Simulation of Non-Gaussian Populations of Images

Blinded

March 10, 2003

Abstract

This paper investigates goodness of fit of the Gaussian distribution in High Dimension, Low Sample Size settings, such as modelling populations of medical images. The driving problem is simulation of synthetic images of 3-d kidney shapes. Independent Component Analysis of a training data set, shows that the population is non-Gaussian. Deeper analysis shows that a power transformation provides reasonable simulation results.

1 Introduction

This paper illustrates the use of non-standard statistical tools, including Independent Component Analysis, and graphically enhanced Q-Q plots, for investigating the goodness of fit of the Gaussian distribution in High Dimension Low Sample Size (HDLSS) contexts. The driving problem is the need for simulation of a population of kidney images, based on a small training sample. It is seen that the Gaussian distribution gives a poor approximation to this population. The analysis suggests an appropriate transformation, which is seen to give a much better approximation, resulting in a carefully tuned method for simulating from a population of kidney shape images. The general statistical techniques are expected to be useful in a wide variety of medical imaging contexts, and to other HDLSS situations as well

Current trends in medical image analysis are in the direction of studying populations of images, often in three dimensions. Because such populations often exhibit complicated HDLSS structure, there is a need for development of new statistical tools. Usually the focus is on the shapes of particular organs, so segmentation of each member of the population is important. Current segmentation methods typically require some human intervention with each image. Thus populations of shapes are expensive to acquire, so sample sizes tend to be small. HDLSS data result because typically large numbers of parameters are needed to effectively represent 3-d objects.

For effective development and testing of new methodologies, an efficient scheme is to compare them on a larger population of simulated images, which "has the same behavior" as the original population. An appealing approach

to the simulation of synthetic images is to represent them as vectors, and then simulate the pseudo population from a suitable multivariate Gaussian distribution. However, this leans heavily on the Gaussian distribution, which it is advisable to check carefully.

These issues are studied here in the context of a population of CT images of human kidneys, provided by the Department of Radiation Oncology at the University of North Carolina, and previously studied by Blinded. There are 36 healthy and normal looking samples with no obvious outliers. The kidney shapes are used as a starting point of a more complex longer term project with the ultimate goal of generating a large number of synthetic medical images and shapes for segmentation performance characterization. The human kidney was chosen as a first step in this program because of its relatively simple shape. Humans normally have a pair of kidneys. For this analysis only the right kidney of each individual is used.

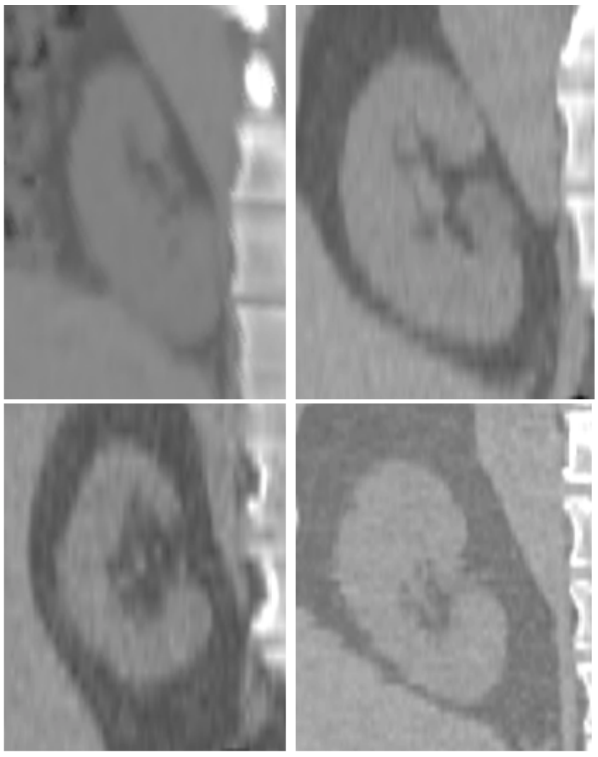

Figure 1: *Coronal view of four kidneys from training data sets.*

Figure 1 shows a coronal view of CT images of four different human kidneys. The variation between these kidneys is apparent in this figure. We shall not be directly concerned with this variation in the raw data, as the kidneys will be

registered first as described briefly in Section 2. Of interest here is the variation between kidneys that remains after registration.

HDLSS settings such as this arise commonly in the statistical analysis of populations of medical images, and provide special statistical challenges. For example, most classical multivariate analysis methods are often nearly useless, because it is impossible to "sphere the data" (since the covariance is not of full rank).

Alternate methods are developed to carefully check the assumption of Gaussianity from several viewpoints in this paper. Standard Principal Component Analysis, done in Section 3, shows that marginal distributions of the projections appear to be roughly Gaussian. But for high dimensional data, it is very dangerous to consider only marginal distributions.

An appealing approach to "searching for directions of non-Gaussianity" is Independent Component Analysis, motivated in Section 4. We extend the ICA methodology to give a formal statistical hypothesis test of Gaussianity. For the kidney data the ICA based test in Section 4.2 reveals strongly significant non-Gaussian behavior. In particular, a number of outliers are revealed. There are too many for outlier deletion to be sensible, see Sections 4.1 and 4.3, so some modification of the Gaussian model is a feasible approach to generating data with these characteristics.

General high dimensional modelling is impossible with so few observations, but an approach based on power transformation appears useful. The preliminary choice of power is done in Section 5. A useful tool in the selection of the appropriate power is a graphically enhanced version of the Q-Q plot.

The simulation properties of our Gaussian power transformation model are studied in Section 6, where we show that the IC analysis gives results similar to those for the real data.

2 Kidney Shape Representations

Human kidneys consist of a pair of kidneys. A single kidney, the right one, is used here. Because of its relatively simple shape it is a good first step in the development of models for creating synthetic images based on a few samples. In this paper we are interested in modelling the shape of the kidney. We characterize kidney shape by a set of "fiducial" points, selected as in Blinded. Figure 2 displays a typical kidney boundary in three dimensions, with highlighted fiducial points. Fiducial points are mostly associated with salient geometric features on a surface. The location and number of the fiducial points is determined via an iterative procedure, starting with a seed set containing a small number of points, with large surface curvature. Based on these points the shape is reconstructed and volume comparisons are made between the reconstructed and the original shape. The number of fiducial points is increased until the difference between model reconstruction and original shape, averaged over the entire training set, is smaller than a predefined threshold. Typically the discrepancy obtained from inter-user manual segmentation by medical experts is used as such a threshold.

For the kidney samples 88 fiducial points were found to adequately describe its shape in this sense. The (x, y, z) coordinates of these 88 fiducial points are put into a single 264 dimensional vector, representing the kidney shape.

