Reducing the Dimensionality of Plant Spectral Databases

Ian E. Bell * Gladimir V. G. Baranoski †

Technical Report CS2002-33

Abstract. This study investigates the application of Principal Components Analysis (PCA) in the storage and reconstruction of plant spectral data. A new Piecewise PCA approach (PPCA), which takes into account the biological factors that affect the interaction of solar radiation with plants, is also proposed. These techniques are examined through experiments involving the reconstruction of reflectance and transmittance curves for herbaceous and and woody specimens. The spectral data used in these experiments was obtained from the LOPEX (Leaf Optical Properties Experiment) database. The reconstructions were performed aiming at a root mean square error (*rmse*) lower than 1%. The results of these experiments indicate that PCA can effectively reduce the dimensionality of plant spectral databases from the visible to the infrared regions of the light spectrum, and that the PPCA approach can further maximize the accuracy/cost ratio of the storage and reconstruction of plant spectral reflectance and transmittance data.

1 Introduction

Vegetation is arguably the most important remote sensing target, since it is a natural resource on which all human and animal life depends [8]. The applications of plant remote sensing includes projects oriented towards agriculture, forestry and ecology. Viewed in this context, ground based measurements of

*School of Computer Science, University of Waterloo, Waterloo, Ontario, Canada
†School of Computer Science, University of Waterloo, Waterloo, Ontario, Canada

plant reflectance and transmittance measurements are essential, since they provide crucial data for the remote estimation of leaf biochemical content from space-borne platforms and for the investigation of photosynthesis and plant growth [11, 15, 4].

These measurements, however, may involve several dimensions, *e.g.*, spectrum, angle of incidence and specimen's maturation stage. These measurements can be stored in a database, with multiple entries for each species. The number of entries in this database grows as new measurements for different species become available. Eventually, however, the size of such database could become impractical for many applications due to the large number of species found in nature. Plant spectral data can also be extracted from computer models of plant reflectance and transmittance [10]. For applications that require high interactivity rates this data needs to be precomputed offline and reconstructed online. A measured or modeled species database can expand even further when combinations of plant parameters are considered. Clearly the key aspects in the design of such databases are compactness and low reconstruction error. Principal Components Analysis (PCA) [5] techniques can be used to achieve these goals.

PCA involves a mathematical procedure that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called principal components [12]. In some application areas, it is also called the (discrete) Karhunen-Love transform, or the Hotelling transform. PCA is widely used in signal processing, statistics, and neural computing. In remote sensing it is mainly applied to the classification and interpretation of satellite images and spectral data associated with different targets [7, 13, 14], including regions of vegetations.

In this study we examine the application of PCA to the reduction of the dimensionality of plant leaf spectral databases. To the best of our knowledge, a broad analysis of the application of PCA techniques in this context has not been done yet. Vrhel *et al.* [17] analyzed the possibility of representing with a set of basis functions, determined using PCA, spectral data of Munsell chips and various natural materials (including plant leaves). Their investigations, however, were aimed at colorimetry applications and were limited to reflectance data in the 390 nm to 730 nm range. Grossman et al., [6] used PCA in conjunction with stepwise multiple linear regression techniques to examine the statistical relationships among a number of foliar biochemicals. They indicated a reduction of the dimensionality of plant reflectance data in the 800 nm to 2498 nm range using PCA.

The investigation presented in this paper extends the observations made by Grossman *et al.* [6], focusing on the compactness and low reconstruction error of plant spectral databases. It covers the visible, near-infrared and infrared regions of the light spectrum, and includes reflectance and transmittance data. Moreover, besides the application of PCA techniques, a piecewise PCA approach (PPCA) is proposed. This biologically-based algorithm allows us to further maximize the storage savings and to minimize the reconstruction errors associated with plant spectral data. These findings are supported by experiments in which measured and reconstructed spectral curves for different specimens are compared.

2 Principal Components Analysis

The mathematical technique usually used in PCA is called eigen-analysis, described in standard texts for statistics and remote sensing [2]. Assuming a square symmetric $m \times m$ data matrix M, one solves for the eigenvalues and eigenvectors with sums of squares and cross products. The eigenvector associated with the largest eigenvalue has the same direction as the first principal component. The eigenvector associated with the second largest eigenvalue determines the direction of the second principal component. The sum of the eigenvalues equals the trace of the square matrix and the maximum number of eigenvectors equals the number of rows (or columns) of this matrix. If, conversely, M is a non-square or non-symmetric $m \times n$ matrix, one does not have to form explicitly the matrix with sums of squares and cross products, $M^T M$. Instead, one applies a numerically more stable method, and forms the singular value decomposition (SVD) [5] of M:

$$M = USV^T \tag{1}$$

where $U[m \times n]$ and $V[n \times n]$ are orthogonal matrices, $S[m \times n]$ is a real, nonnegative, and diagonal matrix. The columns u_i of $U = [u_1, ..., u_n]$ are the left singular vectors, and the columns v_i of $V = [v_1, ..., v_n]$ are the right singular vectors. The diagonal of S diagonal contains the so called singular values σ_i, where $\sigma_1 \geq ... \geq \sigma_n \geq 0$. U When M is symmetric and semidefinite, the singular values σ_i are the eigenvalues of M; otherwise, the singular values are the square roots of the eigenvalues of $M^T M$.

