Some remarks on the R2 for clustering

Article *in* Statistical Analysis and Data Mining · July 2017

DOI: 10.1002/sam.11378

CITATIONS
5

READS
2,426

2 authors:

Nicola Loperfido
University of Urbino
73 PUBLICATIONS **1,521** CITATIONS
SEE PROFILE

Thaddeus Tarpey
Wright State University
91 PUBLICATIONS **1,623** CITATIONS
SEE PROFILE

Some Remarks on the R^2 for Clustering

Nicola Loperfido
Dipartimento di Economia, Società e Politica,
Università degli Studi di Urbino "Carlo Bo"
and
Thaddeus Tarpey *
Department of Mathematics and Statistics, Wright State University
email: thaddeus.tarpey@wright.edu

January 30, 2018

Abstract

A common descriptive statistic in cluster analysis is the R^2 that measures the overall proportion of variance explained by the cluster means. This note highlights properties of the R^2 for clustering. In particular, we show that generally the R^2 can be artificially inflated by linearly transforming the data by "stretching" and by projecting. Also, the R^2 for clustering will often be a poor measure of clustering quality in high-dimensional settings. We also investigate the R^2 for clustering for misspecified models. Several simulation illustrations are provided highlighting weaknesses in the clustering R^2, especially in high-dimensional settings. A functional data example is given showing how that R^2 for clustering can vary dramatically depending on how the curves are estimated.

Keywords: high dimensional data, k-means clustering, multiple regression, skewness.

*We would like to thank the editors and the reviewers for very helpful suggestions that have greatly improved this paper.

1 Introduction

The quality of statistical models is often measured by the proportion of variability explained by the model while maintaining low complexity. One of the most popular descriptive statistics for measuring this quality is often referred to as the "R^2". A model's merit is often measured by an R^2-type statistic, with larger values being preferred. Perhaps the two most common examples of this are the coefficient of determination in multiple regression and the proportion of variation explained by a subset of principal components. This paper examines the R^2 statistic in the framework of cluster analysis, which is one of the most popular methods of unsupervised learning. An appeal of the clustering R^2 is that it gives a relative measure of how close together points are within the same cluster (i.e., the within sum-of-squares) relative to the distance between points and the overall mean; in this sense the R^2 also represents a proportion of total variance explained by a clustering partition.

The R^2 for clustering is typically defined as the between cluster (group) deviance divided by total deviance, see Section 2. Optimal clustering has been associated with maximizing the R^2 for clustering which implies the maximization of the between-cluster variability, or, equivalently, the minimization of the within cluster variability. The latter criterion was first proposed on intuitive grounds [1, 2] and later interpreted within the framework of likelihood-based inference [3]. The R^2 for clustering has been used to determine the number of groups in the data [4], and [5] used it to show the connection between cluster analysis and linear regression, which was further investigated using self-consistency [6]. To date, minimization of the within-cluster deviance and hence maximization of the clustering R^2 is the most popular clustering measure [7].

There are numerous methods for clustering a set of data (or the support of a probability distribution). The clustering of interest for this paper refers to partitioning of the data into k non-overlapping groups (i.e., clusters). Often the criterion of clustering is to find a partition so that the clusters defining the partition are as homogeneous as possible. One common goal of clustering is to help discover hypothesized, but latent subpopulations. Even if distinct subpopulations do not exist, clustering results can be useful for descriptive purposes or in cases where boundaries need to be drawn in order to make a decision (e.g., for medical treatment decisions).

The R^2 statistic is a popular measure in regression and in principal component analysis (PCA), but the R^2 statistic has well-known shortcomings [e.g., 8]. In particular, R^2 can be inflated by simply making the model more complex (adding more predictors in regression and extracting more principal components in PCA). The same shortcomings are also shared by the R^2 for clustering — increasing k, the number of clusters, will increase the clustering R^2 which makes the R^2 problematic for deciding how many clusters to extract from a data set. Often data are linearly transformed prior to performing statistical analyses (e.g., changing the scale of certain variables) including cluster analysis [9, 10]. These types of transformations can also make the clustering R^2 increase when in fact the quality of the clustering deteriorates. One of the main points of this paper is to provide a cautionary note that transformations that increase the clustering R^2 do not necessarily improve the clustering results.

In this paper, we focus on clustering partitioning (into k clusters) that results in a set of self-consistent cluster means and self-consistent partitions. The notion of *self-consistency* [6], discussed further in Section 2, provides a unifying framework for regression, principal component analysis and clustering. Applying the well-known k-means algorithm [e.g., 11] to a data set will always produce cluster means that are self-consistent points for the empirical distribution. Also, closely related to the k-means algorithm is the hierarchical Ward's algorithm [12]. Like the k-means algorithm, Ward's algorithm is also based on a sum-of-squares criterion to minimize the within cluster variability, as noted by [13].

This paper is organized as follows. The underlying framework for this paper and the definition of the clustering R^2 is given in Section 2. Results and illustrations of the clustering R^2 under linear transformations that "stretch" a distribution in a given direction are given in Section 3 and this is followed with results for projections in Section 4 which can be regarded as an extreme form of stretching. Data projections are used in high-dimensional settings and Section 5 gives results on the clustering R^2 when the dimension of the distribution is high. An examination of the clustering R^2 for misspecified models is provided in Section 6. Section 7 describes an application that provides an example that ties together theoretical results from previous sections. The paper is concluded with a discussion in Section 8 and a description of future research problems. Proofs of all results are collected

in the Appendix.

2 Defining the R^2 for Clustering

The framework in this paper is to work with some underlying p-dimensional distribution $\mathcal{F}$ with finite second moments. In some cases $\mathcal{F}$ will refer to the empirical distribution generated by a data set, but in other cases $\mathcal{F}$ may correspond to some underlying continuous, multivariate distribution. Without loss of generality, we assume that the mean of $\mathcal{F}$ is zero (i.e., the distribution has been centered). Let $\boldsymbol{y}$ be a random vector with distribution $\mathcal{F}$ and let $\boldsymbol{\Psi}$ denote the covariance matrix of $\boldsymbol{y}$. The trace of $\boldsymbol{\Psi}$, $tr(\boldsymbol{\Psi})$, will be taken as the measure of the total variance.

Let $D_1, \ldots, D_k$ denote a partition or clustering of the support of $\mathcal{F}$. Denote the cluster means by

$$\boldsymbol{\mu}_j = E[\boldsymbol{y}|\boldsymbol{y} \in D_j]. \tag{1}$$

Most clustering algorithms tend to group observations that are "close" to one another and consequently the resulting clusters are often Voronoi partitions where each set in the partition consists of all points that are closest to a given cluster mean relative to the other cluster means, [14, page 8]. Although there exist clustering methods that can produce non-Voronoi partitions, such as finite mixture models [15] and kernel k-means [e.g., 16], the focus of this paper is on clustering procedures that form Voronoi partitions.

We shall use the Euclidean norm, denoted $\|\cdot\|$, on $\Re^p$ as the measure of closeness. Thus, the partitions under consideration here have the property that $\boldsymbol{y} \in D_j$ if $\|\boldsymbol{y} - \boldsymbol{\mu}_j\| < \|\boldsymbol{y} - \boldsymbol{\mu}_j'\|$, $j \neq j'$. Define a discrete random vector $\boldsymbol{u}$ as

$$\boldsymbol{u} = \boldsymbol{\mu}_j \text{ if } \boldsymbol{y} \in D_j. \tag{2}$$

The random vector $\boldsymbol{u}$ is said to be self-consistent for $\boldsymbol{y}$ if $E[\boldsymbol{y}|\boldsymbol{u}] = \boldsymbol{u}$ [6]. In these situations, the set of cluster means $\boldsymbol{\mu}_j, j = 1, \ldots, k$, is called a set of k *self-consistent points* [17]. Using the Euclidean norm guarantees that the sets D_j are convex sets [18].

A distribution can have more than one set of k self-consistent points; the set of k points

that minimizes the total within cluster variance

$$\sum_{j=1}^{k} \int_{D_j} \|\boldsymbol{y} - \boldsymbol{\mu}_j\|^2 d\mathcal{F}(\boldsymbol{y}) \tag{3}$$

are called the k **principal points** of the distribution [19]. In the signal processing literature, (3) is called the k-quantization error for $\mathcal{F}$. A set of k principal points must be self-consistent points [17].

Letting $\pi_j = P(\boldsymbol{y} \in D_j)$, the total variation can be written as

$$\begin{aligned} tr(\boldsymbol{\Psi}) &= \int \boldsymbol{y}'\boldsymbol{y} d\mathcal{F}(\boldsymbol{y}) \\ &= \sum_{j=1}^{k} \int_{D_j} (\boldsymbol{y} - \boldsymbol{\mu}_j + \boldsymbol{\mu}_j)'(\boldsymbol{y} - \boldsymbol{\mu}_j + \boldsymbol{\mu}_j) d\mathcal{F}(\boldsymbol{y}) \\ &= \sum_{j=1}^{k} \int_{D_j} \|\boldsymbol{y} - \boldsymbol{\mu}_j\|^2 d\mathcal{F}(\boldsymbol{y}) + \sum_{j=1}^{k} \pi_j \|\boldsymbol{\mu}_j\|^2 \\ &= W_{\boldsymbol{y}}(k) + B_{\boldsymbol{y}}(k), \end{aligned}$$

where $W_{\boldsymbol{y}}(k)$ and $B_{\boldsymbol{y}}(k)$ are the within and between cluster variances with respect to the distribution of $\boldsymbol{y}$ partitioned into k clusters. The R^2 for clustering, which we shall denote by R_c^2, is defined by

$$R_c^2 = \frac{B_{\boldsymbol{y}}(k)}{tr(\boldsymbol{\Psi})} = 1 - \frac{W_{\boldsymbol{y}}(k)}{tr(\boldsymbol{\Psi})}, \tag{4}$$

and represents the proportion of total variance explained by the clustering partition. The definition of the R_c^2 for an empirical distribution (i.e., for a set of sample data) is

$$R_c^2 = \frac{\text{between sum-of-squares}}{\text{total sum-of-squares}} = 1 - \frac{\text{within sum-of-squares}}{\text{total sum-of-squares}}. \tag{5}$$

We can relate R_c^2 to the coefficient of determination in multiple regression as follows. Let $d_{ij} = 1$ if the ith observation is in cluster j and zero otherwise; set regression coefficients $\boldsymbol{\beta}_j$ equal to population cluster means $\boldsymbol{\mu}_j$ and set the regression error $\boldsymbol{\epsilon}_i = \boldsymbol{y}_i - \boldsymbol{\mu}_j$ (where j is such that $d_{ij} = 1$). Then the coefficient of determination R^2 for a no-intercept regression model

$$\boldsymbol{y}_i = \sum_{j=1}^{k} \boldsymbol{\beta}_j d_{ij} + \boldsymbol{\epsilon}_i \tag{6}$$

is the same as (5) with the least-squares estimators $\hat{\boldsymbol{\beta}}_j = \bar{\boldsymbol{y}}_j$ given by

$$\bar{\boldsymbol{y}}_j = \frac{1}{|D_j|} \sum \boldsymbol{y} I(\boldsymbol{y} \in D_j),$$

where D_j is the jth cluster and $|\cdot|$ denotes the cardinality of the set.

The k-means algorithm can often be generalized to use dissimilarity measures besides Euclidean distance. In multivariate analysis, a common dissimilarity measure is the Mahalanobis distance which generalizes Euclidean distance. One of the focuses of this paper is to examine the R_c^2 for linear transformations of the data and we note that Mahalanobis distance corresponds to a Euclidean distance after a suitable linear transformation of the data. Thus, results on R_c^2 for clustering under a Mahalanobis distance are a special case of results presented in this paper. In the discussion, we comment further on non-Euclidean dissimilarity measures.