Figure 2: *Single kidney with fiducial points.*

As expected from the variation seen in Figure 1, normal kidneys differ in size and shape. A registration procedure was applied between a kidney "template" and the remaining kidneys in the sample. The first step in this procedure is an affine scaling transformation which equalizes the size of the objects of interest in the template and the training images prior to the non-linear registration step. The 88 fiducial points were determined on the template and, via the registration function, mapped onto fiducial points of the remaining 35 kidneys. This registered data is referred to as the "scaled data" and is used when one is primarily interested in "pure shape". Scaling can introduce outliers in the data. For the purpose of generating synthetic samples it is therefore preferable to work with the "unscaled" data, that is, data for which the effect of scaling has been removed as a final step in the registration process. For more details, see Blinded. In this analysis we shall only be using the *unscaled* registered data consisting of the 88 fiducial points.

3 Principal Component Analysis

The 36 vectors, representing the locations of the 88 fiducial points for each kidney, are organized so that the 88 x coordinates appear first, the 88 y coordinates second, and finally the 88 z coordinates. This organization allows useful insights, using the parallel coordinates view shown in Figure 3. Parallel coordinates were proposed by Inselberg (1985) and Inselberg and Dimsdale (1987) as a means of visualizing high dimensional data. In this view, the entries of each data vector are plotted as a function of the coordinate number (thus $1, ..., 264$ for these data), with different colors indicating the different data vectors.

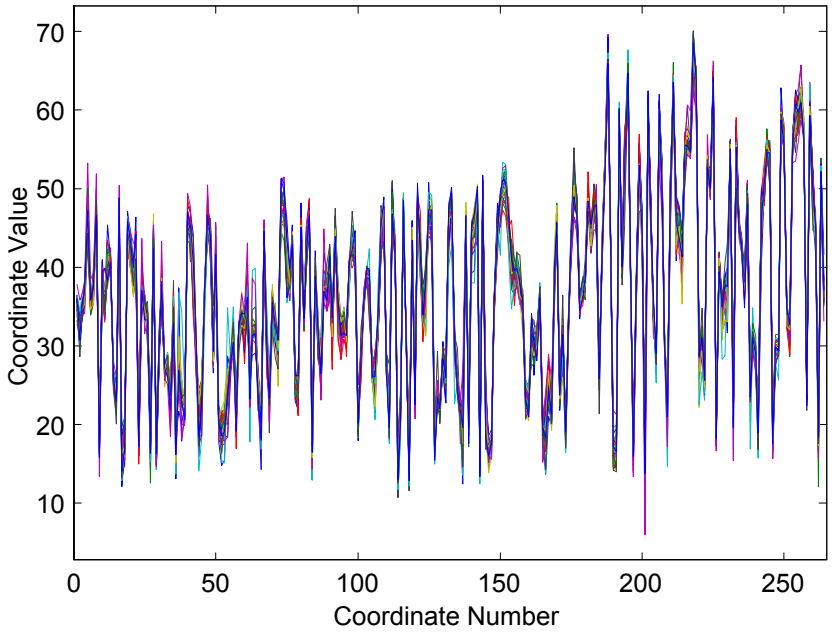

Figure 3: *Parallel coordinate view of kidney data*

A preliminary investigation, based on principal component analysis, determined the number of useful PC directions to be 7, because the magnitude of the remaining components was less than the inter-user manual segmentation error, see Blinded. Our analysis is therefore based on these first 7 PC directions only, thus reducing the data from 264×36 to 7×36. Figure 4 shows the first three directions of the PCA.

The three left panels of Figure 4 give intuitive insight into the PC directions (eigenvectors), by plotting the projections of the data onto each eigenvector. This display is a parallel coordinate plot as in Figure 3. The variability explained by PC1 (21.5% of the mean residual sum of squares, i.e. on the usual R^2 scale) appears uniformly spread across coordinates. Considering Figure 3

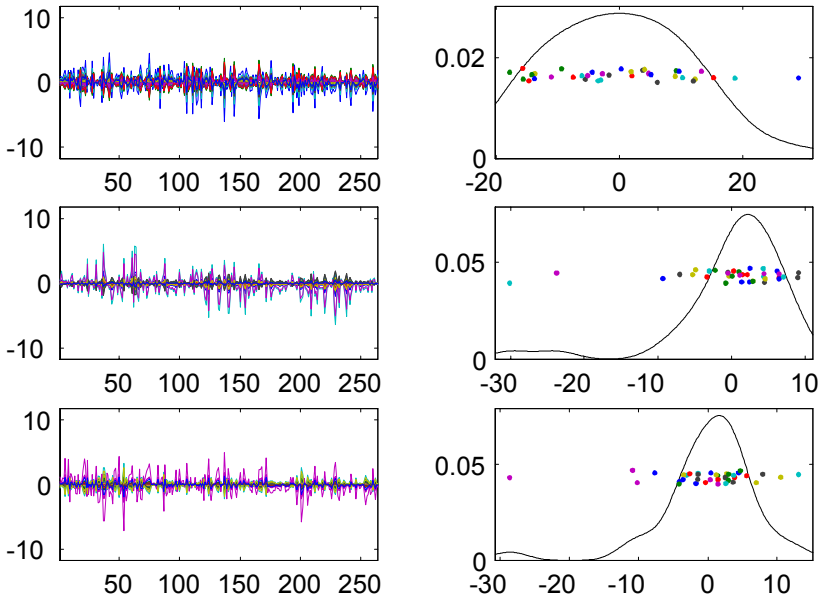

Figure 4: *First three PC directions and PC scores*

one might expect the z directions, given by the last third of the data, to dominate PC1, as they appear to have more variation than the x and y directions. However this greater variation disappears when the mean is subtracted from the original data, and it is this mean corrected data which forms the basis for the PC analysis. In PC2 two records are more noticeable, indicated by the blue and purple data records. These records are picked up as outliers in the middle plot of the right panel, but they are not noticeable in PC1. PC2 accounts for 10.7% of the mean residual sum of squares. PC3 (8.5% of the mean residual sum of squares), and the other components did not show much useful structure. The three right panels in Figure 4 present another view of the data, this time showing the one-dimensional distributions of the coefficients of the projections (also called PC scores). In these plots, the distributions are represented in two ways. The curves are essentially "smooth histograms", constructed as kernel density estimates, see Wand and Jones (1995). The second representation is a "jitter plot" (see e.g. Cleveland (1993)), where the data points are shown as colored dots (useful for identification across plots), with a random height added for visual separation. These marginal distributions are a natural place to check for Gaussianity. An enhanced Q-Q plot analysis of the PC scores, as in Section 5, showed that the PC2 marginal distribution was not quite Gaussian, because of the two outliers apparent in the middle right panel of Figure 4. PC1 and the remaining PC marginal distributions cannot be distinguished from the Gaussian. It is tempting to conclude on the basis of this analysis, that the

data are reasonably Gaussian. However, this view is naive, because it considers only a very few directions, while many more directions (having possibly non-Gaussian behavior) are present in high dimensional space. While PCA was not adequate in determining the distributional properties of our data, it provided a valuable tool in the reduction of the dimension of the data, used in the rest of this paper, to a relatively small number of useful components. Our notation for PCA is carefully developed in Appendix 7.1.

