ROBUST PRINCIPAL COMPONENT ANALYSIS

Matthew Partridge and Marwan Jabri
School of Electrical and Information Engineering
University of Sydney
Email: mp,marwan@sedal.usyd.edu.au
Web: www.sedal.usyd.edu.au

Abstract. Principal component analysis (PCA) is a technique used to reduce the dimensionality of data. In particular, it may be used to reduce the noise component of a signal. However, traditional PCA techniques may themselves be sensitive to noise. Some robust techniques have been developed, but these tend not to work so well in high dimensional spaces. This paper discusses the robustness properties of a recent PCA algorithm, $SPCA_1$. It shows theoretically and experimentally that this algorithm is less sensitive to the presence of outliers.

INTRODUCTION

Principal component analysis (PCA) transforms a set of n variables (or dimensions) into another set of $d \leqslant n$ uncorrelated variables, maintaining as much of the variance of the original data as possible. Techniques for performing PCA include singular value decomposition (SVD) [10], Hotelling's power method [7], and, more recently, Hebbian techniques such as Oja's rule [6]. They are all designed to find principal components that maximise the variance.

Principal component analysis is useful in compression, feature extraction and in classification applications. In classification applications, the components found by PCA may help classifiers by making the dimensionality of the problem smaller—hence faster to process and in some ways simpler. It also means that the data may be "cleaned" in the sense that the first few principal components are the most important—they are the "signal"—and the rest are unimportant—they are the "noise"—and may be discarded.

As PCA is used in noisy environments, it is important that it is robust. For a technique to be robust, it is required to degrade gradually in the presence of outliers. It is often hard to tell which item is an outlier and which is part of the underlying distribution which one is trying to model. Nonetheless, it would be desirable for the technique not to be too dependent on any particular items.

0-7803-6278-0/00$10.00 (C) IEEE

Earlier work on robust techniques is described in the next section. Then the L_1 and L_2 norms of projection are defined and the convergence of the simple PCA (SPCA) algorithms [9] is shown in these terms. This allows the robustness of SPCA_1 to be compared with other algorithms that maximise the explained variance. The stability of the principal components found using SPCA_1 is then demonstrated in several examples and experimentally shown using real data.

EARLIER WORK ON ROBUST TECHNIQUES

Approaches to making PCA robust against non-normal outliers are described:

> "Three broad approaches can be taken to increase the robustness of PCAs to outliers. First, outlying observations can be eliminated from the sample; second, outliers can be modified by replacing them with more appropriate values; third, robust versions of covariance/correlation matrices can be used." [1, p243].

The first two of these approaches edit the data. The last modifies the covariance matrix, which [1] then describes. The data from which the covariance matrix is constructed may be weighted such that samples far from the mean have less importance. This is the idea behind the "M-estimators" approach which weights the data according to:

$$\omega(d) = \begin{cases} d & \text{if } d \leqslant d_0 \\ d_0 e^{-\frac{1}{2}(d-d_0)^2/b_2^2} & \text{otherwise} \end{cases}$$

where, for n-dimensional data,

$$d_0 = n^{\frac{1}{2}} + b_1/2^{\frac{1}{2}}$$

and $d = (\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})$ is the Mahalanobis distance, for some constants b_1 and b_2 which dictate the threshold of the "good"/"bad" data boundary and the degree to which the weights of the outliers decay, respectively. [2] recommend $(b_1, b_2) = (2, \infty)$ or $(2, 1.25)$. [5] report that M-estimators breakdown for asymmetric outliers or in high dimensional spaces.

M-estimators are considered in more depth and the break-down regions for M-estimators as well as for other robust approaches have been quantified [11]. In particular, they find that the fraction of outliers that M-estimators can tolerate is at most $1/(n+1)$ for n dimensional data; minimum volume ellipsoid estimators (MVE) only for $m/n > 5$. This is not useful in high dimensional datasets some of which may have very few samples relative to the dimensionality. Furthermore, the computation of M-estimators and other robust approaches is demanding—firstly the covariance structure must be estimated, secondly the weights must be determined (using the Mahalanobis distance from the estimated covariance structure) and then the covariance

structure re-estimated based on the weighted data; this procedure being repeated until convergence. For high dimensional data, such an approach can be computationally prohibitive.