As in regression where the coefficient of determination increases as additional variables are added to the model, the R_c^2 is non-decreasing as the number of clusters k increases. In the following we examine how R_c^2 can change when k is held fixed, but the distribution is transformed.

3 R_c^2 under Linear Transformations

Often the data will undergo a linear transformation before analysis. In clustering, a common linear transformation is to standardize or weight individual variables [e.g., 10, 9, 20]. It is well known that certain statistical methods, like PCA, are not invariant to linear transformations [e.g., 21]. This is also true for clustering results. However, translations and rotations do not effect the R_c^2 since a clustering solution depends on the relative distances between points. Also, a scalar transformation of the form $c\boldsymbol{y}$ for a non-null $c \in \Re$ does not change the R_c^2 which can be easily seen from (4).

Clustering results can be improved by utilizing particular linear transformations. A clustering solution, in which the cluster means tend to line up in a particular direction, can be obtained by "stretching" the distribution in that direction via a linear transformation. For example, if indeed there are real, distinct groups in the data and the means of these groups are collinear, as in an allometric extension model [e.g., 22, 23], then "stretching"

the distribution sufficiently in the direction of cluster differences will force the estimated cluster means from the k-means algorithm to lie in this direction [24]. Also, the existence of distinct symmetric groups in a population (e.g., a normal mixture) may cause the overall distribution to appear skewed; by estimating the direction of skewness and stretching the distribution in this direction via projection pursuit [e.g., 25, 26] prior to clustering will yield cluster means that better align with this direction of skewness. Note also that cluster results based on standardizing the variables to unit variance first (similar to PCA performed on the correlation matrix) will not necessarily improve cluster results. The linear transformation to a correlation matrix corresponds to either a stretching or constriction of each individual variable depending on whether or not the individual variances are greater or less then one respectively and this transformation may have no bearing on underlying latent cluster structure in the data.

Our first result deals with the allometric extension scenario and linear transformations corresponding to stretching in the corresponding 1-dimensional direction. If the true cluster means of a random vector $\boldsymbol{y}$ are collinear, then assume this line coincides with the first component $\boldsymbol{y}$ (otherwise, we can merely rotate the distribution to obtain this correspondence). Let y_1 denote this first component and write $\boldsymbol{y} = (y_1, \boldsymbol{y}_2')'$. The notion of stretching (in the y_1 direction say) is a linear transformation $(y_1, \boldsymbol{y}_2')' \longrightarrow (cy_1, \boldsymbol{y}_2')'$ for some $c > 1$. Stretching the distribution of $\boldsymbol{y}$ in the direction of y_1 will eventually (as c increases) force cluster means estimated from a sample to also lie along the y_1 axis. In this scenario, the R_c^2 will increase as the degree of stretching increases and thus corresponds to an actual improvement in the clustering quality.

Proposition 1 *If k self-consistent points of a random vector $\boldsymbol{y}$ lie along a line and the distribution of $\boldsymbol{y}$ is stretched in the direction of this line by some stretching factor $c > 1$, then the R_c^2 will increase.*

Proposition 1 provides an example in the realm of clustering whereby inflating variability in the direction of cluster differences will increase R_c^2. In practice, if true cluster means exist, one usually does not know where these true cluster means lie, and in particular, whether or not these cluster means will be collinear. We now illustrate that stretching

linear transformations can lead to a deterioration in cluster quality even though the transformation may cause the R_c^2 to increase. For this illustration, the number of clusters k and the dimension p are held fixed. Let $\boldsymbol{\alpha}_1 \in \Re^p$ denote a unit-length vector specifying the direction of stretching ($c > 1$) or constricting ($c < 1$). Define $\boldsymbol{A}_2 \sim p \times (p-1)$ so that $[\boldsymbol{\alpha}_1 : \boldsymbol{A}_2]$ is an orthogonal matrix. Stretching (or constricting) the distribution of $\boldsymbol{y}$ in the direction of $\boldsymbol{\alpha}_1$ is achieved via the linear transformation

$$\boldsymbol{A}_c'\boldsymbol{y} = \begin{pmatrix} c & \boldsymbol{0}' \\ \boldsymbol{0} & \boldsymbol{I} \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha}_1' \\ \boldsymbol{A}_2' \end{pmatrix} \boldsymbol{y}. \tag{7}$$

When $c = 0$, then $\boldsymbol{A}_c\boldsymbol{y} = \boldsymbol{y}_2$ and the R_c^2 is

$$R_c^2 = 1 - \frac{W_{\boldsymbol{A}_2'\boldsymbol{y}}(k)}{tr(\boldsymbol{A}_2'\boldsymbol{\Psi}\boldsymbol{A}_2)}. \tag{8}$$

As c initially increases from 0, most of the variability in $\boldsymbol{A}_c'\boldsymbol{y}$ remains in the subspace spanned by the columns of $\boldsymbol{A}_2$ and the cluster means will typically remain in this subspace since the k-means algorithm places cluster means where most of the variation occurs. When $c > 0$ is small and the cluster means remain in the subspace spanned the columns of $\boldsymbol{A}_2$, then

$$R_c^2 = 1 - \frac{c^2\boldsymbol{\alpha}_1'\boldsymbol{\Psi}\boldsymbol{\alpha}_1 + W_{\boldsymbol{A}_2'\boldsymbol{y}}(k)}{c^2\boldsymbol{\alpha}_1'\boldsymbol{\Psi}\boldsymbol{\alpha}_1 + tr(\boldsymbol{A}_2'\boldsymbol{\Psi}\boldsymbol{A}_2)}. \tag{9}$$

From Lemma 4.12 in [14] we have $W_{\boldsymbol{A}_2'\boldsymbol{y}}(k) - tr(\boldsymbol{A}_2'\boldsymbol{\Psi}\boldsymbol{A}_2) < 0$, and thus the derivative of (9) with respect to c will be negative. In other words, increasing the stretching factor c from zero adds variability into the distribution and the R_c^2 will decrease.

Now, as c grows larger and the distribution is stretched in the direction of $\boldsymbol{\alpha}_1$, the total probability mass becomes more concentrated in the $\boldsymbol{\alpha}_1'\boldsymbol{y}$ direction. Typically then, the k cluster means will eventually lie on the $\boldsymbol{\alpha}_1$ axis. If this condition holds, then as c increases, the R_c^2 becomes

$$R_c^2 = 1 - \frac{c^2 W_{\boldsymbol{\alpha}_1'\boldsymbol{y}}(k) + tr(\boldsymbol{A}_2'\boldsymbol{\Psi}\boldsymbol{A}_2)}{c^2\boldsymbol{\alpha}_1'\boldsymbol{\Psi}\boldsymbol{\alpha}_1 + tr(\boldsymbol{A}_2'\boldsymbol{\Psi}\boldsymbol{A}_2)}. \tag{10}$$

The derivative of this expression with respect to c is positive because $\boldsymbol{\alpha}_1'\boldsymbol{\Psi}\boldsymbol{\alpha}_1 - W_{\boldsymbol{\alpha}_1'\boldsymbol{y}}(k) > 0$ and hence the R_c^2 increases as the distribution is stretched further and further into the $\boldsymbol{\alpha}_1$ direction.

Figure 1 illustrates this result with a 3-dimensional distribution $\boldsymbol{y} = (y_1, \boldsymbol{y}_2')'$ where $y_1 \sim N(0,1)$ which is independent of $\boldsymbol{y}_2$, a 4-component bivariate normal mixture distribution

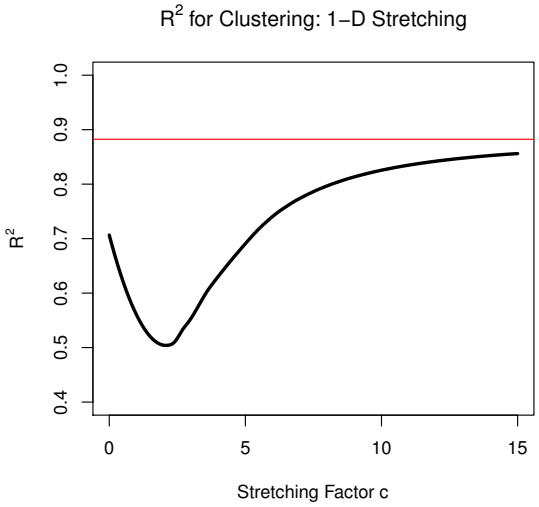

Figure 1: A plot R_c^2 versus a stretching factor $c > 0$ for a 3-dimensional distribution where the first component is $N(0, 1)$ and the distribution of the 2nd and 3rd dimensions is a 4-component normal mixture with equal mixture weights. The horizontal line marks R_c^2 for the distribution projected onto the y_1 axis.

where each bivariate component has the common covariance matrix $\begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix}$ and the mean vectors are

$$\begin{pmatrix} -1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1.5 \end{pmatrix}, \begin{pmatrix} 0 \\ -1.5 \end{pmatrix},$$

with equal mixture weights. When $c = 0$, the distribution is projected onto the y_2 and y_3 plane. As c increases from zero, the R_c^2 initially decreases as noted above. However, as c continues to increase, the R_c^2 will begin to increase. Even though the $k = 4$ true cluster means lie in the space spanned by $(0, 1, 0)'$ and $(0, 0, 1)'$, the estimated cluster means end up lying along the y_1 axis for large values of the stretching factor c. Eventually, the R_c^2 converges to the R_c^2 from partitioning the distribution of y_1 into $k = 4$ clusters (indicated by the horizontal line in Figure 1).

This example illustrates the poor behavior of R_c^2: stretching the distribution via a linear transformation can increase the R_c^2 while steering the cluster results away from the true underlying cluster structure.

To generate the curve in Figure 1, $k = 4$ cluster means were estimated (using the k-

means algorithm [11]) from a very large simulated sample ($n = 8000$, from the distribution described above) with $k = 4$ true cluster means using the R software [27]. The same curve shape results using other values for k (e.g., $k = 2, 3, 5, 6, \ldots$) in the k-means algorithm.

4 R_c^2 for Projections

The previous section looked at the effect of stretching a distribution in a given direction on the R_c^2. This section explores the limiting case of stretching which is projection. Projecting data prior to analysis is common in practice. Examples of this include projecting onto a principal component subspace. If there is an outcome variable as in a regression setting, then a more natural type of projection to consider may be based on partial least-squares. Discarding variables using variable selection algorithms such as the lasso [e.g., 28], can also be regarded as projections. Projection has been used previously in the context of clustering [e.g., 29].

The first result in this section shows that projecting a distribution into the subspace spanned by a set of cluster means will increase the R_c^2.

