

# *PCANOVA: Combining principal components with analysis of variance to assess group structure*

Kevin R. Coombes

Department of Biostatistics, UT M.D. Anderson Cancer Center,  
1515 Holcombe Blvd., Houston TX 77030 USA

## **Abstract**

The development of high-density microarrays has led biologists to collect data sets that simultaneously measure the expression of thousands of genes on tens of samples. In many cases, the samples are known to belong to several distinct groups. Recent work of Landgrebe et al. constructs statistics to test differential gene expression based on a principal components analysis (PCA) of the matrix of group means. In this paper, we investigate the relationships between the usual PCA on all the data, a second PCA on the matrix of group means, and a third PCA on the matrix of residuals obtained after removing the group means. We show that the group means PCA can exhibit significant apparent structure even on randomly generated data, and we compare the resulting plots to ideas suggested by Krzanowski in the context of chemometrics. Next, we introduce a method for quantifying the difference between two sets of orthogonal vectors. By applying our method to the principal component vectors from three different PCAs, we can perform analysis of variance (ANOVA) on the principal component vectors; we call this “PCANOVA”. We use simulated data to explore the distribution of the PCANOVA statistics. Finally, we apply PCANOVA to a real-world data set: a set of gene expression data collected by Ross et al. by performing microarray experiments on the NCI-60 cell lines, which represent 8 different kinds of cancer.

**Keywords:** principal components analysis, analysis of variance, ANOVA, microarray

## Introduction

The development of high-density cDNA or oligonucleotide microarrays has led biologists to collect numerous data sets that simultaneously measure thousands of variables (gene expressions) on tens or at most hundreds of samples. (For examples, see [Alizadeh et al., 2000] or [Ross et al., 2000].) Many methods have been used to elicit structure from this data, including hierarchical clustering, self-organizing maps, and principal components analysis (PCA). In many experimental designs, however, the samples are already known to belong to distinct groups. In this situation, purely unsupervised methods are not the ideal choice for performing the analysis. One would prefer to perform a discriminant analysis, particularly one built around the canonical variates. Unfortunately, such methods founder on the singularity of the matrices when the number of variables far exceeds the number of independent samples.

Nguyen and Rocke suggested a multi-step process to circumvent this difficulty in the context of microarrays [Nguyen and Rocke, 2002]. To start, they borrowed the idea of partial least squares (PLS) as a data reduction technique from the field of chemometrics. They then followed the data reduction step with a classification method such as logistic discrimination or quadratic discriminant analysis. PLS is analogous to PCA; the difference is that PCA chooses orthogonal linear combinations of the predictor variables that maximize the variance, while PLS chooses orthogonal linear combinations that maximize the covariance between the predictors and a response variable. The PLS-based method has two weaknesses. First, if the data reduction step is preceded by a gene selection step (for example, performing gene-by-gene t-tests and selecting the genes with the most extreme t-values), then the results obtained using PLS are little better than those achieved using PCA. Second, and more importantly, the PLS-based method is optimized for binary classification. In order to apply the method when there are more than two groups, one must replace the single binary-valued response vector either with a many-valued response vector or a set of binary contrast vectors. It is not clear how the final results will depend on the chosen contrasts.

The problem of improving on PCA for classification when there are more than two groups has been considered by Krzanowski in the context of chemometrics [Krzanowski, 1992]; he suggested combining two methods. Specifically, he advocated ranking the principal components in the order suggested by a canonical variate analysis. This reordering replaces the goal of maximizing the total variance explained with the goal of maximizing

the ratio of between-group to within-group variance.

Landgrebe and co-workers first applied Krzanowski’s idea to microarray data [Landgrebe et al., 2002], extending the method so it could be used to select genes that distinguished between groups. First, they perform PCA on all the data. They then use Krzanowski’s method to rank the principal components. They perform one-way analyses of variance on the loadings for each component and use the F-test value to find the number,  $k$ , of components that are significant at a fixed  $p$ -value. Next, they perform PCA on the matrix of group means, retaining only the first  $k$  components. Finally, they define a test statistic that uses these components to measure the differential expression of each gene, and use a permutation test to evaluate the significance of these test statistics.

The sudden leap by Landgrebe and co-workers from the PCA on all the data to a different PCA on the matrix of group means deserves closer investigation. There is an implicit assumption here that the latter PCA more truly reflects the group structure in the data. While plausible, this assumption has not been tested.

In this paper, we investigate the relationship between three different PCAs: the usual PCA on all the data, a second PCA on the matrix of group means, and a third PCA on the matrix of residuals obtained after removing the group means. We refer to these computations as the “full PCA”, the “group means PCA”, and the “residual PCA”, respectively. We show first that the leading principal components from the group means PCA do a better job of separating the samples into groups than the components chosen using Krzanowski’s criterion. We also show that the group means PCA can exhibit significant apparent structure even on randomly generated data. Next, we introduce a method for quantifying the difference between two sets of orthogonal vectors. This method is constructed from the angles between principal component vectors; a discussion of the relevance of the geometry of vectors to the analysis of microarray data has recently appeared [Kuruville et al., 2002]. By applying our method to the principal component vectors from three different PCAs, we can perform something like analysis of variance (ANOVA); we refer to this application as “PCANOVA”. We show that PCANOVA can be used to assess how well the data truly reflects the group structure, and we use simulated data to explore the distribution of the PCANOVA statistics. Finally, we apply these ideas to a real-world data set: a set of gene expression data collected by performing microarray experiments on the NCI-60 cell lines, which represent 8 different kinds of cancer [Ross et al., 2000].

## 1: Manipulating matrices

We are interested in multivariate experiments that measure  $p$  variables on a total of  $n$  samples chosen from  $g$  different groups. We assume that  $p \gg n$ . The set of observed data arrives in two parts: a  $p \times n$  data matrix  $X$  and an  $n$ -dimensional vector of group labels. We replace the vector of group labels by a matrix, encoding the group information in an  $n \times g$  matrix  $Q$ , which has the property that every row contains exactly one entry equal to 1, with the remaining entries equal to 0.