Another simple test of Gaussianity is to study the one dimensional marginal distributions of the individual coordinates of the raw data (as shown in Figure 3). We applied the Bera-Jarque moment test and the Lilliefors CDF test, implemented in Matlab, and found that 37.5% (respectively 29%) of the coordinates were significant at the 5% level. This is some suggestion that the data may not be Gaussian, but does not account for the multiple comparison issue, and also seems impossible to interpret because strong correlation is expected between these variables. Also it is not clear how to usefully exploit this information to obtain an improved simulation.

Because of these problems with classical tests of Gaussianity, in Section 4 we propose using Independent Component Analysis as a new and powerful approach. A major benefit of this approach is that it points the way towards a useful simulation model.

4 Independent Component Analysis

The Independent Component Analysis method comes from the signal processing literature, where it was developed as a powerful method of "blind source extraction". Good detailed discussion of this method can be found in Hyvärinen, Karhunen and Oja (2001). A quick and accessible introduction, with access to Matlab software (that was used in this paper) is available in Hyvärinen and Oja (1999).

Our application of ICA is nonstandard, in that we simply use it as an algorithm for finding directions that are "maximally non-Gaussian". This behavior is the key to its excellent blind source extraction properties. Non-Gaussianity was also studied much earlier than ICA, in the context of Projection Pursuit, see Friedman and Tukey (1974), Friedman (1987) and Jones and Sibson (1987). But our application is different, using the maximal non-Gaussianity principal instead as the basis of our statistical test of multivariate Gaussianity.

Generally ICA is an iterative algorithm that attempts to produce an entire new coordinate system, with the first IC direction the "most non-Gaussian", so we base our test only on that. An important aspect in this iterative search is the choice of criterion of non-Gaussianity of the one dimensional projections. Typical candidates are skewness and kurtosis, but other criteria are also available, and have interesting relationships between each other, see Hyvärinen, Karhunen and Oja (2001). We considered skewness and kurtosis because of their statistical interpretability. We found very similar results for both, but to save space only report results for skewness here.

4.1 Maximally Non-Gaussian IC1 Directions

The first step in the ICA is a "whitening" or "sphering" of the data, that is, uncorrelating the data by multiplying by the root inverse covariance matrix. The full 264×36 kidney data set cannot be sphered, because the covariance matrix is not of full rank. Hence we sphere the data after reduction to the first 7 PCs, as described in Appendix 7.1.

Figure 5 illustrates the application of ICA to the kidney data. The top left panel shows the univariate distribution of the projection coefficients onto the first IC direction, in the same smoothed histogram format as shown in the three right panels of Figure 4. One obvious difference between the two figures is the scale on the horizontal axis which is much smaller in Figure 5, because the data have been normalized in the sphering process. The distribution in the top left panel of Figure 5 is similar to the Figure 4 PC2 distribution in looking largely Gaussian, but with a notable outlier. However, the outlier is now farther out (i.e. more standard deviations from the mean), not surprising, since this direction is "most non-Gaussian". This shows the potential of ICA, as a more powerful method than PCA, for finding non-Gaussian directions.

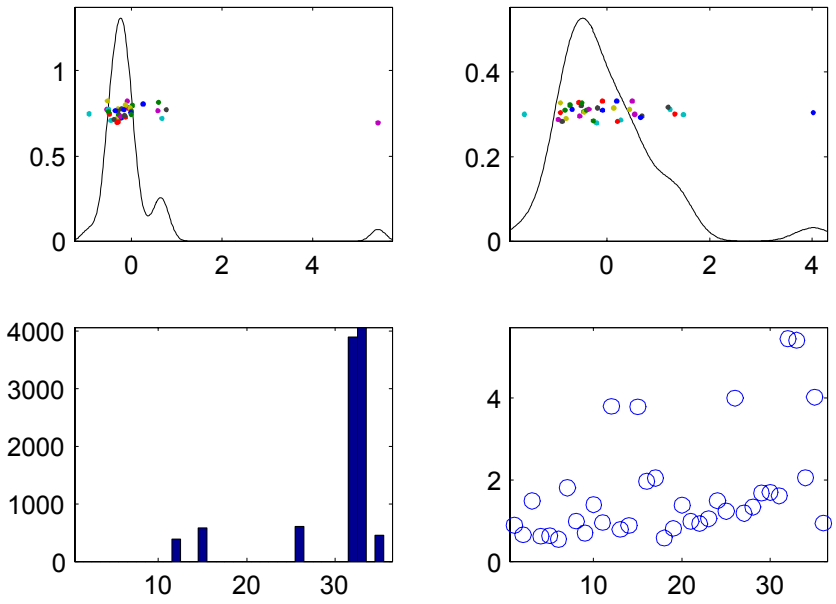

Figure 5: *Maximally non-Gaussian IC1 directions*

Because it is an iterative method, different runs of ICA (even for the same input data) can give different answers (since it uses a randomly chosen starting value). This is caused by the ICA optimization problem having several local optima, with different answers found with different starting values. The top

right panel shows the first ICA direction for another run (thus different starting value). This distribution also has an outlier, but the two outliers in the two top panels have different colors. The outliers correspond to cases number 32 (in the left panel) and 35 (in the right top panel).

To analyze this ambiguity in the ICA directions, we calculated the first independent component from the sphered data 10000 times and indexed each run by the case number of the coefficient with largest absolute value. Six distinct directions were found, driven by different outliers, and two of these directions are much larger than the other four. The results are summarized in the bottom panels of Figure 5. The lower left panel shows the relative frequencies of these six cases $(12, 15, 26, 32, 33$ and $35)$ with case number on the horizontal axis. The lower right panel provides a different view. For each IC1 run we calculate the projection coefficients onto IC1 as in the top two panels of the figure. We display case number on the horizontal axis, and for each case number we find the maximum absolute value of the projection coefficients over the 10000 IC1 runs. This maximum value is displayed on the vertical axis. This analysis shows quite clearly that two case numbers (namely 32 and 33) are separated by their absolute value above 5, and that another four case numbers also have much larger values (about 4) than the rest of the sample. It further shows that the six outliers are four to five standard deviations from the mean, in some of the IC1 directions. It is not surprising that the larger the outlier, the more frequently it is found by ICA (as shown in the bottom left panel). The distribution of case 33 looks very similar to that of case 32 which is shown in the top left panel of the figure. The main difference is the outlier as indicated by a different color in the jitter plot. Similarly the distributions of the smaller outlier cases $12, 15,$ and 26 have shapes similar to that of case 35 which is shown in the top right panel of the figure.