Consider a plant spectral reflectance data set given by a matrix M_R, where the rows represent reflectance spectra, $\rho(\lambda)$, for different species, and the columns represent the sample wavelengths. The goal is to reduce the dimensionality of M_R by applying a PCA technique, namely SVD:

$$\begin{aligned}[][u_R, s_R, v_R] &= svd(M_R) \\ d_R &= diag(s_R) \\ c_R &= M_R * v_R\end{aligned}$$

where d_R represents the singular values, v_R the basis of principal components (as column vectors), and c_R the coordinates relative to the basis (as row vectors). The same procedure can be applied to a plant spectral transmittance data set represented by a matrix M_T, where the rows represent transmittance spectra, $\rho(\lambda)$, for different species, and the columns represent the sample wavelengths.

Once the data matrices have been decomposed, the n basis components of v_R can be stored. This basis matrix may be as large as the original data set M_R, but since the goal is to reduce the dimension of the data, one usually chooses a reduced number of components, $n'_R << n_R$, and stores the smaller basis v'_R (the first n'_R columns of v_R). The selection of n'_R can be made according to the fraction of the singular value sum accounted for by the first n'_R singular values (see Figure 4). The reconstruction of the spectral data, *i.e.*, the formation of a $m \times n'_R$ matrix $\hat{M}_R$, is given by:

$$\hat{M}_R = c'_R * (v'_R)^T$$

It can be proven that the representation given by PCA is an optimal linear dimension reduction technique in the mean-square sense [12], although of course once the basis is chosen, any other basis spanning the same PCA subspace will have the same data compression or noise reduction capabilities.

3 Piecewise Principal Components Analysis

Experiments [16, 18] show that the reflectance and transmittance of plant leaves are relatively low in the visible portion of the light spectrum (400 nm to 700 nm), and that absorptance dominates. Most of the absorption in this region is caused by pigments presented in the leaf tissue. As pointed out by Devlin and Baker [3], the chlorophylls are by far the most important and abundant of these pigments. The near-infrared region (700 nm to 1300 nm), however, is characterized by a lack of absorption and high values for reflectance and transmittance [16], while in the infrared region (beyond 1300 $nm$to 2500 nm) the absorption is controlled by the water content.

These well defined regions suggest that the use of an adaptive PCA approach may result in more compactness and lower reconstruction errors. In other words, instead of an integral application of PCA over the whole light spectrum, we can perform a piecewise application of PCA, in which each piece corresponds to one of the three regions defined above. Since the effectiveness of the PCA method depends on the representativity of the reflectance or transmittance data used in the singular value decomposition, applying PPCA to matrices for the visible, near-infrared, and infrared regions yields different components than applying PCA to a matrix for all three regions. After eliminating components which contribute little to the overall reflectance or transmittance curves, reconstructing a particular curve with PPCA therefore gives pieces for each region which may not join smoothly at region boundaries. Although it would be possible to use a constrained PPCA method to ensure smooth joins, we found the the discontinuities to be noticeable only with some species, and only when using minimal numbers of components.

4 Data

In our experiments we considered specimens whose spectral data is available in the LOPEX (Leaf Optical Properties Experiment) database [9]. This database consists of 70 leaf samples representing 50 woody and herbaceous species that were obtained from trees and crops near the Joint Research Centre in Ispra, Italy. Spectra were originally scanned in 1 nm steps, but the wavelength interval was averaged over 5 nm to reduce noise.

The matrices M_R and M_T described earlier were formed using respectively the reflectance and transmittance spectra of all individual fresh leaves stored in the LOPEX database. For the reconstruction experiments we considered a soy (*Soja Hispida*) specimen representing an herbaceous species, and a poplar specimen (*Populus Canadensis*) representing a woody species. In order to increase our scope of observations we also considered an iris specimen (*Iris Germanica*) whose reflectance and transmittance curves have magnitudes markedly different from soy and poplar.

For the sake of consistency with the plant remote sensing literature we use the root mean square error (RMSE) measure in our experiments. As one increases the number of coefficients used in the reconstruction of the spectral curves, the RMSE of these curves gets lower. The goal is to obtain the best compromise between the number of coefficients and the RMSE. Therefore, we need to select a threshold RMSE value to be used as reference in the evaluation our results. Jacquemoud *et al.* [11] state that in terms of plant reflectance and transmittance reconstruction a RMSE less than or equal to 0.01 is considered to be very low and indicates an accurate reconstruction. Hence, we select this value as the reference for our comparisons.

