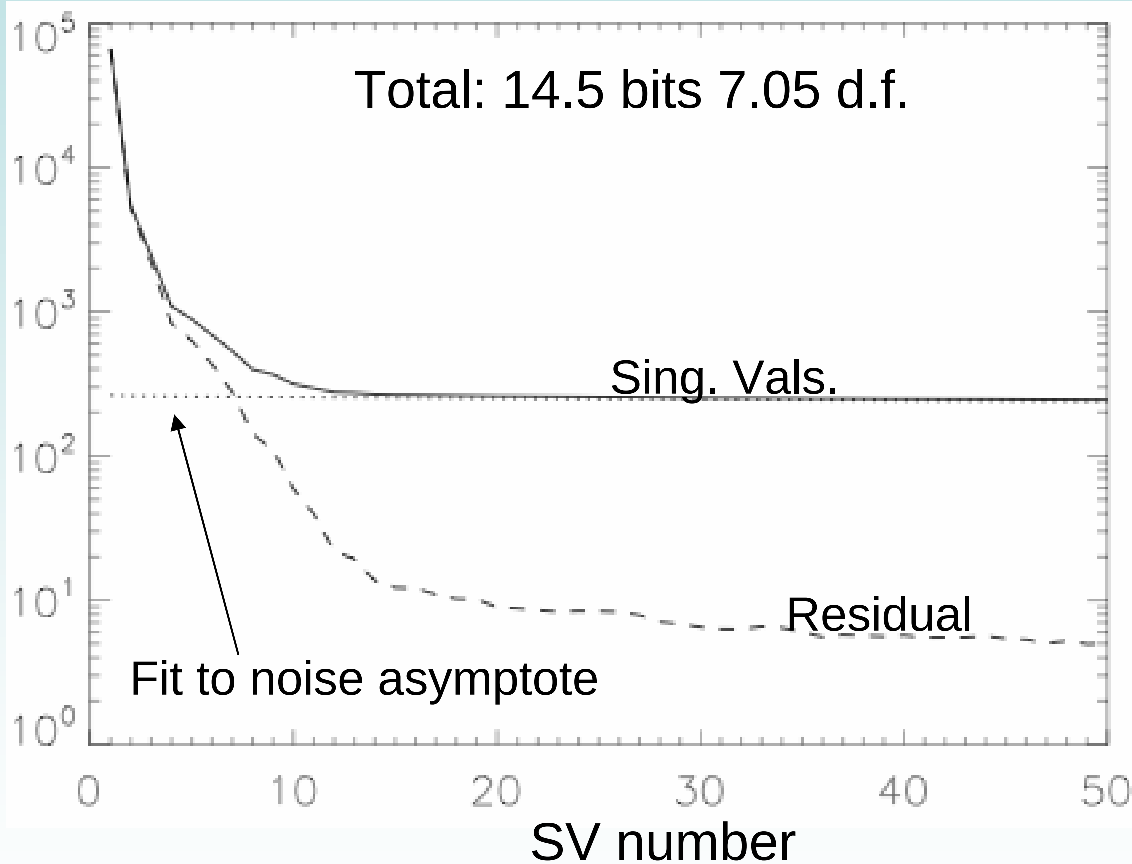


An example of Information Content from Channel 1B2

Using one orbit, 1152 spectra each with 3951 elements.

923-1160 cm^{-1}

n	Lambda	Information		d.f.s	
			cum.		cum.
0	253.315	5.535	5.53	1.000	1.000
1	19.940	2.994	8.53	1.997	0.997
2	8.496	2.147	10.68	2.984	0.986
3	3.226	1.217	11.89	3.896	0.912
4	2.446	0.972	12.86	4.753	0.857
5	1.675	0.668	13.53	5.490	0.737
6	1.072	0.382	13.91	6.025	0.535
7	0.553	0.133	14.05	6.259	0.234
8	0.433	0.086	14.13	6.417	0.158
9	0.232	0.026	14.16	6.468	0.051
10	0.156	0.012	14.17	6.492	0.024
11	0.086	0.004	14.18	6.499	0.007
12	0.076	0.003	14.18	6.505	0.006
13	0.054	0.001	14.18	6.508	0.003
14	0.048	0.001	14.18	6.510	0.002



How do you fit noise to the asymptote?

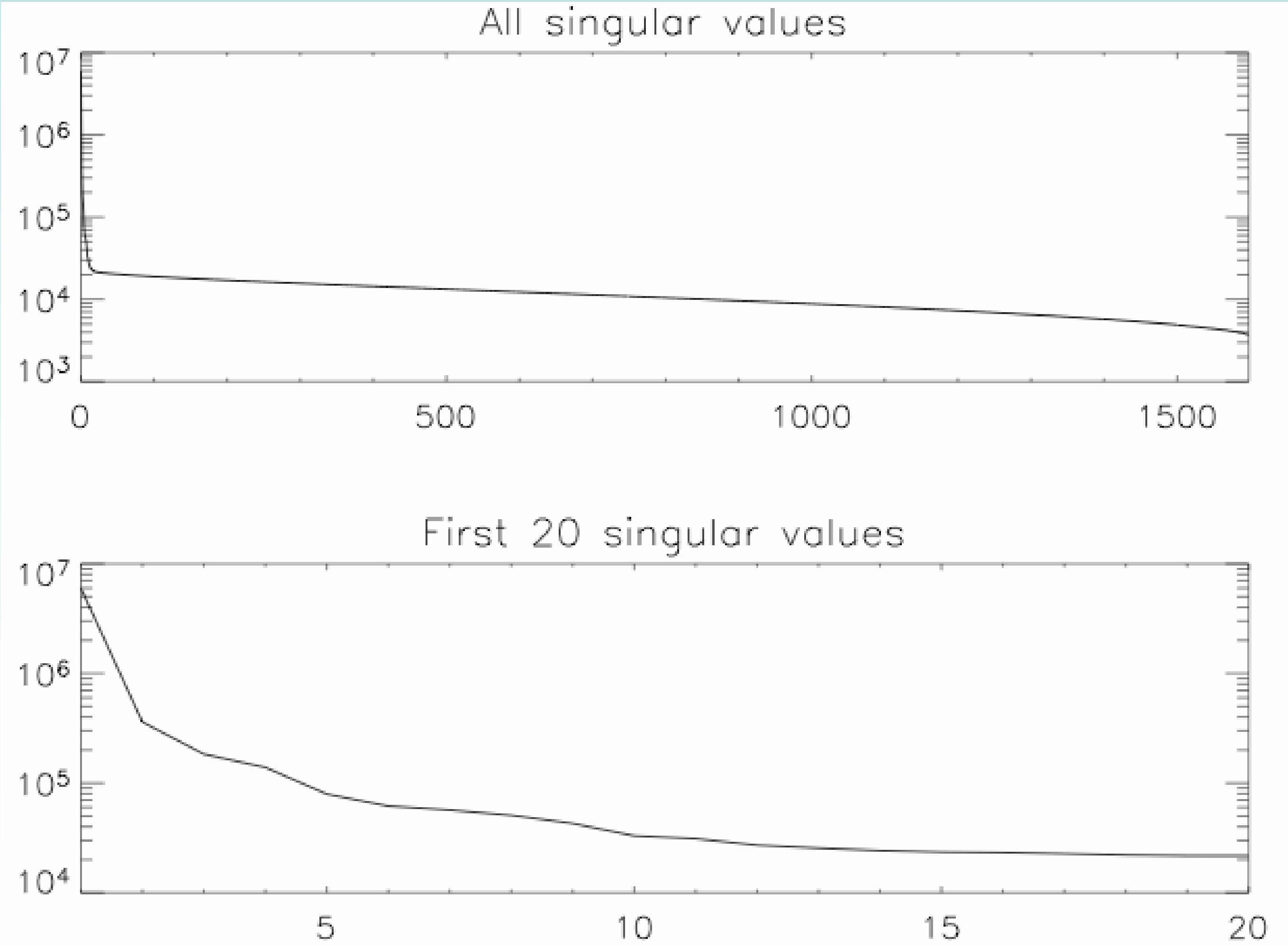
- I havn't used enough spectra…
- Expected distribution of singular values of independent gaussian noise for a finite number of samples
  - Simulation
  - Theory: there is an explicit formula
- Fit expected distribution to the long tail

