P1.15 Applications of Principal Component Analysis (PCA) to AIRS Data

Mitchell D. Goldberg[1], Lihang Zhou[2], and Walter Wolf[2]
NOAA/NESDIS/Office of Research and Applications, Camp Springs, MD[1]
QSS Group Inc.[2]

1 Introduction

The Atmospheric InfraRed Sounder (AIRS) (Aumann et al. 2003) is the first of a new generation of high spectral resolution infrared sounder with 2378 channels measuring outgoing radiance between 650 cm-1 and 2675 cm-[1]. High spectral resolution in the infrared region allows for the derivation of atmospheric soundings of temperature, moisture, ozone and trace gases with higher accuracy and higher vertical resolution. Even though AIRS has more than 2000 channels, the information is not independent. In fact, many of the channels are highly correlated. However the use of highly correlated data also reduces the effects of instrumental noise. Principal component analysis (PCA), also called eigenvector decomposition, is often used to approximate data vectors having many elements with a new set of data vectors having fewer elements, while retaining most of the variability and information of the original data. The new data vectors are called principal component score vectors, and because they consist of the components of the original data vector in an orthogonal coordinate system, the elements of a given principal component score vector are independent of each other (unlike the original spectrum).

For AIRS, PCA is used for a) data compression, b) reconstructing radiances with the properties of reduced noise, c) independent instrument noise estimation, d) quality control, and e) deriving geophysical parameters.

2 Principal Component Analysis

Principal component analysis for high spectral resolution sounders is described by Huang and Antonelli (2001) and Goldberg et al. (2003). Elements of a principal component score vector are projections of the spectrum onto each of the orthogonal basis vectors, which are the eigenvectors (principal components) of the radiance covariance matrix. The total number, n, of eigenvectors is equal to the total number of channels. However, it can be shown that a much smaller set of k eigenvectors (< 100), ordered from largest to smallest eigenvalues, is sufficient to explain most of the variance in the original spectra. The covariance matrix is derived from an ensemble of AIRS normalized spectra, i.e. radiance divided by the instrument noise. The matrix of eigenvectors, **E,** is related to the covariance matrix, **S**, by:

$$\mathbf{S} = \mathbf{E}\,\boldsymbol{\lambda}\mathbf{E}^{T} \qquad (1)$$

where **S** , **E** and **λ** are all dimensioned n x n, and **λ** is a diagonal matrix of eigenvalues. The principal component scores vector **p** is computed from:

$$\mathbf{p} = \mathbf{E}^{T}\,\mathbf{r} \qquad (2)$$

where **r** is the vector of centered (departure from the mean) normalized radiances. The next equation is used to reconstruct the radiances from a truncated set of k eigenvectors **E*** and a vector of principal component scores **p***. (The symbol * indicated that the matrix or the result of a matrix operation is due to truncated set of vectors).

$$\mathbf{r}^{*} = \mathbf{E}^{*}\mathbf{p}^{*} \qquad (3)$$

The normalized reconstructed radiance vector is **r***, **E*** has dimension n x k, and the vector **p*** has length k. To obtain the un-scaled radiance, one must add the ensemble mean normalized radiance used in generating the covariance matrix and multiply the sum by the noise used in constructing the normalized radiances

The number of principal components needed to reproduce the signal in the original radiances is determined by examining the magnitude of the eigenvalues and examining the spatial correlation of the principal component scores. Since we are using normalized radiances, the square root of the eigenvalues can be interpreted as signal to noise. Principal component scores (PCS) can be thought of as superchannels since each one is a linear combination of all channels. The first score contains the largest signal to noise ratio, which as shown in Table 1 is very large. When the eigenvalues fall below unity, the noise has larger contribution than the signal. Based on Table 1, this transition occurs near the 60th eigenvalue. However when we examined the PCS spatial correlations, an additional 25 principal components are needed. Ideally the spatial correlation should be near zero, otherwise the PCs are not capturing all of the signal. Note that the 85 PCs used in our AIRS

---

* *Corresponding author address*: Mitchell Goldberg, NOAA/NESDIS/ORA, 5200 Auth Road, Room 712, Camp Springs, MD 20746,; email: mitch.goldberg@noaa.gov

processing system is based on a global ensemble of observations.

