

Variable Partitioning with Representation for Dimension Reduction

Dennis D. Cox ^{*}
Rice University

E. Neely Atkinson [†]
M. D. Anderson Cancer
Research Center

Iouri Boiko [‡]
M. D. Anderson
Cancer Center

Calum MacAulay [§]
British Columbia
Cancer Research Center

Rebecca R. Richards-Kortum [¶]
University of
Texas at Austin.

Michele Follen ^{||}
M. D. Anderson
Cancer Center

Abstract

A very interpretable method for dimension reduction is to partition the collection of variables into subsets, then from each element of the partition select a single variable to represent that subset. We introduce two general approaches to finding the partition and representatives: correlation cliques and variable clustering. The former is based on finding maximal subsets of variables with a specified lower bound on correlation, and the latter is based on optimizing a general criterion for dimension reduction. This general criterion is based on two maps: a dimension reduction map from the full set of variables to the reduced set, and an approximation map from the reduced set of variables back to the original full set of variables. The objective function is the sum of squared errors of the approximation of the full set by the reduced set. Examples are given including one with a spatial structure which illustrate the methods and their utility for data analysis.

^{*}Address: Department of Statistics, Rice University, 6100 S. Main, Houston, Texas 77005; Email: dcox@rice.edu. Research supported by NSF Grant DMS 9971797.

[†]Address: Department of Biomathematics, M. D. Anderson Cancer Research Center, 1515 Holcombe Boulevard, Houston, TX, 77030; Email: neely@biomath.mdacc.tmc.edu. Research supported by NIH Program Project Award CA82710.

[‡]Address: Biomedical Engineering Center, M. D. Anderson Cancer Center, 1515 Holcombe Boulevard, Houston, TX, 77030; Email: Iouri.Boiko@uth.tmc.edu. Research supported by NIH Program Project Award CA82710.

[§]Address: Department of Cancer Imaging, British Columbia Cancer Research Center, 601 West 10th Avenue, Vancouver, British Columbia, Canada V5Z 1L3; Email: cmacaula@bccancer.bc.ca Research supported by NIH Program Project Award CA82710.

[¶]Address: Biomedical Engineering Program, University of Texas, Austin, TX 78712; Email: kortum@mail.utexas.edu. Research supported by NIH Program Project Award CA82710.

^{||}Address: Department of Gynecologic Oncology, M. D. Anderson Cancer Center, 1515 Holcombe Boulevard, Houston, TX, 77030; Email: mfolle@notes.mdacc.tmc.edu. Research supported by NIH Program Project Award CA82710.

1 Introduction

Statisticians are increasingly coming into contact with data sets having a huge number of variables. Data sets with hundreds of variables are not at all uncommon, and there certainly exist data sets with thousands or more variables. Reducing the number of variables or dimensions is often of interest even if there are only a “few” variables, say 10. Principal Components Analysis (PCA) has held a virtual monopoly on dimension reduction in statistics. The PCA computations are fast and there is an abundance of good software for doing them. It also has nice optimality properties: for a given number of new variables constructed from linear combinations of the original variables, PCA maximizes the percentage of “variance explained.” However, despite the facts that there are much jargon and several graphical presentations associated with the output of PCA, the fact remains that it is difficult to interpret the results in even the best of circumstances. Further, PCA is basically limited to numeric variables, and the mathematical optimality properties may have little connection with the objectives of a data analysis. Also, PCA does not reduce the actual number of variables needed since every variable will almost surely be needed to compute the actual components. In the end, one has reduced the number of dimensions (if one selects a subset of principal components) but not necessarily the complexity – indeed, the complexity is increased in some sense. Given these shortcomings of PCA, it would be nice to have some alternatives.

In this paper we consider an alternative, namely *Variable Clustering with Representatives* (VCR), which consists of partitioning the whole set of variables into disjoint subsets and then selecting one variable from each subset to represent that subset. The basic idea is that within each subset in the partition, all the variables are highly “related”, and so there is not much loss of information in using the representative. This certainly can simplify an exploratory data analysis if the statistician need only examine the representative instead of all the variables in the subset it represents.

Another, related alternative, which is perhaps even more interpretable, is what we call *Correlation Cliques*. For this procedure, the variables are partitioned so that within each subset each variable has a correlation of at least c in absolute value with each remaining variable, where c is a user selected cutoff with $0 \leq c \leq 1$. Unlike VCR, the number of subsets is not specified in advance, but is determined by the data and the cutoff c . Again, once the partitions have been constructed, a representative can be selected for each partition. The name *Correlation Cliques* arises from a connection to graph theory and is explained below.

In Section 2, we discuss some generalities of dimension reduction and describe in some detail the VCR and Correlation Clique methods. Section 3 presents some examples. The first example is an artificial data set with a particular structure and the results show that both VCR and Correlation Cliques quickly identify this structure while PCA does not. The second example involves spectroscopic data which is essentially bivariate functional data, similar to spatial data. There are 704 variables in these data. The methods discussed here do seem to provide some insights. The final example concerns data measured by a Quantitative Pathology (QP) device. There are 113 variables in this data set, which do not have any particular structural relationships as in the previous example. Again, the results of our methods do suggest particular features of interest about the variables which would not appear in a PCA.