Validation – an early example

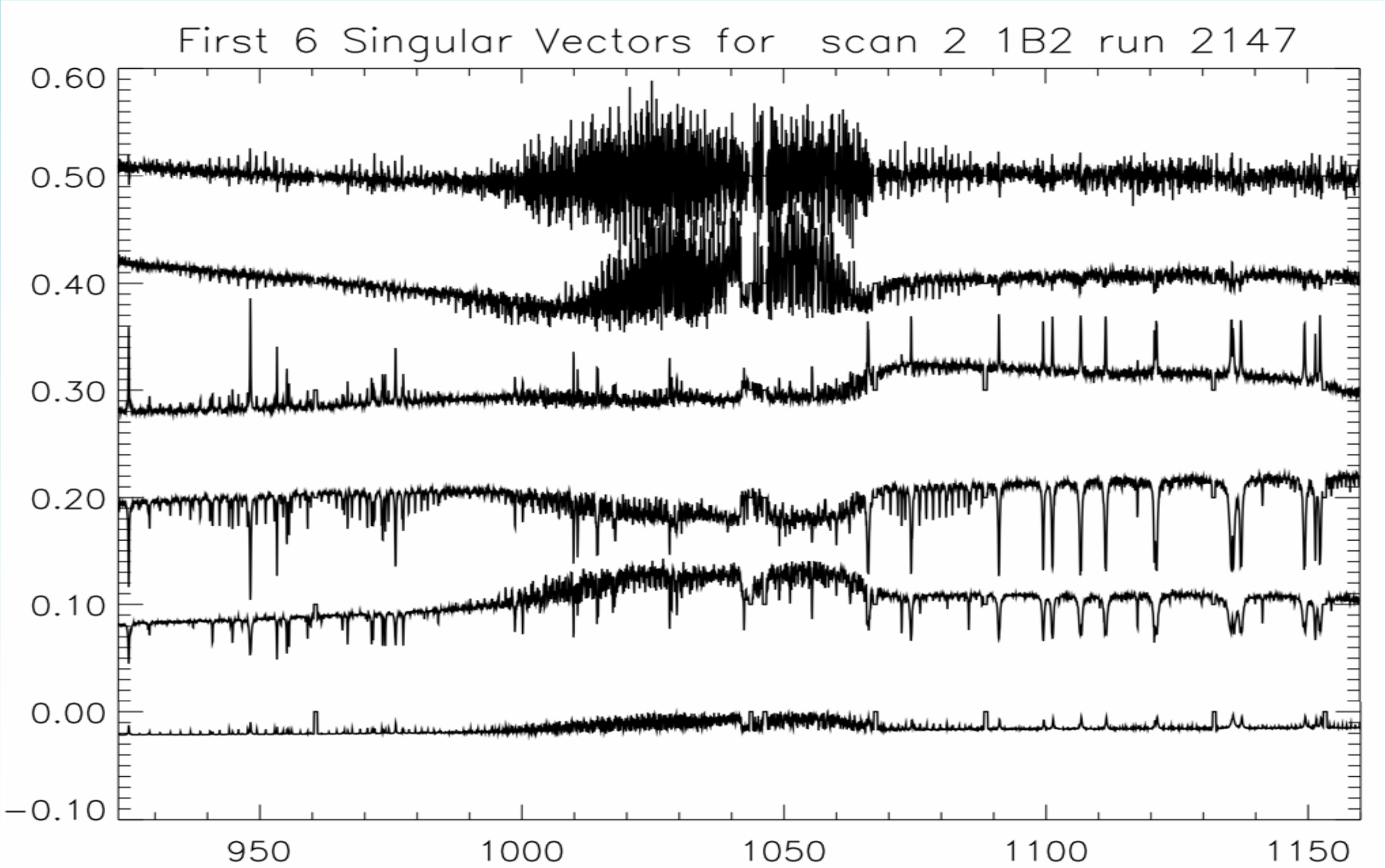
Run 2147
21 Sept 2004?

Channel 1B2
923-1160 cm^{-1}

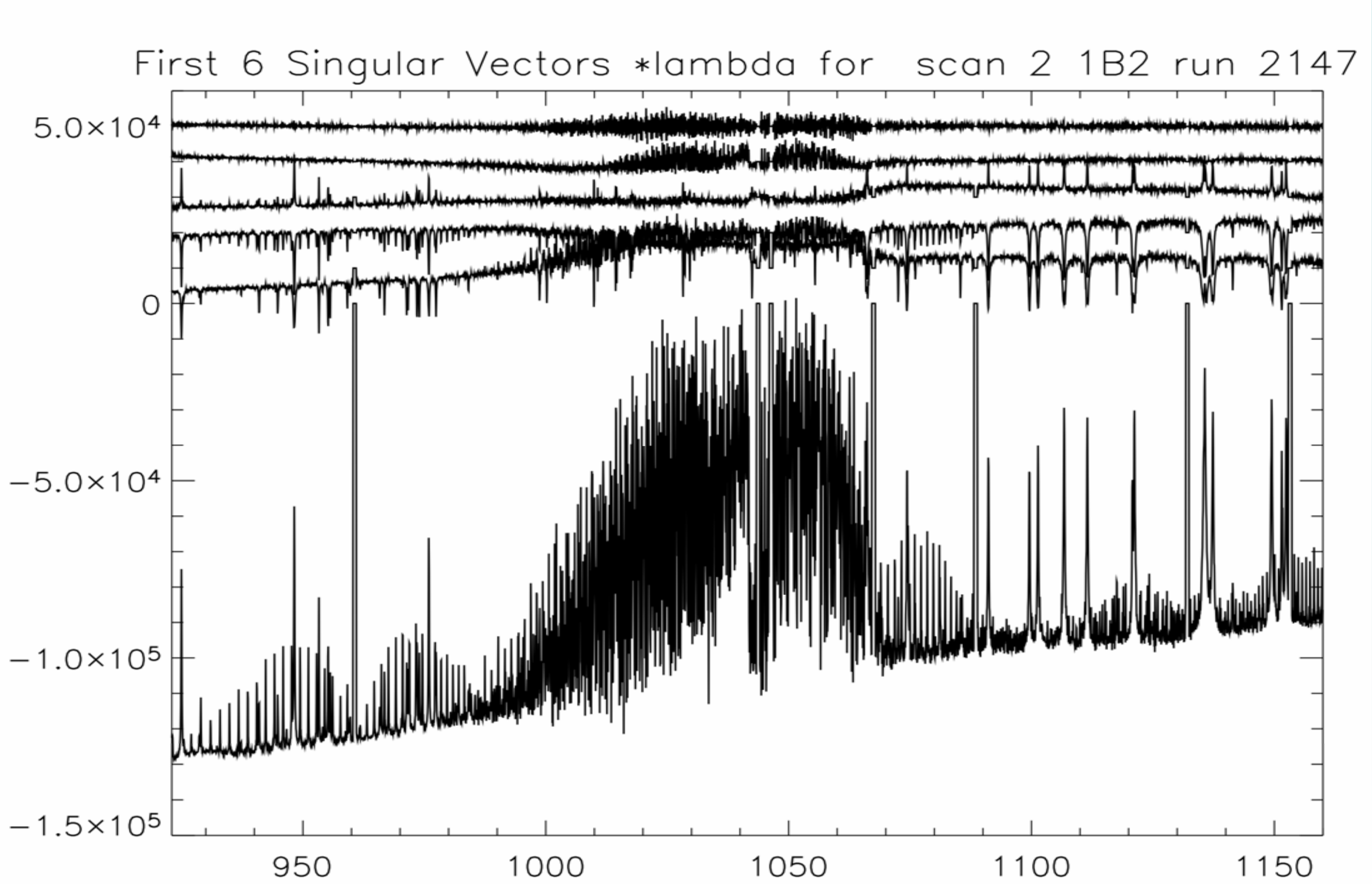
100 views * 16 detectors



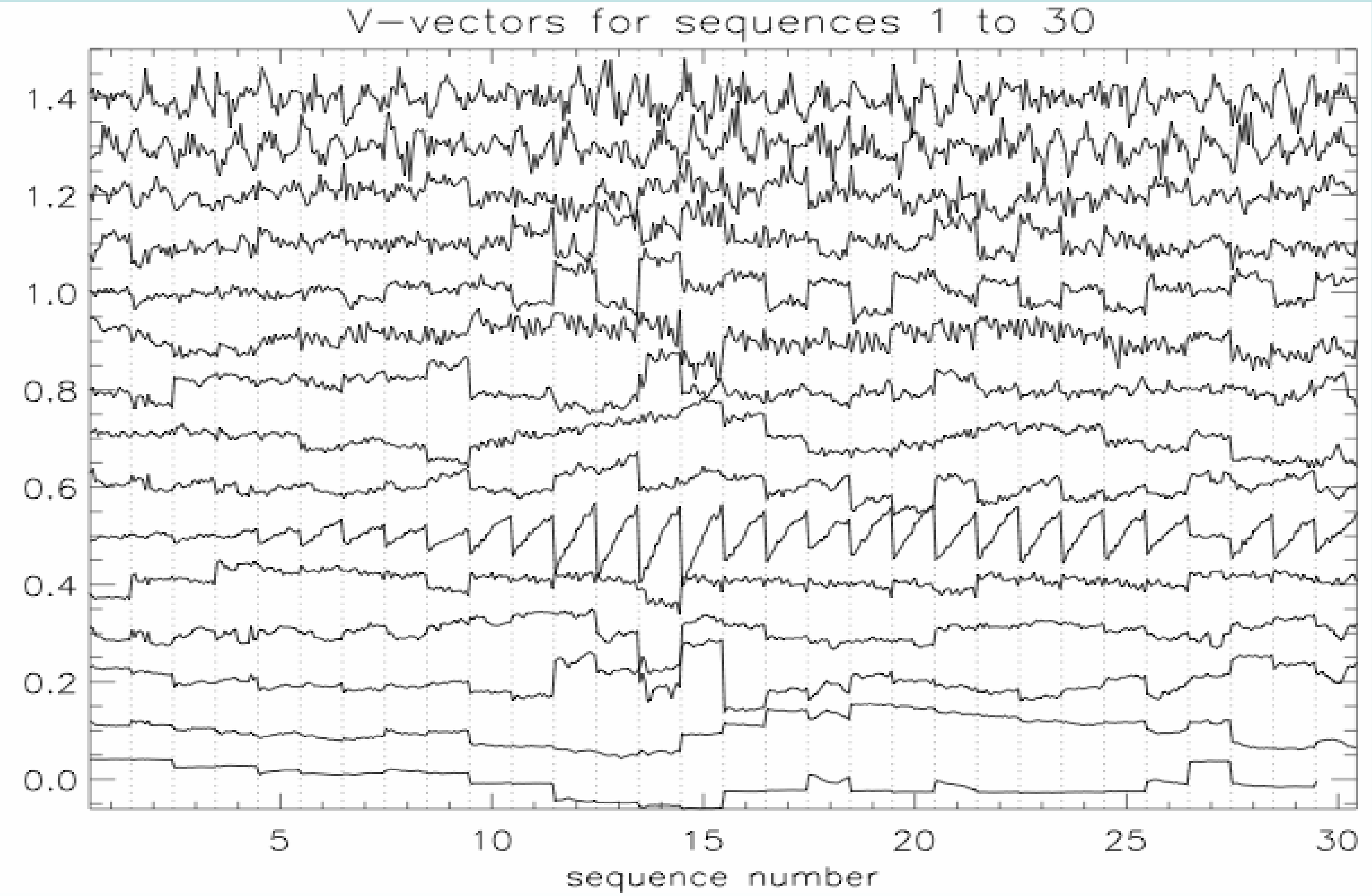
Singular Vectors



Singular vectors * Lambda



V-vectors


V–vectors for sequences 1 to 30
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
5
10
15
20
25
30
sequence number