Another consequence of the iterative search algorithm is that ICA may fail to find a non-Gaussian direction. We observed this occasionally. It happened most frequently for simulated Gaussian data, which might be expected to sometimes give a vague "direction of maximal non-Gaussianity".

4.2 ICA Based Test of Gaussianity

The ICA direction vectors are determined sequentially starting with the most non-Gaussian direction. To test for Gaussianity using ICA, it therefore suffices to consider the first independent component. ICA is a powerful method and will usually find something non-Gaussian even in Gaussian data. These observations form the basis for our tests of Gaussianity.

For the kidney data six distinct IC1 directions were found as shown in the bottom left panel of Figure 5. The directions 32 and 33 have an absolute skewness value above 4.65 and 4.46 respectively, and the remaining four outlier cases $12, 15,$ 26 and 35 have absolute skewness values of 1.41, 1.87, 1.86 and 1.88 respectively. These last four outliers may not be clearly distinguishable from the Gaussian case. To eliminate this situation where IC1 converges on one of these four smaller directions, we have found it sufficient to run IC1 ten times

and to pick the direction which has maximum absolute skewness. We use this maximum IC1 direction as our test statistic and compare it with the corresponding maximum direction obtained for data generated from the Gaussian distribution $N_{7\times 36}(0, I)$. For 1000 drawings from the Gaussian distribution we calculate the p-value as the proportion of times that the simulated Gaussian maximum directions exceed the test statistic. Figure 6 shows the result of this test. On the horizontal axis we have displayed absolute skewness. The results for the 1000 runs are given in a smoothed histogram with the individual values displayed as points at random heights – similar to the jitter plots in Figure 3.

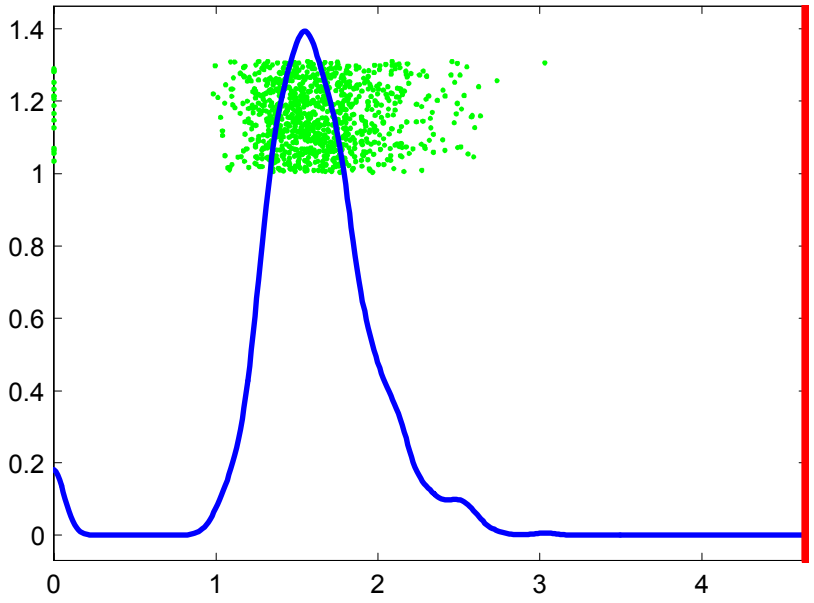

Figure 6: *Skewness hypothesis test of IC1 direction*

The p-value of this test is zero. The red vertical line at 4.65 marks the maximum absolute skewness obtained from the kidney data. For Gaussian data skewness is close to zero. ICA, however, finds non-Gaussian directions even in Gaussian data, and the maximum over 10 IC1 directions will therefore usually result in a strictly positive absolute skewness value . As can be seen in Figure 6, some of the runs have a maximum absolute skewness value of zero. The reason for this is that in 16 of the 1000 runs ICA did not find *any* non-Gaussian directions and failed to converge in 1000 iterations for all 10 starting values. For these runs we define the maximum absolute skewness to be zero.

The results obtained from the corresponding runs based on maximizing kurtosis are very similar and also produce a p-value of zero.

4.3 Analysis of Outliers

The non-Gaussian directions are connected to the outlying cases 32, 33, 12, 15, 26 and 35. To examine whether these particular kidneys represent outliers, we removed the two largest outliers from the original kidney data and carried out the PC and IC1 analysis described above on the reduced data for the remaining 34 records. The first seven principal components were calculated, the PCA data was sphered, and IC1 was calculated for this data. The hypothesis test on the reduced data – similar to that leading to Figure 6 above – has a test statistic with maximum absolute skewness of 3.38, and a p-value of 0.0014, showing that the reduced data is also non-Gaussian. In this test 1000 drawings from the normal distribution $N_{7\times 34}(0, I)$ were used.

The top two panels of Figure 7 show the smoothed histograms of two distinct IC1 directions, corresponding to case numbers 15 (on the top left) and 7 (on the top right), in a similar form to the top two panels in Figure 5. As in Figure 5, the distributions contain distinct outliers, marked by the different colors in the two jitter plots. The lower two panels of Figure 7 contain similar information to those of Figure 5. In both figures case numbers are shown in the horizontal direction. The bottom left panel contains the frequency for each of three IC1 directions that were found in 10000 repetitions of IC1 for the reduced data, and the bottom right panel shows the maximum absolute value by case number as in Figure 5. For convenience of notation we use the original case numbers, with purple vertical lines indicating where data records have been deleted. As can be seen the IC1 analysis found four outliers, two of them (cases 15 and 35) were present in the original analysis and are shown in Figure 5, and two new ones (cases 20 and 7) were not apparent in the full data. Although the absolute skewness of the new outlier case 20 is comparable to that of cases 15 and 35 it was not found in the original analysis. In the presence of the very strong outlier cases 32 and 33, these new outliers were apparently not 'strong enough' to be found in the original ICA – see also the bottom panel of Figure 5. The outlying records 12 and 26 which also appear in the original analysis – see Figure 5 – are not observed as outliers in the bottom left panel of Figure 7, however their maximum absolute values are considerably larger than that of the majority of records as can be seen in the bottom right panel. It is worth noting that the maximum absolute value of outliers 12, 15, 26 and 35 are all comparable in the original analysis, but the absolute skewness of outlier 12 in particular is considerably smaller than that of the other three outliers.