The desirability of the maximisation of the L_1 norm for PCA is identified in [12], but they instead decide also to consider an M-estimator cost function approach:

> "It is well known that Principal Component Analysis (PCA) is optimal in the sense of Mean Square Error (MSE). However, the estimation based on MSE is sensitive to noise or outliers, therefore, it is not a robust estimator. In order to get a robust estimation, absolute error criterion (L_1 norm) could be used, but it is not differentiable at the origin point; and minimax criterion (L_∞ norm) could also be applied, but only for batch learning. ... " [12, p120]

It should be noted that the degree of robustness observed depends on the type of distribution, as [1] observes. However, we believe that many common problems exhibit non-normal outliers which lie in the tails of the distribution, which is a suitable precondition for the application of a method that maximises the L_1 norm.

SPCA ALGORITHMS

More recently, simple PCA (SPCA) has been developed [9]. It is a fast PCA technique. In batch mode, it works on mean-corrected data; it comes in two variants and may be expressed in terms of the data $\mathbf{x}_i$ as:

$$\hat{\boldsymbol{\alpha}} \leftarrow \text{unit}\left(\sum_{\hat{\boldsymbol{\alpha}}'\mathbf{x}_i \geqslant 0} \hat{\boldsymbol{\alpha}}\right) \qquad (\text{SPCA}_1)$$

$$\hat{\boldsymbol{\alpha}} \leftarrow \text{unit}\left(\sum_{\mathbf{x}_i} (\mathbf{x}_i'\hat{\boldsymbol{\alpha}})\mathbf{x}_i\right) \qquad (\text{SPCA}_2)$$

where $\hat{\boldsymbol{\alpha}}$ is the current approximation to the principal eigenvector and $\text{unit}(\mathbf{x}) = \mathbf{x}/\|\mathbf{x}\|$ normalises the vector to unit length. The principal component is the dot product: $\hat{\boldsymbol{\alpha}}'\mathbf{x}_i$, so its accuracy is determined by the accuracy of the eigenvector.

Further principal eigenvectors (and components) may be found by "deflating" the data, a technique analogous to the deflation of the covariance matrix in Hotelling's power method, and which is described in [9]. However, this deflation does not currently concern us.

In this section, the L_1 and L_2 norm of projections are defined, then it is shown that SPCA_1 maximises the L_1 norm of projections and that SPCA_2 maximises the L_2 norm of projections. The explained variance of $\hat{\boldsymbol{\alpha}}$ on a

dataset $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$, is given by

$$\sigma^2_{\hat{\boldsymbol{\alpha}}} = \sum_{\mathbf{x}_i \in \mathbf{X}} \left\| \hat{\boldsymbol{\alpha}}' \mathbf{x}_i \right\|^2 \tag{1}$$

Similarly, we define the projection length (PL) of $\mathbf{X}$ projected onto $\hat{\boldsymbol{\alpha}}$, by

$$\mathrm{PL}(\mathbf{X}, \hat{\boldsymbol{\alpha}}) \equiv \sum_{\mathbf{x}_i \in \mathbf{X}} \left| \hat{\boldsymbol{\alpha}}' \mathbf{x}_i \right| \tag{2}$$

The explained variance may be seen as the L_2 norm of projections; the projection length as the L_1 norm of projections.