2 Description of Methods.

2.1 A General Framework for Dimension Reduction

For now, we will consider a p -dimensional random vector X . We will shortly indicate how the discussion applies to observed multivariate data. A *dimension reduction system* is a pair of maps $\eta : \mathbb{R}^p \rightarrow \mathbb{R}^q$ with $q \leq p$ and $\xi : \mathbb{R}^q \rightarrow \mathbb{R}^p$. Here, $Y = \eta(X)$ is the “reduced” data, and $\hat{X} = \xi(Y)$ is the “approximation” to X . Thus, given an observed value x , we can construct the corresponding dimensionally reduced approximant $y = \eta(x)$, and the actual approximation would be $\hat{x} = \xi(\eta(x))$. In general, one may only be interested in the reduction map η which goes from the higher dimensional space to the lower dimensional space, but it is of interest to consider ξ in order to evaluate the quality of a particular dimension reduction system. In particular, if $d(x_1, x_2)$ is a measure of inaccuracy on \mathbb{R}^p , then we can consider the expected value of the approximation error

$$\lambda(\xi, \eta) = E[d(X, \xi(\eta(X)))] . \quad (1)$$

Thus, for a given $q < p$, we may seek a dimension reduction system which minimizes λ . Even with $q = 1$, one can find maps (η, ξ) for which $x = \xi(\eta(x))$ for all x , so in a purely mathematical sense the problem is trivial. The problem becomes interesting once we place constraints on the maps η and ξ . For instance, if we require that both η and ξ be linear, and if $d(x_1, x_2) = \|x_1 - x_2\|^2$, then the optimal solution is to select the first q principal components (Mardia, Kent, and Bibby, 1979).

For this paper, the dimension reduction map η will be a coordinate projection onto a subset of q of the coordinates of X , i.e.

$$\eta(X_1, \dots, X_p) = (X_{i_1}, \dots, X_{i_q}) . \quad (2)$$

The approximation map will be determined by a variable partitioning. Suppose that the index set $1, 2, \dots, p$ is partitioned into q subsets, say C_1, \dots, C_q . For each j , there is an index $i_j \in C_j$ such that the variable X_{i_j} “represents” the all other variables with indices in C_j in the sense that for

$$\xi_i(X_{i_1}, \dots, X_{i_q}) = h_i(X_{i_j}), \text{ for } i \in C_j . \quad (3)$$

That is, the i 'th component of the approximation map is a function of the single variable X_{i_j} that represents C_j . The functions h_i can be either linear functions (which are estimated from data by simple linear regression) or nonlinear (in which case they can be estimated from data via nonparametric regression). The general idea of splitting the entire set of variables into disjoint subsets and selecting one from each subset to represent all variables within the subset will be called *Variable Partitioning with Representatives* or VPR. We will consider two special methods for constructing a VPR.

There are at least two advantages to VPR as a dimension reduction system. One is that by only considering univariate functions of a single variable for the approximation maps in (3), we can in practice (i.e., when estimating the maps from data) consider a wider class of functions. It is very difficult to fit high dimensional nonparametric regressions, but fitting one dimensional nonparametric regressions is easy. Secondly, we would argue that variable partitioning is highly interpretable. Having each variable associated with a single representative means, for instance, that if we find something interesting in the representative then we can further explore the subset which it represents to see if the other variables exhibit the same behavior.

2.2 Variable Clustering with Representatives (VCR).

This method has already been introduced in Cox, et.al. (1999), where the basic algorithm is presented. It is a procedure which attempts to find the optimal VPR in the sense of maximizing the amount of variance explained for a given number of subsets in the partition. It is referred to as “variable clustering” since the algorithm utilized is very similar to the k -means clustering algorithm (Hartigan, 1975). The algorithm consists of alternating two steps to convergence: for a given set of k representatives, it is easy to find the corresponding best partition by assigning each variable to the representative with which it has the highest correlation. For a given partition, it is easy to find the best set of representatives by searching each variable within a given subset in the partition to see how much total variance within the subset is explained by that variable. We restart the procedure several hundred times with a different randomly selected set of k representatives and select the result which gives the highest overall proportion of variance explained.

2.3 Correlation Cliques.

The goal of this technique is to partition the variables of interest into subsets such that each variable within a given subset has a correlation of at least c in absolute value with every other variable in that subset, where c is a prespecified parameter. The total number of subsets is not specified but rather the correlation threshold c . Once the subsets have been constructed, a representative is selected for each subset. The variable selected as representative is that which explains the greatest proportion of variance of all the variables in the subset.

This selection method can be couched in the terminology of graph theory. Let each variable be represented by a node in a graph. Let those nodes corresponding to variables whose correlation is at least c be joined by edges. Then we wish to partition the graph into subgraphs such that all nodes in each subgraph are connected. A completely connected subgraph is known as a *clique* and the partition we have described is a clique partition, hence the terminology *Correlation Cliques*. Ideally, we wish to find the minimum clique partition, i.e. the clique partition with the fewest number of cliques. This partition leads to the smallest number of representative variables and the greatest reduction in the dimension of the data.

For general graphs, the computation of the minimum clique partition is known to be an NP hard problem (p. 192 of Garey and Johnson, 1979). We therefore take a modified approach. The algorithm begins with a single clique containing one randomly selected variable; a single variable is, of course, a clique since it has correlation of 1 with every member of the clique, i.e. itself. The remaining variables are then examined one by one. If a variable has correlation at least c with every other variable already in the clique, it is entered into the clique. When all variables have been examined, the clique so constructed is added to the list of cliques in the partition and the variables in the clique removed from further consideration. The process then repeats with a new clique initially containing a single variable. The process continues until all variables have been placed into cliques. This process can produce different results depending on the order in which variables are examined; the algorithm may be run on permutations of the data in order to partially overcome this difficulty. Even if all permutations of the variables are considered, this algorithm may not find the minimum clique partition.