Proposition 2 *Suppose the random vector $\boldsymbol{y}$ has a set of k self-consistent points that span a linear subspace of dimension $q < p$. Then the R_c^2 corresponding to these cluster means will increase if the distribution is projected onto the subspace spanned by the cluster means.*

Proposition 2 indicates that projecting the distribution onto the line containing the true cluster means will increase the R_c^2. However, in general the R_c^2 behaves poorly under projections. In order to illustrate this point, we introduce another index for cluster quality: the *variation of information* (VI) [30], which is a measure of how well two clusterings of a data set coincide with each other. This measure is particularly useful in simulations where we know the true cluster memberships of data points and we can then use VI to determine how well a clustering result coincides with the true clustering. Given a clustering $\mathcal{C}$, let $P(j) = n_j/n$ where n_j are the number of observations classified to cluster j and n is the total sample size. Then the entropy associated with $\mathcal{C}$ can be defined as

$$H(\mathcal{C}) = -\sum_{j=1}^{k} P(j) \log(P(j)),$$

which is always nonnegative. Now given two clusterings of the same data, $\mathcal{C}_1$ and $\mathcal{C}_2$, let

$$P(j, j') = \frac{|C_j \cap C_{j'}|}{n},$$

for cluster C_j in $\mathcal{C}_1$ and cluster $C_{j'}$ in $\mathcal{C}_2$. Define the mutual information as

$$I(\mathcal{C}_1, \mathcal{C}_2) = \sum_{j=1}^{k} \sum_{j'=1}^{k} P(j, j') \log(\frac{P(j, j')}{P_1(j) P_2(j')}).$$

Then, the variation of information is defined to be

$$VI(\mathcal{C}_1, \mathcal{C}_2) = H(\mathcal{C}_1) + H(\mathcal{C}_2) - 2I(\mathcal{C}_1, \mathcal{C}_2). \tag{11}$$

$VI = 0$ if the two clusterings produce identical clusters (up to a re-labeling); otherwise $VI > 0$. One nice feature of VI is that it is a metric on the space of clusterings [30] and smaller values of VI correspond to a better agreement between two clusterings of the same distribution. Other indices of clustering quality could also be considered such as the Rand index [31] or the adjusted Rand index [32]. We chose to use VI though due to its alignment with information theoretic principles as discussed in [30].

We now present a simulation showing that although R_c^2 tends to increase when clustering lower-dimensional projections of a distribution, the actual quality of the clustering tends to deteriorate, as measured by VI.

Data from a normal mixture distribution with $K = 5$ true clusters and dimension $p = 50$ was simulated using randomly generated means and covariance matrices. $n = 100$ observations were simulated from each mixture component. 1000 random projections onto lower dimensional subspaces of dimensions ranging from $q = 1$ to $q = p - 1$ were also generated and the k-means algorithm was run on each projection specifying $k = 5$ cluster means. The random projections were obtained using a matrix of eigenvectors from an eigen-decomposition of a covariance matrix obtained from an independent, large sample of a p-variate standard normal variates. For each dimension $q = 1$ to 49, 1000 q-dimensional projections were randomly generated and the k-means algorithm was run on each producing the plots in the top row of Figure 2. For each projection, the k-means algorithm was run 100 times in order to find a cluster solution close to the global optimal. The top-left panel of this figure shows the cluster R_c^2 (averaged over the 1000 projections $\pm$ 2 s.d. of the R_c^2).

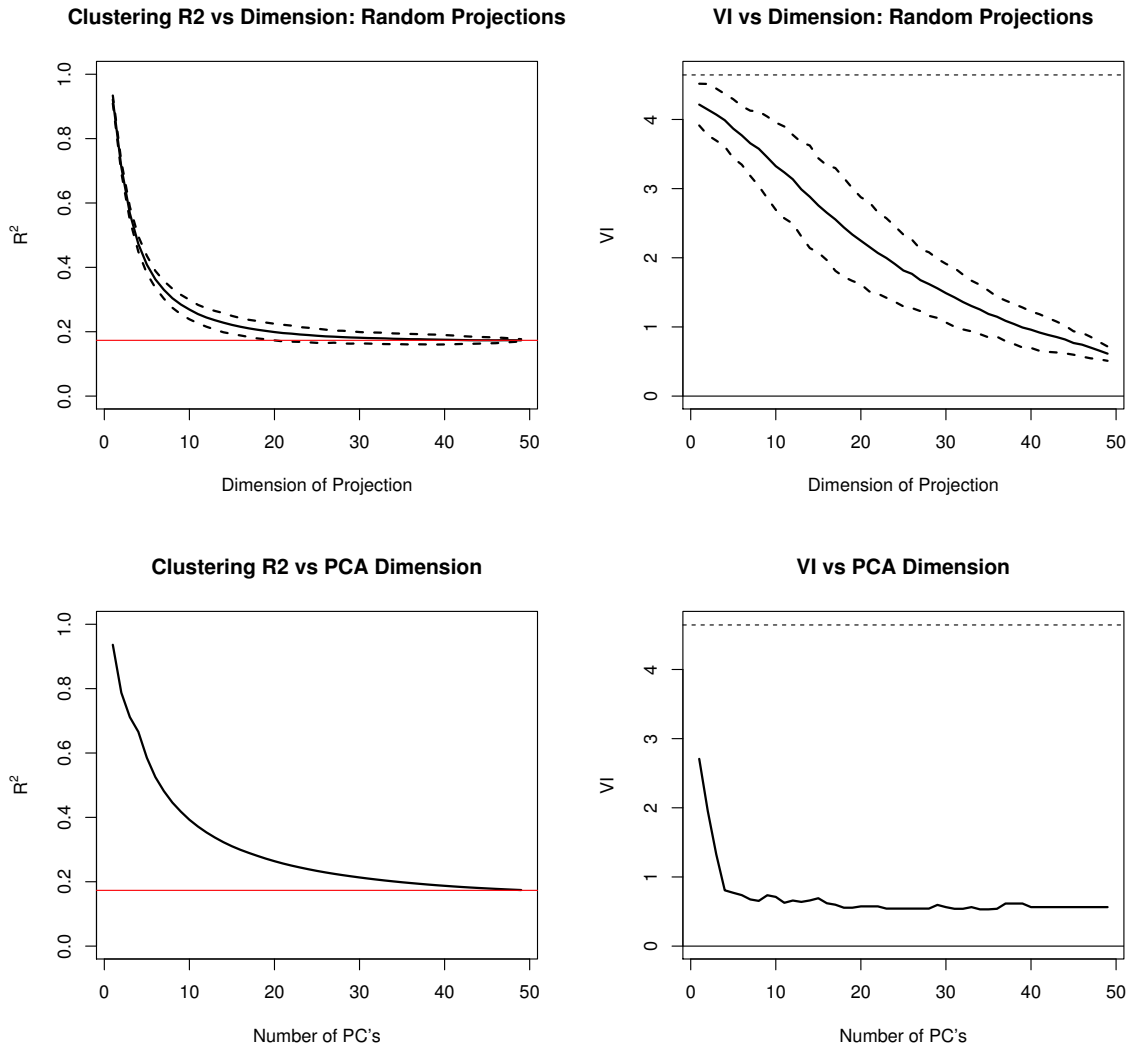

Figure 2: Top left panel: R_c^2 versus the dimension of the projected data for 1000 random projections of $p = 50$ dimensional data from the $K = 5$ normal mixture. Top right panel: the corresponding Variation of information (VI) versus the dimension of the projected data. (The dashed bands in the top two panels correspond to $\pm$ 2 s.d. of the 1000 R_c^2 and VI's). The bottom left and right panels are R_c^2 and VI respectively for projections onto principal component subspaces of dimensions 1 to $p - 1$.

The horizontal line is the R_c^2 from the original un-projected data. This top-left panel makes it appear that the optimal projection for clustering in terms of maximizing the R_c^2 is to project the data onto a line. However, in terms of the variation of information metric shown in the top-right panel of Figure 2, an opposite relation occurs between the dimension of the projection and the quality of the clustering in terms of VI. The top-right panel shows the average VI (over the 1000 random projections, ± 2 s.d. band); the theoretical maximum of the VI, $2\log(k)$ [30], is shown by the dashed horizontal line. As the top-right panel of Figure 2 shows, the VI increases as the dimension of the projection decreases, contrary to relationship for R_c^2. Thus, the best clustering quality occurs with no projection at all, but R_c^2 gives the wrong impression that clustering quality improves as the distribution is projected into lower and lower dimensional subspaces. It is interesting to note that the width of the error band in the top-left panel of Figure 2 for R_c^2 is very narrow and only changes marginally based on the number of random starts of the k-means algorithm. Thus, for random projections onto a subspace of a given dimension, the R_c^2 is almost constant. However, the error band for the VI in the top-right panel of Figure 2 remains relatively wide (regardless of how many random starts are used for the k-means algorithm) showing that VI can vary substantially depending on how the data are projected.

The bottom two panels of Figure 2 show the R_c^2 and VI for projections onto the principal component subspaces of varying dimension (instead of random subspaces as in the top two panels). The behavior of the clustering R^2 for projections onto principal component subspaces is very similar to the projections onto randomly generated subspaces shown, which is not surprising since the R_c^2 varies little regardless of which projection is used. However, the VI for the principal component projections drops quite rapidly as the dimension of the projected data increases and then levels off as the dimension continues to increase. This is an indication that a large component of the overall variability in the data is due to differences between the cluster means. Consequently, projecting into much lower-dimensional principal component subspaces hurts the clustering quality. The VI for projections into the principal component subspaces (bottom-right panel) are considerably less than the corresponding VI in randomly chosen subspaces (top-right panel).

Figure 2 shows that VI improves (decreases) using principal component projections in-

stead of random projections whereas the behavior of R_c^2 remains roughly the same whether random projections or principal component projections are used. In this simulation scenario, differences in the true clusters were exhibited in all $p = 50$ variables which helps to explain why VI improves for high-dimensional random projections. We return to this example in the next section.

5 R_c^2 in High Dimensions

Many modern statistical analyses involve high dimensional data ($p >> n$). In this section, we examine the behavior of the R_c^2 as the dimension increases. The asymptotic theory on rate distortion from coding theory is usually given in terms of $k \to \infty$. However, from a statistical point of view, even as the dimension of the data increases (e.g., more and more variables are considered for analysis), the number of clusters to use for partitioning often stays fixed. Thus, in this section, we are primarily interested in examining R_c^2 for clustering for a fixed value of k as the dimension $p \to \infty$. Here, we show that as the dimension p of the distribution increases, R_c^2 will typically grow smaller (approach zero) when k is held fixed if the overall variability increases without bound. Intuitively, the reason this happens has to do with the *curse of dimensionality*. Recall that R_c^2 is a relative measure of how close points are to their cluster means relative to how close they are to the overall mean.

Theorem 3 *Let $\mathbf{\Psi}(p)$ denote the covariance matrix for a p-dimensional distribution and suppose the total variance goes to infinity as the dimension of the data increases, $\lim_{p\to\infty} tr(\mathbf{\Psi}(p)) = \infty$, while the largest eigenvalue remains bounded. Then, for a fixed number of clusters k, $R_c^2(p) \to 0$, where $R_c^2(p)$ is the R^2 for clustering of the p-dimensional distribution into k clusters. If the overall variability remains bounded, then the R_c^2 will not converge to zero.*

In the quantization literature, asymptotic results have been obtained when k goes to infinity. A thorough overview of these results can be found in [14]. The within cluster variability is referred to as the kth quantization error. Asymptotic results for $R_c^2(p)$ can be extended as both k and p go to infinity. Although the quantization error goes to zero as k goes to infinity, when scaled by a factor of $k^{2/p}$ (using an L^2 norm for the distance), the

limit is finite and positive. Details on the quantization error for when both k and p go to infinity can be found in Section 9.3 of [14].

In some high-dimensional applications, large numbers of variables are measured and many (or most) may be completely unrelated to the underlying cluster structure. For instance, the recent depression study [33] measured hundreds of variables derived from brain imaging modalities in order to discover moderators for treatment response in depression and it was not known whether or not many of these variables would provide moderating effects. In such situations, many variables are likely to be pure noise in terms of distinguishing clusters.