$\mathbf{SPCA_1}$ Maximises L_1 Norm of Projections

Let $J_1 = \mathrm{PL}(\mathbf{X}, \hat{\boldsymbol{\alpha}})$ be the objective function, which we would like to show that SPCA_1 maximises. Denote $\mathbf{X}_+ = \{\mathbf{x}_i \in \mathbf{X} : \hat{\boldsymbol{\alpha}}' \mathbf{x}_i \geqslant 0\}$ and $\mathbf{X}_- = \{\mathbf{x}_i \in \mathbf{X} : \hat{\boldsymbol{\alpha}}' \mathbf{x}_i < 0\} = \mathbf{X} \backslash \mathbf{X}_+$.

$$J_1 = \sum_{\mathbf{x}_i \in \mathbf{X}} \left| \hat{\boldsymbol{\alpha}}' \mathbf{x}_i \right| = \left(\sum_{\mathbf{x}_i \in \mathbf{X}_+} \hat{\boldsymbol{\alpha}}' \mathbf{x}_i - \sum_{\mathbf{x}_i \in \mathbf{X}_-} \hat{\boldsymbol{\alpha}}' \mathbf{x}_i \right) = \hat{\boldsymbol{\alpha}}' \left(\sum_{\mathbf{x}_i \in \mathbf{X}_+} \mathbf{x}_i - \sum_{\mathbf{x}_i \in \mathbf{X}_-} \mathbf{x}_i \right)$$

Now for mean corrected data, $\sum_{\mathbf{x}_i \in \mathbf{X}} \mathbf{x}_i = \mathbf{0}$, so

$$\begin{aligned} \sum_{\mathbf{x}_i \in \mathbf{X}_+} \mathbf{x}_i &= - \sum_{\mathbf{x}_i \in \mathbf{X}_-} \mathbf{x}_i \\ 2 \sum_{\mathbf{x}_i \in \mathbf{X}_+} \mathbf{x}_i &= \sum_{\mathbf{x}_i \in \mathbf{X}_+} \mathbf{x}_i - \sum_{\mathbf{x}_i \in \mathbf{X}_-} \mathbf{x}_i \\ \therefore J_1 &= 2 \hat{\boldsymbol{\alpha}}' \sum_{\mathbf{x}_i \in \mathbf{X}_+} \mathbf{x}_i \end{aligned}$$

For any unit vector $\mathbf{u}$, $\arg\max_{\mathbf{u}} \mathbf{u}'\mathbf{y}$ is parallel to $\mathbf{y}$, so to maximise J_1, $\hat{\boldsymbol{\alpha}}$ is parallel to $\sum_{\mathbf{x}_i \in \mathbf{X}_+} \mathbf{x}_i$; so $\hat{\boldsymbol{\alpha}} = \pm \mathrm{unit} \left(\sum_{\mathbf{x}_i \in \mathbf{X}_+} \mathbf{x}_i \right)$ maximises J_1.

So SPCA_1 is an expectation-maximisation (EM) algorithm [4] for maximising $\hat{\boldsymbol{\alpha}}$. In the E-step, the partitions $\mathbf{X}_+$ and $\mathbf{X}_-$ are estimated using the current value of $\hat{\boldsymbol{\alpha}}$. In the M-step, J_1 is maximised, determining $\hat{\boldsymbol{\alpha}}$, using the current partitioning of $\mathbf{X}_+$ and $\mathbf{X}_-$.

$\mathbf{SPCA_2}$ Maximises L_2 Norm of Projections

Note that this proof includes the proof showing that Hotelling's Power Method maximises the explained variance (see [8] for example). Let $J_2 = \sigma^2_{\hat{\boldsymbol{\alpha}}}$ be the objective function, which we would like to show that SPCA_2 maximises. Consider the sample covariance matrix $\boldsymbol{\Sigma} = 1/(m-1) \sum_i^m \mathbf{x}_i \mathbf{x}_i'$. Note that any vector is expressible as a linear combination of the eigenvectors—so suppose

initially $\hat{\boldsymbol{\alpha}}^{(0)} = \sum_i^n \kappa_i \boldsymbol{\alpha}_i$ for some scalars κ_i and the eigenvectors $\boldsymbol{\alpha}_i$ (with corresponding eigenvalues λ_i). Then after one iteration of SPCA$_2$,