As a notational convenience, we write  $\mathbf{1}_k$  for the  $k \times 1$  (column) vector each of whose entries equals one. In many contexts, matrix multiplication using these vectors replaces summation signs with multiple indices. For instance, the property of  $Q$  referred to above reduces to the identity  $Q\mathbf{1}_g = \mathbf{1}_n$ . One can also easily check that  $\mathbf{1}_n^T \mathbf{1}_n = n$ , and that  $\mathbf{1}_n \mathbf{1}_n^T$  is an  $n \times n$  matrix with all entries equal to 1.

Define  $N = Q^T \mathbf{1}_n$ . Then  $N = (n_1, n_2, \dots, n_g)^T$  is a  $g \times 1$  (column) vector whose entries  $n_i$  count the number of samples belonging to the  $i^{\text{th}}$  group. Furthermore, one has  $\mathbf{1}_g^T N = n$ , the total number of samples.

The model underlying our analysis views the columns of the data matrix  $X$  as independent realizations of a random  $p$ -dimensional vector  $\mathbf{X}$ , with the group classifications as an auxiliary categorical covariate. Let  $\mu = E(\mathbf{X})$  be the vector of expected values over the entire population, and let  $\mu_i = E(\mathbf{X} | \text{group } i)$  be the vector of expected values over that part of the population belonging to the  $i^{\text{th}}$  group. Finally, let  $\tau_i = \mu - \mu_i$  be the group effect vectors. Numbering the columns of  $X$  in a way that reflects the group structure, we take the vector model

$$X_{ij} = \mu + \tau_i + \epsilon_{ij} \quad \text{for } i = 1, \dots, g; \quad j = 1, \dots, n_i.$$

Here the residuals  $\epsilon_{ij}$  are themselves column vectors of length  $p$ . The simplest model would take the residuals to be independent and normally distributed for each gene, with mean zero and gene-specific variance. Since our interest at the moment is in a different aspect of the problem, we will not further pursue the role of these distributions.

Our fundamental idea is to combine PCA with ANOVA in order to understand the structure of this data. We start from the ANOVA point of view by trying to decompose the variance into between-group and within-group pieces. We focus on descriptive, sample-oriented statistics and replace  $\mu$  and  $\mu_i$  by their estimates derived from the data matrix  $X$ . In particular, we will abuse notation and use  $\mu$  to denote its estimate,  $(1/n)X\mathbf{1}_n$ . If

we now write  $Z = X - \mu \mathbf{1}_n^T$ , then  $Z$  is the matrix of row-centered data. (In other words, the columns of  $Z$  are given by  $Z_{ij} = X_{ij} - \mu$ .) Now the matrix  $\mathbf{T} = ZZ^T$  is the “total sum of squares” matrix; up to a scalar that we will be more careful about shortly, this is essentially just the covariance matrix. It is the product  $\mathbf{T}$  that we must decompose.

Next, we observe that  $XQ$  is the  $p \times g$  matrix whose columns are the group sums. So, the matrix of group means is given by  $C = XQ/\Delta(N)$ , where we write  $\Delta(N)$  for the diagonal matrix whose diagonal is composed of the entries from the vector  $N$ . In most cases, we will use this identity in the form  $C\Delta(N) = XQ$ . One should also note that  $Q^T Q = \Delta(N)$ , and so the previous identity can also be written as  $CQ^T Q = XQ$ .

Now  $C$  itself is a  $p \times g$  matrix. In order to subtract the group means from the  $p \times n$  data matrix  $X$ , we need to copy the  $i^{\text{th}}$  column of  $C$  into  $n_i$  specific locations. Fortunately, this operation also has a matrix-theoretic description as  $CQ^T$ . Thus, we can compute the residuals in our model as the columns of the matrix  $R = X - CQ^T$ .

Finally, take  $S = Z - R = (X - \mu \mathbf{1}_n^T) - (X - CQ^T) = CQ^T - \mu \mathbf{1}_n^T$ . Because  $Q$  counts each group with the correct weight (based on the number of times it appears in our samples), the product  $\mathbf{B} = SS^T$  is the between-groups sum of squares matrix, just as  $\mathbf{T} = ZZ^T$  is the total sum of squares matrix and  $\mathbf{W} = RR^T$  is the within-group sum of squares matrix.

Although it is a well-known fact that  $\mathbf{T} = \mathbf{B} + \mathbf{W}$  (or equivalently  $ZZ^T = SS^T + RR^T$ ), it is worth reviewing how this computation plays out in the present matrix-theoretic context. First, we have

$$\begin{aligned} ZZ^T &= (X - \mu \mathbf{1}_n^T)(X - \mu \mathbf{1}_n^T)^T \\ &= (X - \mu \mathbf{1}_n^T)(X^T - \mathbf{1}_n \mu^T) \\ &= XX^T - \mu \mathbf{1}_n^T X^T - X \mathbf{1}_n \mu^T + \mu \mathbf{1}_n^T \mathbf{1}_n \mu^T. \end{aligned}$$

Since  $\mathbf{1}_n^T \mathbf{1}_n = n$  and  $X \mathbf{1}_n = n \mu$ , this simplifies to  $ZZ^T = XX^T - n \mu \mu^T$ . In exactly the same fashion, we can use  $Q^T Q = \Delta(N)$  and  $XQ = C\Delta(N)$  to show that  $RR^T = XX^T - C\Delta(N)C^T$ . Combining the same method with the fact that  $CN = n \mu$ , we learn that  $SS^T = C\Delta(N)C^T - n \mu \mu^T$ . The decomposition of the total sum of squares matrix into between-group and within-group parts follows.