The analysis with the reduced data set has uncovered two further outliers. This raises the question whether the data is a mixture of distributions and the seven outliers form a cluster. We resolve this by calculating the pairwise angles for all eight outlying directions. As above, we have identified the IC1 direction with the case number of the coefficient with largest absolute value. Each of the eight outlier direction vectors is of dimension seven, corresponding to the seven PC directions. To calculate the angle between two such vectors, we first normalize them and then use the fact that the dot or scalar product of two vectors of unit length equals the cosine of the angle between them. For the

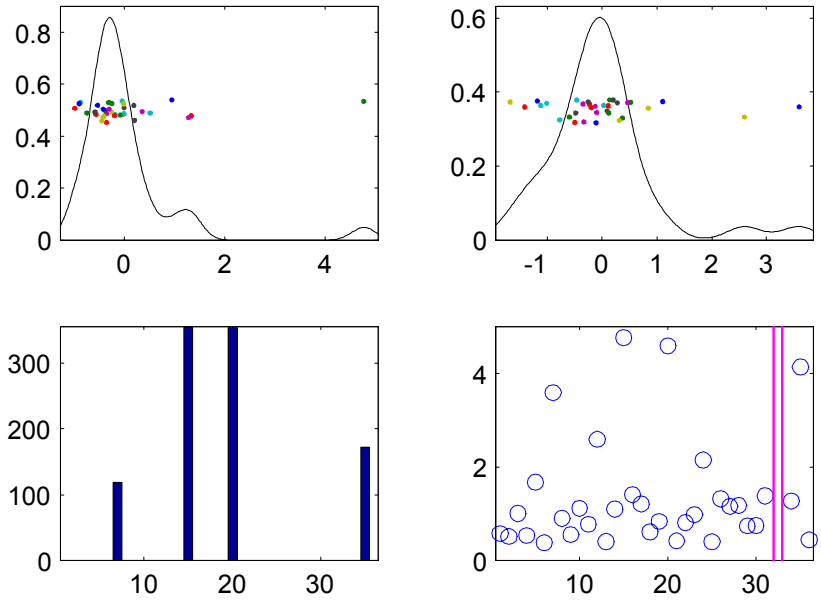

Figure 7: *IC1 directions for reduced data*

six outliers found in the original analysis described in Section 4.1, the pairwise angles are nearly orthogonal, and standardized to a range of 0 to 90 degrees the smallest pairwise angle of 76 degrees occurs between case numbers 15 and 35 and between 26 and 35. For all eight outlier directions – which include the two new ones – both new outlier cases 7 and 20 result in pairwise angles as low as around 60 degrees with some of the smaller original outliers. This shows that the outliers do not form a cluster, and suggests that the data is not a mixture of two separated distributions.

Since six outliers appeared in the original analysis presented in Figure 5, we also examined deleting more than the largest two outliers and repeating the IC analysis and Gaussianity test for those reduced data sets. In all cases, more, and sometimes new, outliers appeared, with the results similar to the ones presented above.

An inspection of the original kidney records for these "outlier" cases was inconclusive. These kidneys did not appear to be clinically different from the other normal kidneys and could therefore not be regarded as outliers in any meaningful medical sense. Removal of six or more records out of 36 therefore is clearly not suitable. A more appropriate interpretation is that these records demonstrate the natural variation inherent in normal kidneys. This variation needs to be taken into account when we generate synthetic kidneys.

5 Power Transformation

The analysis of the previous sections provides strong evidence against the normal distribution as appropriate for these data. Our goal is to generate synthetic data which appropriately models the variability inherent in healthy kidney populations. Since the Gaussian model is not adequate for this purpose, we will proceed by suitably modifying a Gaussian model using parametric transformations. Our goal is a modified Gaussian model that exhibits the same properties with respect to the IC1 analysis as the kidney population. This suggests that the modification should be done at the level of the sphered PC data as this provides the input to the IC1 analysis. Because the training sample size is small, the modification must be rather simple in nature, yet capture the "heavier than Gaussian" tails apparent in these data. We chose to approach this using a power transform in the radial direction.

This simple parametric transformation involves a number of steps: We first reduce the columns of the sphered PC matrix $Q = P_{sph}$ to the squared lengths of the data vectors s_j, $j = 1, .., n$. We determine the parametric transformation, indexed by its power α, which gives the closest fit of the s_j to the Gaussian. The parameter α_0 which results in the *best* one-dimensional fit induces a matrix transformation $Q \to Q_{\alpha_0}$ at the level of the sphered PC data. The transformed data matrix Q_{α_0} is intended to be more Gaussian than Q. We will apply the IC1 analysis to Q_{α_0} to determine its deviation from the Gaussian. These two steps are based on the actual kidney data. The final step is concerned with the simulation of synthetic data. For this, an appropriate inverse transform will be used to generate the synthetic kidney populations starting from Gaussian random variables. The first two steps which involve the actual kidney data will be described in this section, while the generation of the synthetic data and the simulation verifications will be the topic of the following section.

First consider the sphered or normalized PC data $Q = P_{sph}$, using the notation defined in eq (6) in Appendix 7.1. For each data (or column) vector $\underrightarrow{q}_j$ $(j = 1, ..., n)$, in Q we consider its squared length $s_j = s(\underrightarrow{q}_j)$, see eq (7) in Appendix 7.2. If the data – here Q – have a Gaussian distribution, then the s_j, appropriately normalized, follow a $\chi^2(k)$ distribution with $k = 7$ degrees of freedom. By applying the inverse probability integral transform to the distribution of these squared lengths, we can compare them with the quantiles of the uniform distribution. The top left panel in Figure 8 shows this comparison with the uniform distribution for the kidney data in the form of a smoothed quantile or Q-Q plot. The red line shows the quantiles of the data and the green line indicates the quantiles of the uniform distribution. This plot is enhanced by the blue curves which arise from 100 simulations from the uniform distribution and which show the variability that exists in a uniform random sample. This visual device for understanding the sampling variation in a Q-Q plot was also used by Hernández-Campos, Marron, Samorodnitsky and Smith (2002). Since the kidney data are not Gaussian, it is not surprising that the red line deviates significantly from the blue envelope representing the uniform