Features

- Most of variation is in the first singular vector. First six are:

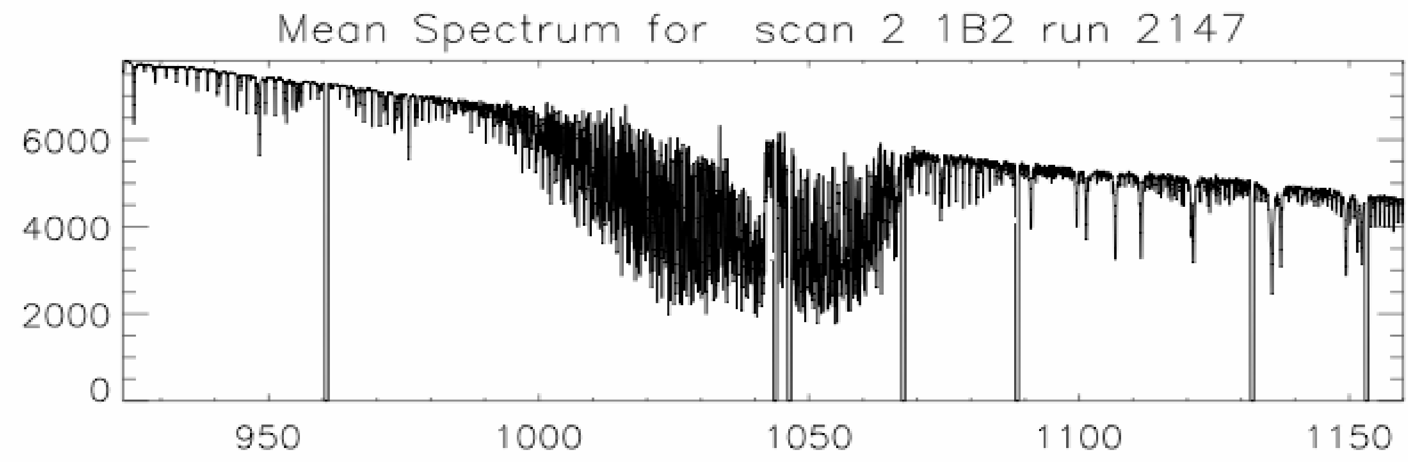
$$5.96\times10^6 \quad 3.6\times10^5 \quad 1.83\times10^5 \quad 1.39\times10^4 \quad 7.93\times10^4 \quad 6.16\times10^4$$

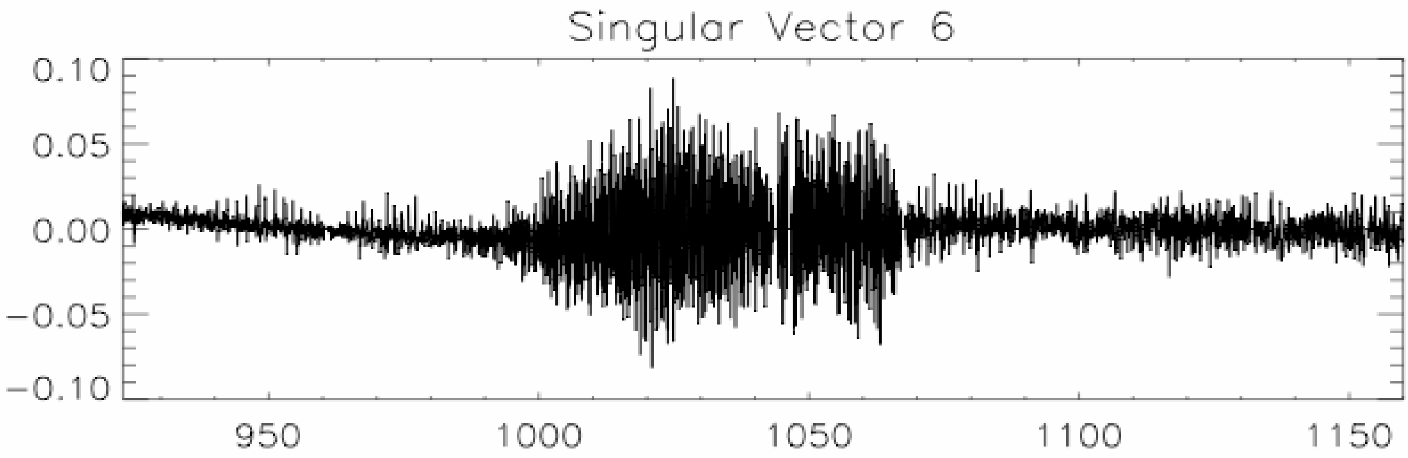
- Data spikes - identified
- Data spikes - unidentified
- Pixel-dependent variation in the spectra
- Scan-direction dependence

Singular Vector 6

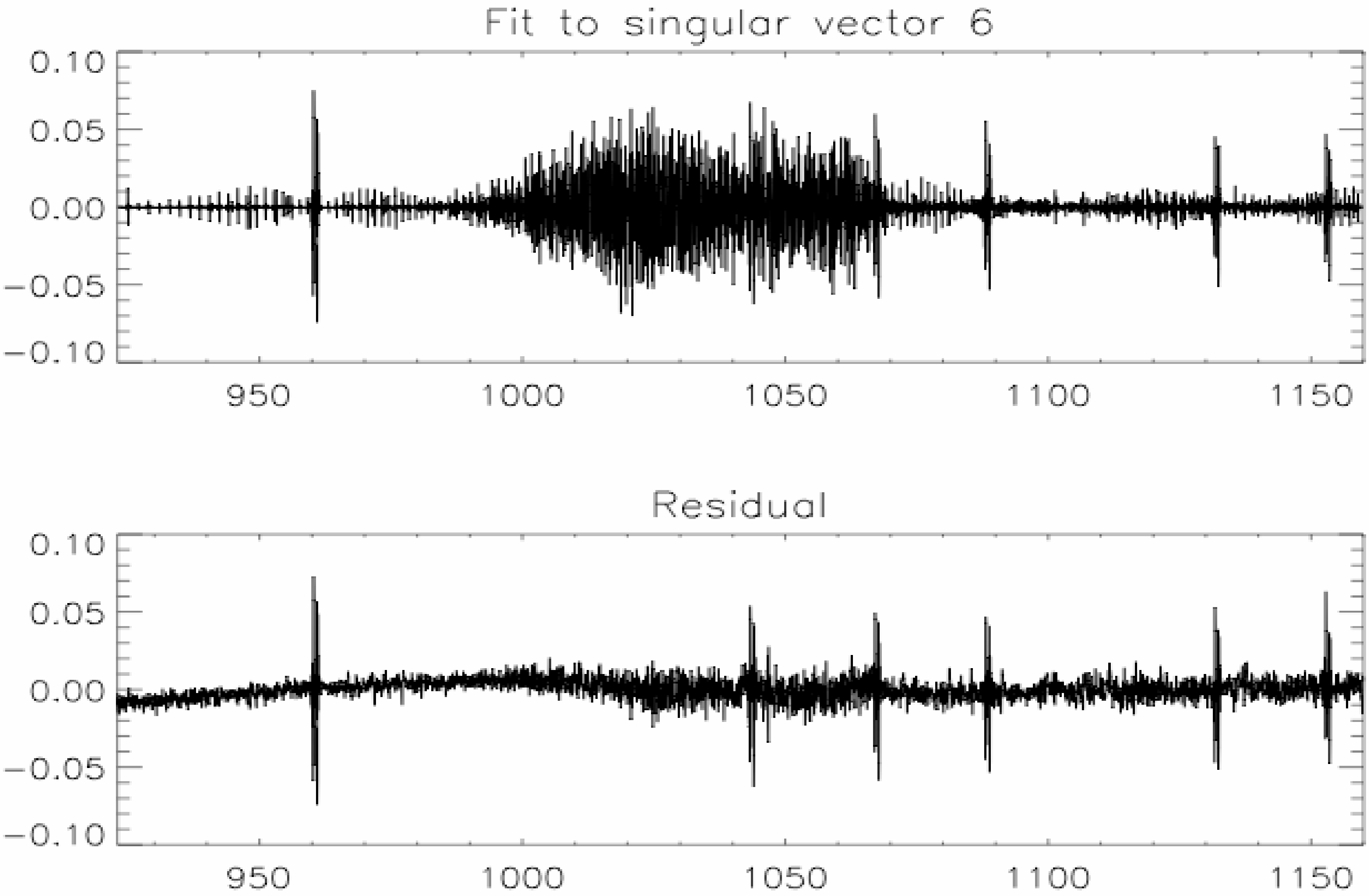
- Systematic variation across the detector array
- Must be an artefact
- Suggests systematic error in ILS
- How is it related to mean spectrum?
- Least squares fit to find function that when convolved with mean spectrum gives SV6