An alternative approach is to apply hierarchical clustering, such as is implemented in the Splus/R function `hclust`. See Venables and Ripley (1999) for more discussion of this method. In this case, we consider the data points to be clustered

to be the variables and the distance between two data points to be $1 - |\rho(x_1, x_2)|$. The algorithm begins by placing each data point into a separate cluster. Then, at each iteration the two clusters which are closest to each other are merged into a single cluster. For our purposes, we take the distance between two clusters as the maximum of all possible pairwise distances between their members; this is the “complete” or “compact” method which is the default selection for `hclust`. We can trim the tree at any level of distance to select the appropriate cliques for a given c . This approach is not effected by the ordering of the variables. Further, if `hclust` is used to construct the entire cluster tree, then repeated calls the `cutree` can be used to construct the cliques for a variety of values of c . In the greedy algorithm described in the previous paragraph, if the value of c is changed, the cliques must be recomputed *de novo*. The partition computed using this method may not be optimal.

The structure of the correlation matrix places restrictions on the graphs which can arise from representing the matrix. We are currently investigating whether these restriction can be exploited to accelerate the algorithm or perhaps even guarantee the selection of a minimum partition.

2.4 Discussion of Variance Explained.

These methods require only a “variance explained matrix” in order to operate. The Variance Explained for Y by X from the function class \mathcal{H} is

$$VE(Y|X) = \inf_{h \in \mathcal{H}} E[(Y - h(X))^2]. \quad (4)$$

For the case where h is restricted to be a linear function of the form

$$h(x) = a + bx,$$

we have

$$VE_l(Y|X) = \rho^2(X, Y)Var(Y),$$

where $\rho(X, Y)$ is the correlation between X and Y . If our variables are standardized, then we may simply use $\rho^2(X, Y)$. We refer to these as linear correlations.

If we allow arbitrary measurable functions for h above, then the Variance Explained is

$$VE_{nl}(Y|X) = E[(Y - E[Y|X])^2].$$

The subscript *nl* refers to “nonlinear.” We may define a nonlinear correlation as

$$\rho_{nl}(X, Y) = [VE_{nl}(Y|X)/Var(Y)]^{1/2}. \quad (5)$$

Note that this correlation is asymmetric and always positive.

To estimate $\rho_{nl}(X, Y)$ from data, one replaces the conditional expectation with a nonparametric regression estimate and the expectations with sample averages. The result is

$$\hat{\rho}^2(X, Y) = \frac{1}{(n-1)s_y^2} \sum_i [y_i - \hat{h}(x_i)]^2,$$

where (x_i, y_i) denotes a value in the sample from the joint distribution of (X, Y) , n is the sample size, s_y^2 is the sample variance of the y_i ’, and \hat{h} is the estimated nonparametric regression function. Our choice for nonparametric regression estimate is a generalized cross-validated smoothing spline, but other choices are possible. See e.g. Venables and Ripley (1999) for further discussion of nonparametric regression.

For VCR, one works with the matrix with (i, j) entry $VE(X_i, X_j)$. The Correlation Clique method may be with Pearson's, Spearman's, or other correlations. If the correlation used is not symmetric in its arguments, the minimum of $-\text{cor}(x,y)$ — and $-\text{cor}(y,x)$ — is used.

3 Examples.

In this section, we consider both VCR and Correlation Cliques as applied to some examples. The first example is with artificial data which has a certain correlation structure. The second and third examples are based on actual data.

3.1 Artificial Example.

We generated an artificial data set to test the behaviours of the algorithms in a case in which we understood the covariance structure of the data. The generated data set consists of 50 cases, each with 16 variables. We first constructed four 50 by 4 matrices with elements uniformly distributed in $(0,1)$. We then computed the singular value decomposition of each test matrix, reset the singular values to be 1.0, 0.01, 0.0001, and 0.000001 in each decomposition, and reconstructed the data matrices using the new singular values. This procedure produced 4 data sets, with each data set having a high degree of correlation for all variables within that data set and a low degree of correlation with all variables in the remaining three data sets. We then merged the 4 sets to produce our test data. The test data have a correlation matrix which is block structured; there are 4 blocks of variables such that all variables in each block are highly correlated and variables in different blocks have low correlations.

In Figure 1, we show a plot of the Proportion Variance Explained versus the number of variables for the three general methods. We see that all methods essentially identify the data set as four dimensional. We did not include the `hclust` version of correlation cliques as it adds nothing new for this example. Clearly, there is relatively little difference between the methods, and they stand in their correct mathematical ordering based on their respective optimality properties.

In Figure 2 is shown the dendrogram from the hierarchical clustering with the compact linkage. This clearly shows the structure in the data – we see that the 4 groups of variables that we constructed to be highly correlated are merged at a very low height, and it is some height above this before any of the 4 clusters are joined. There is no obvious way PCA could have helped to discover this structure.

3.2 Fluorescence Excitation Emission Matrix Data.

The next data set we consider is the Excitation Emission Matrix (EEM) produced by a fluorescence spectroscopy measurement of the cervical tissue as described in Cox, et.al. (2001). These data have essentially a two dimensional spatial structure. One spatial index variable is the excitation wavelength of the light going into the tissue. The other is the emission wavelength of the fluoresced light coming back out of the tissue. The measurement variable is the logarithm of the intensity of the emitted light, normalized to an estimated peak intensity. As the excitation wavelength is always shorter than the emission wavelength by the nature of fluorescence, the matrix is essentially lower triangular. In fact, measurements were made at a subset of the possible emission wavelengths, so the EEM is even sparser. The data set considered here consists of 848 EEMs from measurements on the cervixes of 10