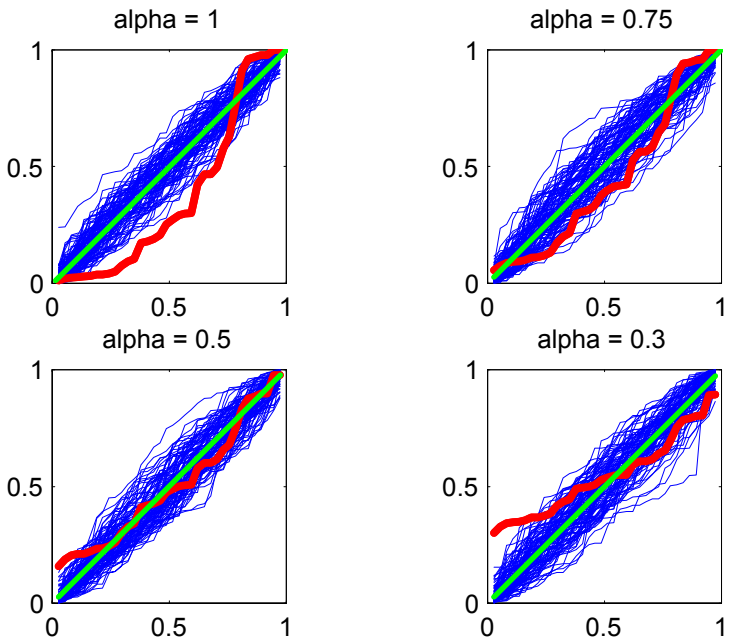

Figure 8: *Power transformed sums of squares data*

samples.

Our approach is to map the data into an approximately Gaussian shape by applying a simple radial power transformation, which is based on the one-dimensional level of the squared lengths s_j. This is a simple way to generate behavior of the type observed in this data.. For the squared lengths $s_j = s(\underrightarrow{q}_j)$ ($j = 1, .., n$) with $\underrightarrow{q}_j$ in Q, and for $0 < \alpha \leq 1$, define the power transform $s_j \rightarrow s_j^\alpha$ as in eq (9) in Appendix 7.2. Note that $\alpha = 1$ corresponds to the identity transform. The distribution of the transformed squared lengths s_j^α, $j = 1, ..., n$, is compared to the uniform distribution under the assumption that the data have a $\chi^2(k)$ distribution. The results can be seen in Figure 8 for a number of values of α. The top right panel shows the case $\alpha = 0.75$; in the bottom panels $\alpha = 0.5$ (left) and $\alpha = 0.3$ (right) were used. The value $\alpha_0 = 0.5$ shows the best agreement with the uniform distribution; $\alpha = 0.3$ shows a bimodal structure, indicating that the transformation has gone too far. Values of $\alpha > 1$ remove the data even further from the Gaussian shape and are therefore not displayed here.

The power transform is applied to the one-dimensional squared lengths s_j, since it is simpler to compare univariate data with Gaussians than high dimensional data. For high dimensional data the results obtained in the Q-Q plot can be misleading, and what appears to be Gaussian at the one-dimensional level

of the s_j may not be as Gaussian when examined with the more powerful ICA. For each α the transform $s_j \to s_j^\alpha$ induces a natural matrix transform $Q \to Q_\alpha$ of the sphered data Q to the α-transformed $k \times n$ matrix Q_α. Mathematical details of this induced transformation are given in eq (11) in Appendix 7.2. To examine the effect of the power transform at the level of the PC data, we applied IC1 and the Gaussianity test to Q_α. The optimal value $\alpha_0 = 0.5$ led to the *most* Gaussian Q_α, and we shall therefore report the results of the IC1 analysis for $Q_{0.5}$ only. Observe that re-sphering is necessary before applying IC1 to Q_α, since the induced power transform changes the covariance structure, see Appendix 7.2. Results of the IC1 analysis of $Q_{0.5}$ showed that the maximum absolute skewness is reduced to 3.81 compared to 4.65 of the original test. Only some of the original outliers, namely 32, 33 and 15 are still present, and their maximum values are decreased. The other three original outliers have no longer been detected. The IC1 hypothesis test remains unchanged, resulting in a p-value of 0. Thus from this viewpoint, our simple radial power transformation has not resulted in an exactly Gaussian result.

However, the IC1 test is very sensitive, and exact Gaussianity of this type is a lot to request from a model based on such a small training set. A perhaps better way of seeing that our result is reasonable is to study the characteristics of the resulting simulated data, which we do in the next Section.

6 Simulation Verification

The transformed kidney data $Q_{0.5}$ with the best parameter $\alpha_0 = 0.5$ still contains significant non-Gaussian directions. This is further indication of the power of the IC1 test of Gaussianity and the variability inherent to the kidney data. Although this simple transformation did not result in a precisely Gaussian distribution, the performance of backtransformed simulated data is more important. This section describes how we generate such non-Gaussian samples, and shows that the resulting synthetic data is reasonable by means of the IC1 analysis.

First recall the steps leading from the sphered kidney data Q to *uncorrelated almost Gaussian* data. Schematically these steps can be represented by

$$Q \to Q_\alpha \to R_\alpha(Q_\alpha) \simeq Z, \tag{1}$$

where the first transform is the induced power transform of the previous section. The next arrow, leading to $R_\alpha(Q_\alpha)$, denotes the re-sphering of Q_α required for the IC1 analysis and described in Appendix 7.2. As discussed in the previous section $R_\alpha(Q_\alpha)$, the transformed and re-sphered data, is closer to uncorrelated Gaussians than the original sphered data Q. This fact is indicated by $R_\alpha(Q_\alpha) \simeq Z$, where $Z \sim N_{k\times n}(0, I)$.

For simulation of synthetic kidney data we reverse the process described in eq (1). We start with a Gaussian generator matrix, also denoted by $Z \sim N_{k\times n}(0, I)$, and obtain a non-Gaussian synthetic population, Q_{sim}, as follows:

$$Z \to R_\alpha^{-1}(Z) \to [R_\alpha^{-1}(Z)]_{1/\alpha} = Q_{sim}, \tag{2}$$

where R_α^{-1} denotes the "un-sphering" of the Gaussian generator matrix Z. The next arrow indicates the application of an appropriate inverse power transform with power $1/\alpha$ in order to obtain the non-Gaussian population Q_{sim}, our synthetic model of the sphered kidney data Q. Mathematical details of the power transformed data and its covariance structure are given in Appendix 7.2, and details of the inverse transformation are given in eq (16) in Appendix 7.4.