In order to illustrate such a scenario and Theorem 3, we return to the simulation example of Section 4. In addition to the original $p = 50$ variables from the mixture distribution, another 50 pure noise variables were appended to the data. These pure noise variables were simulated from $N(\mathbf{0}, \sigma^2 \boldsymbol{I})$ for a range of noise variances σ^2 with $\sigma = 1, 5, 10, 15, 20$. The results for projections into the PCA subspaces are shown in Figure 3 where each curve corresponds to each of the σ values. The R_c^2 in the left frame of the figure behaves just the same as when there is no extra-added noise, as in Figure 2. From Theorem 3, we see that the R_c^2 goes to zero as the dimension increases. For VI in the right panel, when the noise level is low ($\sigma = 1, 5, 10$), one obtains lower VI as the dimension increases as in Figure 2, and then the VI levels off as the dimension continues to grow. However, as the noise variance grows bigger and dominates the PCA, the VI jumps up to close its maximal value and remains essentially constant for higher and higher dimensional projection subspaces. In other words, the k-means algorithm is unable to recover the true clusters as the dimension and the degree of noise in the data swamps the signal. The R_c^2 is oblivious to this deterioration in cluster quality for low dimensional projections.

In high-dimensional clustering problems, an attractive dimension reduction approach is to use an L^1 penalty, as in the lasso [28], to select variables (or features) and produce a sparse solution. The above simulation (with the added 50 noise variables) was run using a sparse k-means clustering algorithm [34]. In the sparse clustering, in order to improve results, weights are estimated for each feature depending on how well the feature contributes to differentiating clusters. Using an L^1 penalty, some weights are shrunk to zero thereby

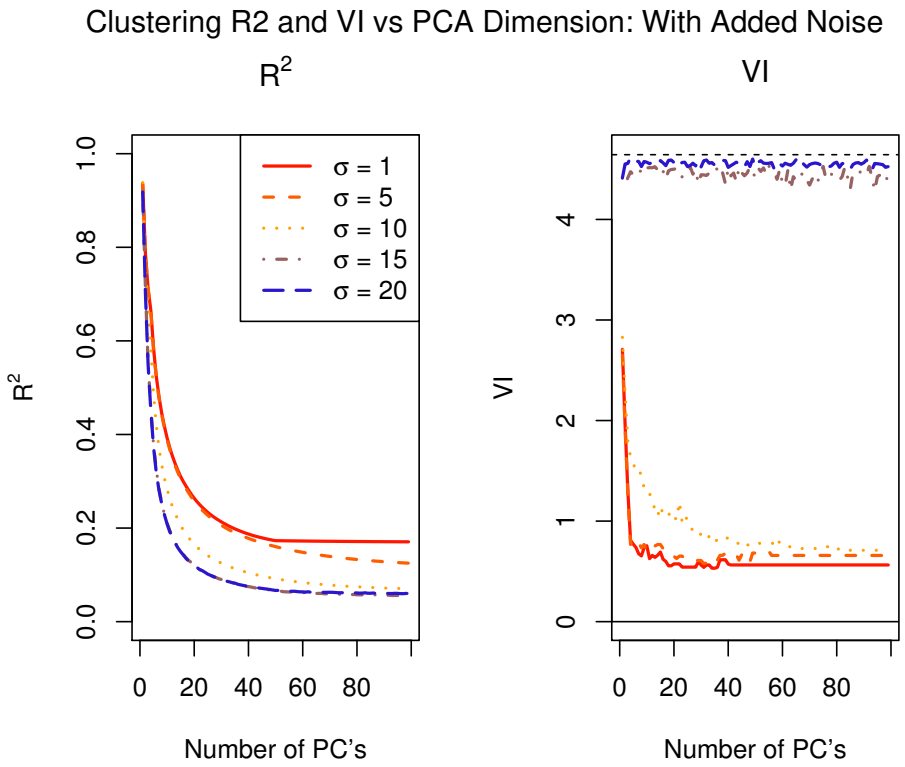

Figure 3: **Added noise**: R_c^2 (left panel) and VI (right panel) versus dimension of principal component projected data similar to Figure 2 except 50 pure noise variables were added to the data with increasing levels of variability (σ).

eliminating the corresponding feature. Figure 4 shows simulation results using a sparse k-means clustering from the *sparcl* package [35] in R. A range of tuning parameters was used so that the number of features selected ranged from very few (1–3 say) to all 100 features being selected. The left panel of Figure 4 shows the R_c^2 versus the number of features selected for the different levels of noise σ; the right panel shows the corresponding curves for VI. The left panel of this figure is very similar to the left panel of Figure 3 where the R_c^2 curve decreases as the number of selected features increases (for each value of the noise σ). As in the PCA case above, the R_c^2 is oblivious to determining how many features usefully contribute to the clustering. Thus, even using a sparse clustering approach when there are true noise variables, the R_c^2 incorrectly indicates that the clustering is optimized using the fewest possible features. On the other hand, the VI shows that the sparse clustering tends to perform quite well when the noise level was moderate to small – the VI initially decreased as the number of features selected increased and then showed a big drop when most of the 50 non-noise variables had been selected, and ultimately the VI curve flattened out as the remaining purely noise variables were included.

6 R_c^2 for Elliptical and Skew-Elliptical Distributions

A $p-$dimensional random vector $\mathbf{y} \sim ELL_p(\boldsymbol{\mu}, \boldsymbol{\Psi}, \phi)$ is said to be elliptical with location vector $\boldsymbol{\mu} \in \Re^p$, scatter matrix $\boldsymbol{\Psi} = \boldsymbol{\Psi}' \in \Re^p \times \Re^p$ and characteristic function ϕ if it may be represented as $\mathbf{y} = \boldsymbol{\mu} + \mathbf{A}\mathbf{w}$, where $\mathbf{A}\mathbf{A}' = \boldsymbol{\Psi}$ and the distribution of the $p-$dimensional random vector $\mathbf{w}$ is invariant with respect to orthogonal transformations. Let $\boldsymbol{y}$ denote a p-variate with mean $\boldsymbol{\mu} = \mathbf{0}$ elliptically symmetric random vector with covariance matrix $\boldsymbol{\Psi}$. Let λ_1 and $\boldsymbol{v}_1$ denote the largest eigenvalue and corresponding eigenvector of $\boldsymbol{\Psi}$. Then $k = 2$ principal points of $\boldsymbol{y}$ must lie on the first principal component axis [36] and have the form $\boldsymbol{\mu} + c_1\boldsymbol{v}_1$ and $\boldsymbol{\mu} + c_2\boldsymbol{v}_1$ for two constants c_1, c_2. Univariate marginal distributions of $\boldsymbol{y}$ are symmetric since $\boldsymbol{y}$ is elliptical, but two principal points of a symmetric univariate distribution need not be symmetric about the mean. All univariate marginal distributions of $\boldsymbol{y}$ have the same distributional form [e.g., 37,Theorem 2.9, page 36]. Let $y_1 = \boldsymbol{v}'\boldsymbol{y}/\sqrt{\lambda_1}$. If the principal points are indeed symmetric about the mean, then the 2 principal points must be of the form $\pm\{E|\boldsymbol{v}'\boldsymbol{y}|\}\boldsymbol{v}$ [19]. From (4) and also (15) (in the Appendix), the R_c^2

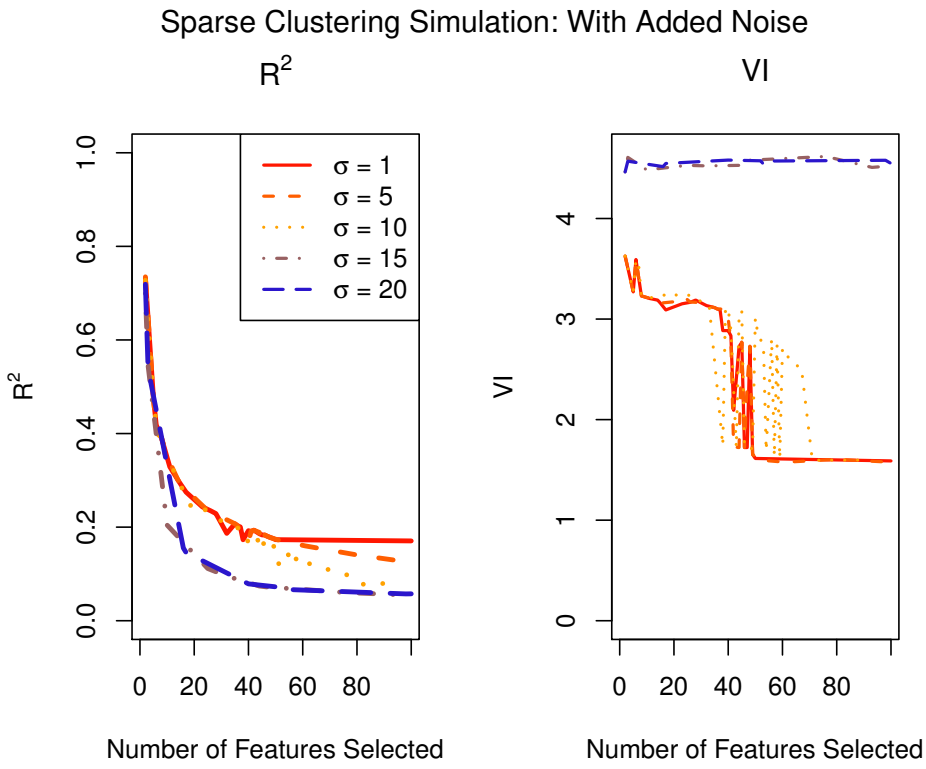

Figure 4: **Sparse clustering simulation results**: R_c^2 (left panel) and VI (right panel) versus number of features selected by the sparse k-means clustering similar to Figure 2 except 50 pure noise variables were added to the data with increasing levels of variability (σ).

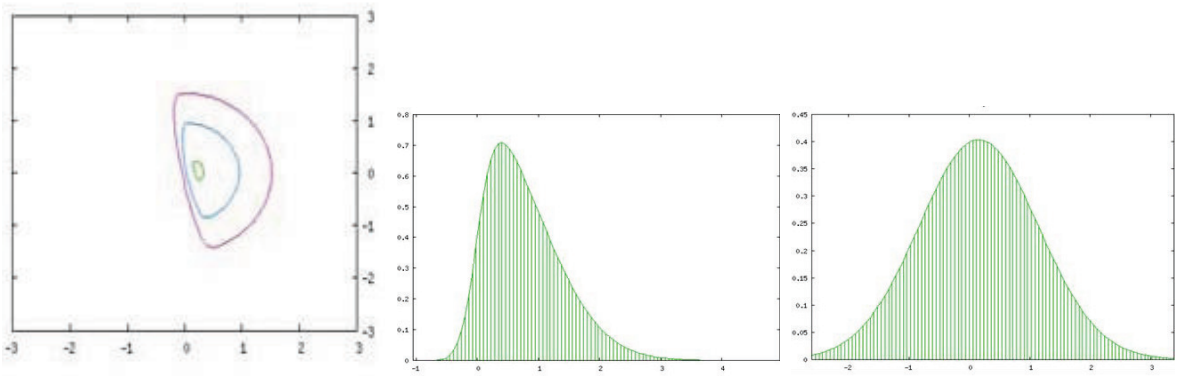

Figure 5: Contour plot for a bivariate skew-elliptical distribution (left panel) and the corresponding univariate marginal density curves (center and right panels).

for the 2 principal points, using the symmetry of the univariate projection of $\boldsymbol{y}$ onto $\boldsymbol{v}$, is

$$R_c^2 = \frac{(E|\boldsymbol{v}'\boldsymbol{y}|)^2}{tr(\boldsymbol{\Psi})} = \frac{\lambda_1(E|y_1|)^2}{tr(\boldsymbol{\Psi})}. \tag{12}$$

Elliptical distributions are symmetric, but in practice observed data are often skewed and we now explore the R_c^2 in the presence of skewness. First, we review the skew-elliptical distributon [e.g., 38]. A p-dimensional skew-elliptical ($SELL_p$) distribution can be obtained from a $(p+1)-$dimensional elliptical distribution, by conditioning on the event that one component is greater than its location. Implications are twofold. In the first place it provides a very straightforward interpretation of skew-elliptical distributions in terms of non-random sampling: they arise when an elliptical random vector is included in the sample if and only if one of its components is greater than its location. In the second place, the same conditioning argument provides an efficient way to generate skew-elliptical random vectors using elliptical ones. For illustration, Figure 5 shows a contour plot (left panel) of a bivariate skew-elliptical distribution. Contours of equal density for elliptical distributions are ellipsoids, but the skew-elliptical contours become stretched into non-elliptical shapes, as in the left panel of Figure 5. The center and right panels of Figure 5 show the univariate (skewed) density curves for this illustration.