$$\begin{aligned}
\hat{\boldsymbol{\alpha}}^{(1)} &= \frac{1}{m-1}\sum_i^m (\mathbf{x}_i'\hat{\boldsymbol{\alpha}}^{(0)})\mathbf{x}_i = \frac{1}{m-1}\sum_i^m (\mathbf{x}_i\mathbf{x}_i')\hat{\boldsymbol{\alpha}}^{(0)} \\
&= \boldsymbol{\Sigma}\hat{\boldsymbol{\alpha}}^{(0)} = \boldsymbol{\Sigma}\sum_i^N \kappa_i\boldsymbol{\alpha}_i \\
&= \sum_i^N \kappa_i \boldsymbol{\Sigma}\boldsymbol{\alpha}_i \\
&= \sum_i^N \kappa_i\lambda_i\boldsymbol{\alpha}_i \qquad (\text{as } \boldsymbol{\Sigma}\boldsymbol{\alpha}_i = \lambda_i\boldsymbol{\alpha}_i) \\
\frac{\hat{\boldsymbol{\alpha}}^{(1)}}{\lambda_1} &= \kappa_1\boldsymbol{\alpha}_1 + \sum_{i=2}^N \frac{\lambda_i}{\lambda_1}\kappa_i\boldsymbol{\alpha}_i
\end{aligned}$$

Similarly, after the p^{th} pass,

$$\frac{\hat{\boldsymbol{\alpha}}^{(p)}}{\lambda_1^p} = \kappa_1\boldsymbol{\alpha}_1 + \sum_{i=2}^N \left(\frac{\lambda_i}{\lambda_1}\right)^p \kappa_i\boldsymbol{\alpha}_i$$

When $\hat{\boldsymbol{\alpha}}$ is normalised, one sees that $\hat{\boldsymbol{\alpha}} \to \boldsymbol{\alpha}_1$ as $p \to \infty$, at the rate $O\left(\frac{\lambda_1}{\lambda_2}\right)^p$. Also

$$J_2 = \sum_{\mathbf{x}_i \in \mathbf{X}} \left\|\hat{\boldsymbol{\alpha}}'\mathbf{x}_i\right\|^2 = \sum_{\mathbf{x}_i \in \mathbf{X}} \left(\hat{\boldsymbol{\alpha}}'\mathbf{x}_i\right)\left(\mathbf{x}_i'\hat{\boldsymbol{\alpha}}\right) = \hat{\boldsymbol{\alpha}}'\boldsymbol{\Sigma}\hat{\boldsymbol{\alpha}}$$

But $\hat{\boldsymbol{\alpha}} = \boldsymbol{\alpha}_1$ maximises J_2 as $\boldsymbol{\alpha}_1'\boldsymbol{\Sigma}\boldsymbol{\alpha}_1 = \lambda_1$ is a maximum.

SPCA$_1$ and SPCA$_2$ are fast PCA techniques [9]. SPCA$_1$ which maximises the L_1 norm is more robust than those techniques that maximise the L_2 norm. These techniques include SVD, Householder-QR, Hotellings power method, Oja's Hebbian rule and SPCA$_2$. The next sections will demonstrate through simulations and experiments that the L_1 norm is more robust than the L_2 norm. SPCA$_2$ will be used representitively of these L_2 maximisation rules for speed reasons.

EXPERIMENTS

For a technique to be robust, it is required to degrade gradually in the presence of outliers. It is often hard to tell which item is an outlier and which is part of the underlying distribution which one is trying to model. Nonetheless, it would be desirable for the technique not to be too dependent on any

particular items. Concretely, points which are far from the mean (which often are outliers coming from a different distribution) should not be overly weighted.

For the central value, minimising the L_1 norm (finding the median value) is more robust in this sense than minimising the L_2 norm (finding the mean value). For PCA the same holds. Intuitively, the L_1 norm weights distant points less than the L_2 norm.