5 Results

Figures 1 to 3 illustrate the application of PCA to the reconstruction of spectral reflectance and transmittance curves for the three specimens considered. The original curve, and reconstructions with 1, 5, and 9 principal components are shown (intermediate numbers of components are possible, but omitted here to avoid clutter). A visual inspection of these curves shows that there is good qualitative agreement between the original and the reconstructed curves. Although reconstruction errors vary from one specimen to another, by increasing the number of principal components one can obtain an asymptotical reduction of the error as illustrated by the graphs presented in Figure 4.

Tables 1 and 2 show the number of components required to reconstruct spectral curves within specified RMSE bounds. For PCA, numbers of components are given for the whole spectrum of interest, *i.e.*, from visible to infrared (400 nm to 2500 nm); for PPCA, numbers of components are given for the three characteristic sub-regions, namely visible (400 nm to 700 nm), near-infrared (700 nm to 1300 nm) and infrared (400 nm to 2500 nm). From the figures presented in these tables it can be clearly observed that the number of principal components varies from one species to another and from one region to another, and more principal components are required when one consider the whole region of interest, which indicates that the proposed PPCA algorithm requires a smaller number of principal components for a given error bound.

Figures 5 and 6 present comparisons involving the reconstruction of spectral reflectance and transmittance curves for the iris specimen applying PCA and PPCA, both using 5 components. We select this species because, from the species considered in our experiments, it presented the lower convergence rate (Figure 4). A visual inspection of these graphs shows that considerably lower reconstruction errors can be obtained using the proposed PPCA algorithm, *i.e.*, a smaller number of components are required to achieve the same accuracy level using PPCA. For example, 4 instead of 5 for RMSE less than or equal to 0.01, and 12 instead of 20 for RMSE less than or equal to 0.001. Similar improvements were observed for the other two specimens as indicated by the figures presented in Tables 1 and 2.

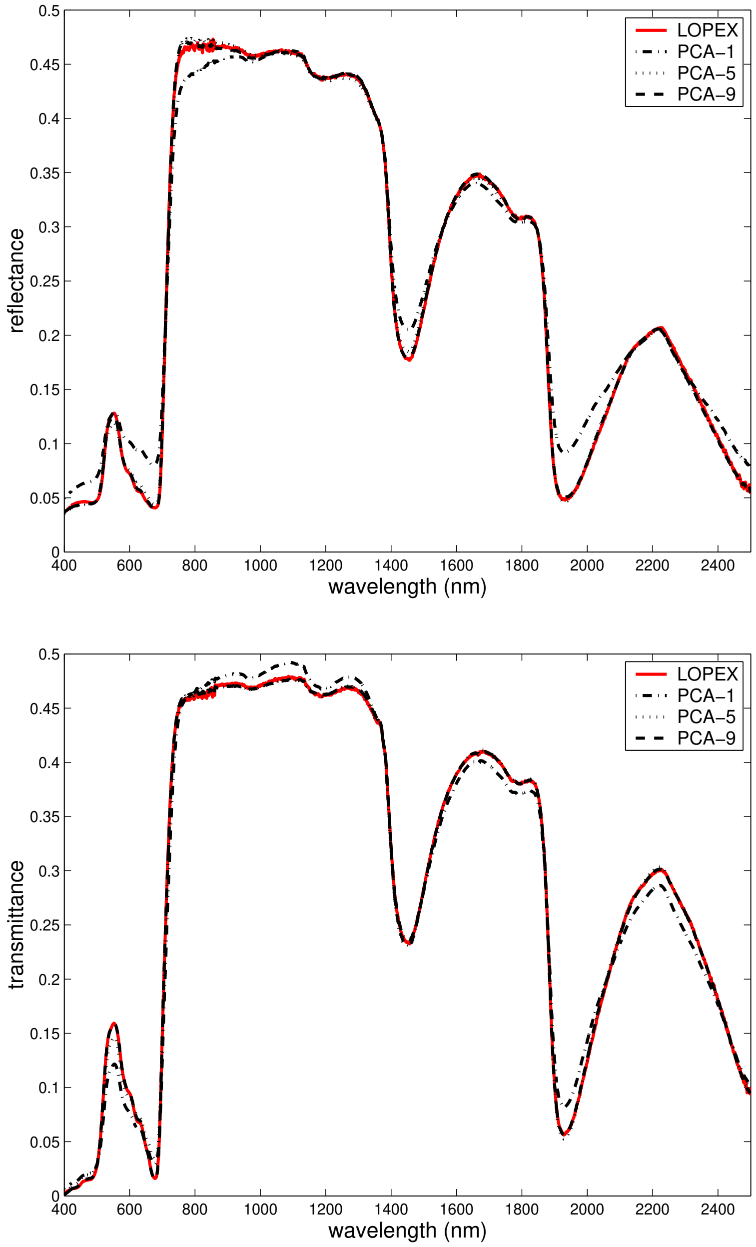

Figure 1: Measured (LOPEX) and reconstructed (PCA) spectral curves for a fresh soy leaf (*Soja hispida*). Curves were reconstructed using 1, 5 and 9 components.