SV6


Mean Spectrum for scan 2 1B2 run 2147
6000
4000
2000
0
950
1000
1050
1100
1150

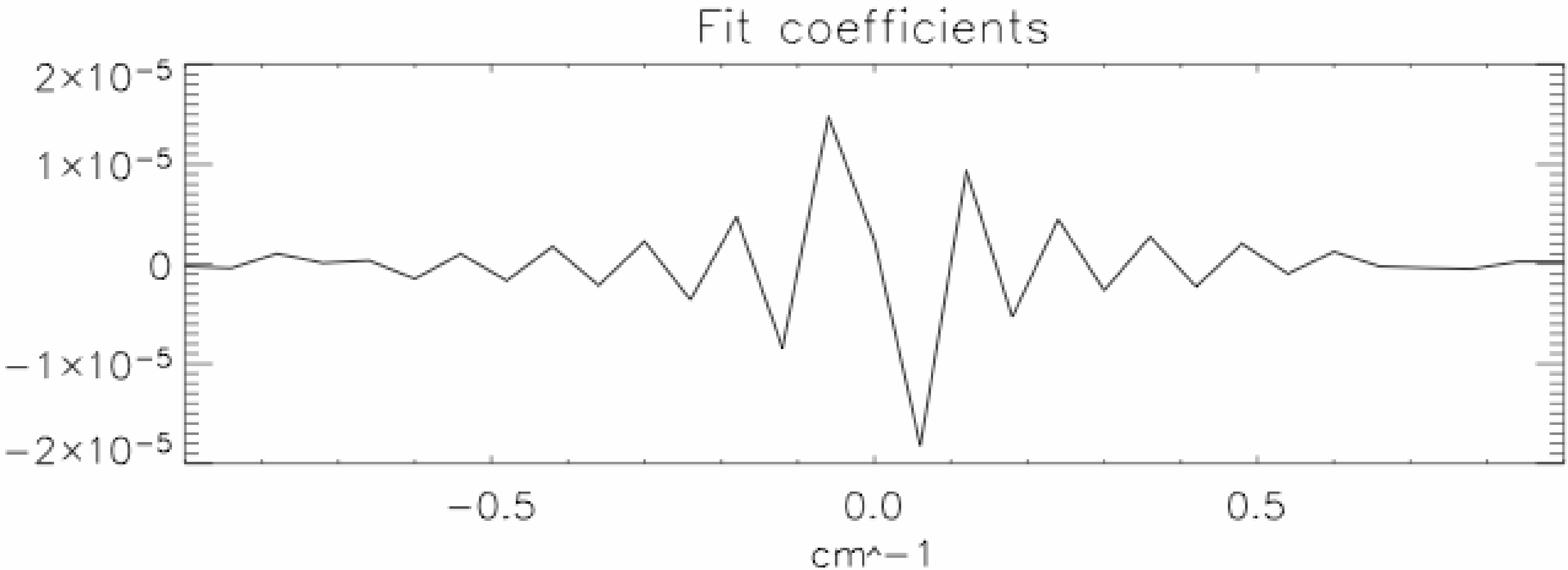

Singular Vector 6
0.10
0.05
0.00
−0.05
−0.10
950
1000
1050
1100
1150

SV6



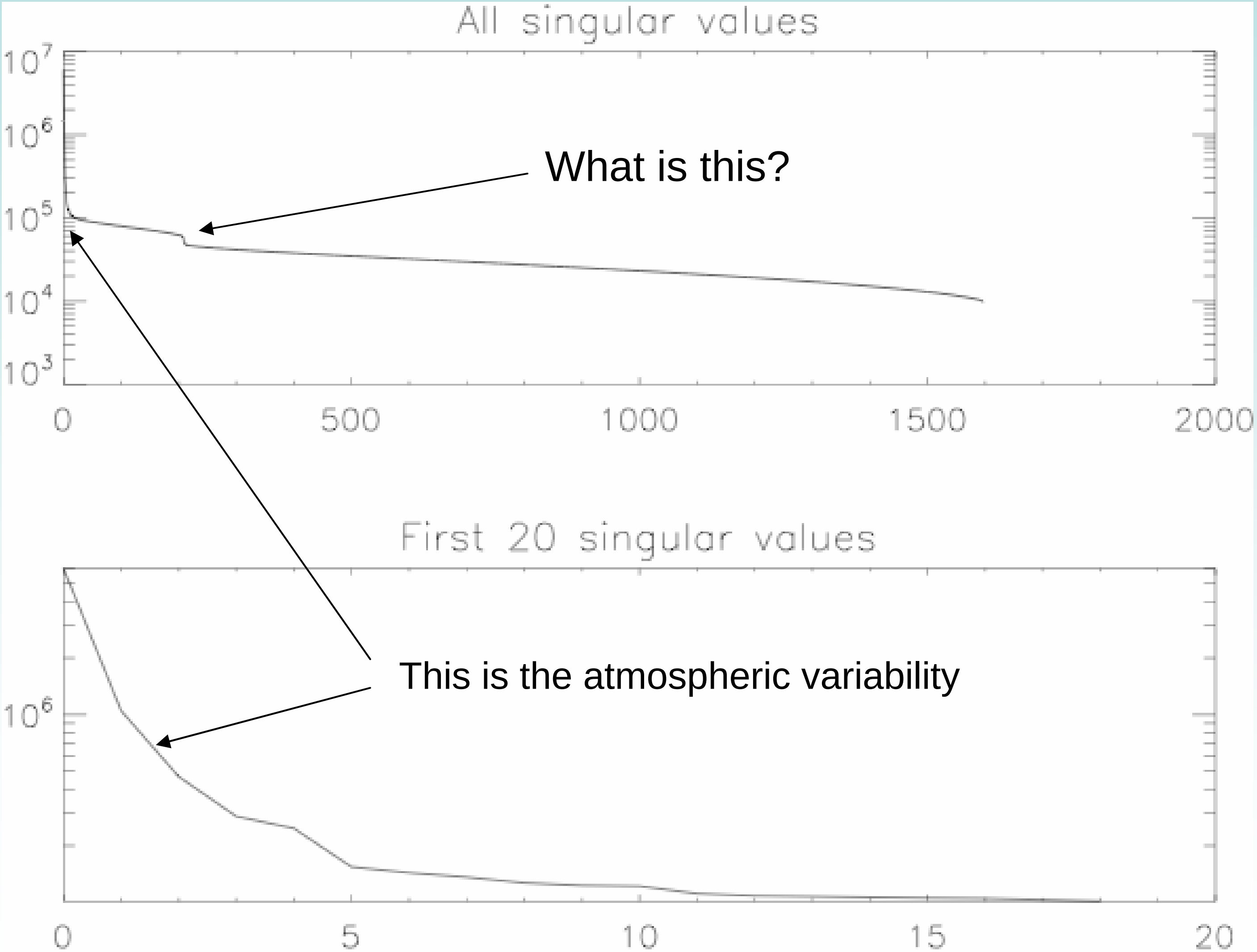
Fit to singular vector 6
0.10
0.05
0.00
−0.05
−0.10
950
1000
1050
1100
1150
Residual
0.10
0.05
0.00
−0.05
−0.10
950
1000
1050
1100
1150

SV6



Suggests the derivative of the ILS

Curious behaviour of the singular values of an ensemble of 1600 2B1 spectra

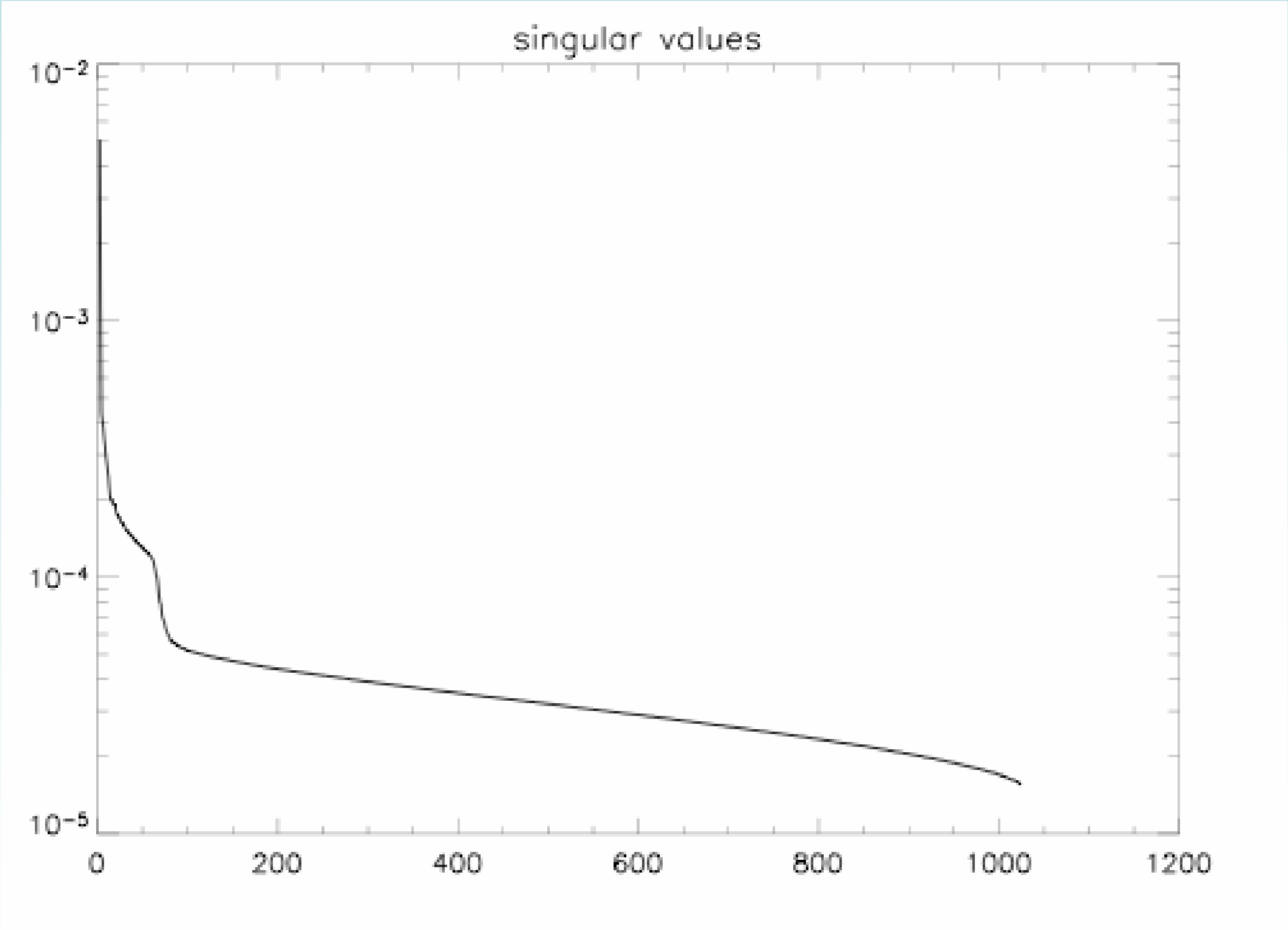


650-900 cm^{-1}

Features

- It occurs at around SV 200. I have used 100 sequences, 1600 spectra.
- Implies two of these SV's per sequence. Confirmed by trying other numbers of sequences
- The vec   tors look like noise
- Only in 2B1 does this

