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
  \[ y = x + \varepsilon \]
  \( x \) is true spectrum, \( y \) is measurement, \( \varepsilon \) is noise

• Maximum a posteriori estimate of \( x \) is
  \[ x_r = x_a + S_a(S_a + S_\varepsilon)^{-1}(y - x_a) \]
Continued…

• If we have a large sample of spectra:
  - Expect that \( \mathbf{x}_a = \langle \mathbf{x} + \mathbf{\varepsilon} \rangle = \langle \mathbf{y} \rangle \)
  - 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 \((y - \langle y \rangle)\) be \(N\) columns of a matrix \(Y\)

- Represent \(Y\) as its singular vector decomposition:
  \[
  Y = U \Lambda V^T
  \]
  where \(\Lambda\) is diagonal, \(U^T U = I\) and \(V^T V = I\)

- The \(j\)'th individual spectrum \(y_j\) is then
  \[
  y_j = \langle y \rangle + \sum_i u_i \lambda_i v_{ij}^T
  \]

- The spectrum is represented as a sum of columns \(u_i\) of \(U\), with coefficients \(\lambda_i v_{ij}^T\).

- Because \(U^T U = I\), we can compute \(\lambda_i v_{ij}^T\) for any spectrum as
  \(U^T (y_j - \langle y \rangle)\).
What do we expect?

• $\frac{1}{N}YY^T$ is the covariance matrix $S_y$ of the spectra

• Left singular vectors $U$ are the same as eigenvectors of $YY^T$, singular values are the square roots of its eigenvalues

• In the linear case with independent constant noise, $S_y$ would be
  
  $$S_y = S_a + \sigma_e^2 I$$

• $S_a$ has rank $\leq n$, $I$ is of dimension $m \gg n$, where $n$ is the degrees of freedom of the atmosphere

• The eigenvalues of $S_y$ are $\lambda_i^2/N$

• The eigenvalues of $S_a$ should be $\lambda_i^2/N - \sigma_e^2$
Reconstructing Spectra

- We can drop terms with $\frac{\lambda_i^2}{N} \sim \sigma_e^2$, i.e. $i > n$, without significant loss
  - they correspond to noise only
  - Better, multiply retained terms by something like
    $\left(\frac{\lambda_i^2 - N\sigma_e^2}{\lambda_i^2}\right)$

- 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

\[ H = \frac{1}{2} \sum_i \ln(\frac{\lambda_i^2}{N \sigma_\varepsilon^2}) \]

• The degrees of freedom for signal is

\[ d_s = \sum_i (1 - \frac{N \sigma_\varepsilon^2}{\lambda_i^2}) \]

• 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\(^{-1}\)

<table>
<thead>
<tr>
<th>n</th>
<th>Lambda</th>
<th>Information</th>
<th>d.f.s</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>cum.</td>
<td>cum.</td>
</tr>
<tr>
<td>0</td>
<td>253.315</td>
<td>5.535</td>
<td>5.53</td>
</tr>
<tr>
<td>1</td>
<td>19.940</td>
<td>2.994</td>
<td>8.53</td>
</tr>
<tr>
<td>2</td>
<td>8.496</td>
<td>2.147</td>
<td>10.68</td>
</tr>
<tr>
<td>3</td>
<td>3.226</td>
<td>1.217</td>
<td>11.89</td>
</tr>
<tr>
<td>4</td>
<td>2.446</td>
<td>0.972</td>
<td>12.86</td>
</tr>
<tr>
<td>5</td>
<td>1.675</td>
<td>0.668</td>
<td>13.53</td>
</tr>
<tr>
<td>6</td>
<td>1.072</td>
<td>0.382</td>
<td>13.91</td>
</tr>
<tr>
<td>7</td>
<td>0.553</td>
<td>0.133</td>
<td>14.05</td>
</tr>
<tr>
<td>8</td>
<td>0.433</td>
<td>0.086</td>
<td>14.13</td>
</tr>
<tr>
<td>9</td>
<td>0.232</td>
<td>0.026</td>
<td>14.16</td>
</tr>
<tr>
<td>10</td>
<td>0.156</td>
<td>0.012</td>
<td>14.17</td>
</tr>
<tr>
<td>11</td>
<td>0.086</td>
<td>0.004</td>
<td>14.18</td>
</tr>
<tr>
<td>12</td>
<td>0.076</td>
<td>0.003</td>
<td>14.18</td>
</tr>
<tr>
<td>13</td>
<td>0.054</td>
<td>0.001</td>
<td>14.18</td>
</tr>
<tr>
<td>14</td>
<td>0.048</td>
<td>0.001</td>
<td>14.18</td>
</tr>
</tbody>
</table>

Fit to noise asymptote

Total: 14.5 bits 7.05 d.f.
How do you fit noise to the asymptote?