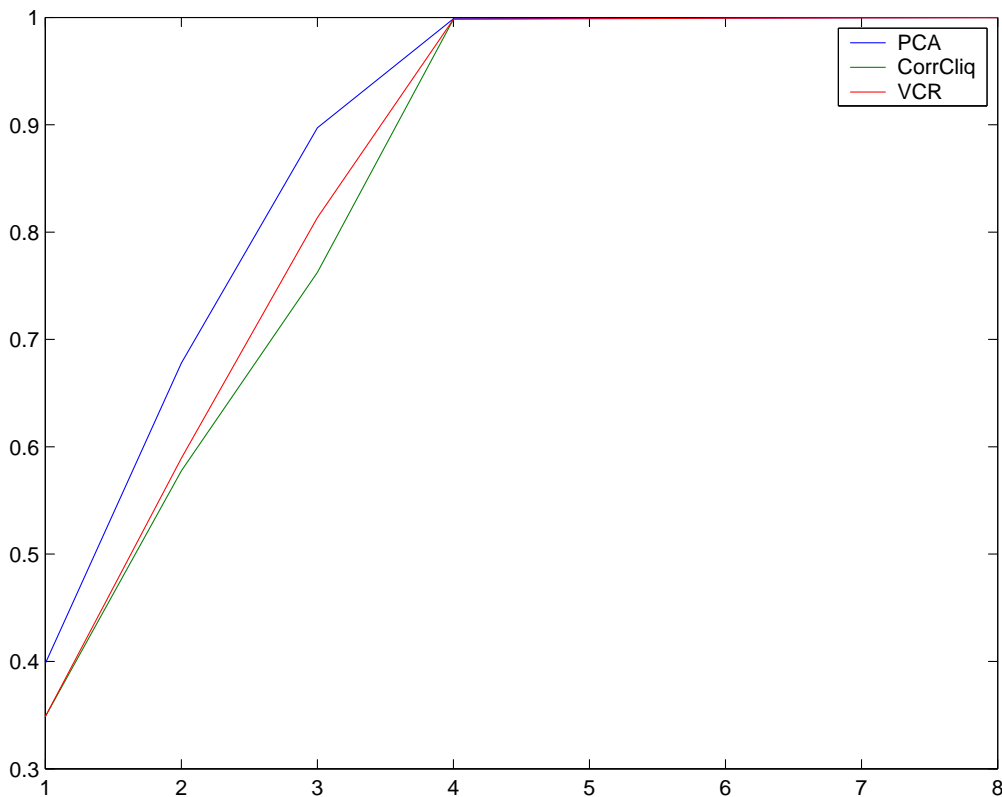


Figure 1: Proportion of Variance Explained as a function of number of dimensions retained for PCA, VCR, and Correlation Cliques applied to the synthetic data.

subjects who were part of a study to determine the effects of the menstrual cycle on the fluorescence spectroscopy of the cervix. The women were measured at 3 cervical sites for approximately 30 days. Some EEMs had problems and were discarded from the analysis, which explains the sample size of 848. There were a total of 703 excitation-emission wavelength combinations where measurements were made. Further discussion may be found in Cox, et. al. (2002). Figure 3 shows the mean of the 848 EEMs. The peak intensity that was used to normalize the data was also included as a variable, making a total of 704 variables. It is however not depicted in the figures that follow.

Plot 4 shows the Proportion of Variance Explained for each of the different methods as a function of the number of variables retained. We see considerable differences now. PCA is far superior to any of the Variable Partitioning methods. Also, the VCR is considerably better than the Correlation Clique methods. There is no real difference between our greedy algorithm and the hierarchical clustering result in this example. All of the variable partitioning methods are essentially equivalent

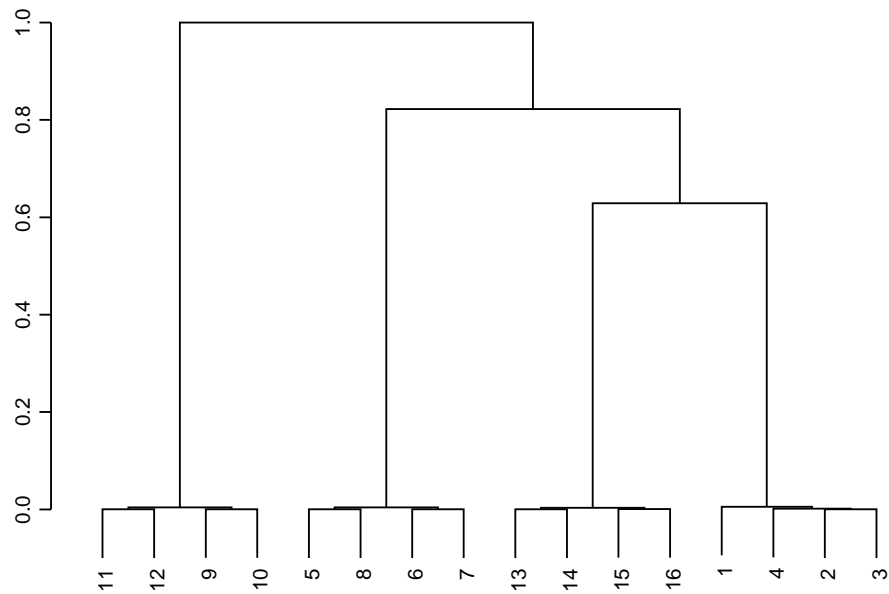


Figure 2: Dendrogram from hierarchical clustering applied to synthetic data set.

from about dimension $q = 200$ onward.