Here is an algorithm for generating a random vector $\mathbf{z} \sim SELL_p\,(\boldsymbol{\xi}, \boldsymbol{\Psi}, \boldsymbol{\alpha}, \phi)$. Note that

we are using $\boldsymbol{\xi}$ to denote the location parameter here. First, define the vector $\boldsymbol{\delta}$ as follows:

$$\boldsymbol{\delta} = \frac{\boldsymbol{\Psi\alpha}}{\sqrt{1+\boldsymbol{\alpha}'\boldsymbol{\Psi\alpha}}}.$$

Second, generate a random variable y and a random vector $\mathbf{x}$ such that

$$\begin{pmatrix} y \\ \mathbf{x} \end{pmatrix} \sim ELL_{p+1}\left[\begin{pmatrix} 0 \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} 1 & \boldsymbol{\delta}' \\ \boldsymbol{\delta} & \boldsymbol{\Psi} \end{pmatrix}, \phi\right].$$

Third, set $\mathbf{w}$ equal to $\mathbf{x}$ $(-\mathbf{x})$ if y is positive (negative):

$$\mathbf{w} = \begin{cases} \mathbf{x} & y > 0 \\ \\ -\mathbf{x} & y \leq 0 \end{cases}.$$

Fourth, add the vector $\boldsymbol{\xi}$ to the vector $\mathbf{w}$ and name the sum $\mathbf{z}$: $\mathbf{z} = \mathbf{w}+\boldsymbol{\xi} \sim SELL_p(\boldsymbol{\xi}, \boldsymbol{\Psi}, \boldsymbol{\alpha}, \phi)$.

In many applications, an assumption of an elliptical distribution is often not tenable due to existing skewness. For the skew-elliptical distribution, the shape parameter $\boldsymbol{\alpha}$ controls the skewness of $SELL_p(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\alpha}, \phi)$ and its divergence from $ELL_p(\boldsymbol{\xi}, \boldsymbol{\Omega}, \phi)$. In particular, when $\boldsymbol{\alpha} = \mathbf{0}_p$, the two distributions coincide. It is then interesting to understand the effect of $\boldsymbol{\alpha}$, and therefore of skewness, on the R_c^2 , from the previous example. The following theorem takes a step in this direction.

Theorem 4 *Let $\boldsymbol{y}$ and u be a random vector and a random variable whose joint distribution is*

$$\begin{pmatrix} \boldsymbol{y} \\ u \end{pmatrix} \sim ELL_{p+1}\left[\begin{pmatrix} \mathbf{0}_p \\ 0 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Psi} & \mathbf{0}_p \\ \mathbf{0}_p' & 1 \end{pmatrix}, \phi\right].$$

Also, let $\widetilde{\boldsymbol{y}} \sim SELL_p(\mathbf{0}_p, \boldsymbol{\Psi}, \boldsymbol{\alpha}, \boldsymbol{\phi})$ be a skew-elliptical random vector where the shape parameter $\boldsymbol{\alpha}$ belongs to the linear space spanned by the nondominant eigenvectors of $\boldsymbol{\Psi}$. Then $\boldsymbol{\Psi}$ and $cov(\widetilde{\boldsymbol{y}})$ have the same dominant eigenvector. Furthermore, if $k = 2$ principal points of $\boldsymbol{y}$ are symmetric about the mean, then the R_c^2 of the symmetric, self-consistent points of $\widetilde{\boldsymbol{y}}$ associated with the Voronoi regions of $k = 2$ symmetric principal points of $\boldsymbol{y}$ is

$$\frac{\lambda_1 \cdot E^2(|u|) + \boldsymbol{\mu}'\boldsymbol{\mu}}{E(u^2) \cdot tr(\boldsymbol{\Psi}) - \boldsymbol{\mu}'\boldsymbol{\mu}}, \tag{13}$$

where $\boldsymbol{\mu}$ and λ_1 are the expectation of $\widetilde{\boldsymbol{y}}$ and the dominant eigenvalue of $\boldsymbol{\Psi}$ respectively.

Note that this expression for the R_c^2 can be re-expressed independently of $\boldsymbol{\mu}$:

$$\frac{\lambda_1 \cdot E^2(|u|) + \boldsymbol{\mu}'\boldsymbol{\mu}}{E(u^2) \cdot tr(\boldsymbol{\Psi}) - \boldsymbol{\mu}'\boldsymbol{\mu}} = \frac{E^2(|u|) \cdot (\lambda_1 + \boldsymbol{\delta}'\boldsymbol{\delta})}{E(u^2) \cdot tr(\boldsymbol{\Psi}) - E^2(|u|)\,\boldsymbol{\delta}'\boldsymbol{\delta}} = \frac{E^2(|u|) \cdot (\lambda_1 + \boldsymbol{\delta}'\boldsymbol{\delta})}{tr[cov(\tilde{\mathbf{x}})]},$$

where $\boldsymbol{\delta} = \boldsymbol{\Psi}\boldsymbol{\alpha}/\sqrt{1 + \boldsymbol{\alpha}'\boldsymbol{\Psi}\boldsymbol{\alpha}}$.

Self-consistent points of $\widetilde{\boldsymbol{y}}$ characterized by the Voronoi regions of two symmetric principal points of $\boldsymbol{y}$ will not in general be either symmetric nor principal. However, their R_c^2 will be higher than the R_c^2 of the principal points of $\boldsymbol{y}$, under the above mentioned assumptions. In that sense, we can say that skewness increases clustering efficiency, as measured by R_c^2.

The above result sheds some light on the role of ellipticity in principal points clustering. When ellipticity is replaced with the weaker assumption of skew-ellipticity, the two symmetric principal points lose their optimal clustering property. On the other hand, the R_c^2 for clustering associated with the Voronoi regions associated with the $k = 2$ symmetric principal points increases. This simple argument shows that two intuitively appealing criteria for clustering, i.e., principal points and the R^2, might conflict with each other.

When the shape parameter $\boldsymbol{\alpha}$ is the null vector, the distributions of $\boldsymbol{y}$ and $\widetilde{\boldsymbol{y}}$ coincide, thus leading to the following corollary.

Corollary 5 *Let $\boldsymbol{y}$ and u be a random vector and a random variable whose joint distribution is*

$$\begin{pmatrix} \boldsymbol{y} \\ u \end{pmatrix} \sim ELL_{p+1}\left[\begin{pmatrix} \mathbf{0}_p \\ 0 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Psi} & \mathbf{0}_p \\ \mathbf{0}_p' & 1 \end{pmatrix}, \phi\right].$$

Then the R_c^2 of two symmetric principal points of $\boldsymbol{y}$ is

$$\frac{\lambda_1 \cdot E^2(|u|)}{E(u^2) \cdot tr(\boldsymbol{\Psi})},$$

where λ_1 is the dominant eigenvalue of $\boldsymbol{\Psi}$.

In particular, $R_c^2 = (2/\pi)\,\lambda_1/tr(\boldsymbol{\Psi})$ if $\boldsymbol{y}$ is normally distributed. The corollary might be derived without referring to the above theorem, using an argument which also gives some insight into principal points of elliptical distributions. Assume that two principal points of the centered and elliptically distributed $\boldsymbol{y}$ are symmetric about the mean which is the origin. Let $\boldsymbol{v}$ denote the leading eigenvector of the covariance matrix of $\boldsymbol{y}$. Then by [36] (Theorem 1), the two principal points must be of the form: $\pm E(|\boldsymbol{v}'\boldsymbol{x}|)\,\boldsymbol{v}$. From (4), the

R_c^2 for the two principal points, using the symmetry of the univariate projection of $\boldsymbol{y}$ onto $\boldsymbol{v}$, is

$$R_c^2 = \frac{E^2\left(|\boldsymbol{v}'\boldsymbol{y}|\right)}{tr\left[cov(\boldsymbol{y})\right]} = \frac{\lambda_1^2 E^2\left(|y_1|\right)}{tr(\boldsymbol{\Psi})}. \tag{14}$$

Finally, we can relate this last result to the R_c^2 in the high dimensional setting described in Section 5. It is easy to see from (14) that R_c^2 for $k = 2$ for a family of elliptical distributions defined over an increasing sequence of dimensions p, goes to zero if the proportion of variance explained by the first principal component goes to zero:

$$\text{if } \lim_{p\to\infty} \frac{1}{\lambda_1} \sum_{j=2}^{p} \lambda_j = 0 \text{ then } R_c^2 \to 0.$$

Practical relevance of the above theoretical results strictly depends on the appropriateness of skew-elliptical distributions for skewness modeling. The problem of whether skewness is due to the presence of different subpopulations in the sampled population dates back to [39] and has been thoroughly reviewed by [40]. They also showed that data generated from a multivariate skew-normal distribution (i.e., the best known example of skew-elliptical distribution) are well approximated by a mixture of two or three multivariate normal distributions. However, skew-elliptical distributions are motivated by either subject-matter considerations or large-sample results. In the former case, skew-elliptical distributions are appropriate when the observed data are a non-random sample from an elliptical distribution [41]. In the latter case, the distribution of the sample mean of i.i.d. skewed random vectors (as for example the mixture of two multivariate normal distributions with the same covariance matrix and unequal weights) may be well approximated by the multivariate skew-normal distribution [42]. [43] and [44] illustrated the practical relevance of this result with air pollution data and with discrete distributions, respectively.

7 Application: R_c^2 for Clustering Curves

This section provides an illustration of R_c^2 using data from a clinical study of depression. If the points in a data set correspond to curves, i.e., functional data [45], it is common to represent the curves using a set of basis functions. In such cases, the basis representation of the curves is obtained from a linear transformation of the raw data, similar to Section 3 and

the results of Section 5 apply as the richness of the basis representation grows (i.e., as the number of basis functions increases). Cluster means obtained from clustering curves can be useful for identifying representative curve shapes [e.g., 46, 47, 48, 24, 49]. Figure 6 gives an illustration, showing curves fitted using $p = 5$ B-spline basis functions (black curves) [50] and $p = 5$ Fourier basis functions (red dashed curves). These curves are obtained from a 12-week depression clinical trial where all participants were treated with fluoxetine [51]. For clarity, only a subset of 25 from a total of $n = 414$ curves are shown. The curves correspond to a smoothed Hamilton Rating Scale for Depression (HRSD) where lower scores correspond to less severe depression symptoms over the 12 week period (with time scaled to take values from zero to one). As the figure shows, the fitted curves using B-splines and Fourier basis functions are very similar to one another. The curves were clustered into $k = 4$ groups by running the k-means algorithm on the 5-dimensional basis coefficients. The R_c^2 from the B-spline results is $R_c^2 = 0.534$ whereas for the Fourier coefficients we obtain a much higher $R_c^2 = 0.876$. This might indicate that the Fourier basis representation leads to a better clustering result (i.e., higher R_c^2). However, the first principal component of the Fourier basis representation accounts for 99.8% of the coefficient variability but the first principal component of the B-spline coefficient representation accounts for only 51.1% of the variability. That is, the Fourier coefficients are concentrated almost entirely in a 1-dimensional subspace and consequently, based on results above, the R_c^2 using the Fourier coefficients is higher compared to the B-spline R_c^2 (where the first three principal components account for only 87.8% of the total variability). Therefore, the higher R_c^2 from clustering the Fourier basis curves compared to the B-spline curves does not necessarily mean the Fourier cluster results are better.