With 64 sequences, plotted in 9/04

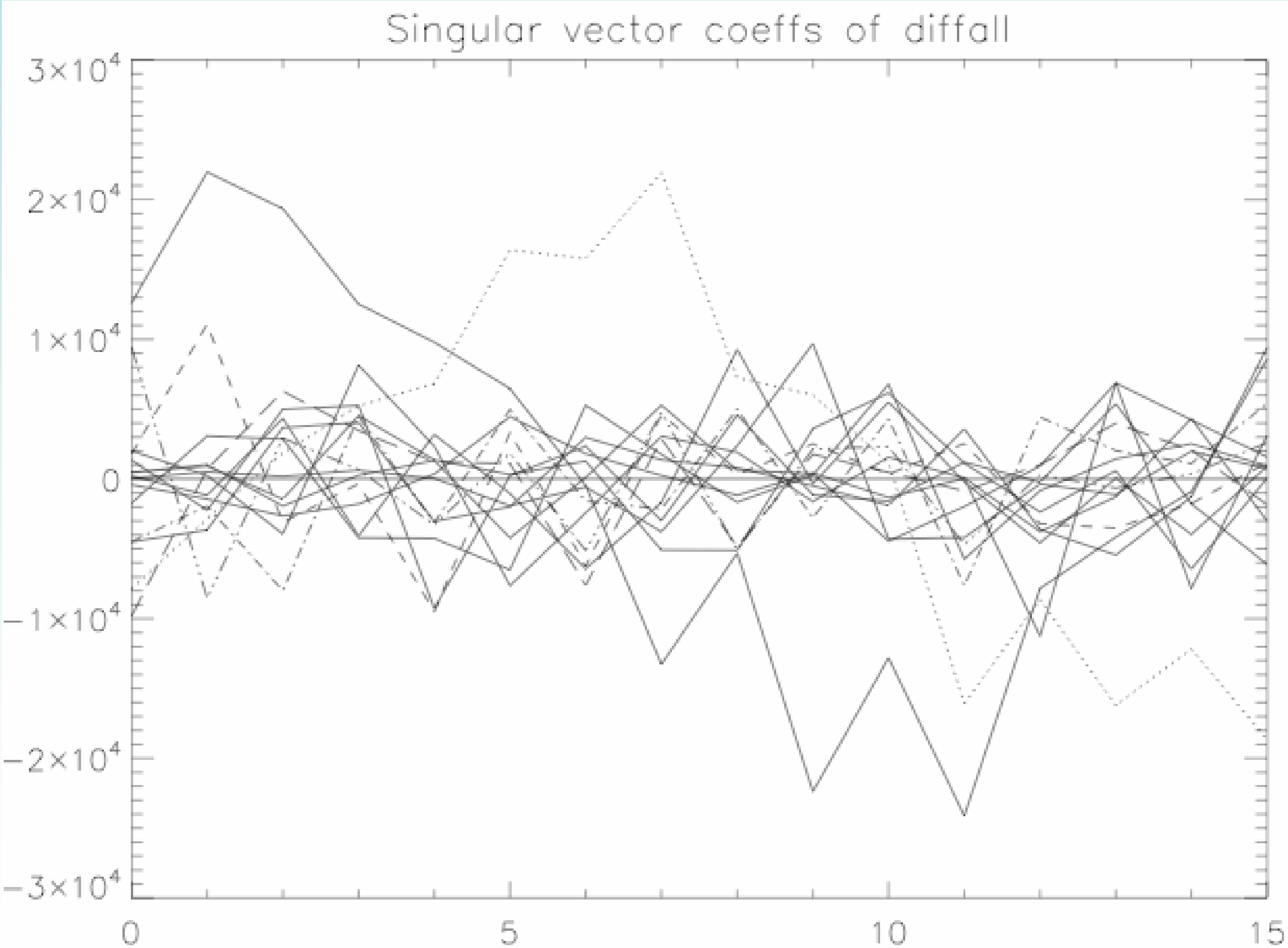


The shoulder here is at 64 – there was only 1 per sequence then

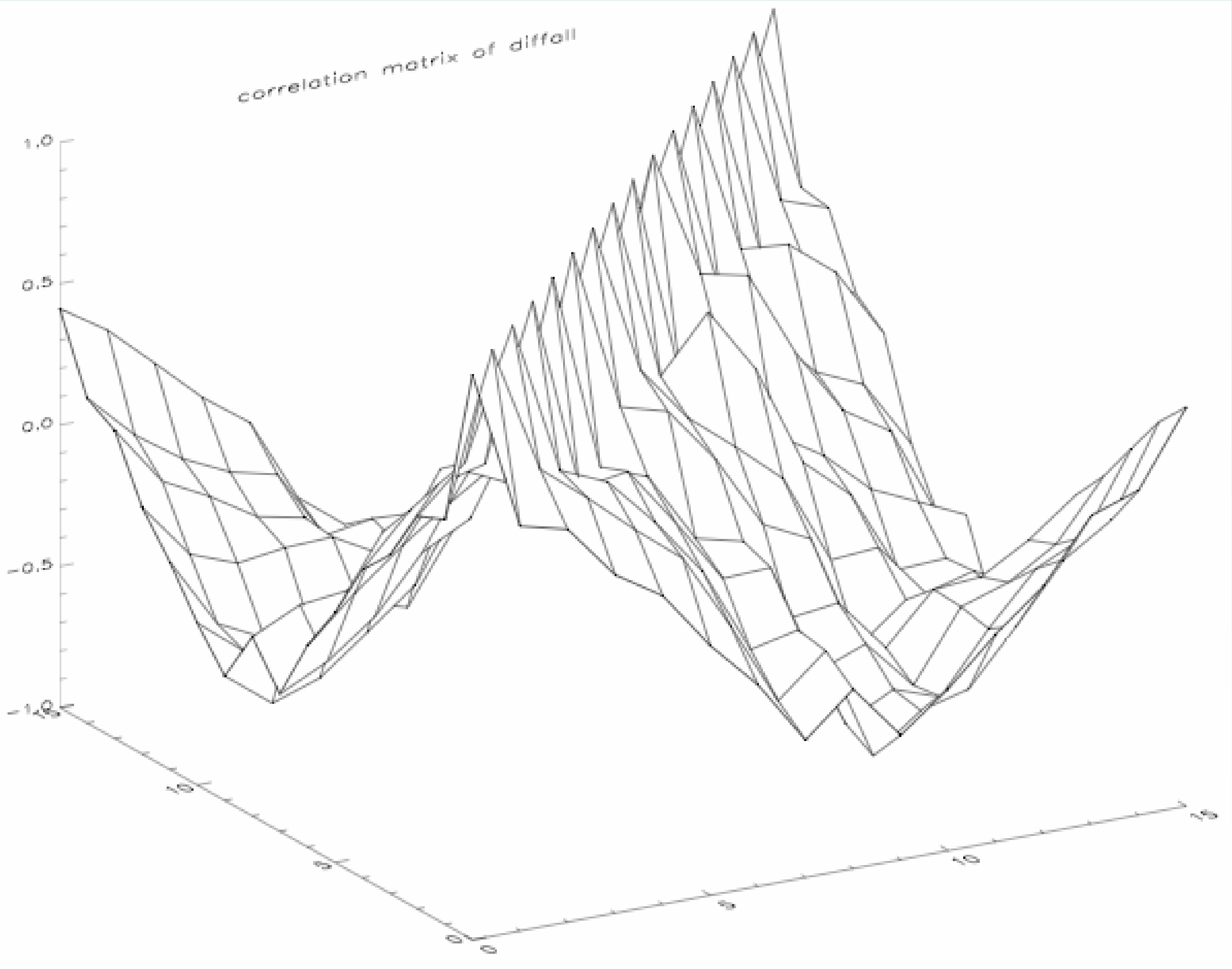
Examine pixel noise

- Look at the spectrum for each pixel minus the mean over all pixels
  - removes atmospheric structure,
  - but differences will be slightly correlated
- Singular vectors of the difference: look like noise
- Coefficients of the singular vectors for one set of 16: don't look like noise for two of the SV's
- Covariance & correlation matrices