Figure 5 shows a plot of the clusters when the VCR algorithm was applied to these data with 30 clusters. This gives 92.56% of the variance explained. To produce this plot, a value between 1 and 30 was assigned to all the variables with the same value for those variables within the same cluster. The values of the clusters were randomly varied until a plot was produced that separated contiguous clusters well. Several features are evident in this plot. For emission wavelengths between 450 and 600, the clusters run primarily horizontally, lining up along the excitation wavelengths. This may be an artifact of the measurement process: the measurements are made by scanning through each excitation wavelength individually and measuring the emission spectrum that is produced. Thus, there may be effects that are consistent for a given excitation wavelength, or there may be time varying effects that become confounded with the excitation wavelength. The irregular clusters about emission wavelength 600 may be due to the fact that the measurements are inherently noisier in this region as the tissue is more transparent to such wavelengths and so light leaks in from sources unrelated to the fluorescence process. An interesting feature for

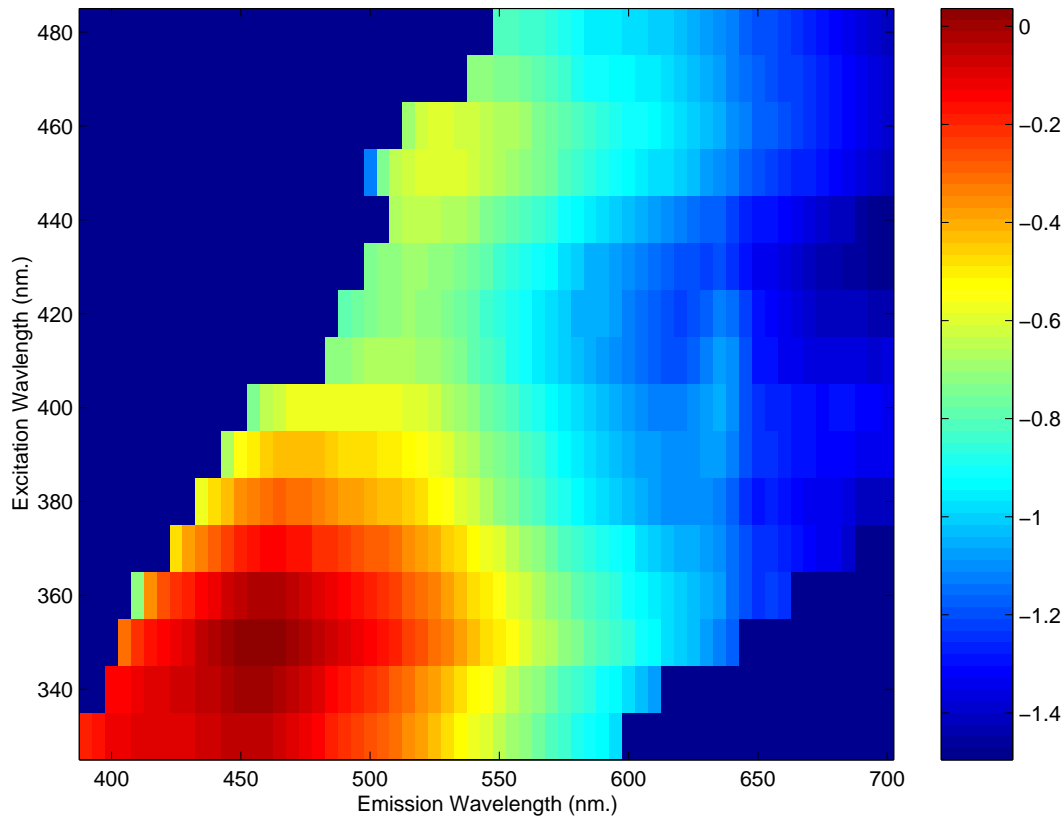


Figure 3: Mean EEM for menstrual cycle study. The vertical bar on the right depicts the scale corresponding to the spectral color map. Note the the peak is at about emission 460 and excitation 350.

which we have no explanation are the appearance of two large clusters for emission wavelength less than 450.

Figures 6 and 7 show the Correlation Cliques for the EEM data with the threshold $c = 0.7$. This value was chosen as it gives 30 cliques and is thus comparable to the VCR results in Figure 5. Two plots are given to help resolve some of the ambiguities of clique membership. The general features are quite different from corresponding plot from the VCR method. One sees rather large blocks of high correlation above excitation wavelength 380, and cliques are rather jumbled near the peak of the average intensity and also near the edges. The latter is not surprising as the edges are more susceptible to noise in many cases.