As another illustration, if the number of B-spline basis functions used to smooth the curves is increased from 5 to 6 to 7 basis functions, the R_c^2 from the B-spline coefficients decreases respectively from .534 to .501 to .434. Again, the lowering of the R_c^2 does not necessarily mean that the clustering quality decreases as the basis representation grows richer, but it is a consequence of the R_c^2 decreasing as the dimension of the basis representation increases, as seen in Section 5.

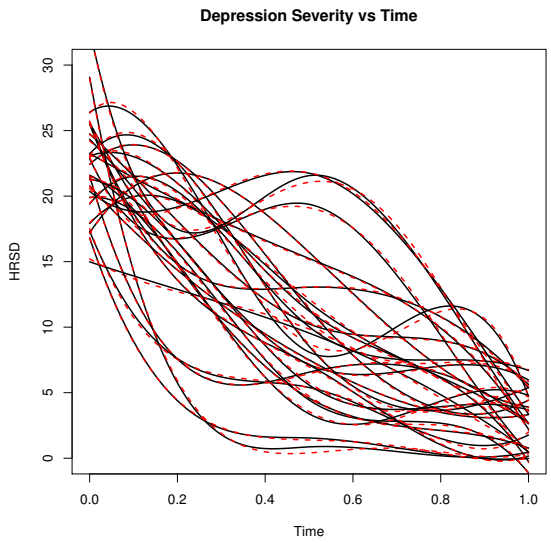

Figure 6: Estimated B-spline (solid) and Fourier basis (dashed) curves for HRSD response over a 12 week period.

8 Discussion

R^2-type summary statistics, that measure the proportion of variability explained by a statistical model, are very common in practice. We have provided an overview of R_c^2, the R^2 for clustering, with illustrations. A major drawback of the general R^2 is that it can be artificially inflated while the quality of the statistical model deteriorates. It is well known that increasing the number of predictors in a regression model, the number of components extracted in a principal component analysis and the number of clusters estimated in cluster analysis all inflate the R^2. This note has highlighted that in cluster analysis, R_c^2 can also be inflated via linear transformations or by projecting the data onto lower-dimensional subspaces; alternatively R_c^2 can be deflated by increasing the dimensionality of the data, e.g., adding additional variables. These increases (or decreases) in R_c^2 do not necessarily reflect an increase or decrease in clustering quality. This paper has also showed that model misspecifications may lead to overestimated R_c^2.

An alternative clustering index to the R_c^2, that is also a function of the within and total deviance, is $tr\left(\mathbf{W}\mathbf{T}^{-1}\right)$ which was introduced on intuitive grounds by [2]. A relative merit of $tr\left(\mathbf{W}\mathbf{T}^{-1}\right)$ over R_c^2 is its invariance with respect to nonsingular linear transformations,

thus potentially avoiding the problems with "stretching" discussed in Section 3.

Many clustering algorithms, such as k-medoids [52], use a matrix of pairwise dissimilarities between data points as their input. The k-means algorithm described here is equivalent to using a dissimilarity matrix based on squared Euclidean distances [e.g., see 53, page 509]. For other non-Euclidean dissimilarity measures, an R^2 can be defined similarly to (5) based on partitioning the total sum-of-squares into a within and between group sum-of-squares. This approach has been described in [54, 55] in the context of MANOVA with permutation testing. For non-Euclidean dissimilarities, [54, 55] propose a "pseudo" F-test statistic that coincides with the usual the MANOVA F-test statistic when the dissimilarity measure is Euclidean distance. Similarly, for general dissimilarities measures, the corresponding clustering R^2 could be called a "pseudo"-R^2 for clustering. We have highlighted weaknesses in the R_c^2 when clustering based on Euclidean distances between points. It would be interesting to investigate the corresponding pseudo-R^2 for clustering when other dissimilarity measures are used, such as the Bray-Curtis measure of ecological distance [54].

In summary, besides providing a useful summary of how much of the total variability is explained by a clustering of the data, R_c^2 has very little practical utility and can be quite misleading. R_c^2 can be inflated or deflated based on how many variables are used and what types of transformations are used on the data. Clustering high-dimensional data is a recurrent problem in many fields of science [56], and often requires dimension reduction, especially in the presence of variables that are not helpful in discriminating groups [e.g., 57]. As we have seen, the R_c^2 is not very informative in these high-dimensional settings. Also, the R_c^2 is not very useful for determining the number of clusters in a data set and other methods geared specifically to this problem are preferable, [e.g., 58, 59].

Our interest in this line of work was motivated by the problem of finding transformations of the data that can lead to useful cluster partitions, particularly in applications involving functional data [24]. Future work related to cluster analysis is in the area of medical research, specifically (1) precision medicine to discover patient subgroups that benefit most from particular treatments and (2) developing data-driven methods for the discovery and diagnosis of psychiatric disorders using classification based on biological measures [60]. Current medical technology produces very high-dimensional data (e.g., MRI and fMRI

brain scans) and clustering methods need to be advanced in order to handle these type of modern data modalities. One approach is to develop the notion of "pre-conditioning" the data via linear transformations (or "stretching") to optimize clustering algorithms; this would parallel recent work related to pre-conditioning for variable selection via the lasso [61]. Additionally, we are interested in extending the results of Section 3 by exploring linear stretching transformations in more than one direction. An optimality criterion one can consider in the context of precision medicine with two treatments is to determine stretching transformations of outcome variables that will minimize the VI computed from (i) cluster labels from clustering the data pooled across treatments with (ii) the two treatment labels. It is anticipated that this approach will be useful in identifying outcomes specific to particular treatments while acknowledging that there will often be groups with similar outcomes for different treatments (e.g., placebo responders). The R_c^2, in the form described in this paper, is not likely to be a useful summary statistic for this future work; we anticipate that these research endeavors will motivate the development of alternative summary measures in the realm of clustering.

9 Appendix: Proofs of results

Proof of Proposition 1

Proof. Let $\xi_1, \ldots, \xi_k$ denote these k cluster means for the standardized (unit variance) y_1/σ_1 distribution where σ_1^2 is the variance of y_1; let $\mathbf{\Psi}_2$ be the covariance matrix of all the components of $\boldsymbol{y}$ besides y_1. Then, by (4), the R_c^2 for the collinear pattern of cluster means along the y_1 axis is

$$R_c^2 = \frac{\sigma_1^2 \sum_{j=1}^k \xi_j^2}{\sigma_1^2 + tr(\mathbf{\Psi}_2)}.$$

If D_j is the jth cluster (or Voronoi region) corresponding to the jth self-consistent point, then by self-consistency, $E[\boldsymbol{y}|\boldsymbol{y} \in D_j] = (\sigma_1 \xi_j, 0, \ldots, 0)'$. Now stretching the distribution in the y_1 direction via a stretching factor $c > 1$ transforms $\boldsymbol{y}$ to $\tilde{\boldsymbol{y}} = (cy_1, \boldsymbol{y}_2')'$ where $\boldsymbol{y}_2$ is the $p-1$ vector of components of $\boldsymbol{y}$ besides y_1. From the self-consistency property, the points $(c\sigma_1 \xi_j, \mathbf{0}')'$ are self-consistent points for $\tilde{\boldsymbol{y}}$ and the R_c^2 becomes

$$R_c^2 = \frac{c^2 \sigma_1^2 \sum_{j=1}^k \xi_j^2}{c^2 \sigma_1^2 + tr(\mathbf{\Psi}_2)}. \tag{15}$$

The derivative of this expression with respect to c is positive and hence, stretching the distribution in the y_1 direction increases R_c^2. ■

Proof of Proposition 2.

Proof. Let $\boldsymbol{A}_1$ be a $p \times q$ matrix with orthonormal columns that span the space spanned by the k self-consistent points and let $\boldsymbol{A}_2$ be a $p \times (p-q)$ matrix with orthonormal columns so that $\boldsymbol{A} = [\boldsymbol{A}_1 : \boldsymbol{A}_2]$ is orthogonal. Let the k cluster means be denoted $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k$ and note that $\boldsymbol{A}'\boldsymbol{\mu}_j = (\boldsymbol{A}_1'\boldsymbol{\mu}_j, \boldsymbol{0})'$. By Lemma 2.3 in [36], the points $\boldsymbol{A}_1'\boldsymbol{\mu}_j, j = 1, \ldots, k$, represents a set of self-consistent points for $\boldsymbol{A}_1'\boldsymbol{y}$. Also, because $\boldsymbol{A}$ is orthogonal $\|\boldsymbol{\mu}_j\|^2 = \boldsymbol{\mu}_j'\boldsymbol{\mu}_j = (\boldsymbol{A}'\boldsymbol{\mu}_j)'\boldsymbol{A}'\boldsymbol{\mu}_j = \|\boldsymbol{A}_1'\boldsymbol{\mu}_j\|^2$. Furthermore, $tr(\boldsymbol{\Psi}) = tr(\boldsymbol{A}'\boldsymbol{\Psi}\boldsymbol{A}) = tr(\boldsymbol{A}_1'\boldsymbol{\Psi}\boldsymbol{A}_1) + tr(\boldsymbol{A}_2'\boldsymbol{\Psi}\boldsymbol{A}_2) \geq tr(\boldsymbol{A}_1'\boldsymbol{\Psi}\boldsymbol{A}_1)$. From (4), the R_c^2 is

$$R_c^2 = \frac{\sum_{j=1}^k \pi_j \|\boldsymbol{\mu}_j\|^2}{tr(\boldsymbol{\Psi})} = \frac{\sum_{j=1}^k \pi_j \|\boldsymbol{A}_1'\boldsymbol{\mu}_j\|^2}{tr(\boldsymbol{\Psi})} \leq \frac{\sum_{j=1}^k \pi_j \|\boldsymbol{A}_1'\boldsymbol{\mu}_j\|^2}{tr(\boldsymbol{A}_1'\boldsymbol{\Psi}\boldsymbol{A}_1)},$$

and this last expression is the R_c^2 for $\boldsymbol{A}_1'\boldsymbol{y}$ with self-consistent points $\boldsymbol{A}_1'\boldsymbol{\mu}_j, j = 1, \ldots, k$. ■