Note that eq (1) made use of the parameter α. For the inverse path, described in eq (2) we therefore use the inverse parameter $1/\alpha$. The best parameter was found to be $\alpha_0 = 0.5$, we therefore report this case only. In particular, the synthetic data Q_{sim} referred to below is generated using $1/\alpha_0 = 2$.

It remains to determine whether the synthetic data Q_{sim} exhibits behavior similar to the kidney data. As for the sphered data, Q_{sim} consists of 36 data vectors each of length 7. To examine the variability and the existence of outlying data vectors in Q_{sim}, we use IC1. We calculate the most non-Gaussian directions for these data, and carry out the IC1 test described in 4.2. Figure 9 illustrates the results of the IC1 analysis for three independent runs of generating Q_{sim}, where each column corresponds to one particular Q_{sim}.

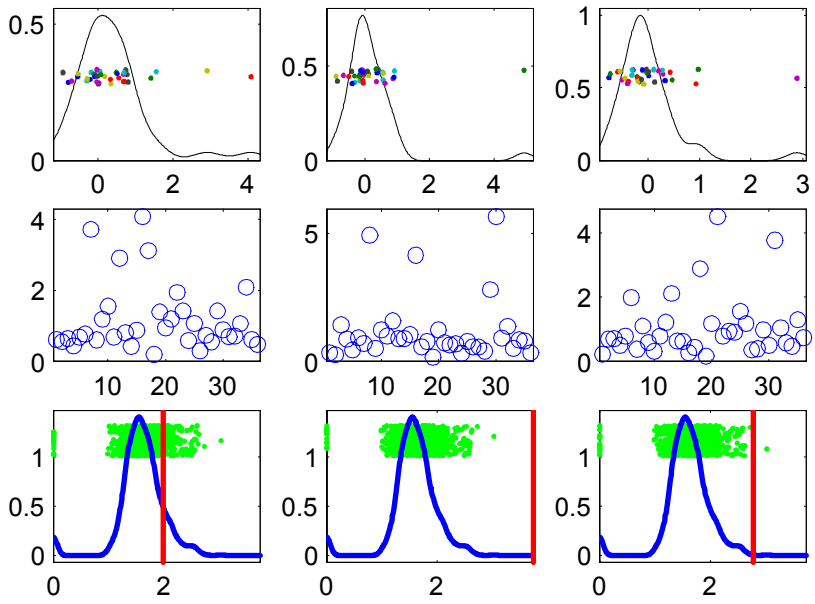

Figure 9: *IC1 results for synthetic data*

The top row shows the univariate distribution of the projection coefficients onto the IC1 direction which gives rise to the largest absolute skewness. These distributions are similar to the top left panel of Figure 5 which displays the corresponding distribution for the kidney data. All three simulated distributions

have notable outliers which are comparable in size to those of Figure 5. As in the case of the kidney data, the simulated data should have a number of outliers which can be found in repeated application of IC1, see Section 4.1. For each simulated data set Q_{sim} we calculated the first IC direction 1000 times. Similar to the bottom right panel in Figure 5, the middle row shows the maximum (over the 1000 ICA runs) of the absolute value of the projection coefficients for each case number. This view of the simulated data allows us to assess the number and size of the outliers. The case number of the outliers is not important for the simulated data, however the size and number of outliers is relevant for a comparison with the kidney data. It is worth noting that all simulated data sets had at least one large outlier, but more commonly had three or four outliers whose sizes are comparable to that of the kidney data.

The bottom row displays the results of the IC1 tests, based on 1000 drawings from the Gaussian distribution, which are similar to the tests described in Section 4.2. We observe that there is a reasonable amount of variability in the p-values obtained in these tests. The p-values in these three data sets are from left to right 0.124, 0, and 0.001.

A comparison of Figures 5 and 6 for the kidney data with the corresponding parts of Figure 9 shows that our approach to generating non-Gaussian data has been successful. The simulated data appear to have the same properties as the kidney data when examined with the powerful IC tools, in particular all synthetic data sets have a number of outlying samples which appear to characterize the kidney data.

A last step in the modelling process consists in the generation of the synthetic high-dimensional data at the level of the fiducial points - see Figure 2, as this is the data used for segmentation purposes and medical investigations. Our analysis is based on the $k \times n$ dimensional synthetic sphered data Q_{sim} which exists at the PC level with $k = 7$. The generation of such $d \times n$ dimensional data (with $d = 264$) is given in eq (17) in Appendix 7.4. Parallel coordinate views of such synthetic data are very similar to that of Figure 3.

7 Appendix

7.1 PCA Notation

Let $X = (X_{i,j})$ denote the $d \times n$ data matrix consisting of n (column) data vectors, each of length d. We shall assume that d is much bigger than n. We assume that X has mean 0. Let Σ denote the $d \times d$ covariance matrix of X, given by its "outer product" $\Sigma = XX^t$ where X^t denotes the transpose of X. Using the eigenvalue decomposition of Σ into $\Sigma = UDU^t$, where U is a unitary transformation consisting of the eigenvectors of Σ, and D is a diagonal matrix consisting of the eigenvalues of Σ, the column vectors $\underrightarrow{X}_i$ of X can be represented as linear combinations of the eigenvectors $\underrightarrow{u}_j$, the column vectors in U:

$$\underset{\rightarrow}{X}_i = \sum_{j=1}^{d} c_{i,j}\, \underset{\rightarrow}{u}_j = \sum_{j=1}^{r} c_{i,j}\, \underset{\rightarrow}{u}_j. \tag{3}$$

Since $d >> n$, the covariance matrix does not have full rank, in particular $r = rank(X) \leq n$. Here the $c_{i,j}$ are the projections of the data vectors onto the eigenvectors, $c_{i,j} = \underset{\rightarrow}{X}_i^t \underset{\rightarrow}{u}_j$, and $c_{i,j} = 0$ for $j > r$, since D has only r non-zero entries.

As $d >> n$, the first step in the data analysis is a reduction in dimension. For $k, m < d$ let $U_{1:k,\cdot}$ denote the first k rows of U (and all columns), $U_{\cdot,1:m}$ the first m columns of U (and all rows), and let $U_{1:k,1:m}$ denote the matrix consisting of the first k rows and first m columns of U. As appropriate we will regard $U_{1:k,1:m}$ as an $k \times m$ matrix or a $d \times d$ matrix with zeros filled in.