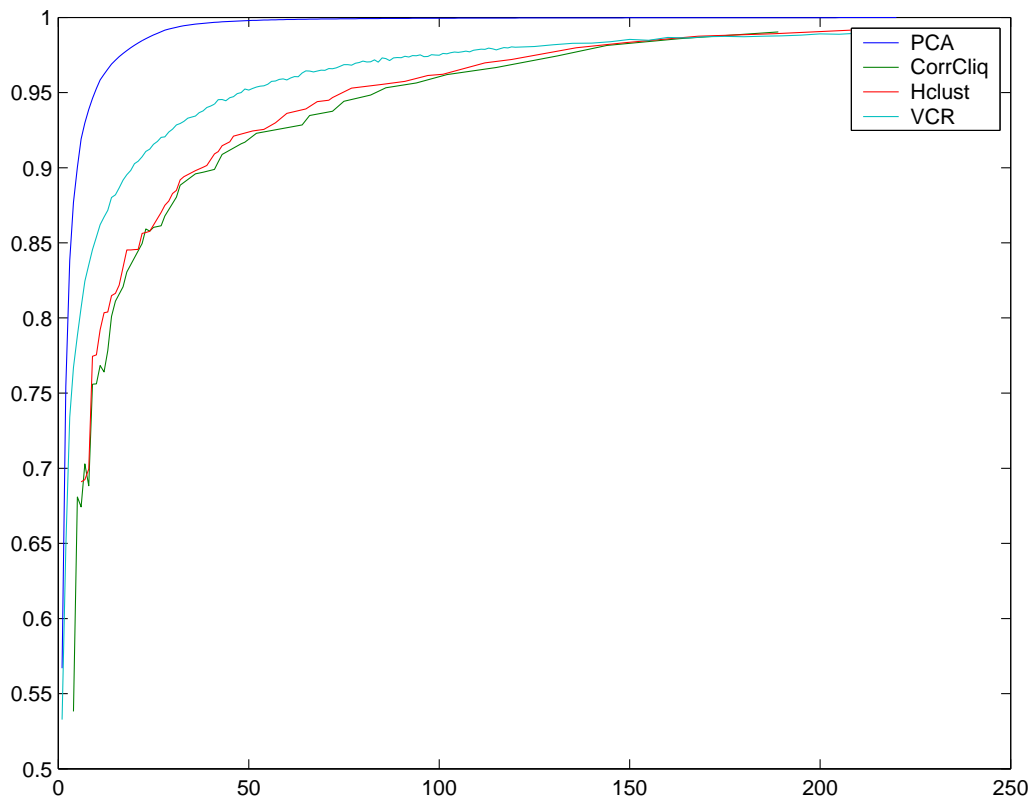


Figure 4: Proportion of Variance Explained as a function of number of dimensions retained for PCA, VCR, and Correlation Cliques with both the original greedy algorithm and the `hclust` algorithm applied to the EEM data.

3.3 Quantitative Pathology Data.

There are a number of systems in commercial use which extract features from measurements of cellular nuclei for quantification of pathological diagnoses. We will use the features measured on Cyto-SavantTM computer-assisted image analysis system (Cancer Imaging, Vancouver, BC, Canada). This system produces 112 features, and the patient number was kept as a 113th feature. The data set we consider consists of 31 cervical tissue samples from different patients, and between 105 and 320 cells were measured on each patient with a mean of 225.1. There are a total of 6978 observations. There are 112 measured variables and a patient number. We retained the patient number as a variable to see what might be correlated with it. As all of the variables are measured on different scales, etc., we standardized them to have mean 0 and variance 1. Thus, the total variance in all analyses is 113. For these data, it makes sense to consider the nonlinear correlation as well. Many of the variables are based on variants of the same basic notion, such as capturing texture

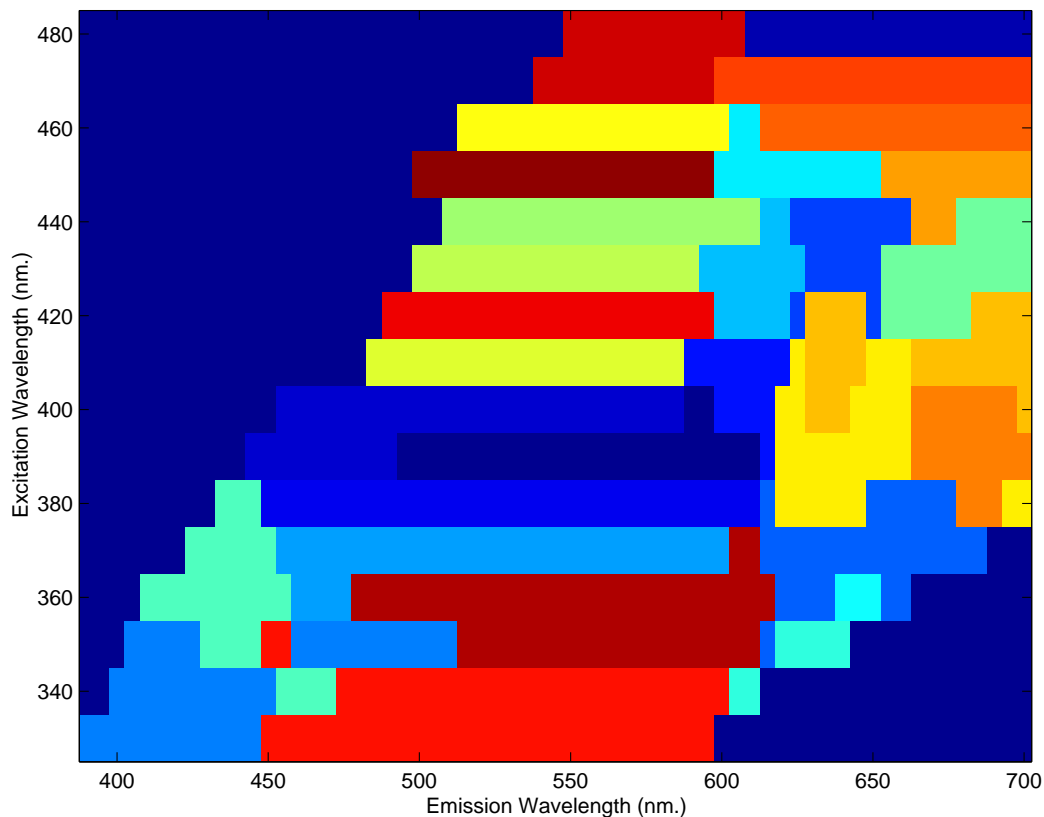


Figure 5: Color coded plot to show the clusters from VCR applied to the EEM data.

of the nuclear material, so we may expect to see high nonlinear correlation.