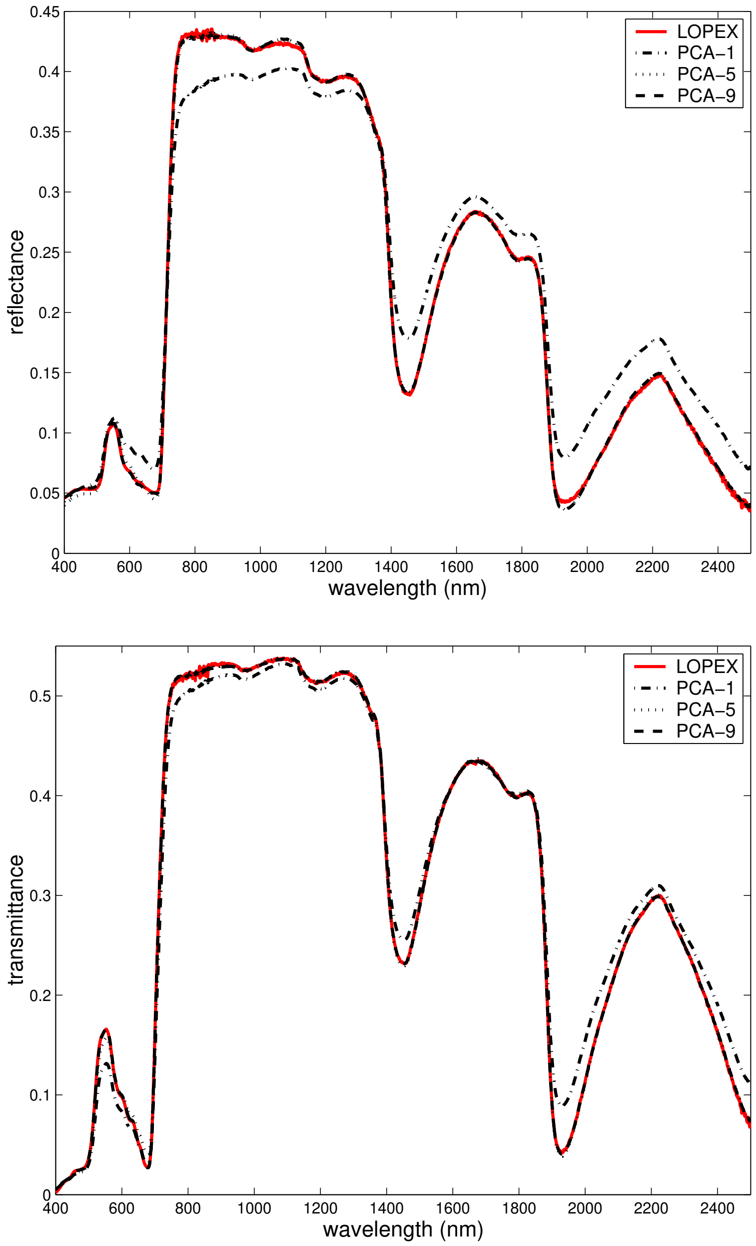

Figure 2: Measured (LOPEX) and reconstructed (PCA) spectral curves for a fresh poplar leaf (*Populus canadensis*). Curves were reconstructed using 1, 5 and 9 components.