Square Root of Eigenvalues (first 72)

1	7497.60	19	14.68	37	3.38	55	1.25
2	1670.40	20	13.49	38	3.11	56	1.19
3	945.52	21	12.28	39	2.82	57	1.16
4	496.01	22	11.32	40	2.53	58	1.15
5	284.01	23	10.70	41	2.41	59	1.09
6	266.30	24	9.08	42	2.39	60	1.05
7	156.95	25	8.24	43	2.34	61	1.02
8	139.67	26	7.85	44	2.24	62	0.98
9	88.27	27	6.77	45	2.03	63	0.90
10	72.83	28	5.98	46	1.86	64	0.86
11	60.03	29	5.83	47	1.78	65	0.81
12	53.42	30	5.39	48	1.71	66	0.80
13	45.01	31	5.34	49	1.65	67	0.78
14	39.72	32	4.98	50	1.61	68	0.77
15	34.54	33	4.34	51	1.54	69	0.73
16	26.57	34	4.09	52	1.52	70	0.72
17	22.62	35	3.62	53	1.35	71	0.70
18	17.60	36	3.48	54	1.34	72	0.66

Table 1 Square root of the first 72 eigenvalues

3 Applications

a) Noise Filtering

Because reconstructed radiances are derived from the principal components containing most of the signal as opposed to noise, the reconstructed radiances are nearly noise-free. The reconstructed radiances can be used in the retrieval process or directly assimilated. Fig. 1 shows the AIRS instrument noise at scene brightness temperature, and the root mean square (rms) difference between reconstructed brightness temperatures, from 60 principal component scores, and noise-free simulated brightness temperatures. To compute these results, we simulated brightness temperatures from a global ensemble with and without expected instrument noise. The reconstructed brightness temperatures are computed from the instrument noise-contaminated data. The original noise curve in Fig. 1 is simply the rms error of the two datasets (noise and noise-free). The rms difference between the reconstructed brightness temperatures and the noise-free simulated brightness temperatures is extremely small in comparison to the instrument noise. The reconstructed data are more similar to noise-free observations. Since the reconstructed rms error is very small, we can use the reconstructed data to estimate the noise. This is done by simply computing the rms difference between the reconstructed brightness temperatures and the original noisy data. The difference between the original noise curve in Fig. 1 and the noise estimate using PCA is shown in Fig. 2. The difference is extremely small and we use this technique as an independent approach for estimating instrumental noise. Furthermore, when we find an occasional large difference between reconstructed and the original radiances it is often due to a problem in the original radiances. So PCA is also used for quality control. Another advantage of using reconstructed radiances is that a reduced channel set can be used in a retrieval algorithm or in radiance assimilation with the benefits of using information from the entire spectrum, since each reconstructed radiance is a linear combination of all channels.

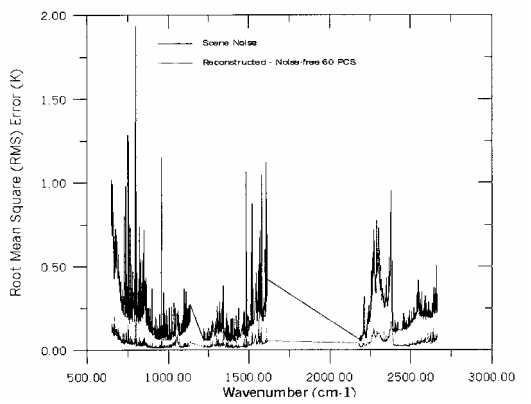

Fig. 1 Root mean square (rms) error of noise minus noise-free brightness temperatures (scene noise) and rms of reconstructed brightness temperatures from 60 principal component scores minus noise free brightness temperatures

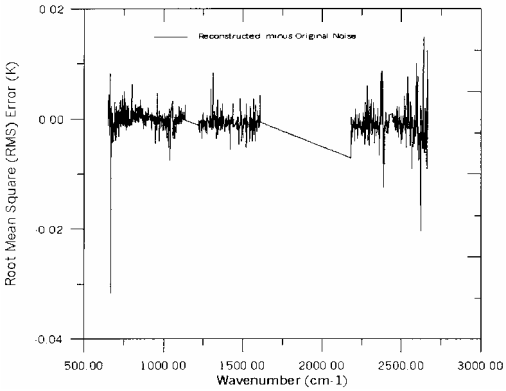

Fig. 2 Difference between the rms of reconstructed brightness temperatures from 60 principal component scores minus noise contaminated brightness temperature and the scene noise (Fig. 1).

b) Data compression

Instead of distributing 2378 AIRS channels, a data producer can distribute 85 PCS. Thereby, reducing the amount of data to be distributed or archived by a factor of 30. The user can reconstruct all or a subset of the channels.

c) Geophysical Retrievals using PC Regression

Another application is the use of PC scores in least squares regression to derive geophysical retrievals. For AIRS, we use 85 principal component scores for predictors and solve for atmospheric temperature, moisture, ozone profiles and surface temperature and surface emissivity. With 2000+ channels, many of the channels are similar to each other, making the covariance matrix nearly collinear. A significant advantage for using 85

principal component scores instead of all 2000+ channels is that the inverse of the predictor matrix is more stable and less collinear. Another advantage is that the regression solution is computationally fast. In matrix notation the form of the regression coefficients C, dimensioned *m* number of parameters by the *k* number of principal component scores, is

$$\mathbf{C} = \mathbf{X}\mathbf{P}^{*T}(\mathbf{P}^{*}\mathbf{P}^{*T})^{-1} \qquad (5)$$

where **X** is a training dependent predictand ensemble matrix, of dimension *m* by sample size *s*. **P*,** the training predictor ensemble matrix, is dimensioned *k* by *S*. On independent data the *m*-dimensioned solution vector is obtained from the matrix multiplication of **C p*** , where **p*** is the independent vector of principal component scores of length *k*.