The results of Proportion of Variance Explained for each of the methods is shown in Figure 8. We see that for small numbers of dimensions, up to about $q = 5$, VCR with the nonlinear correlation is a serious competitor to PCA. At about $q = 50$, the hierarchical clustering version of the correlation cliques with nonlinear correlation becomes the dominant procedure among the variable partitioning methods. We surmise that it is doing better than the computed VCR with the nonlinear correlations since we did not restart the VCR enough times. To achieve 80% variance explained, PCA requires 28 variables while the best of the variable partitioning methods (VCR with nonlinear correlation) requires 51 variables.

4 Discussion and Conclusions.

The ideas on variable partitioning presented here are not particularly new. There have been many other researchers who have had similar or even identical notions.

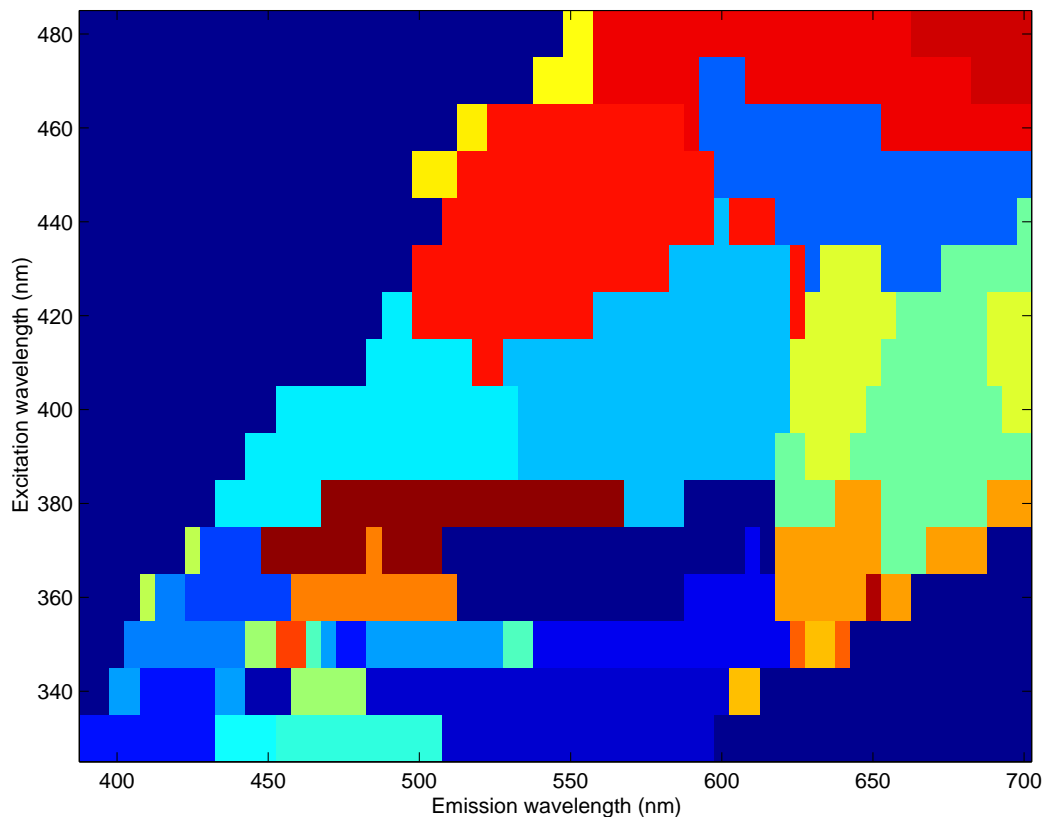


Figure 6: Color coded plot to show the correlation cliques applied to the EEM data.

The Correlation Cliques notion appears in the Masters Thesis of Eijssen. We expect there are many other citations we have missed, and invite the interested reader to let us know of other work on these methods.

Perhaps the most interesting result here is the surprising performance of the hierarchical clustering as an heuristic for finding correlation cliques. In all three examples, it did basically as well as our original greedy algorithm in terms of the PVE with representatives selected. As the `hclust` algorithm is very fast and produces results which are very interpretable, we recommend it for further use.

References

Cox, D.D., Follen, M., Pandey, D., Atkinson, E.N., Poulin, N., MacAulay, C., and Richards-Kortum, R. (1999) "Fluorescence Spectroscopy, Quantitative Pathology, and Classification of Tissue," *Computing Science and Statistics*, Vol. 31, pp. 110-117.