We are now prepared to perform principal components analysis. Because the sum of squares matrices have size  $p \times p$  but only have rank at most equal to  $n$ , methods for computing their eigenvalues and eigenvectors tend to be numerically unstable. We will

work instead with a singular value decomposition. Thus, we decompose  $Z$  as a product  $UDV^T$  where

- (i)  $U$  is a  $p \times n$  matrix with orthonormal columns, so that  $U^T U = I_n$ ;
- (ii)  $V$  is an orthogonal  $n \times n$  matrix, so that  $V^T V = I_n = V V^T$ ; and
- (iii)  $D = \Delta(\delta_1, \delta_2, \dots, \delta_n)$  is a diagonal matrix with

$$|\delta_1| \geq |\delta_2| \geq \dots \geq |\delta_n| \geq 0.$$

Using this singular value decomposition, we can write

$$Y = ZZ^T = (UDV^T)(UDV^T)^T = UV D^T V D U^T = U D^2 U^T.$$

Right multiplication by  $U$  yields the identity  $YU = UD^2$ , and we conclude that each column of  $U$  is an eigenvector for  $Y$  with corresponding eigenvalue equal to  $\delta_i^2$ .

Note, however, that  $Y$  is a  $p \times p$  matrix, and we have only found  $n$  eigenvalue-eigenvector pairs with  $n < p$ . Because the rank of  $Y$  is at most  $n$ , this is not really a problem: the remaining eigenvalues all equal 0, and the remaining eigenvectors can be found by choosing a matrix  $W$  whose columns are orthonormal and span the orthogonal complement of the column space of  $U$ .

The actual covariance matrix of  $X$  is (estimated by)  $\Sigma = ZZ^T/(n-1) = Y/(n-1)$ . Dividing a matrix by a constant leaves the eigenvectors unchanged and divides all the eigenvalues by the same constant. So, we write  $\Gamma = (U \ W)$  for the  $p \times p$  matrix obtained by adjoining the columns of  $U$  and  $W$ , and we write  $\Lambda = \Delta(\delta_1^2/(n-1), \dots, \delta_n^2/(n-1), 0, \dots, 0)$ . Then we have

$$\begin{aligned} \Gamma \Lambda \Gamma^T &= (U \ W) \begin{pmatrix} D^2/(n-1) & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} U^T \\ W^T \end{pmatrix} \\ &= (UD^2/(n-1) \ 0) \begin{pmatrix} U^T \\ W^T \end{pmatrix} \\ &= UD^2U^T/(n-1) = ZZ^T/(n-1) = Y/(n-1) = \Sigma. \end{aligned}$$

Thus, using the singular value decomposition, we have found a principal components decomposition of the (total) covariance matrix. The components themselves are the columns of  $U$  (and of  $W$ , but those hardly matter). The amount of variance explained by each component is given by the diagonal terms of  $D^2/(n-1)$ . Finally, since  $X - \mu \mathbf{1}_n^T = Z = UDV^T$ , the coefficients describing the projection of the columns of  $X$  as a linear combination of the columns of  $U$  can be read directly from the matrix  $DV^T$ .

Krzanowski's idea is that the principal components that explain the most variance may not be the ones that best explain the group differences [Krzanowski, 1992]. Letting  $u_i$  denote the  $i^{th}$  principal component, then the eigenvalues  $\lambda_i = \delta_i^2$  of  $\mathbf{T}$  can be rewritten as  $\lambda_i = u_i^T \mathbf{T} u_i$ . The between-group and within-group sums of squares are given, respectively, by  $u_i^T \mathbf{B} u_i$  and  $u_i^T \mathbf{W} u_i$ . Krzanowski suggests ordering the principal components by the canonical variate criterion  $u_i^T \mathbf{B} u_i / u_i^T \mathbf{W} u_i$ . The principal components that maximize the ratio of between-group to within-group variances should do the best job of distinguishing the groups.

After opening up the possibility of reordering the principal components, it is hard to understand why one should be restricted to selecting the principal components themselves. Suppose, for example, that the canonical variates pick out the ninth and fifteenth principal components as the top two choices. Surely some information about the difference between groups lies hidden in the 13 principal components that were skipped. Perhaps we can combine some of those components to get an even better separation of the groups.

Toward this end, we can perform a separate PCA (the group means PCA) on the between-groups sum of squares matrix. To accomplish this goal, we take a singular value decomposition of  $S = U_0 D_0 V_0^T$ , so that the between-groups covariance matrix becomes  $SS^T / (n-1) = U_0 (D_0^2 / (n-1)) U_0^T$ . We can write the group centers in terms of the columns of  $U_0$  by taking  $CQ^T - \mu \mathbf{1}_n^T = S = U_0 D_0 V_0^T$ . This time the coefficients can be extracted from  $D_0 V_0^T$ .

It is particularly instructive to write the original sample vectors in terms of the columns of the between-groups principal components  $U_0$ . To do this, we must find a matrix  $M$  such that  $X - \mu \mathbf{1}_n^T = Z = U_0 M$ . Using the fact that  $U_0$  has orthonormal columns, we find that the best solution is given by

$$M = U_0^T Z.$$

One has to be careful about this computation, because it is not really the case that  $Z = U_0 M = U_0 U_0^T Z$ . The problem, of course, is that  $U_0$  only accounts for part of the full  $p$ -dimensional space; the remainder is accounted for by the columns of the matrix  $W_0$  whose columns span the orthogonal complement of the column space of  $U_0$ . The complete decomposition is given by the true statement that  $Z = U_0 U_0^T Z + W_0 W_0^T Z$ , and so the difference lies completely in the orthogonal complement.

In another sense, an implementation of this strategy must be careful for another reason. The matrix  $S$  is a  $p \times n$  matrix that only contains  $g$  distinct columns. Implementations

of the singular value decomposition are numerically unstable in this situation. Fortunately, the component vectors corresponding to nonzero eigenvalues can be found using a matrix that only contains the  $g$  distinct columns; only the weights on the amount of variance explained need to be adjusted if this alternate computational strategy is adopted.

Finally, we can perform yet another principal components analysis on the within-group differences. We take a singular value decomposition on the matrix of residuals  $R = U_1 D_1 V_1^T$ . The corresponding covariance matrix satisfies  $RR^T/(n-1) = U_1(D_1^2/(n-1))U_1^T$ . In this case, we can directly write  $X - CQ^T = R = U_1 D_1 V_1^T$ , and we can read the coefficients from the columns of  $D_1 V_1^T$ .