Retrieval rms errors (differences between the retrieval and collocated radiosondes) based on the AIRS PC regression are shown in Fig. 3. Also shown in this figure are retrieval errors from the NESDIS ATOVS system (Reale,. 2002). The AIRS retrieval errors (dashed curve), including the systematic bias are significantly lower than ATOVS. The larger errors in the lower tropospheric temperature are probably due to uncertainties arising from collocation temporal and spatial differences. However, the difference between the ATOVS and AIRS retrieval remains large. Previous simulation studies have found that AIRS generally reduces the retrieval error by about 0.5K, and this appears to be holding for this radiosonde comparison. For moisture, the retrieval errors are significantly smaller than ATOVS. The large natural variability of water vapor combined with uncertainties in radiosonde-observed water vapor will prevent demonstrating the 10-15% accuracies often reported in simulated studies

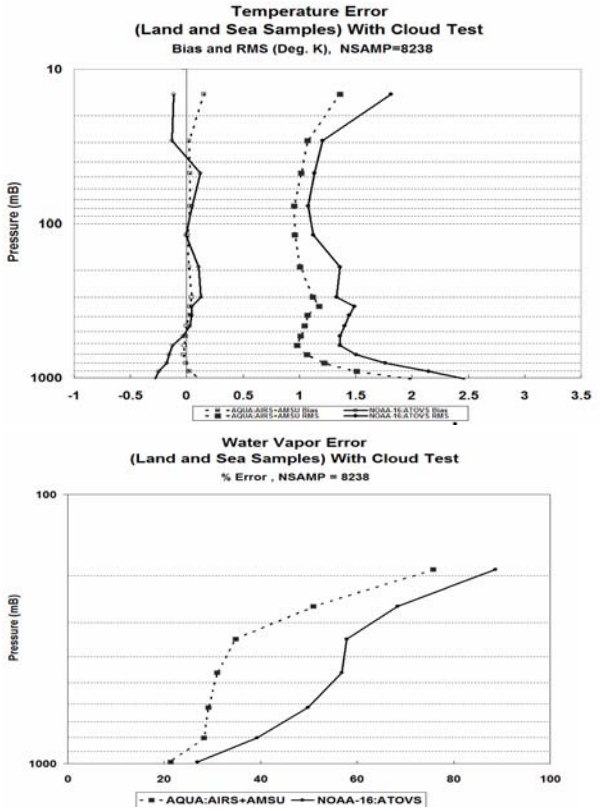

Fig. 3: Temperature (in K, top) and moisture (in %, bottom) RMS differences between the regression retrieval and the collocated radiosondes. The dashed curves are the AIRS errors, NOAA 16 errors are the solid curve

4 Acknowledgements

The views, opinions, and findings contained in this report are those of the author(s) and should not be construed as an official National Oceanic and Atmospheric Administration or U.S. Government position, policy, or decision. This work was funded by the NESDIS Office of Research and Applications, and the NASA AIRS Science Team.

5 References

Aumann, H. M.T. Chahine, C. Gautier, M. Goldberg, E. Kalnay, L. McMillin, H. Revercomb, P.W. Rosenkranz, W.L. Smith, D. Staelin, L. Strow, and J. Susskind,"AIRS/AMSU/HSB on the AQUA mission: Design, science objectives, data products and processing systems," IEEE Transaction on Geoscience and Remote Sensing, *IEEE Trans. Geosci. Remote Sensing*, Vol. **41**, pp 253-264, Feb. 2003

Goldberg, M.D., Y. Qu, L.M. McMillin, W. Wolf, L. Zhou, and M. Divakarla, 2003: AIRS near-real-time products and algorithms in support of operational numerical weather prediction, *IEEE Trans. Geosci. Remote Sensing*, Vol. **41**, pp 379-389, Feb. 2003

Huang, H-L and P. Antonelli, Application of principal component analysis to high-resolution infrared measurement compression and retrieval. J. Appl. Meteor., 40, 365-388, 2001.

Reale, A.L. 2002. NOAA operational sounding products for advanced-TOVS. *NOAA Technical Report NESDIS 107*. U.S. Dept. of Commerce, Washington DC, 29 pp.