SV coefficients



2B1 noise is significantly correlated between pixels



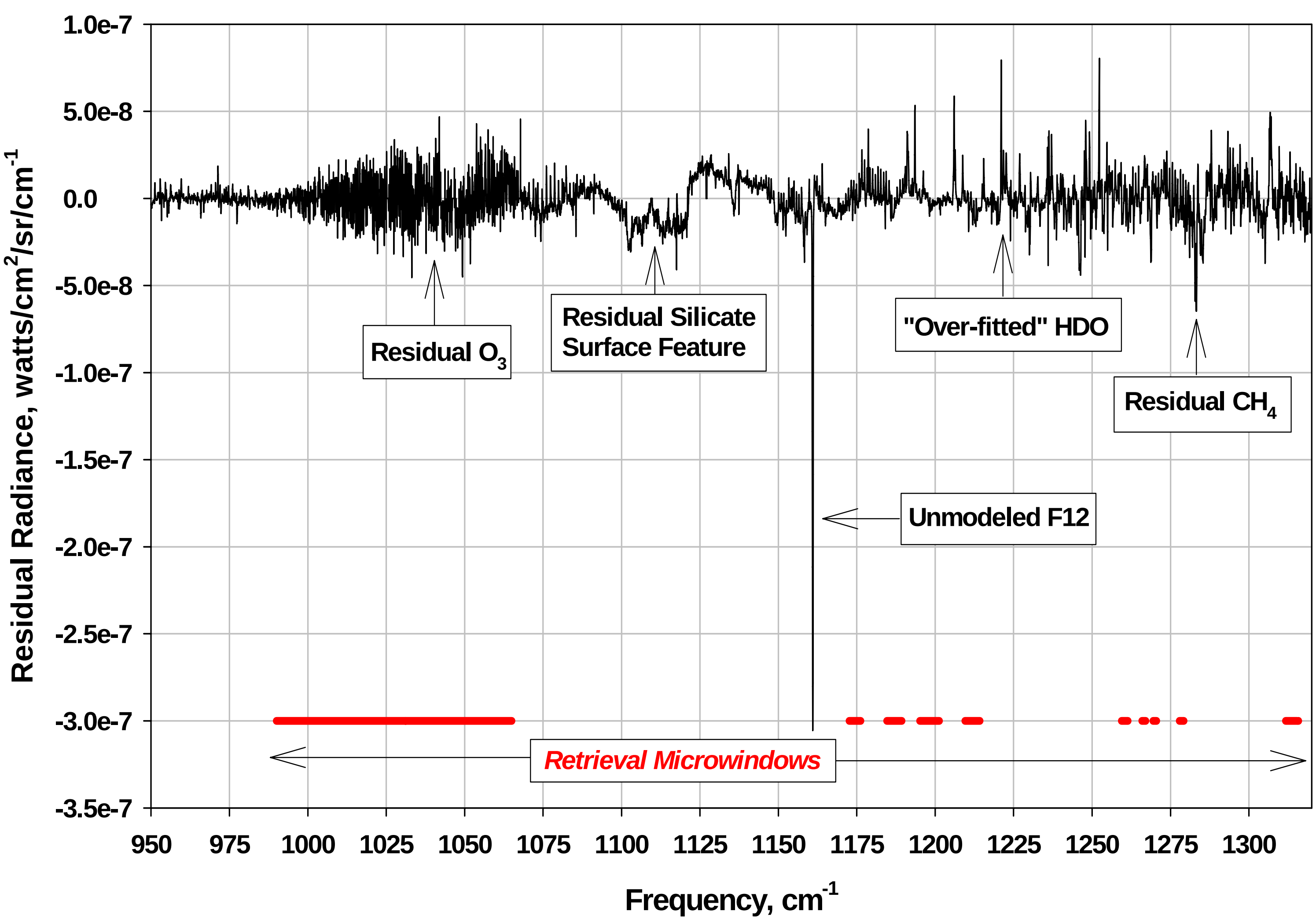
Deductions

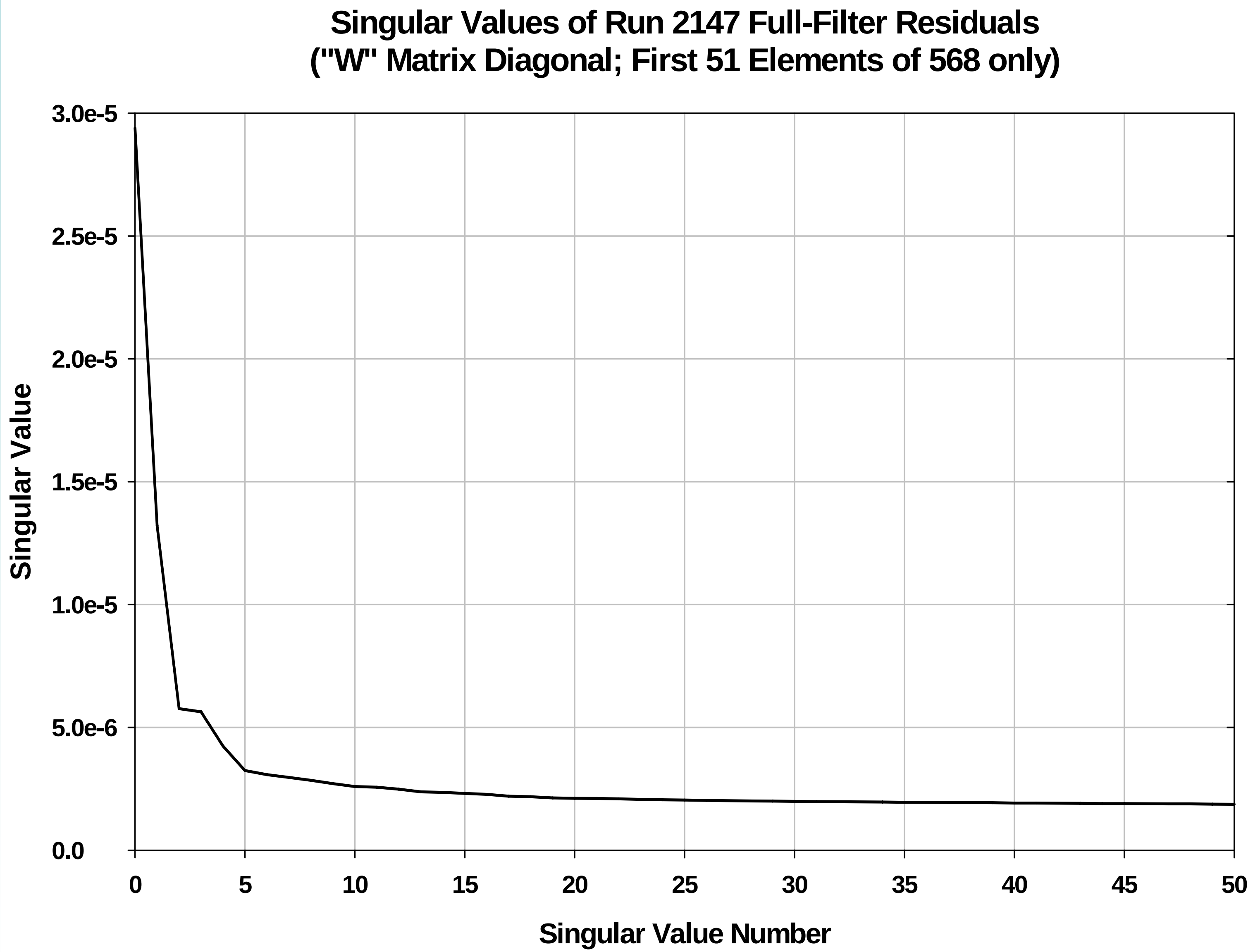
- The change from 1 to 2 SVs per sequence implies something in the processing rather than the instrument
- Each pixel has its own independent noise
- Plus noise correlated between all the pixels of each sequence
- The correlations depend on distance between the pixels, quasi-sinusoidally
- Not an odd pixel/even pixel effect