## 2: On the ubiquitous existence of spurious structure

In this section, we apply these methods to two sets of simulated data. In both cases, we simulated the values of 1000 variables for 100 samples from 6 different groups. The vector of group labels was the same for the two simulations. In the first simulation, we generated a completely random matrix; more precisely, all entries were simulated as independent samples from a standard normal distribution. In the second simulation, we imposed a great deal of structure: The matrix of population group means was first generated by independent samples from a standard normal distribution, and then random noise was added from normal distributions with mean zero and standard deviation three to create the simulated samples. The simulations and implementation of the methods were carried out in S-Plus; the code is available from our web site<sup>1</sup>.

[\*\* Figure 1 goes approximately here \*\*]

Figure 1 shows the results of plotting the two sets of simulated data against selected principal components. The left-hand panels show the results for the completely random matrix; the right-hand panels show the results for the highly structured matrix. The top panel on each half plots the first two principal components from the full PCA on the original data matrix. The middle panel on each half plots the two principal components ranked highest by the canonical variates criterion, as suggested by Krzanowski. The bottom panel on each half plots the group means and the sample vectors against the first two principal components selected from the group means PCA.

We'll start with the right-hand panels, where the results more closely match our preconceived notions. In this case, the usual first two principal components do a good job

---

<sup>1</sup> <http://www.mdanderson.org/depts/cancergenomics/pcanova.html>

of distinguishing the groups. The highest ranked components by Krzanowski's method are numbers 1, 4, 3, and 2 among the original principal components. Krzanowski's method is only slightly better at separating the groups. Finally, the first two principal components arising from the group means separate the data along different axes. Using the group means components yields more compact groups, and a fairly clean separation between pairs of groups with the exception of groups A and F. These results make sense since the only structure we imposed on the data was the classification into groups.

The results shown in the left-hand panels are more striking. No group structure is reflected in the principal components from the full data matrix. With work, it may be possible to convince oneself that the principal components selected by Krzanowski's method do a slightly better job of finding a group structure. Since the full data matrix was completely random, it is unsurprising that no structure shows in the top two plots. The truly shocking result, however, lies in the plot of the principal components chosen from the matrix of group means. Based solely on this visual evidence, without knowing the origin of the data, most analysts would conclude that there is significant group structure in the data. We, however, know that this apparent structure is spurious. Moreover, we could randomly scramble the group labels (and we have; data not shown) and compute a new pair of principal components that would dramatically demonstrate a different spurious group structure just as strongly.

Before proceeding, it is worth examining why this result is so far out of line with our intuition. Fundamentally, the existence of this kind of spurious structure is inherent in data where the number of variables far exceeds the number of samples. For instance, suppose we selected 100 variables at random from the full set of 1000 variables. In all likelihood, the  $100 \times 100$ -dimensional matrix of data arising from these variables would be nonsingular. Thus, given any desired set of values (such as a set of integers representing group membership), we could find a linear combination of those variables that yields those values. By choosing a disjoint set of 100 variables, we could even do this with two different arbitrary sets of values. Consequently, we can create arbitrary two-dimensional portraits of the data. (In our simulated examples, we can even create arbitrary ten-dimensional portraits of the data.) Of course, these portraits don't help understand the true structure of the data, since they are built by overfitting the noise.

We can draw a number of preliminary conclusions. First, it is stunningly easy to overfit data when the number of variables exceeds the number of observations; spurious

structure abounds. Second, conclusions drawn from the group means PCA cannot be accepted at face value. At the very least, they must be supported by permutation tests [Landgrebe et al., 2002], bootstrap estimates [Kerr and Churchill, 2001], or other measures of statistical significance. Third, there is a compelling need for a quantitative method to decide how much apparent structure should be believed.

### 3: Angular ANOVA

As mentioned previously, there is a third PCA we can perform in addition to the full PCA and the group means PCA: we can subtract out an estimate of the group means and perform PCA on the matrix of residuals. We have performed residual PCA on our two simulated data sets; the results are shown in Figure 2. In the case of the highly structured data matrix, the full PCA and group means PCA are very similar and show a strong group structure, while the residual PCA looks random. In the case of the completely random data matrix, only the group means PCA is structured, and the full PCA and residual PCA show a lack of structure to similar degrees.

[\*\* Figure 2 goes approximately here \*\*]

This observation suggests that we can perform a kind of ANOVA on PCA by finding a quantitative way to decide if the full PCA (on the total sum of squares matrix) is closer to the group means PCA (on the between-groups sum of squares matrix) or to the residual PCA (on the within-groups sum of squares matrix). The remainder of this section is devoted to describing such a measure.

The output of PCA is an ordered set of orthonormal vectors. Each of the three PCAs being considered produces vectors in the same  $p$ -dimensional space of observed variables, and so the sets of vectors can be compared in a single ambient space. Our goal, then, is to quantify the similarity between two ordered sets of vectors in  $p$ -space.

To achieve this goal, we proceed by induction. When each set only contains one vector, we must determine the similarity between two vectors. Either the angle between them or some simple trigonometric function (such as the cosine, which yields the Pearson correlation coefficient) of that angle will provide a reasonable measure of the difference, which we can convert to a similarity measure upon demand. The concrete interpretation of the angle, of course, is that it is the amount by which we must rotate the first vector in order to make it point in the same direction as the second vector. (Since our vectors form orthonormal bases, they all have length one; thus, we only need to perform rotations to make them congruent.)

This interpretation in terms of rotations suggests a way to proceed in order to define the differences between larger sets of vectors. First, by focusing on rotations, it is reasonable to use the angle itself as the measure of difference instead of the cosine of the angle (since addition has a natural interpretation for successive rotations). Next, we look at the two-dimensional case in detail. Let  $A = \{a_1, a_2\}$  and  $B = \{b_1, b_2\}$  be two sets of orthonormal vectors in  $p$ -dimensional space. Assume we have already performed a rotation that aligns vector  $b_1$  with vector  $a_1$ . How much further must we rotate in order to make the plane spanned by  $B$  congruent with the plane spanned by  $A$ ? Of course, there are many different rotations that will work. However, the chosen rotation should not change  $b_1$  and should, in some reasonable sense, be minimal. The best solution seems to be to rotate  $b_2$  until it points in the direction of its projection onto the  $A$ -plane. Thus, the angle through which this second rotation must proceed is given by the angle between  $b_2$  and its projection onto the  $A$ -plane.