Proof of Theorem 3

Proof. The mean of the distribution (which we are taking as the origin) always lies in the convex hull of a set of k self-consistent points. Therefore, the linear span of any set of k self-consistent points has dimension at most $k-1$. Let $\lambda_1(p) \geq \lambda_2(p) \geq \cdots \geq \lambda_p(p)$ denote the ordered eigenvalues of $\boldsymbol{\Psi}(p)$. For any given self-consistent partition of the distribution into k clusters, let $\boldsymbol{A}_1(p)$ denote a $p \times (k-1)$ matrix with orthonormal columns that spans the space spanned by a set of k (self-consistent) cluster means $\boldsymbol{\mu}_1(p), \ldots, \boldsymbol{\mu}_k(p)$ and let $\boldsymbol{A}_2(p) \sim p \times (p-k+1)$ be a matrix such that $\boldsymbol{A}(p) = [\boldsymbol{A}_1(p) : \boldsymbol{A}_2(p)]$ is orthogonal. For notational convenience, we shall drop the dependence on the dimension p. Let $D_j \subset \Re^p$

denote the jth cluster. The $R_c^2(p)$ is given by

$$\begin{aligned}
R_c^2(p) &= \frac{tr(\boldsymbol{\Psi}) - \sum_{j=1}^{k} \int_{D_j} \|\boldsymbol{y} - \boldsymbol{\mu}_j\|^2 dF(\boldsymbol{y})}{tr(\boldsymbol{\Psi})} \\
&= \frac{tr(\boldsymbol{A}_1'\boldsymbol{\Psi}\boldsymbol{A}_1) + tr(\boldsymbol{A}_2'\boldsymbol{\Psi}\boldsymbol{A}_2) - \sum_{j=1}^{k} \int_{D_j} \|\boldsymbol{A}_1'(\boldsymbol{y} - \boldsymbol{\mu}_j)\|^2 + \|\boldsymbol{A}_2'\boldsymbol{y}\|^2 dF(\boldsymbol{y})}{tr(\boldsymbol{\Psi})} \\
&= \frac{tr(\boldsymbol{A}_1'\boldsymbol{\Psi}\boldsymbol{A}_1) + tr(\boldsymbol{A}_2'\boldsymbol{\Psi}\boldsymbol{A}_2) - \sum_{j=1}^{k} \int_{D_j} \|\boldsymbol{A}_1'(\boldsymbol{y} - \boldsymbol{\mu}_j)\|^2 dF(\boldsymbol{y}) - tr(\boldsymbol{A}_2'\boldsymbol{\Psi}\boldsymbol{A}_2)}{tr(\boldsymbol{\Psi})} \\
&= \frac{tr(\boldsymbol{A}_1'\boldsymbol{\Psi}\boldsymbol{A}_1) - \sum_{j=1}^{k} \int_{D_j} \|\boldsymbol{A}_1'(\boldsymbol{y} - \boldsymbol{\mu}_j)\|^2 dF(\boldsymbol{y})}{tr(\boldsymbol{\Psi})} \\
&\leq \frac{\sum_{i=1}^{k-1} \lambda_i}{\sum_{i=1}^{p} \lambda_i} \\
&\leq \frac{(k-1)\lambda_1}{\sum_{i=1}^{p} \lambda_i} \\
&\to 0 \text{ as } p \to \infty,
\end{aligned}$$

since λ_1 is assumed bounded.

If the overall variability remains bounded, then $tr(\boldsymbol{A}_2'\boldsymbol{\Psi}\boldsymbol{A}_2) < M < \infty$ for some M as $p \to \infty$ and the R_c^2 becomes

$$R_c^2 = \frac{tr(\boldsymbol{A}_1'\boldsymbol{\Psi}\boldsymbol{A}_1) - \sum_{j=1}^{k} \int_{D_j} \|\boldsymbol{A}_1'(\boldsymbol{y} - \boldsymbol{\mu}_j)\|^2 dF(\boldsymbol{y})}{tr(\boldsymbol{A}_1'\boldsymbol{\Psi}\boldsymbol{A}_1) + tr(\boldsymbol{A}_2'\boldsymbol{\Psi}\boldsymbol{A}_2)} > \frac{tr(\boldsymbol{A}_1'\boldsymbol{\Psi}\boldsymbol{A}_1) - \sum_{j=1}^{k} \int_{D_j} \|\boldsymbol{A}_1'(\boldsymbol{y} - \boldsymbol{\mu}_j)\|^2 dF(\boldsymbol{y})}{tr(\boldsymbol{A}_1'\boldsymbol{\Psi}\boldsymbol{A}_1) + M}.$$

In this case, the R_c^2 will not go to zero as the dimension increases. ■

Proof of Theorem 4

Proof. Without loss of generality we can assume that the dominant eigenvalue is simple, so that the dominant eigenvector is uniquely defined, up to a proportionality constant. Let z be a random variable such that the joint distribution of $\boldsymbol{y}$ and z is

$$\begin{pmatrix} \boldsymbol{y} \\ z \end{pmatrix} \sim ELL_{p+1} \left[\begin{pmatrix} \mathbf{0}_p \\ 0 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Psi} & \boldsymbol{\delta} \\ \boldsymbol{\delta}' & 1 \end{pmatrix}, \phi \right],$$

where $\boldsymbol{\delta} = \boldsymbol{\Psi}\boldsymbol{\alpha}/\sqrt{1 + \boldsymbol{\alpha}'\boldsymbol{\Psi}\boldsymbol{\alpha}}$, so that $\boldsymbol{\alpha} = \boldsymbol{\Psi}^{-1}\boldsymbol{\delta}/\sqrt{1 - \boldsymbol{\delta}'\boldsymbol{\Psi}^{-1}\boldsymbol{\delta}}$. Since $\boldsymbol{\alpha}$ also belongs to the subspace generated by the nondominant eigenvectors of $\boldsymbol{\Psi}$, then $\boldsymbol{\delta}$ also belongs to this subspace. Let $y_1 = \boldsymbol{v}'\boldsymbol{y}/\lambda_1$, where $\boldsymbol{v}$ is a unit-length eigenvector associated with λ_1:

$\boldsymbol{\Psi}\boldsymbol{v} = \lambda_1\boldsymbol{v}$ and $\boldsymbol{v}'\boldsymbol{v} = 1$. Then the joint distribution of $\boldsymbol{y}$, y_1 and z is

$$\begin{pmatrix} \boldsymbol{y} \\ y_1 \\ z \end{pmatrix} \sim ELL_{p+2}\left[\begin{pmatrix} \mathbf{0}_p \\ \mathbf{0}_2 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Psi} & \boldsymbol{A} \\ \boldsymbol{A}' & \boldsymbol{I}_2 \end{pmatrix}, \phi\right].$$

where $\boldsymbol{I}_2$ is a 2×2 identity matrix, $\mathbf{0}_2$ is a 2-dimensional null vector and $\boldsymbol{A}$ is a $p\times 2$ matrix whose first and second columns are $\sqrt{\lambda_1}\boldsymbol{v}$ and $\boldsymbol{\delta}$, respectively. Zero correlation between z and y_1 follows from $\boldsymbol{\delta}$ belonging to the linear space generated by eigenvectors of $\boldsymbol{\Psi}$ different from $\boldsymbol{v}$. The joint distribution of $\boldsymbol{w} = \boldsymbol{y} - \sqrt{\lambda_1}\boldsymbol{v}y_1 - \boldsymbol{\delta}z$, y_1 and z is

$$\begin{pmatrix} \boldsymbol{w} \\ y_1 \\ z \end{pmatrix} \sim ELL_{p+2}\left[\begin{pmatrix} \mathbf{0}_p \\ \mathbf{0}_2' \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Psi} - \lambda_1\boldsymbol{v}\boldsymbol{v}' - \boldsymbol{\delta}\boldsymbol{\delta}' & \mathbf{O} \\ \mathbf{O}' & \boldsymbol{I}_2 \end{pmatrix}, \phi\right],$$

where $\mathbf{O}$ is a $p \times 2$ null matrix. The distribution of $(\boldsymbol{w}', y_1, z)'$ depends on y_1 and z only through their squared values y_1^2 and z^2, implying that $(\boldsymbol{w}', y_1, z)'$, $(\boldsymbol{w}', -y_1, z)'$, $(\boldsymbol{w}', -y_1, -z)'$ and $(\boldsymbol{w}', y_1, -z)'$ are identically distributed. As a direct consequence, the vectors $\boldsymbol{w}$, $\boldsymbol{w}|\{y_1 > 0, z > 0\}$, $\boldsymbol{w}|\{y_1 < 0, z > 0\}$, $\boldsymbol{w}|\{y_1 < 0, z < 0\}$ and $\boldsymbol{w}|\{y_1 > 0, z < 0\}$ are identically distributed, too. The expected value $E(\widetilde{\boldsymbol{y}}|y_1 > 0) = E(\boldsymbol{y}|y_1 > 0, z > 0)$ is then

$$\begin{aligned} &E(\boldsymbol{w}|y_1 > 0, z > 0) + \sqrt{\lambda_1}\boldsymbol{v}E(y_1|y_1 > 0, z > 0) + \boldsymbol{\delta}E(z|y_1 > 0, z > 0) \\ &= E(\boldsymbol{w}) + \sqrt{\lambda_1}\boldsymbol{v}E(y_1|y_1 > 0) + \boldsymbol{\delta}E(z|z > 0) = \sqrt{\lambda_1}\boldsymbol{v}E(|y_1|) + \boldsymbol{\delta}E(|z|). \end{aligned}$$

The last identity follows from symmetry of $\boldsymbol{w}$, y_1 and z. Further simplification is achieved by recalling that z, y_1 and u are identically distributed: $E(\widetilde{\boldsymbol{y}}|y_1 > 0) = E(|u|)\left(\boldsymbol{\delta} + \sqrt{\lambda_1}\boldsymbol{v}\right)$. In a similar way we can prove that $E(\widetilde{\boldsymbol{y}}|y_1 < 0) = E(|u|)\left(\boldsymbol{\delta} - \sqrt{\lambda_1}\boldsymbol{v}\right)$.

The same argument implies that $\boldsymbol{y} - \boldsymbol{\delta}z$ and z are mutually independent, thus implying $E(\widetilde{\boldsymbol{y}}) = E(\boldsymbol{y}|z > 0) = \boldsymbol{\delta}E(|u|) = \boldsymbol{\mu}$. The covariance matrix of $\boldsymbol{y}$ is proportional to $\boldsymbol{\Psi}$, and the proportionality constant is $E(u^2)$, due to the elliptical distribution of $\boldsymbol{y}$. Since $\widetilde{\boldsymbol{y}}\widetilde{\boldsymbol{y}}'$ is an even function of $\widetilde{\boldsymbol{y}}$, its distribution is the same as the distribution of $\boldsymbol{y}\boldsymbol{y}'$ [41], and consequently $E\left(\widetilde{\boldsymbol{y}}\widetilde{\boldsymbol{y}}'\right) = E(u^2)\boldsymbol{\Psi}$. Hence the covariance matrix of $\widetilde{\boldsymbol{y}}$ is $cov(\widetilde{\boldsymbol{y}}) = E(u^2)\boldsymbol{\Psi} - E^2(|u|)\boldsymbol{\delta}\boldsymbol{\delta}'$. The vector $\boldsymbol{\delta}$ belongs to the linear subspace spanned by the nondominant eigenvectors of $\boldsymbol{\Psi}$, so that $\boldsymbol{\delta}'\boldsymbol{v} = 0$ and $\boldsymbol{v}$ is an eigenvector of $cov(\widetilde{\boldsymbol{y}})$ associated with the

eigenvalue λ_1, too. Also, the ith largest eigenvalue of $cov\left(\widetilde{\boldsymbol{y}}\right)$ is never larger than the ith largest eigenvalue of $\boldsymbol{\Psi}$. As a direct consequence, λ_1 and $\boldsymbol{v}$ are the dominant eigenvalue and eigenvector of $cov\left(\widetilde{\boldsymbol{y}}\right)$, respectively.