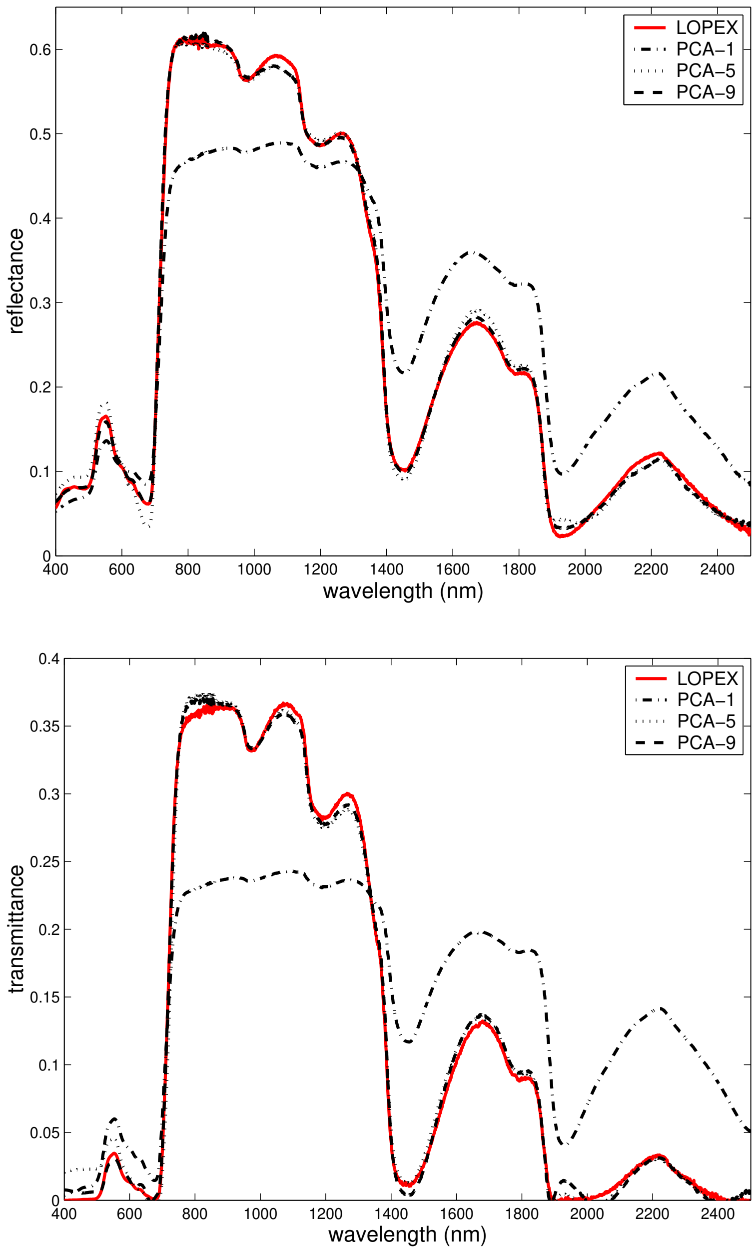

Figure 3: Measured (LOPEX) and reconstructed (PCA) spectral curves for a fresh iris leaf (*Iris germanica L.*). Curves were reconstructed using 1, 5 and 9 components.

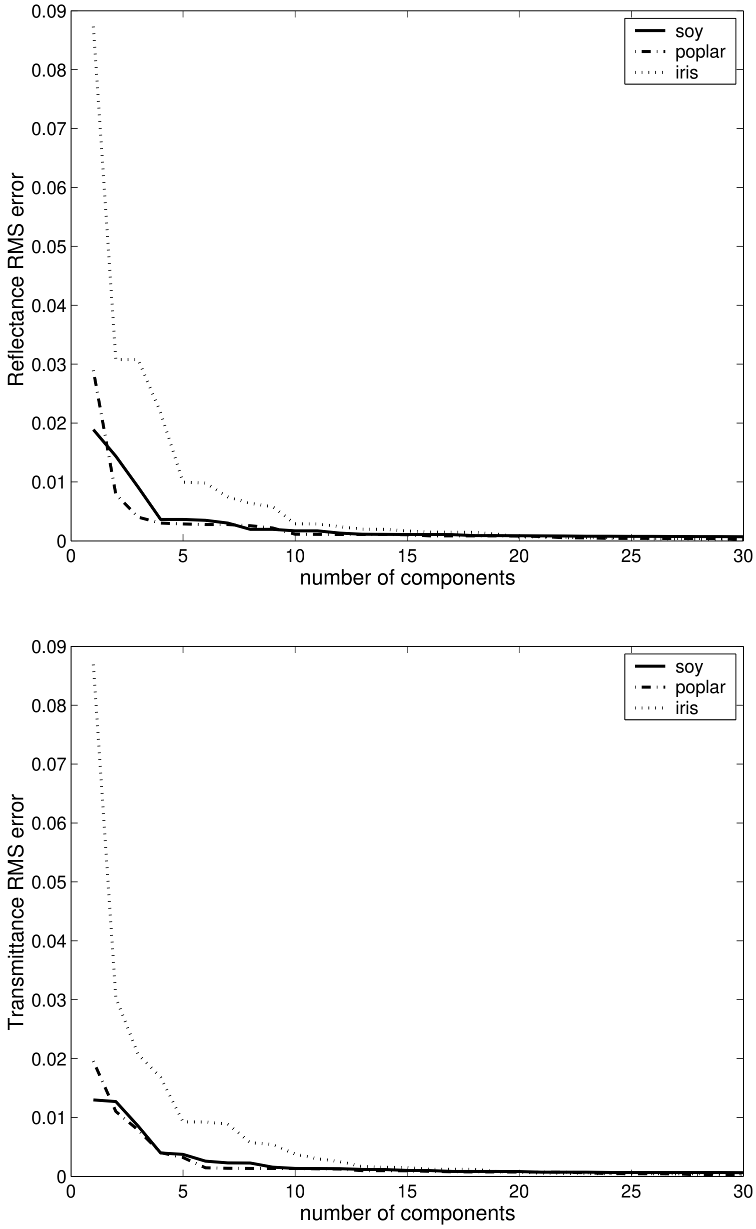

Figure 4: RMS errors associated with the reconstruction (PCA) of spectral reflectance curves (top) and spectral transmittance curves (bottom) for soy, poplar and iris leaves, within the visible to infrared (400 nm to 2500 nm) region of the light spectrum, and using different numbers of PCA components.