There is only one small problem with that idea: it is not symmetric in  $A$  and  $B$ . We would really like a measure of the difference between  $A$  and  $B$  to behave at least somewhat like a distance metric. In this regard, symmetry is a minimal requirement. We circumvent this problem in the usual way: we average the two nonsymmetric differences to produce a symmetric version.

Now given two ordered sets  $A$  and  $B$ , each containing  $k$  orthonormal vectors in  $p$ -dimensional space, we can define a sequence  $d_i$  for  $i = 1, \dots, k$  of measures of their difference as follows. First, write  $\text{proj}_{A_i}(b_i)$  for the projection of the vector  $b_i$  onto the space spanned by  $A_i = \{a_1, \dots, a_i\}$ . For any two vectors  $a$  and  $b$ , let  $\text{angle}(a, b)$  denote the angle between them. Now define

$$d_i = (\text{angle}(a_i, \text{proj}_{B_i}(a_i)) + \text{angle}(b_i, \text{proj}_{A_i}(b_i))) / 2.$$

These values almost give us the measures we want. There are a couple of minor changes that will make interpretation easier. The numbers  $d_i$  range from 0 to  $\pi/2$ . When they are close to zero, then the  $i^{\text{th}}$  vectors in the two sets are close together. When they are close to  $\pi/2$ , then the  $i^{\text{th}}$  vectors are far apart, being nearly orthogonal. Since we are more interested in similarity than in difference, we would prefer it if values close to zero could be interpreted as “not similar”. Since there is nothing particularly meaningful about  $\pi/2$  in such a context, we might as well rescale to make our measurements more like correlation coefficients. Thus, we take similarity measurements in the form  $s_i = 1 - 2d_i/\pi$ , which range from 0 (meaning dissimilar) to 1 (meaning similar).

We make one final adjustment in our measurements. The value  $s_i$  relates primarily to the  $i^{th}$  vector in the set; we are interested in how the entire spaces spanned by the first  $i$  vectors match up. Thus, we should probably take an average measure of similarity by defining  $m_i = (s_1 + \dots + s_i)/i$ . Given our two sets each containing  $k$  orthonormal vectors, then, we obtain the sequence  $(m_1, \dots, m_k)$  as a measurement of their similarity. We refer to the  $m_i$  as “PC correlation” values.

In our context, we want to compare three different sequences of similarity measurements. One sequence compares the full PCA to the group means PCA; one compares the full PCA to the residual PCA; and one compares the group means PCA to the residual PCA. We can plot these three sequences on a common set of axes; our PCANOVA method consists in an interpretation of these plots.

[\*\* Figure 3 goes approximately here \*\*]

We apply this idea to our two simulated data sets; the results are shown in Figure 3. In each panel, we plot the three sequences of similarity measurements. In the panel arising from the completely random data matrix, we see that the similarity measures clearly show that the full PCA is much closer to the residual PCA than to the group means PCA. This result is compatible with what we know about the structure: there is no true group structure in this data. Conversely, applying the similarity measures to the highly structured data matrix shows that the full PCA is much closer to the group means PCA, and is far away from the residual PCA. This result is also fully compatible with the known structure imposed on the simulated data.

#### 4: Empirical distribution of PC correlations

We do not, at present, have a perfect theoretical understanding of the distribution of the PC correlation values described in the previous section. From the single example presented in that section, it appears that the PC correlations have the following properties

1. When there is no true group structure in the data, then the PC correlations between the full PCA and the residual PCA are substantially larger than the PC correlations between the full PCA and the group means PCA.
2. When there is a true group structure in the data, then the PC correlations between the full PCA and the group means PCA are substantially larger than the PC correlations between the full PCA and the residual PC.

Although we would prefer a complete description of the distribution of these statistics,

we are willing to settle for simulations that give a reasonable idea of the distribution. Toward that end, we performed the following experiment. As in the example above, we simulated data sets that measured 1000 variables on 100 samples belonging to 6 different groups. We used the same vector of group labels for all simulations. In total, we simulated 500 pairs of data sets. The first data set in each pair was a random matrix, with all entries drawn from a standard normal distribution. The second data set in each pair was a matrix generated with group structure following the procedure described in the previous section. For each data set, we computed the PC correlations between the full PCA and the residual PCA, and we also computed the PC correlations between the full PCA and the group means PCA.

[\*\* Figure 4 goes approximately here \*\*]

The results are shown in Figure 4. Each panel in this figure displays estimates of two density functions: one for the PC correlations computed on the random matrices and one for the PC correlations computed on the structured matrices. The three panels in the top row display the distributions of the first three PC correlations between the full PCA and the group mean PCA; the three panels in the bottom row display the first three PC correlations between the full PCA and the residual PCA. It is clear in each panel that the statistics have different distributions for the random matrices as compared to the structured matrices. Except for the first PC correlation coefficient on the group means PCA, one can reliably separate the structured matrices from the random matrices using only one PC correlation coefficient.

The extent of the separation, of course, will depend on the number of genes, the number of samples in each group, and the degree of structure present. For practical applications of this method, some kind of simulation or permutation test will be needed to assign a significance level to the PC correlations. Nevertheless, the simulations described here strongly suggest that these statistics can be used to distinguish putative group structure from actual group structure quite effectively.

## 5: Applications to the NCI-60 data set

We now apply PCANOVA to a real world data set, which was collected by performing microarray experiments on 60 cancer cell lines from the National Cancer Institute [Ross et al., 2000]. (The data is available from the original authors web site<sup>2</sup>.) The 60 cell lines used in this study were derived from 9 different kinds of cancer: breast (B), colon (C), leukemia (L), melanoma (M), non-small cell lung (N), ovarian (O), prostate (P), renal (R), and central nervous system (S). Since there were only two samples of prostate cancer, we omit them from our analysis. Hierarchical clustering, as used by the original authors, suggested that the microarray data captured a fair amount of the group structure present in the data. We have also analyzed this data previously [Coombes et al, 2002], applying both hierarchical clustering and PCA using various subsets of genes based on chromosomal location or biological function. Using a semiquantitative method based on the hierarchical cluster diagrams, we found that different subsets of genes recovered different amounts of structure. In this paper, we only apply PCANOVA to the complete data matrix composed of all (well-measured) genes on the microarray. PCANOVA plots for the subsets of genes used in the earliler study are available from our web site<sup>3</sup>. The PCANOVA results confirm—and make more precise—the earlier semiquantitative analysis.