As noted above, two principal points of an elliptical distribution lie on the subspace generated by the dominant eigenvector of its covariance matrix. The two principal points of $\boldsymbol{y}$ are then $E\left(\boldsymbol{y}|y_1 > c\right)$ and $E\left(\boldsymbol{y}|y_1 < c\right)$, for some scalar quantity c. In particular, $c = 0$ when principal points are symmetric. Assuming $c = 0$, the Voronoi regions of the two principal points of $\boldsymbol{y}$ is the set of $p-$dimensional vectors whose projections onto the direction of the dominant eigenvector of $\boldsymbol{\Psi}$ have the same sign. The self-consistent points of $\widetilde{\boldsymbol{y}}$ corresponding to these regions are $E\left(\widetilde{\boldsymbol{y}}|y_1 > 0\right)$ and $E\left(\widetilde{\boldsymbol{y}}|y_1 < 0\right)$. The R_c^2 coefficient of the two self-consistent points of $\widetilde{\boldsymbol{y}}$ is therefore

$$\frac{P\left(\widetilde{\boldsymbol{y}}|y_1 > 0\right) \cdot \left\|E\left(\widetilde{\boldsymbol{y}}|y_1 > 0\right)\right\|^2 + P\left(\widetilde{\boldsymbol{y}}|y_1 < 0\right) \cdot \left\|E\left(\widetilde{\boldsymbol{y}}|y_1 < 0\right)\right\|^2}{E\left(u^2\right) \cdot tr\left(\boldsymbol{\Psi}\right) - \boldsymbol{\mu}'\boldsymbol{\mu}},$$

which simplifies to (13) by recalling that $E\left(\widetilde{\boldsymbol{y}}|y_1 > 0\right) = E\left(|u|\right)\left(\boldsymbol{\delta} + \sqrt{\lambda_1}\boldsymbol{v}\right)$, and $E\left(\widetilde{\boldsymbol{y}}|y_1 < 0\right) = E\left(|u|\right)\left(\boldsymbol{\delta} - \sqrt{\lambda_1}\boldsymbol{v}\right)$ and by noticing that the norms of the two expectations are equal, due to the orthogonality of $\boldsymbol{v}$ and $\boldsymbol{\delta}$: $\left\|E\left(\widetilde{\boldsymbol{y}}|y_1 > 0\right)\right\| = \left\|E\left(\widetilde{\boldsymbol{y}}|y_1 < 0\right)\right\| = E^2\left(|u|\right)\left(\lambda_1 + \boldsymbol{\delta}'\boldsymbol{\delta}\right) = \lambda_1 \cdot E^2\left(|u|\right) + \boldsymbol{\mu}'\boldsymbol{\mu}$. ■

References

[1] A. W. F. Edwards and L. L. Cavalli-Sforza. A method for cluster analysis. *Biometrics*, 21:362–375, 1965.

[2] H. P. Friedman and J. Rubin. On some invariant criteria for grouping data. *Journal of the American Statistical Association*, 62(320):1159–1178, 1967.

[3] A. J. Scott and M. J. Symons. Clustering methods based on likelihood ratio criteria. *Biometrics*, 27:387–397, 1971.

[4] W. J. Krzanowski and Y. T. Lai. A criterion for determining the number of groups in a data set using sum-of-squares clustering. *Biometrics*, 44:23–34, 1988.

[5] A. D. Gordon and J. J. Henderson. An algorithm for Euclidean sum of squares classification. *Biometrics*, 33:355–362, 1977.

[6] T. Tarpey and B. Flury. Self–consistency: A fundamental concept in statistics. *Statistical Science*, 11:229–243, 1996.

[7] D. Steinley. K-means clustering: A half-century synthesis. *British Journal of Mathematical and Statistical Psychology*, 59:1–34, 2006.

[8] T. O. Kvalseth. Cautionary note about R^2. *The American Statistician*, 39:279–285, 1985.

[9] G. W. Milligan. A validation study of a variable weighting algorithm for cluster analysis. *Journal of Classification*, 6:53–71, 1989.

[10] G. W. Milligan and M. Cooper. A study of standardization of variables in cluster analysis. *Journal of Classification*, 5:181–204, 1988.

[11] J. A. Hartigan and M. A. Wong. A K-means clustering algorithm. *Applied Statistics*, 28:100–108, 1979.

[12] J. H. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963.

[13] F. Murtagh and P. Legendre. Ward's hierarchical agglomerative clustering method: Which algorithms implement Ward's criterion? *Journal of Classification*, 31:274–295, 2014.

[14] L. Graf and H. Luschgy. *Foundations of Quantization for Probability Distributions*. Springer, Berlin, 2000.

[15] N. E. Day. Estimating the components of a mixture of normal distributions. *Biometrika*, 56:463–474, 1969.

[16] Inderjit S. Dhillon, Yuqiang Guan, and Brian Kulis. Kernel k-means: Spectral clustering and normalized cuts. In *Proceedings of the Tenth ACM SIGKDD International*

Conference on Knowledge Discovery and Data Mining, KDD '04, pages 551–556, New York, NY, USA, 2004. ACM.

[17] B. Flury. Estimation of principal points. *Applied Statistics*, 42:139–151, 1993.

[18] H. Mann. Untersuchungen über wabenzellen bie allgemeiner minkowski metrik. *Monatshefte für Mathematik*, 42:417–424, 1935.

[19] B. Flury. Principal points. *Biometrika*, 77:33–41, 1990.

[20] R. Gnanadesikan, J. R. Kettenring, and S. L. Tsao. Weighting and selection of variables for cluster analysis. *Journal of Classification*, 12:113–136, 1995.

[21] D. N. Naik and R. Khattree. Revisiting Olympic track records: Some practical considerations in the principal component analysis. *The American Statistician*, 50:140–144, 1996.

[22] S. Bartoletti, B. Flury, and D. G. Nel. Allometric extension. *Biometrika*, 55:1210–1214, 1999.

[23] S. Matsuura and H. Kurata. Principal points for an allometric extension model. *Statistical Papers*, 55:853–870, 2014.

[24] T. Tarpey. Linear transformations and the k-means clustering algorithm: Applications to clustering curves. *The American Statistician*, 61:34–40, 2007.

[25] N. Loperfido. Skewness and the linear discriminant function. *Statistics and Probability Letters*, 83:93–99, 2013.

[26] N. Loperfido. Vector-valued skewness for model-based clustering. *Statistics and Probability Letters*, 99:230–237, 2015.

[27] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016.

[28] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1996.

[29] H. H. Bock. *On the interface between cluster analysis, principal component analysis, and multidimensional scaling*, pages 17–34. D. Reidel Publishing Company, 1987.

[30] M. Meilă. Comparing clusterings: an information based distance. *Journal of Multivariate Analysis*, 98:873–895, 2007.

[31] W.M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.

[32] L. Hubert and P. Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

[33] E. Petkova, R. T. Ogden, T. Tarpey, A. Ciarleglio, B. Jiang, Z. Su, T. Carmody, P. Adams, H. C. Kraemer, B. Grannemann, M. A. Oquendo, R. V. Parsey, M. Weissman, P. J. McGrath, M. Fava, and M. H. Trivedi. Statistical analysis plan for Stage 1 EMBARC (Establishing Moderators and Biosignatures of Antidepressant Response for clinical Care). *Contemporary Clinical Trials Communications*, 6:22–30, 2017. http://dx.doi.org/10.1016/j.conctc.2017.02.007.

[34] Daniela M. Witten and Robert Tibshirani. A framework for feature selection in clustering. *Journal of the American Statistical Association*, 105(490):713–726, 2010.

[35] Daniela M. Witten and Robert Tibshirani. *sparcl: Perform sparse hierarchical clustering and sparse k-means clustering*, 2013. R package version 1.0.3.

[36] T. Tarpey, L. Li, and B. Flury. Principal points and self–consistent points of elliptical distributions. *Annals of Statistics*, 23:103–112, 1995.

[37] K. T. Fang, S. Kotz, and K. W. Ng. *Symmetric Multivariate and Related Distributions*. Chapman and Hall, London, 1990.

[38] M. D. Branco and D. K. Dey. A general class of skew-elliptical distributions. *Journal of Multivariate Analysis*, 79:99–113, 2001.

[39] K. Pearson. Contributions to the mathematical theory of evolution. ii. skew variation in homogeneous material. *Philisophical Transactions of the Royal Society of London*, 186:343–414, 1895.

[40] T. Tarpey, D. Yun, and E. Petkova. Model misspecification: Finite mixture or homogeneous? *Statistical Modelling*, 8:199–218, 2009.

[41] M. G. Genton and N. Loperfido. Generalized skew-elliptical distributions and their quadratic forms. *Annals of the Institute of Statistical Mathematics*, 57:389–401, 2005.

[42] M. Christiansen and N. Loperfido. Improved approximation of the sum of random vectors by the skew-normal distribution. *Journal of Applied Probability*, 51:466–482, 2014.

[43] S. Bartoletti and N. Loperfido. Modelling air pollution data by the skew-normal distribution. *Stochastic Environmental Research & Risk Assessment*, 24:513–517, 2010.

[44] C. H. Chang, J. J. Lin, Pal N, and M. C. Chiang. A note on improved approximation of the binomial distribution by the skew-normal distribution. *The American Statistician*, 62:167–170, 2008.

[45] J. O. Ramsay and B. W. Silverman. *Applied Functional Data Analysis*. Springer, New York, 2002.

[46] J-M. Chiou and P-L. Li. Functional clustering and identifying substructures of longitudinal data. *Journal of the Royal Statistical Society, Series B*, 69:679–699, 2007.

[47] C. Abraham, P. A. Cornillon, E. Matzner-Lober, and N. Molinari. Unsupervised curve clustering using B-splines. *Scandinavian Journal of Statistics*, 30:581–595, 2003.

[48] S. Ray and B. Mallick. Functional clustering by Bayesian wavelet methods. *Journal of the Royal Statistical Society, Series B*, 68:305–332, 2006.

[49] T. Tarpey, E. Petkova, Y. Lu, and U. Govindarajulu. Optimal partitioning for linear mixed effects models: Applications to identifying placebo responders. *Journal of the American Statistical Association*, 105(491):968–977, 2010.

[50] C. de Boor. *A Practical Guide to Splines*. Springer-Verlag, New York, 1978.

[51] P. J. McGrath, J. W. Stewart, E. Petkova, F. M. Quitkin, J. D. Amsterdam, J. Fawcett, F. W. Reimherr, J. F. Rosenbaum, and C. M. Beasley. Predictors of relapse during

fluoxetine continuation or maintenance treatment for major depression. *Journal of Clinical Psychiatry*, 61(7):518–524, 2000.

[52] L. Kaufman and P.J. Rousseeuw. *Finding Groups in Data: an introduction to cluster analysis*. Wiley, 1990.

[53] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edition*. Springer, New York, 2009.

[54] Marti J. Anderson. A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, 26:32–46, 2001.

[55] Brian H. McArdle and Marti J. Anderson. Fitting multivariate models to community data: A comment on distance-based redundancy analysis. *Ecology*, 82(1):290–297, 2001.

[56] S. Bouveyron, C. Girard, and C. Schmid C. High-dimensional data clustering. *Computational Statistics & Data Analysis*, 52:502–519, 2007.

[57] P. D. McNicholas. Model-based clustering. *Journal of Classification*, 33:331–373, 2016.

[58] R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a dataset via the gap statistic. *Journal of the Royal Statistical Society, Series B*, 63:411–423, 2001.

[59] C. Sugar and G. James. Finding the number of clusters in a data set: An information theoretic approach. *Journal of the American Statistical Association*, 98:750–763, 2003.

[60] B. A. Clementz, J. A. Sweeney, J. P. Hamm, E. I. Ivleva, L. E. Ethridge, G. D. Pearlson, M. S. Keshavan, and C. A. Tamminga. Identification of distinct psychosis biotypes using brain-based biomarkers. *American Journal of Psychiatry*, 173(4):373–383, 2016.

[61] Jinzhu Jie and Karl Rohe. Preconditioning the lasso for sign consistency. *Electronic Journal of Statistics*, 9:1935–7524, 2015.