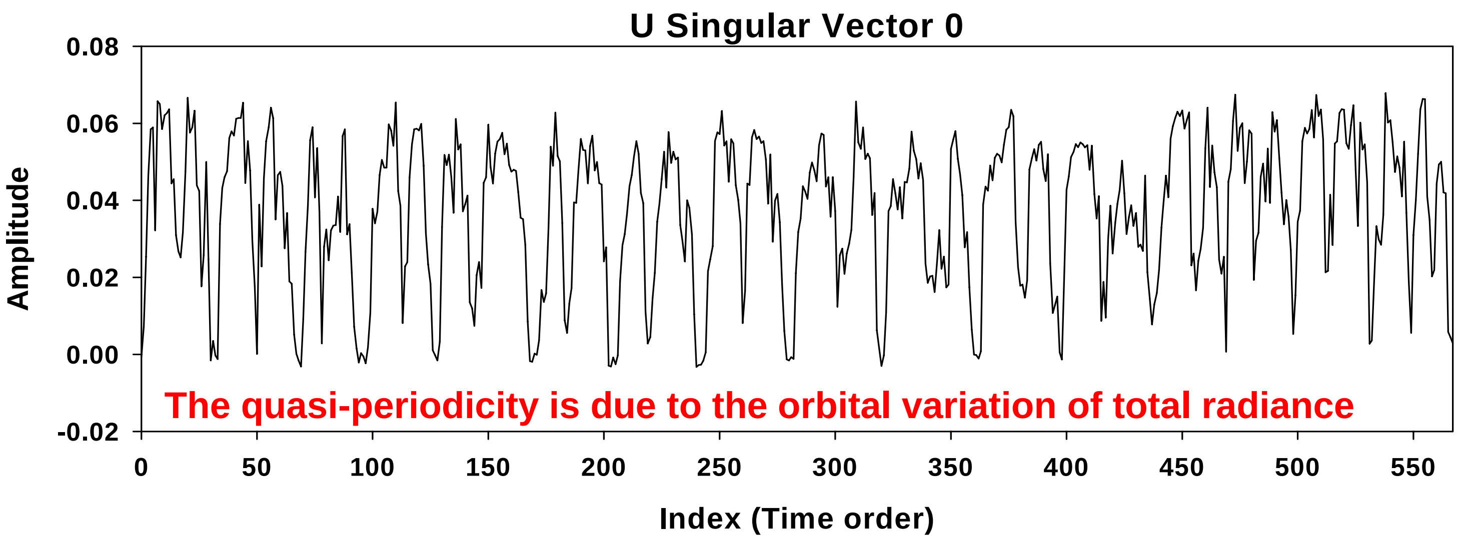
SV's of Residual Spectra

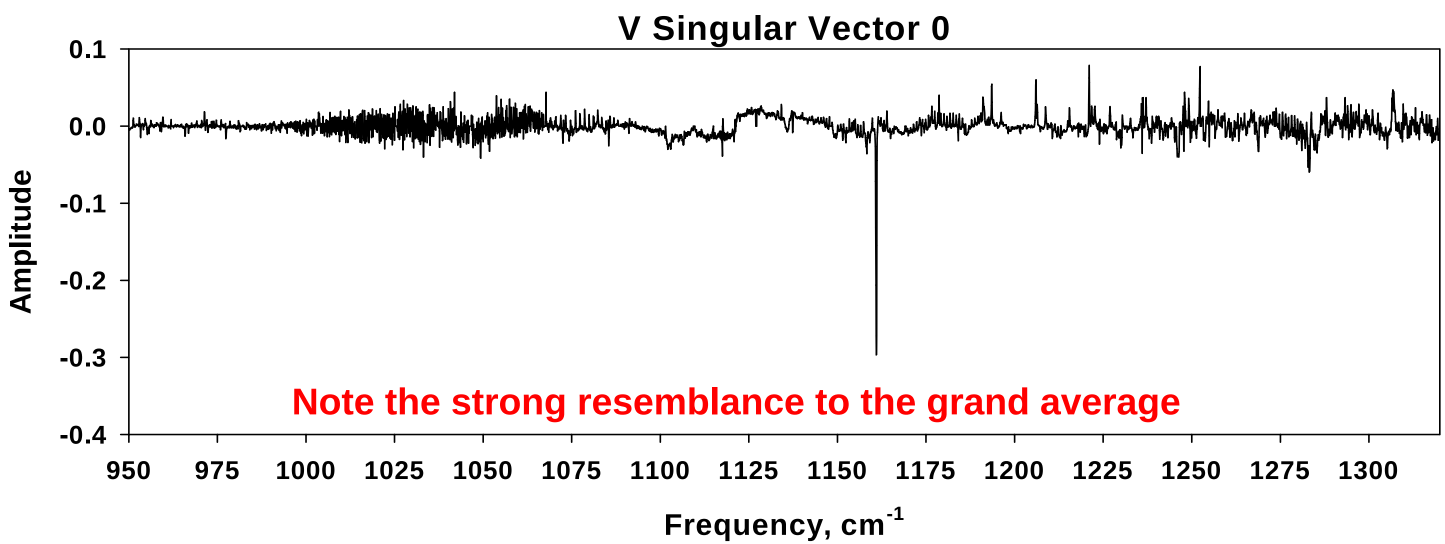
- Difference between measured spectrum and calculated from retrieval
- No just in the microwindows used
- Courtesy Reinhard Beer & Susan S. Kulawik

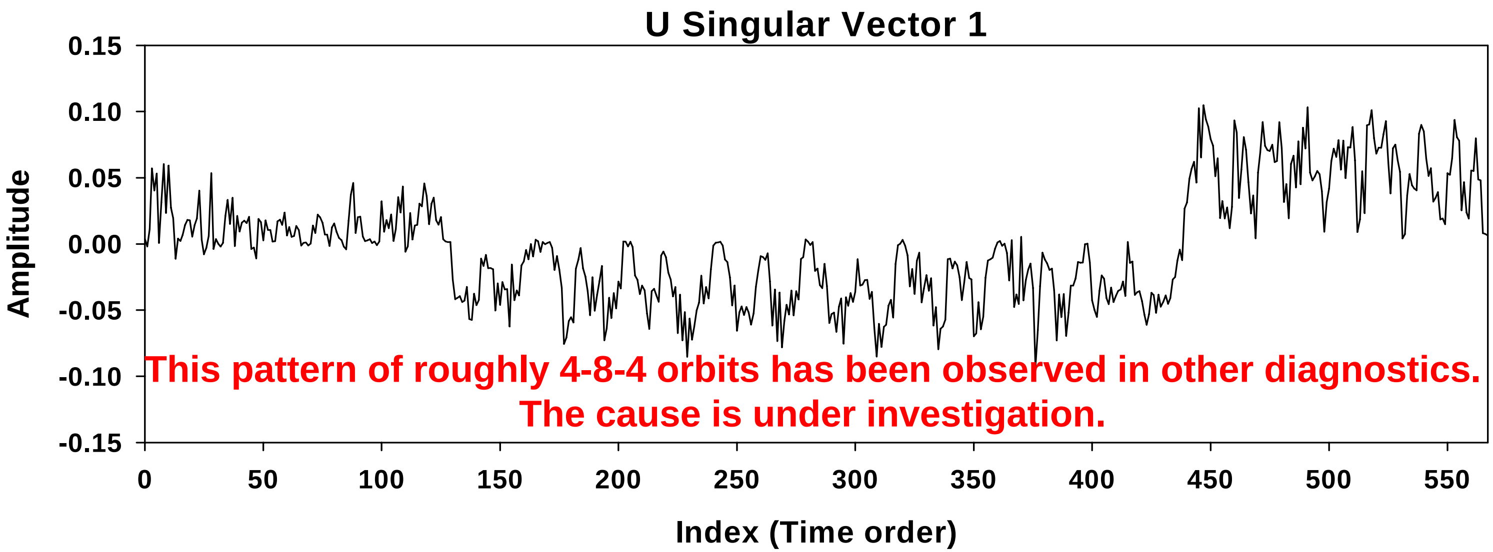
Mean Residual Radiance for Run 2147 (2004 September 21) from 220 Spectra between 30°S & 30°N

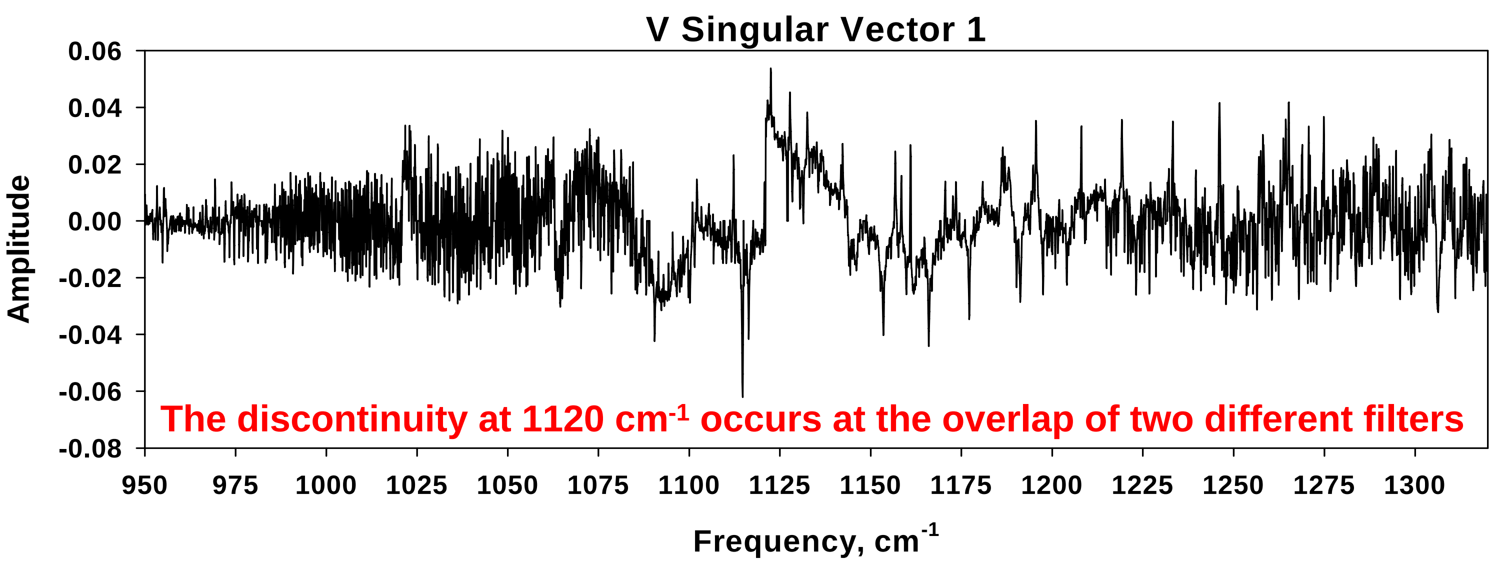


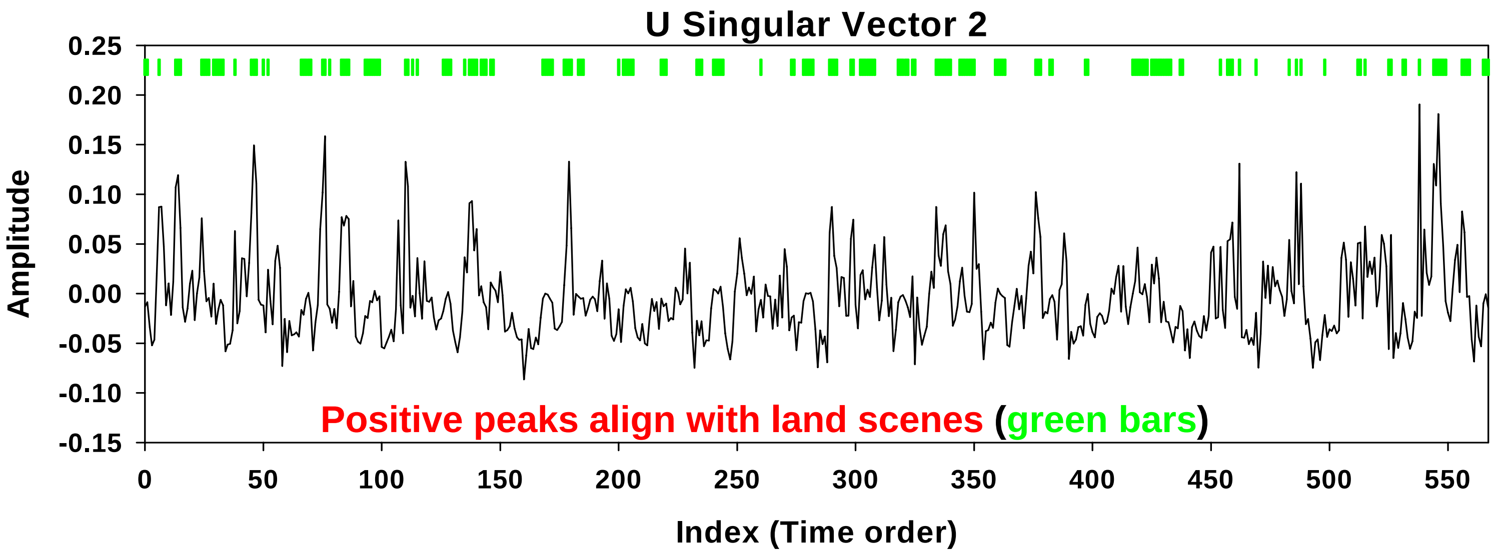

Singular Values of Run 2147 Full-Filter Residuals
("W" Matrix Diagonal; First 51 Elements of 568 only)
3.0e-5
2.5e-5
2.0e-5
1.5e-5
1.0e-5
5.0e-6
0.0
Singular Value
0
5
10
15
20
25
30
35
40
45
50
Singular Value Number