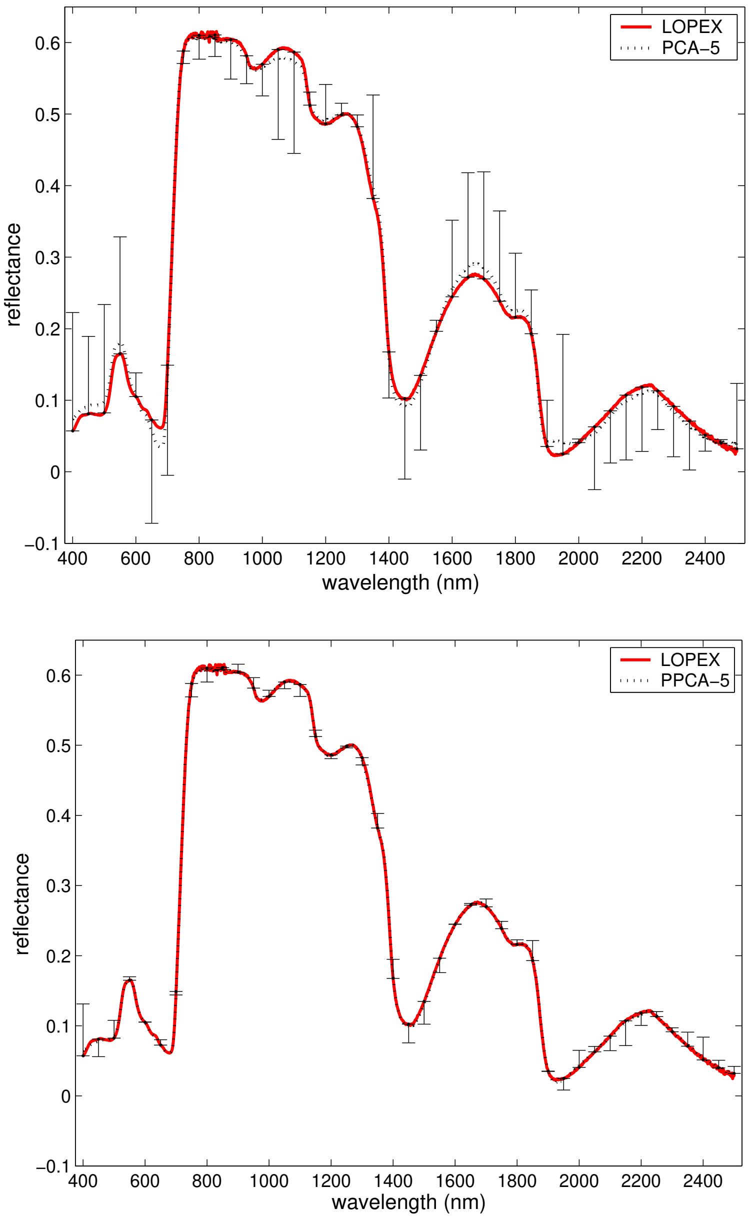

Figure 5: Measured (LOPEX) and reconstructed (PCA (top) and PPCA (bottom)) reflectance curves for a fresh iris leaf (*Iris germanica L.*). Curves were reconstructed using 5 components. Error bars are magnified by a factor of 10.