Two Dimensional Artificial Examples

To demonstrate this, we constructed a zero-mean 2 dimensional artificial dataset. We used SPCA_1 to find the maximal L_1 projection norm and SPCA_2 to find the maximal L_2 projection norm. They found the same direction as depicted in Figure 1a. Next we added a few random zero-mean points which were much further away from the mean. Whilst they were much further from the mean (we had to "zoom out" of the figure), they only represented less than 2% of the total number of data points, so for a robust technique, their impact should be slight. As seen in Figure 1b, the impact on the maximal L_1 projection is much less than that of the maximal L_2 projection. In fact, the correlation between the vector found originally and that found by the maximal L_1 projection is about 99%; whereas between the original direction and the maximal L_2 projection is only about 78%.

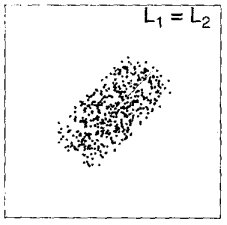

Figure 1a: The L_1 and L_2 projection norms yields the same result.

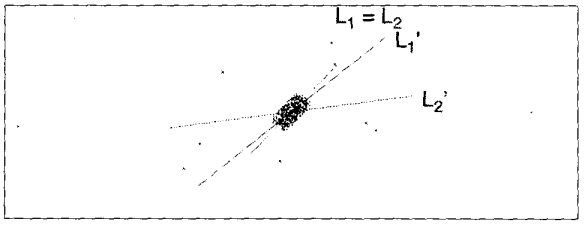

Figure 1b: In the presence of distant points, maximising the L_1 projection norm is less perturbed than the L_2 projection norm.

As another demonstration, we considered the robustness of the two PCA algorithms when a robust centre was used. In particular, the SPCA_1 algorithm was modified to be performed on median-corrected data, as opposed to mean-corrected data.[1] On the original data, the small difference between the mean and median values meant that different solutions were found by the two techniques, as is shown in Figure 2. The correlation between these vectors was found to be 99.8%.

In another experiment we investigate non-centered noise. This time a few

[1]Note that as the data is no longer mean-corrected, to ensure J_1 maximisation, the SPCA_1 algorithm should be modified. $\hat{\alpha} \leftarrow \text{unit}(\sum_{\mathbf{x}_i \in \mathbf{X}_+} \mathbf{x}_i - \sum_{\mathbf{x}_i \in \mathbf{X}_-} \mathbf{x}_i)$.

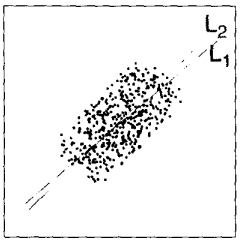

Figure 2: Maximisation of L_1 projection norm with respect to the median yields a slightly different solution to the maximisation of L_2 projection norm with respect to the mean.

(again about 2%) random points were added which had a similar spread, but a very different mean value. In the blow-up in Figure 3a, one can see that the median is a more robust estimate of the average value than the mean. This impact, along with the impact of the distant data points can be seen in Figure 3b. The correlation between two vectors produced by the maximal L_1 projection based on the median was found to be 85%. The correlation between two vectors produced by the maximal L_2 projection based on the mean was found to be 6.3%.

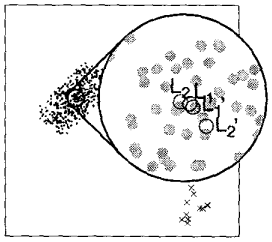

Figure 3a: The median (L_1) is perturbed less (L_1') by the introduction of new data than the mean (L_2, L_2').

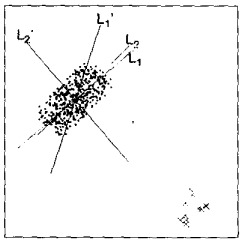

Figure 3b: Maximising the L_1 projection norm using the median is perturbed less than maximising the L_2 projection norm using the mean.