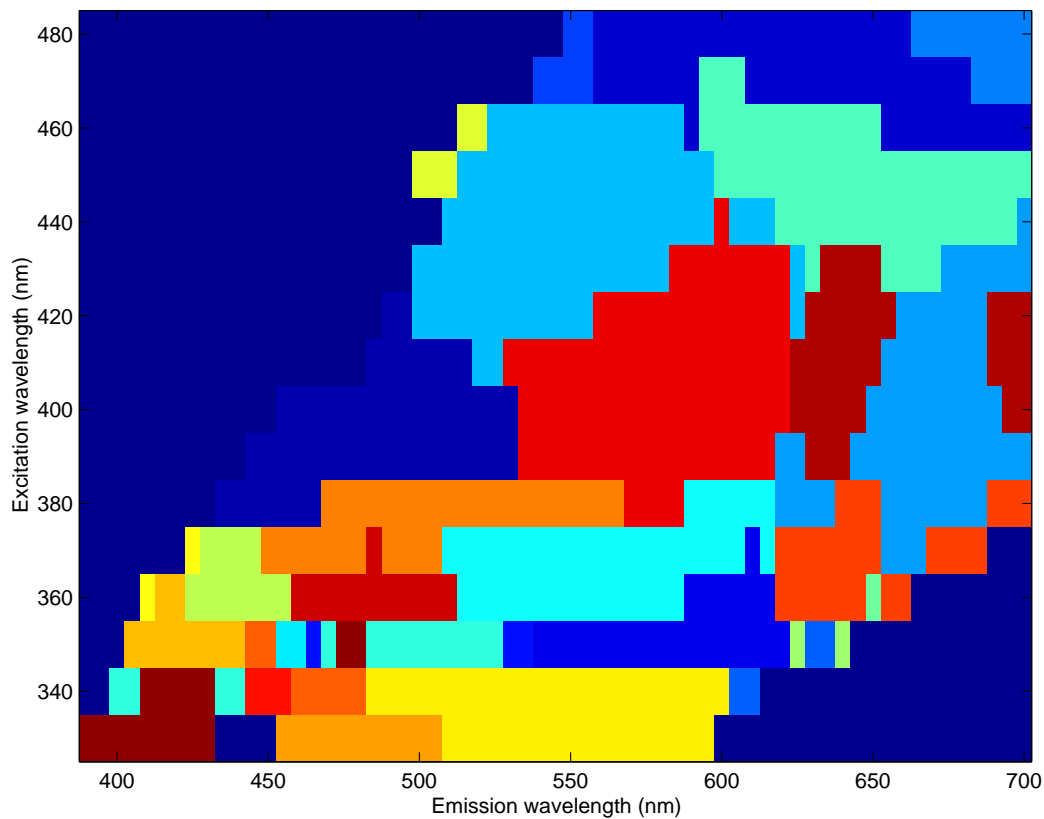


Figure 7: Second color coded plot to show the correlation cliques applied to the EEM data.

Cox, D.D., Chang, S.K., Dawood, M.Y., Staerkel, G., Utzinger, U., Richards-Kortum, R., and Follen, M. (2002) "Detecting a signal from the menstrual cycle in fluorescence spectroscopy of the cervix," submitted to *Journal of Applied Spectroscopy*; revision requested.

Doudkine A., MacAulay C., Poulin N., Palcic B., (1995) "Nuclear texture measurements in image cytometry," *Pathologica*, Vol. 87, pp. 286-299.

Eijssen, L., (2000) *Cluster Analysis of Microarray Gene Expression Data*, unpublished M.S. thesis, Maastricht University, Netherlands.

Garey, M. R., and Johnson, D. S. (1979) *Computers and Intractability*, W. H. Freeman and Co., New York.

Hartigan, J.A. (1975), *Clustering Algorithms*, New York: Wiley.

Mardia, K.V., Kent, J.T., and Bibby, J.M. (1979) *Multivariate Analysis*, Academic Press.

Venables, W.N., and Ripley, B.D. (1999) *Modern Applied Statistics with S-PLUS*,

Third Edition, Springer Verlag, New York.

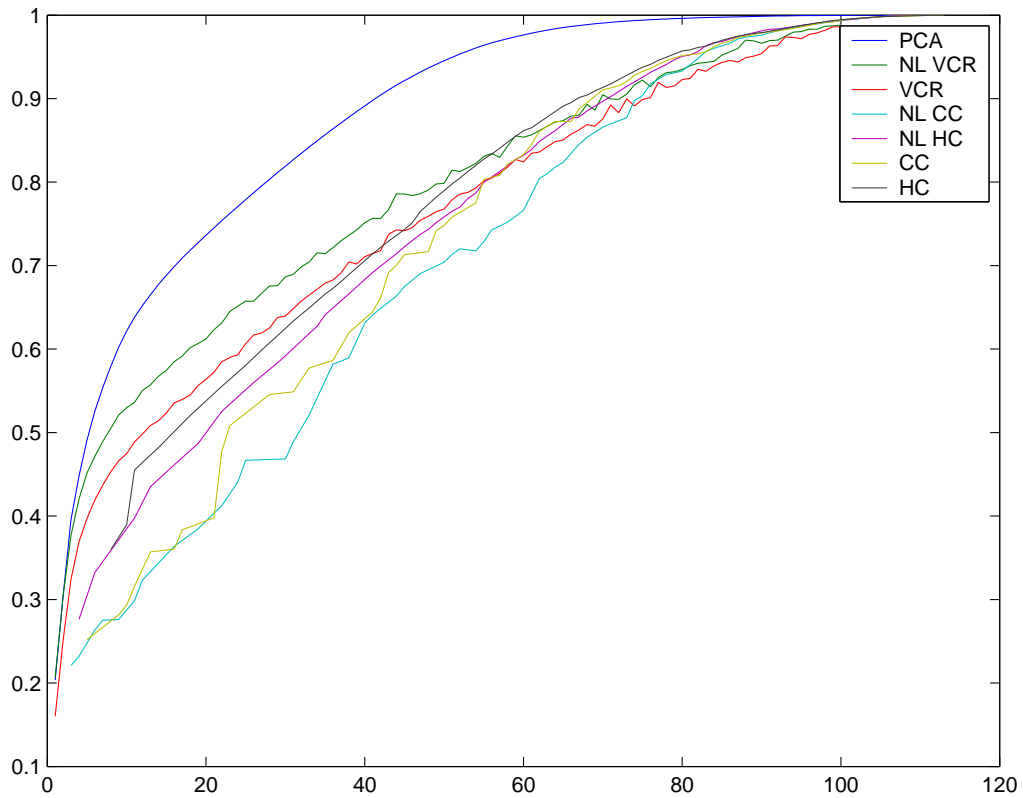


Figure 8: Proportion of Variance Explained as a function of number of dimensions retained for PCA, VCR, and Correlation Cliques (CC) with both the original greedy algorithm and the hierarchical clustering algorithm (HC) applied to the Quantitative Pathology data. For the VCR and both versions of the Correlation Cliques, we also consider the nonlinear correlation, designated “NL” in the legend.