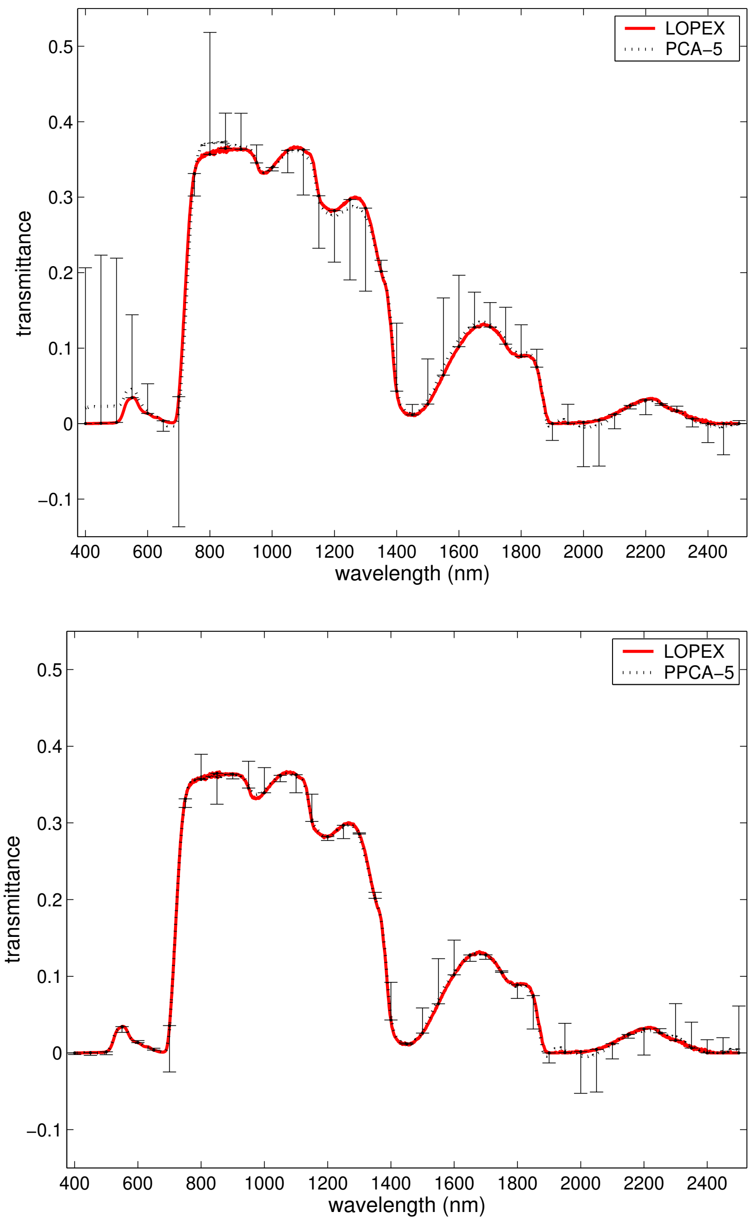

Figure 6: Measured (LOPEX) and reconstructed (PCA (top) and PPCA (bottom)) transmittance curves for a fresh iris leaf (*Iris germanica L.*). Curves were reconstructed using 5 components. Error bars are magnified by a factor of 10.

	visible to infrared			visible			near-infrared			infrared		
$rmse$	s	p	i	s	p	i	s	p	i	s	p	i
0.010	3	2	5	2	2	2	1	2	4	2	2	4
0.009	4	2	7	2	2	2	2	2	4	2	2	4
0.008	4	2	7	2	2	2	2	2	4	3	2	4
0.007	4	3	8	2	2	2	2	2	4	3	2	5
0.006	4	3	9	2	2	2	2	2	5	3	2	5
0.005	4	3	10	2	2	2	2	2	5	3	2	5
0.004	4	4	10	3	2	3	4	2	5	3	3	5
0.003	8	5	10	4	3	3	4	3	5	3	3	5
0.002	8	12	13	4	4	5	4	4	5	4	3	7
0.001	18	18	20	7	4	10	12	6	12	8	10	10

Table 1: Number of principal components required to reconstruct spectral reflectance curves for soy (s), poplar (p) and iris (i) leaves, within the visible to infrared (400 nm to 2500 nm), visible (400 nm to 700 nm), near-infrared (700 nm to 1300 nm) and infrared (1300 nm to 2500 nm) regions of the light spectrum, considering different RMS error (RMSE) bounds. The PCA method applies to the visible-to-infrared data; the PPCA method applies to the three remaining categories. Note that for a given species and RMSE error, the number of components required for PPCA is consistently smaller than for PCA.

	visible to infrared			visible			near-infrared			infrared		
rmse	s	p	i	s	p	i	s	p	i	s	p	i
0.010	3	3	5	2	2	1	3	2	3	2	2	4
0.009	3	3	7	2	2	1	3	3	3	3	2	4
0.008	4	3	8	2	2	1	3	3	3	3	2	4
0.007	4	4	8	2	2	1	3	3	4	3	2	4
0.006	4	4	8	2	2	1	3	3	4	3	2	4
0.005	4	4	10	2	2	2	3	3	5	3	2	5
0.004	4	5	10	3	2	2	3	3	5	3	3	5
0.003	6	6	11	3	2	2	3	3	5	3	4	7
0.002	9	6	13	3	2	3	3	3	6	4	4	7
0.001	16	13	19	6	5	4	9	9	7	7	7	8

Table 2: Number of principal components required to reconstruct spectral transmittance curves for soy (s), poplar (p) and iris (i) leaves, within the visible to infrared (400 nm to 2500 nm), visible (400 nm to 700 nm), near-infrared (700 nm to 1300 nm) and infrared (1300 nm to 2500 nm) regions of the light spectrum, considering different RMS error (RMSE) bounds. The PCA method applies to the visible-to-infrared data; the PPCA method applies to the three remaining categories. Note that for a given species and RMSE error, the number of components required for PPCA is consistently smaller than for PCA.

6 Discussion

In this paper we examined the application of techniques, PCA and PPCA, to reduce the dimensionality of plant spectral databases. Instead of storing the full spectra of plant leaves, one needs to store only a certain number of components and a basis, which requires only a fraction of the memory space. For example, each raw reflectance in the LOPEX database requires 2101 floating-point values; however, table 1 shows that for soy, 3-component PCA suffices for an RMSE of 0.01, thus 3 floating-point values (once the SVD basis has been stored). This is a dramatic reduction of data, especially for databases of many reflectances. Using the PPCA method, soy would require only 2, 1, and 2, components for each region respectively to achieve the same RMSE of 0.01. If all components are stored, this gives 5 floating-point values; a substantial reduction from the original data, but not as small as the 3 floating-point values for PCA. There are, however, alternative interpretations of this table that show PPCA can be the preferred method.