[\*\* Figure 5 goes approximately here \*\*]

The results of the PCANOVA on the complete data set are shown in Figure 5. The upper left panel contains a graph of the first two principal components from the full PCA. This plot appears to show a fair amount of structure. Leukemia samples are grouped on the right near colon cancer samples; ovarian cancers are grouped at the top; melanoma and renal cancer form tight groups near the origin. The upper right panel plots the first two principal components from the group means PCA. This plot accentuates some of the features found in the first plot. The lower left panel plots the first two components from the residual PCA, which strongly identifies two outliers but does retain some structure, including a group of non-small cell lung cancer samples. Finally, the lower right panel plots the PCANOVA similarity measures. We see (as one might expect with real data) that the PCANOVA plots are intermediate between those for the completely unstructured, random data of our first simulated example and the ideal, highly structured group data of our second simulated example. The plot does, however, provide strong evidence that the first

---

<sup>2</sup> <http://genome-www.stanford.edu/nci60>

<sup>3</sup> <http://www.mdanderson.org/depts/cancergenomics/camda.html>

4 principal components of the full PCA reflect true group structure present in the data. Beyond this point, however, the principal components from the full PCA begin to look more like the unstructured residuals. This suggests further that a more detailed discriminant analysis might proceed after a data reduction using the first 4 principal components.

## 6: Conclusions

In this paper, we have looked at data sets where the number of variables far exceeds the number of experimental samples, where the samples have already been classified into groups. Data sets like this have arisen preciously in chemometrics, and they are beginning to occur frequently in gene expression studies using microarrays. In these gene expression studies, the goals include (1) identifying genes that are differentially expressed between certain groups, and (2) developing reliable methods to classify new samples into the existing groups. Both goals typically involve some kind of data reduction; ideally, one wants to reduce the data in a way that reflects the group structure on the data.

We have focused here on developing methods to assess accurately the extent to which the data reflects the known groups. Toward this end, we have looked at the relationships between the full PCA, the group means PCA, and the residual PCA. We saw that the group means PCA can exhibit substantial structure even in the case of completely random data. We introduced a new set of statistics, the principal component correlations, to measure the similarity between these three PCAs. We found, for random data sets, that the PC correlations between the full PCA and group means PCA are small while the PC correlations between the full PCA and residual PCA are large. Conversely, we found, for data sets that were simulated to include substantial group structure, that the PC correlations between the full PCA and group means PCA are large while the PC correlations between the full PCA and residual PCA are small. We conclude that these two sets of PC correlations can tell us if the data truly does reflect the group labels.

We then applied these ideas to the NCI60 microarray data set. The results were consistent with previous analyses of this data set, which found that the microarray data could be used to recover most of the group structure. One advantage of our method, in this context, is that it can be used to give a more quantitative flavor to the word “most” in the previous sentence. In particular, the first four principal component correlations show that the full PCA is closer to the group means PCA than to the residual PCA. This suggests that an appropriate data reduction technique would proceed by using the first four principal components for a linear discriminant analysis, and that a search for differentially

expressed genes should concentrate on those that have large loadings for these components.

We must point out, however, that we can not yet make a completely convincing case that the number of “useful” principal components is given by the point at which the residual PC correlations become as large as the group means PC correlations. At present, this assertion is merely a subjective interpretation of the results, and it will require further research to determine if the PC correlations can be used this way. Nevertheless, we believe that this paper has demonstrated that PC correlations can be used to distinguish spurious group structure from true group structure when the number of variables is large compared with the number of samples.

## 7: Acknowledgements

This work was supported by the Tobacco Settlement Funds as appropriated by the Texas State Legislature, by a generous donation from the Michael and Betty Kadoorie Foundation, and by grant number 003657-0020-2001 from the Texas Advanced Research Program.

## 8: References

- Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Yang L, Marti GE, Moore T, Hudson J Jr., Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Staudt LM, et al. (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**: 503–511.
- Coombes KR, Baggerly KA, Stivers DN, Wang J, Gold D, Sung HG, and Lee SJ (2002) Biology-driven clustering of microarray data: Applications to the NCI60 data set. In: *Methods of Microarray Data Analysis II*, Kluwer Academic Publishers, Boston. In press.
- Kerr MK and Churchill GA (2001) Bootstrapping cluster analysis: Assessing the reliability of conclusions from microarray experiments. *Proc Natl Acad USA* **98**: 8961–8965.
- Krzanowski WJ (1992) Ranking principal components to reflect group structure. *J Chemometrics* **6**: 97–102.
- Kuruvilla FG, Park PJ, and Schreiber SL (2002) Vector algebra in the analysis of genome-wide expression data. *Genome Biology* **3**: research0011.1–0011.11.
- Nguyen DV and Rocke DM (2002) Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics* **18**: 39–50.

Landgrebe J, Wurst W, and Welzl G (2002) Permutation validated principal components analysis of microarray data. *Genome Biology*. **3**: research0019.1–research0019.11.

Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P, Iyer V, Jeffrey SS, Van de Rijn M, Waltham M, Pergamenschikov A, Lee JC, Lashkari D, Myers TG, Weinstein JN, Botstein D, and Brown PO (2000) Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genetics* **24**: 236–244.

## 9: Figure Legends

**Figure 1:** Plots of samples along two principal component axes for two simulated data sets: random (left half) and structured (right half). Top panels use the first two components from a full PCA; middle panels, the two components most highly ranked by the canonical variates criterion; and bottom panels, the first two components from a group means PCA.