- I haven’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

First 6 Singular Vectors for scan 2 1B2 run 2147

950 1000 1050 1100 1150

0.60 0.50 0.40 0.30 0.20 0.10 0.00 -0.10
Singular vectors * Lambda

First 6 Singular Vectors *lambda for scan 2 1B2 run 2147
V-vectors for sequences 1 to 30
Features

- Most of variation is in the first singular vector. First six are:
  \[5.96 \times 10^6 \quad 3.6 \times 10^5 \quad 1.83 \times 10^5 \quad 1.39 \times 10^4 \quad 7.93 \times 10^4 \quad 6.16 \times 10^4\]

- 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
Mean Spectrum for scan 2 1B2 run 2147

Singular Vector 6
Suggests the derivative of the ILS
Curious behaviour of the singular values of an ensemble of 1600 2B1 spectra

What is this?

This is the atmospheric variability

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 vectors 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

![Graph showing SV coefficients](image.png)
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

Residual Radiance, Watts/cm²/sr/cm⁻¹

Frequency, cm⁻¹

-3.5e⁻⁷ -3.0e⁻⁷ -2.5e⁻⁷ -2.0e⁻⁷ -1.5e⁻⁷ -1.0e⁻⁷ -0.5e⁻⁷ 0.0 5.0e⁻⁸ 1.0e⁻⁷ 1.5e⁻⁷ 2.0e⁻⁷ 2.5e⁻⁷ 3.0e⁻⁷ 3.5e⁻⁷

950 975 1000 1025 1050 1075 1100 1125 1150 1175 1200 1225 1250 1275 1300

- Residual O₃
- Residual Silicate Surface Feature
- "Over-fitted" HDO
- Residual CH₄
- Unmodeled F12
- Retrieval Microwindows
Singular Values of Run 2147 Full-Filter Residuals
("W" Matrix Diagonal; First 51 Elements of 568 only)
The quasi-periodicity is due to the orbital variation of total radiance.

Note the strong resemblance to the grand average.
This pattern of roughly 4-8-4 orbits has been observed in other diagnostics. The cause is under investigation.

The discontinuity at 1120 cm\(^{-1}\) occurs at the overlap of two different filters.
Positive peaks align with land scenes (green bars)

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 $y$ with the first $r$ singular vectors, $U_r$, giving $y_r$
  $$y_r = U_r U_r^T y$$

- Measurement function is:
  $$y_r = U_r U_r^T F(x) + U_r U_r^T \varepsilon$$

- We assume that $U_r U_r^T F(x) \sim F(x)$, so
  $$y_r = F(x) + U_r U_r^T \varepsilon = F(x) + \varepsilon_r$$

- Covariance of $\varepsilon_r$ is
  $$S_{\varepsilon_r} = U_r U_r^T S_{\varepsilon} U_r U_r^T$$
  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 $\mathbf{M}$, 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, $\mathbf{y}_m = \mathbf{M}U_r U_r^T\mathbf{y}$, but we assume that we can model it as $\mathbf{M}\mathbf{F}(\mathbf{x})$.

- The error covariance of $\mathbf{e}_m$ is
  
  $S_m = \mathbf{M} U_r U_r^T S_\mathbf{e} U_r U_r^T \mathbf{M}^T$

  which may be singular if $m > r$.  

  \[ 38/41 \]
Do another SVD…

- Use the SVD of $MU_r U_r^T = L \Lambda R^T$ to give
  $L^T y_m = L^T MF(x) + \Lambda R^T \varepsilon$

- Drop elements of $L^T y_m$ with zero (or small) singular values, leaving $p = \min(r, m)$ elements at most. Gives $L_p^T y_m$

- Retrieve from the rest of $L_p^T y_m$ with $L_p^T MF(x)$ as the forward model. $L_p^T$ can be precomputed.

- Error covariance is $\Lambda_p R_p^T S_\varepsilon R_p \Lambda_p$. If $S_\varepsilon = \sigma^2 I$, this reduces to $\sigma^2 \Lambda_p^2$. 
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 haven’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