In applications where very low RMSE tolerances must be met, PPCA can compete with PCA directly in terms of total number of components. Table 1 shows this trend with iris transmittance for an RMSE of 0.001, where 19 components are required for PCA, and an equal 4+7+8=19 for PPCA. As another example in table 1, poplar reflectance with an RMSE tolerance of 0.002 requires 12 components for PCA, but only 4+4+3=11 components for PPCA. Further investigation of quantization error at the bit level may reveal other advantages to the PPCA approach. Even with floating-point values, however, the flexibility of PPCA is an asset.

For some applications, *e.g.*, infrared remote sensing [1], there may be different RMSE requirements for each spectral region. If we consider soy, with an RMSE of 0.001 in the infrared, but only 0.01 in the other regions, the PCA method needs 18 components. The PPCA method, however, can meet the infrared RMSE tolerance with 8 components, and require only 2 and 1 components respectively for the visible and near-infrared regions, a total of 11 components. PPCA is therefore more adaptable when the spectral regions differ in significance.

These storage savings depend on the number of principal components considered, which, in turn depend on the level of accuracy expected for the reconstruction of the original curves. Both techniques examined in this study allow a significant reduction of the dimensionality of plant spectral databases while preserving the information represented by the reflectance and transmittance data. The proposed PPCA technique, however, presents a higher accuracy/cost ratio.

Besides contributing to a reduction in the computational overhead associated with the manipulation of such data sets, such techniques may improve the quality of the stored spectral information stored, since the data not contained in the n selected principal components may be mostly due to noise in the measurements. Another benefit of these methods is that new reflectance or transmittance data may be incorporated into the database without repeating the SVD calculation: it is only necessary to project the new data onto the principal components to determine its coordinates, a simple matrix multiplication.

References

[1] M. Chelle, B. Andrieu, and K. Bouatouch. Nested radiosity for plant canopies. *The Visual Computer*, 14(3):109–125, 1998.

[2] J. C. Davis. *Statistics and Data Analysis in Geology.* John Wiley and Sons, New York, 1986.

[3] R.M. Devlin and A.V. Barker. *Photosynthesis.* Van Nostrand Reinhold Company, New York, 1971.

[4] N. S. Goel. Models of vegetation canopy reflectance and their use in estimation of biophysical parameters from reflectance data. *Remote Sensing Reviews*, 4:1–212, 1988.

[5] G. Golub and C. Van Loan. *Matrix Computations.* John Hopkins University Press, Baltimore, second edition, 1989.

[6] Y.L. Grossman, E.W. Sanderson, and S.L. Ustin. Relationships between canopy chemistry and reflectance for plant species from jasper ridge biological preserve, california. In *14th International Geoscience Remote Sensing Symposium - IGARSS'94*, page 2357, Pasadena, California, August 1994. Volume 4.

[7] R.P. Gupta. *Remote Sensing Geology.* Springer-Verlag, New York, 1991.

[8] H.N. Horler and J. Barber. Principles of remote sensing of plants. In H. Smith, editor, *Plants and the Daylight Spectrum*, pages 43–64, London, January 1981. Academic Press.

[9] B. Hosgood, S. Jacquemoud, G. Andreoli, J. Verdebout, G. Pedrini, and G. Schmuck. Leaf optical properties experiment 93. Technical Report Report EUR 16095 EN, Joint Research Center, European Comission, Institute for Remote Sensing Applications, 1995.

[10] S. Jacquemoud and S.L. Ustin. Leaf optical properties: A state of the art. In *8th International Symposium of Physical Measurements & Signatures in Remote Sensing*, pages 223–332, Aussois, France, 2001. CNES.

[11] S. Jacquemoud, S.L. Ustin, J. Verdebout, G. Schmuck, G. Andreoli, and B. Hosgood. Estimating leaf biochemistry using prospect leaf optical properties model. *Remote Sensing of Environment*, 56:194–202, 1996.

[12] I.T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, New York, 1986.

[13] T.M. Lillesand. *Remote Sensing and Image Interpretation*. Wiley & Sons, New York, 3rd edition, 1994.

[14] P.M. Mather. *Computer Processing of Remotely-Sensed Images, An Introduction*. Wiley, New York, 2nd edition, 1999.

[15] R. B. Myneni and J. Ross (Eds.). *Photon-Vegetation Interactions: Applications in Optical Remote Sensing and Plant Ecology*. Springer-Verlag, Berlin, 1991.

[16] C.J. Tucker and M.W. Garrat. Leaf optical system modeled as a stochastic process. *Applied Optics*, 16(3):635–642, 1977.

[17] M.J. Vrhel, R. Gershon, and L.S. Iwan. Measurement and analysis of object reflectance spectra. *Color Research and Application*, 19(1):4–9, 1994.

[18] J.T. Woolley. Reflectance and transmittance of light by leaves. *Plant Physiology*, 47:656–662, 1971.