**Figure 2:** Plots of samples against the first two principal components from a residual PCA for two simulated data sets: random (left) and structured (right).

**Figure 3:** Plots of the principal component correlation values for two simulated data sets: random (left) and structured (right).

**Figure 4:** Distribution of the first three principal component correlation (PCC) values comparing the full PCA to the group means PCA (top row) and comparing the full PCA to the residual PCA (bottom row). Each distribution is estimated by simulating 500 random and 500 structured data sets.

**Figure 5:** PCANOVA of the NCI60 microarray data set. Samples are plotted against the first two components from the full PCA (top left), group means PCA (top right), and residual PCA (bottom left). Principal component correlations are plotted in the bottom right panel.

Figure 1

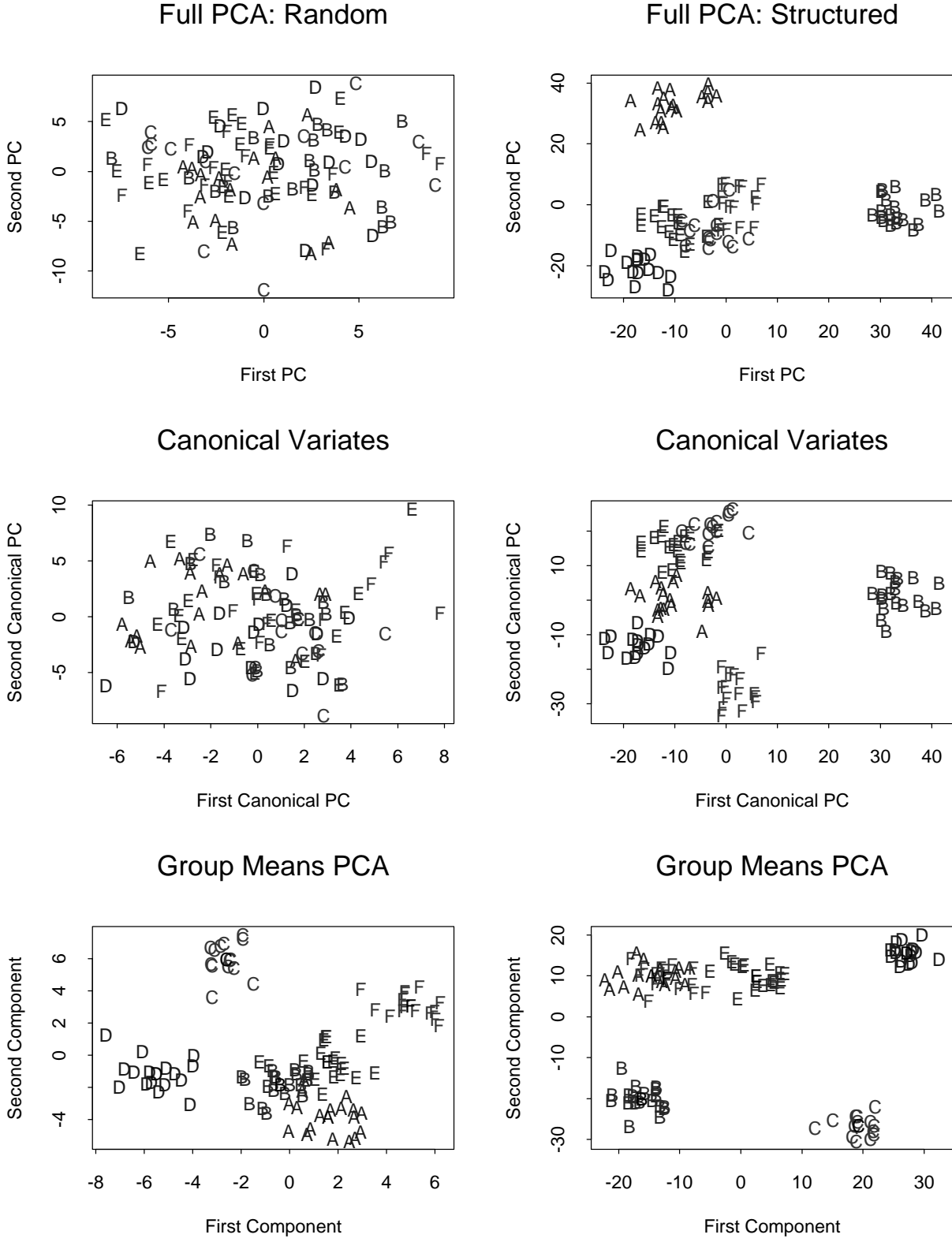


Figure 2

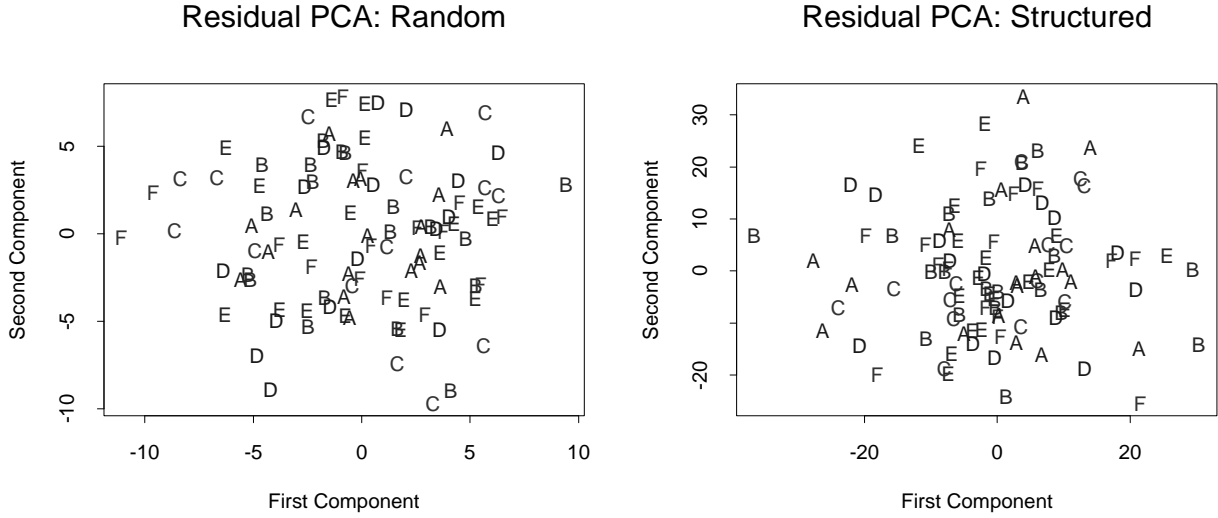


Figure 3

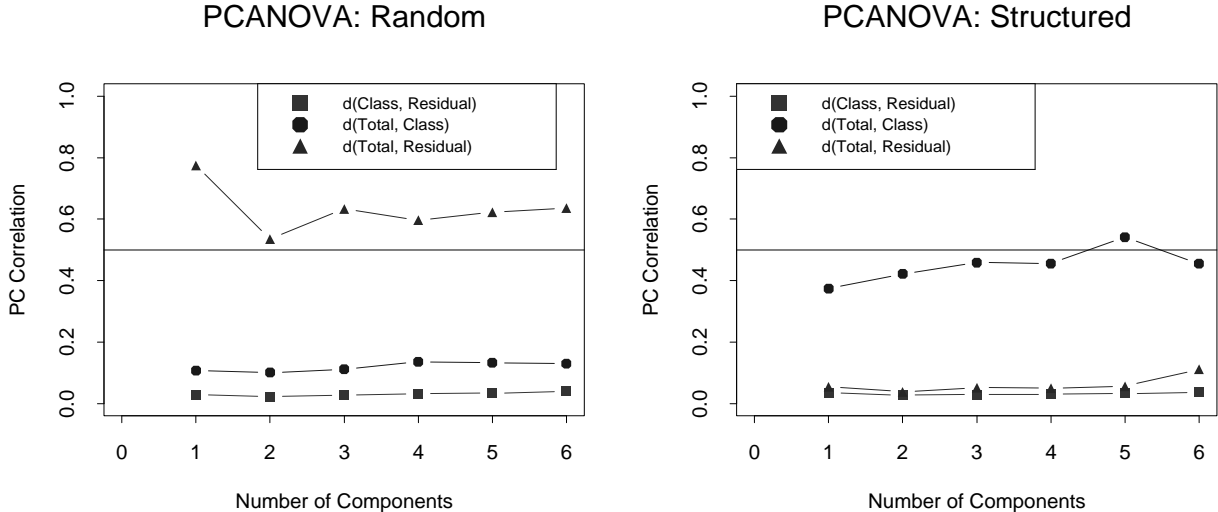


Figure 4

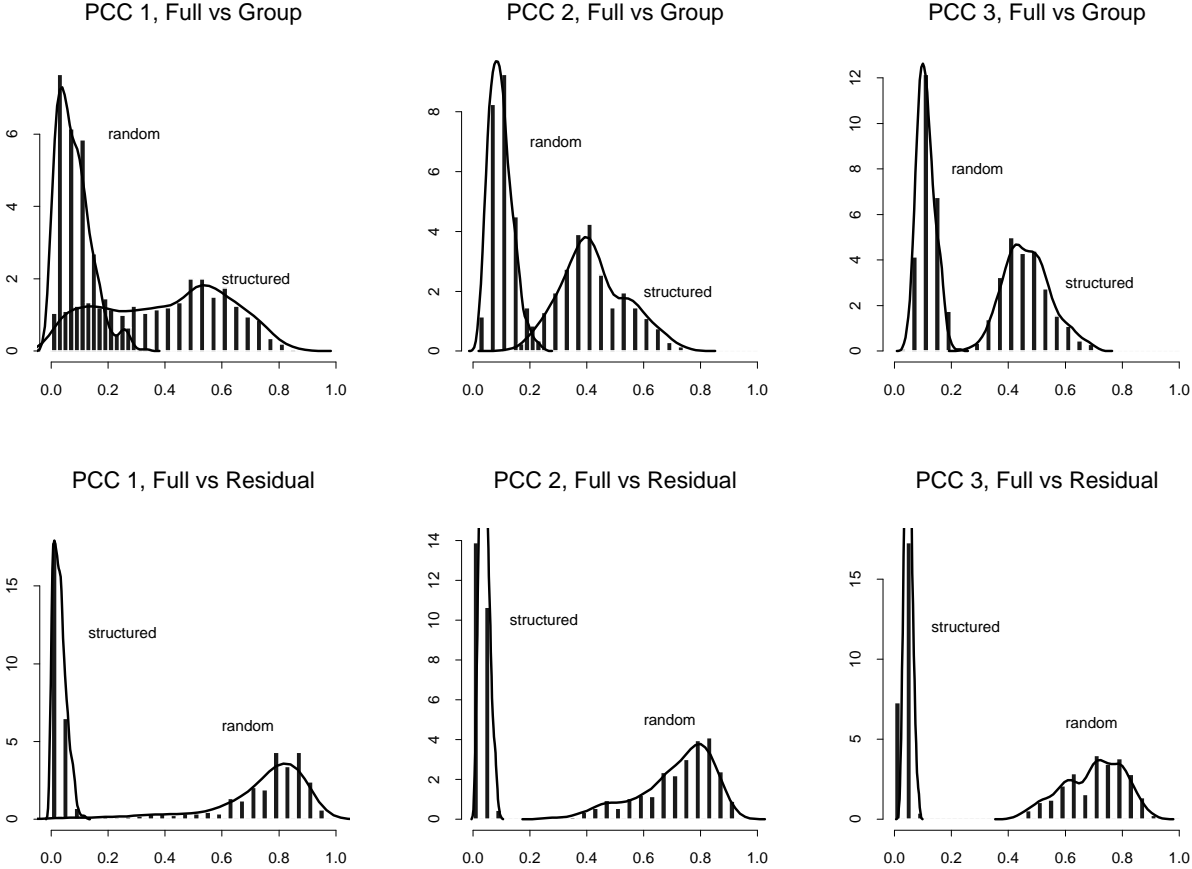


Figure 5