Experiments Using Real Data

Human intra-cardiac electrogram (ICEG) readings were used to experimentally assess the relative robustness of the SPCA$_1$ algorithm. The ICEG dataset was made available by Telectronics Pacing Systems[2] (see [3] for example). It consists of waveforms sampled 30 times per instance. The patient identification numbers label the waveforms; examples are shown in Figure 4.

The ICEG dataset was used to investigate how the robustness depends on whether the L_1 or L_2 norm is used. Results on several classes from this dataset were obtained to assess the sensitivity of the robustness claims. The

[2]Australian manufacturer of implantable defibrillators.

Figure 4: Examples of the ICEG dataset.

first principal eigenvector was determined (separately for each class), by each method. Noise was then added to the dataset and the principal eigenvectors were redetermined, based on the noisy data. The correlation between the principal eigenvector on the "clean" data and the principal eigenvector found on the "dirty" data was then determined. This procedure was repeated for different noise levels. 150 trials at each noise level were performed to determine the average and variance of the results.

The noise was added to the dataset by changing the labels. Noise could have been added to the data samples, but it is unclear what model of noise is appropriate in this case. Each sample had a probability of having its label being reassigned to any class (including potentially back to its original class). The probabilty of this reassignment was used to control the level of noise—it was varied between 0% and 15%. This type of noise meant that the class distributions were altered by the inclusion of data from other classes.

For a robust principal component estimator, the method should degrade "smoothly" in the presence of increasing noise. In this case, this means that the correlation between the "clean" principal eigenvector and the "dirty" principal eigenvector should be close to one, and decrease slowly as more noise is added.

Results. Figure 5 shows the results for the first 9 classes of the ICEG dataset. The correlation is shown as a function of the noise level for each class for SPCA_1 and SPCA_2. The solid curve shows the median result over the 150 trials, and the shaded areas are bounded by upper and lower quartiles of the trials.

In most cases, one can see that there is a good degree of separation between the two methods, and that the method based on the L_1 norm degrades more slowly than that based on the L_2 norm. For class 0711, the results do not appear to be significant, but neither method has degraded to a correlation less than 0.99 on average for the noise range that was added, so both appear to be quite robust. This is probably because the first eigenvector strongly dominates subsequent eigenvectors. For class 0754, the SPCA_1 algorithm appears to be significantly more robust over the initial noise range, but later insignificantly so. In this case, the first two eigenvalues are probably quite close, and hence a small amount of added noise has a large effect on which eigenvector is found. Nonetheless, in all cases, methods based on L_1 norm appear to be more robust to this type of noise for this type of dataset than methods based on the L_2 norm.

CONCLUSION

A robust technique is one which is not too sensitive to any particular data item. The notion of a norm of projection is introduced. The analysis and examples demonstrate that the L_1 norm of projection is more robust to data lying further from the central value than the L_2 norm of projection. This indicates that it may be safer to use a technique which maximises the L_1 norm of projection (such as SPCA_1) when considering data which perhaps comes from more than one source, or if it is expected that there is noise in the data sample. This has been verified through artificial and practical experimentation. Unlike earlier robust PCA techniques, the maximisation of the L_1 norm of projection is a very fast procedure and works for high dimensional data.

REFERENCES

[1] A. Basilevsky, **Statistical Factor Analysis and Related Methods: Theory and Applications**, New York: Wiley-Interscience, 1994.

[2] N. A. Campbell, "Robust Procedures in Multivariate Analysis I: Robust Covariance Estimation," **Applied Statistics**, vol. 29, pp. 231–237, 1980.

[3] R. J. Coggins, M. A. Jabri, "A Low Complexity Intracardiac Electrogram Compression Algorithm," **IEEE Transactions on Biomedical Engineering**, vol. 46, no. 1, pp. 82-91, 1999.

[4] A. P. Demster, N. M. Laird and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," **Journal of the Royal Statistical Society**, vol. 39, no. B, pp. 1–38, 1977.