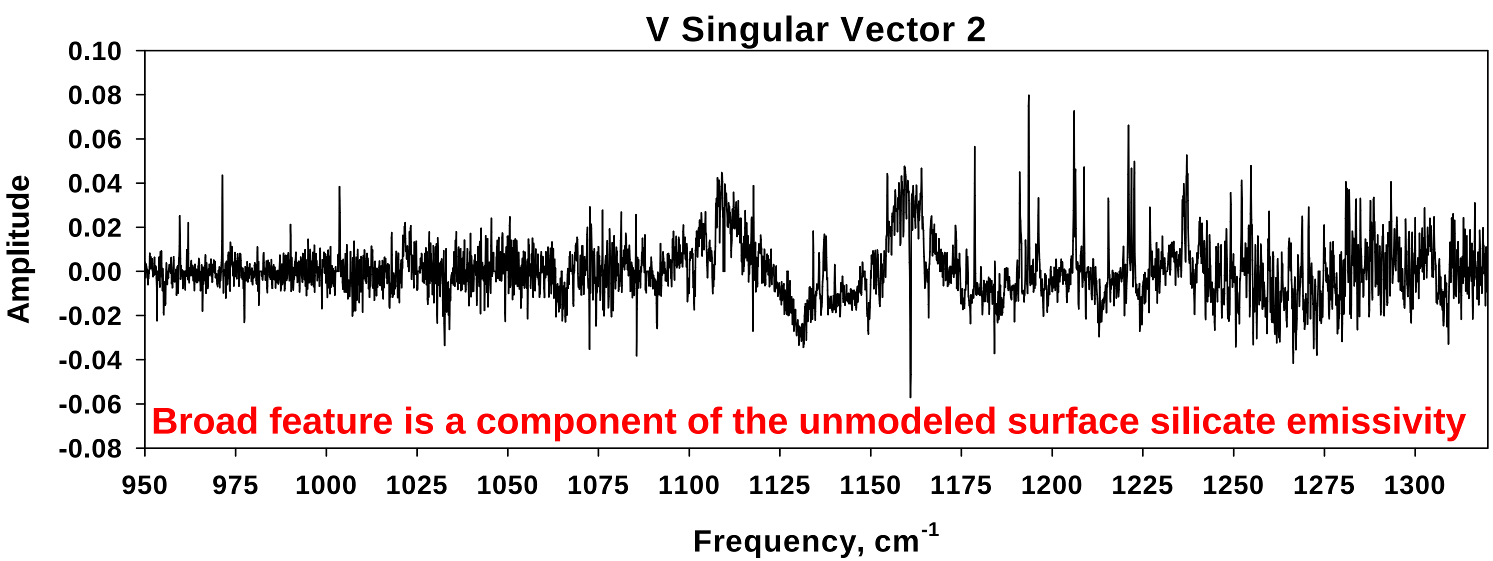

U Singular Vector 0
Amplitude
0.08
0.06
0.04
0.02
0.00
-0.02
0
50
100
150
200
250
300
350
400
450
500
550
Index (Time order)
The quasi-periodicity is due to the orbital variation of total radiance


V Singular Vector 0
Amplitude
0.1
0.0
-0.1
-0.2
-0.3
-0.4
950
975
1000
1025
1050
1075
1100
1125
1150
1175
1200
1225
1250
1275
1300
Frequency, cm^{-1}
Note the strong resemblance to the grand average


U Singular Vector 1
Amplitude
0.15
0.10
0.05
0.00
-0.05
-0.10
-0.15
0
50
100
150
200
250
300
350
400
450
500
550
Index (Time order)
This pattern of roughly 4-8-4 orbits has been observed in other diagnostics.
The cause is under investigation.


V Singular Vector 1
Amplitude
0.06
0.04
0.02
0.00
-0.02
-0.04
-0.06
-0.08
950
975
1000
1025
1050
1075
1100
1125
1150
1175
1200
1225
1250
1275
1300
Frequency, cm-1
The discontinuity at 1120 cm-1 occurs at the overlap of two different filters


U Singular Vector 2
Amplitude
0.25
0.20
0.15
0.10
0.05
0.00
-0.05
-0.10
-0.15
0
50
100
150
200
250
300
350
400
450
500
550
Index (Time order)
Positive peaks align with land scenes (green bars)


V Singular Vector 2
Amplitude
0.10
0.08
0.06
0.04
0.02
0.00
-0.02
-0.04
-0.06
-0.08
950
975
1000
1025
1050
1075
1100
1125
1150
1175
1200
1225
1250
1275
1300
Frequency, cm-1
Broad feature is a component of the unmodeled surface silicate emissivity

Retrieval – some untested thoughts

Possible methods include:

- Retrieve from denoised spectrum in the same way as usual, but with noise covariance reduced in some ad-hoc way
  - not optimal, inefficient
- Ditto, but with the correct error covariance
  - covariance matrix is singular
- Select a subset of channels according to information content
  - straightforward if linear, not otherwise
- Retrieve from SV representation coefficients
  - needs a special forward model for efficiency – PCRTM
  - or a regression or neural net method

Characterising a denoised spectrum

- Reconstruct $\mathbf{y}$ with the first *r* singular vectors, $\mathbf{U}_r$, giving $\mathbf{y}_r$

$$\mathbf{y}_r = \mathbf{U}_r\mathbf{U}_r^T\mathbf{y}$$

- Measurement function is:

$$\mathbf{y}_r = \mathbf{U}_r\mathbf{U}_r^T\mathbf{F}(\mathbf{x}) + \mathbf{U}_r\mathbf{U}_r^T\varepsilon$$

- We assume that $\mathbf{U}_r\mathbf{U}_r^T\mathbf{F}(\mathbf{x}) \sim \mathbf{F}(\mathbf{x})$, so

$$\mathbf{y}_r = \mathbf{F}(\mathbf{x}) + \mathbf{U}_r\mathbf{U}_r^T\varepsilon = \mathbf{F}(\mathbf{x}) + \varepsilon_r$$

- Covariance of ε_r is

$$\mathbf{S}_{\varepsilon r} = \mathbf{U}_r\mathbf{U}_r^T\,\mathbf{S}_\varepsilon\mathbf{U}_r\mathbf{U}_r^T$$

which is of rank *r*

Retrieve from a denoised spectrum

- We can use all or some of the spectrum.
  - Microwindows
  - Channels selected by information content
  - Channels selected by Xu Liu's approach

- Define a selection operator **M**, and retrieve from the *m* element subset $\mathbf{y}_m = \mathbf{M}\mathbf{y}_r$

- The subset depends on the whole of the original spectrum, $\mathbf{y}_m = \mathbf{M}\mathbf{U}_r\mathbf{U}_r^T\mathbf{y}$, but we assume that we can model it as $\mathbf{M}\mathbf{F}(\mathbf{x})$.

- The error covariance of ε_m is

$$\mathbf{S}_m = \mathbf{M}\,\mathbf{U}_r\mathbf{U}_r^T\mathbf{S}_\varepsilon\mathbf{U}_r\mathbf{U}_r^T\mathbf{M}^T$$

which may be singular if $m > r$.

Do another SVD…

- Use the SVD of $\mathbf{M}\mathbf{U}_r\mathbf{U}_r^T = \mathbf{L}\mathbf{\Lambda}\mathbf{R}^T$ to give

$$\mathbf{L}^T\mathbf{y}_m = \mathbf{L}^T\mathbf{M}\mathbf{F}(\mathbf{x}) + \mathbf{\Lambda}\mathbf{R}^T\mathbf{\varepsilon}$$

- Drop elements of $\mathbf{L}^T\mathbf{y}_m$ with zero (or small) singular values, leaving p=min(r,m) elements at most. Gives $\mathbf{L}_p^T\mathbf{y}_m$

- Retrieve from the rest of $\mathbf{L}_p^T\mathbf{y}_m$ with $\mathbf{L}_p^T\mathbf{M}\mathbf{F}(\mathbf{x})$ as the forward model. $\mathbf{L}_p^T$ can be precomputed.

- Error covariance is $\mathbf{\Lambda}_p\mathbf{R}_p^T\mathbf{S}_\varepsilon\mathbf{R}_p\mathbf{\Lambda}_p$. If $\mathbf{S}_\varepsilon = \sigma^2\mathbf{I}$, this reduces to $\sigma^2\mathbf{\Lambda}_p^2$.

Comments

- This allows us to retrieve with the minimum number of evaluations of the forward model
- And the minimum length of measurement vector
- With very little extra matrix manipulation
- I havn't tried it yet, so I don't know what the catch is…

Summary/Comments

- PCA does not improve the information content of a measurement
- But it does allow us to use it more efficiently
- Very useful for validation - separates out independent sources of variation, e.g. artefacts, omitted physics, . . .
- Denoised spectra guide the eye to real features that may otherwise not be seen
- Data compression