[5] S. J. Devlin, R. Gnanadesikan and J. R. Kettenring, "Robust Estimation of Dispersion Matrices and Principal Components," **Journal of the American Statistical Association**, vol. 76, pp. 354–362, 1981.

[6] K. I. Diamantaras and S. Y. Kung, **Principal Component Neural Networks: Theory and Applications**, New York: John Wiley & Sons, 1996.

[7] H. Hotelling, "Analysis of a Complex of Statistical Variables into Principal Components," **Journal of Educational Psychology**, vol. 24, pp. 417–441, 1933.

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

[9] M. G. Partridge and R. A. Calvo, "Fast Dimensionality Reduction and Simple PCA," **Intelligent Data Analysis**, vol. 2, no. 3, pp. 1, 1998.

[10] W. H. Press, S. A. Teukolsky, W. T. Vetterling and B. P. Flannery, **Numerical Recipes in C: The Art of Scientific Computing**, Cambridge: Cambridge University Press, 2nd edn., 1992.

[11] P. J. Rousseeuw and B. C. Van Zomeren, "Unmasking Multivariate Outliers and Leverage Points," **Journal of the American Statistical Association**, vol. 85, pp. 633–639, 1990.

[12] C. Wang, H.-C. Wu and J. C. Principe, "Cost Function for Robust Estimation of PCA," in **Proceedings of SPIE**, Bellingham, WA, USA, 1996, vol. 2760, pp. 120–127.

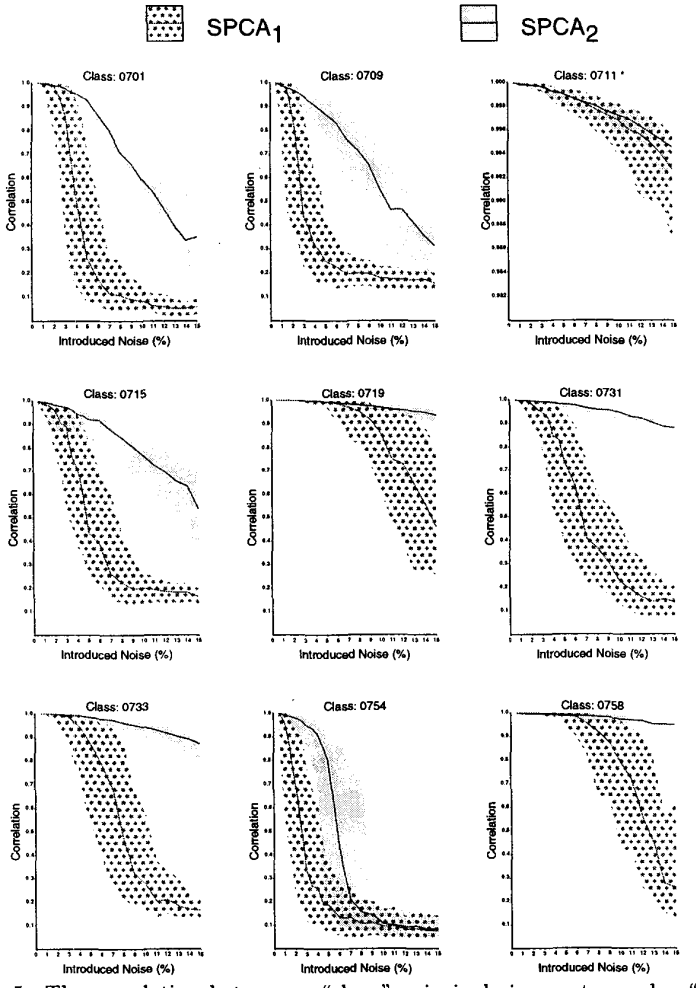

Figure 5: The correlation between a "clean" principal eigenvector and a "dirty" principal eigenvector for various noise levels for the different classes of the ICEG dataset. The median of 150 trials is shown as a solid line; the upper and lower quartiles bound the shaded areas. *Note the different scale for class 0711.