If k denotes the number of significant PC directions, then the PC analysis provides a "reduced principal component representation" $P = P_X$ of X, over the subspace of the first k eigenvectors in U, that is, the PC matrix P is the matrix of projections of X onto the first k eigenvectors, and so

$$P = U_{1:k,\cdot}^t\, X, \tag{4}$$

making P a $k \times n$ matrix with $k < r \leq n$. The $k \times k$ covariance matrix Σ_p of P is given by

$$\Sigma_p = cov(P) = cov(U_{1:k,\cdot}^t X) = U_{1:k,\cdot}^t \Sigma U_{\cdot,1:k} \,. = D_{1:k,1:k}, \tag{5}$$

since $\Sigma = UDU^t$, and since $U_{1:k,\cdot}^t U = 1_{1:k,\cdot}$, the identity matrix with k rows and d columns, and similarly, $U^t U_{\cdot,1:k} = 1_{\cdot,1:k}$.

Normalizing a matrix, say P, means multiplying P by the inverse of the root of its covariance matrix, say Σ_p, and results in the *whitened* or *sphered* matrix, $Q = P_{sph}$,

$$Q = \Sigma_p^{-1/2} P = D_{1:k,1:k}^{-1/2} U_{1:k,\cdot}^t X. \tag{6}$$

Observe that Q has mean 0 and covariance matrix $I_{k \times n}$ and Q is referred to as the *sphered data*. In Sections 4.1 and 4.2 the sphered data is decomposed into independent directions.

7.2 Details of Power Transformation

For the sphered $k \times n$ data matrix Q consider the column vectors $\underset{\rightarrow}{q}_j$ as data vectors. For each $\underset{\rightarrow}{q}_j$ let s_j denote the squared distance of $\underset{\rightarrow}{q}_j$ given by

$$s_j = s(\underset{\rightarrow}{q}_j) = \sum_{i=1}^{k} q_{ij}^2 \qquad j = 1, ..., n, \tag{7}$$

and let $\left\| \underset{\rightarrow}{q}_j \right\|$ denote the distance to the origin given by

$$\left\| \underset{\rightarrow}{q}_j \right\| = \sqrt{s_j}, \qquad j = 1, ..., n. \tag{8}$$

If the data are Gaussian, then the s_j will be $\chi^2(k)$. If the data vectors $\underset{\rightarrow}{q}_j$ in Q do not follow a Gaussian distribution, it might be possible to apply a power transform which makes the transformed vectors approximately Gaussian in the following sense. For $\alpha > 0$, put

$$s_j^\alpha = [\sum_{i=1}^{k} q_{ij}^2]^\alpha, \qquad j = 1, ..., n. \tag{9}$$

Normalize these n distances by their mean and divide by k. Determine that value $\alpha > 0$ which provides the best fit of the normalized squared distances (9) to the $\chi^2(k)$ distribution. Here the best fit is determined experimentally.

To apply this power transformation to the sphered PC data prior to IC1, we transform the sphered data matrix Q in such a way that each transformed vector has length $\left\| \underset{\rightarrow}{q}_j \right\|^\alpha$. The distances $\left\| \underset{\rightarrow}{q}_j \right\|$ in eq (8) form a vector of length n. We use this vector to define the diagonal $n \times n$ matrix V_α with non-zero entries v_j^α given by

$$v_j^\alpha = \left\| \underset{\rightarrow}{q}_j \right\|^{\alpha-1} = [\sqrt{\sum_{i=1}^{k} q_{ij}^2}]^{\alpha-1} \qquad j = 1, ...n. \tag{10}$$

The power transformed matrix Q_α is given by

$$Q_\alpha = QV_\alpha. \tag{11}$$

This is a $k \times n$ matrix, and it can easily be seen that the j^{th} column vector in Q_α has a squared distance s_j^α, and therefore distance to the origin of length $\left\| \underset{\rightarrow}{q}_j \right\|^\alpha$. A consequence of applying this power transform to Q is that the new matrix Q_α is correlated. Let Σ_α denote the covariance matrix of Q_α and let R_α denote the re-sphering, so

$$R_\alpha(Q_\alpha) = \Sigma_\alpha^{-1/2} Q_\alpha \tag{12}$$

is a $k \times n$ matrix. The transformed and sphered matrix $R_\alpha(Q_\alpha)$ is used as the new input to IC1 instead of the previous sphered data matrix Q.

7.3 Simulation of Sphered Gaussian Data

For a given covariance matrix Σ, it is desired to generate a Gaussian random $X \sim N_{d \times n}(0, \Sigma)$. For $k \leq r = rank(\Sigma)$ define the "random generator matrix" $Z \sim N_{k \times n}(0, I)$. Let $S_{\cdot,1:k} = D_{\cdot,1:k}^{1/2}$, where D is the diagonal matrix as in eq (5) consisting of the first k eigenvalues of Σ, and put

$$X = US_{\cdot,1:k}Z. \tag{13}$$

Note that

$$\begin{aligned} cov(X) &= cov(US_{\cdot,1:k}Z) = US_{\cdot,1:k}cov(ZZ^t)(US_{\cdot,1:k})^t \\ &= US_{\cdot,1:k}IS^t_{1:k,\cdot}U^t = UD_{1:k,1:k}U^t. \end{aligned} \tag{14}$$

For $k = r$ it follows that $UD_{1:r,1:r}U^t = \Sigma$, and thus $X \sim N_{d\times n}(0, \Sigma)$.

Some interpretations are that the random matrix X can be generated from only $r \cdot n$ independent Gaussians, i.e. has $r \cdot n$ "degrees of freedom". Also $X = US_{\cdot,1:r}Z$, can be viewed as "a rescaling by S", followed by a "rotation by U".

If $k < r$, then the Gaussian random variable X can be more efficiently computed by $X = U_{\cdot,1:k}\, S_{1:k,1:k}Z$ with $Z \sim N_{k\times n}(0, I)$. Observe that $UD_{1:k,1:k}U^t = U_{\cdot,1:k}\, D_{1:k,1:k}U^t_{1:k,\cdot}$, and therefore the covariance matrix $cov(X)$ of X is the same as that obtained in eq (14). If k denotes the number of significant PC directions, then $cov(X)$ has the same eigenvalues as the covariance matrix Σ_p of P in eq (5). Here it is more convenient to regard $cov(X)$ as a $d\times d$ matrix with non-zero entries only in the top $k \times k$ part, while Σ_p in eq (5) is a $k \times k$ matrix.

Since the number of significant PCs is generally much smaller than r, $X = U_{\cdot,1:k}\, S_{1:k,1:k}Z$ produces an efficient computational approach for simulating Gaussian random X. From eq (6) it follows that

$$Q = Z, \tag{15}$$

that is, the random generator matrix is a model for the sphered data.

7.4 Simulation of Sphered Non-Gaussian Data

By eq (15) the random generator matrix is a model for the sphered data matrix if the data is Gaussian. To generate the sphered data when the process is non-Gaussian, we make use of the transformation described in eqs (11) and (12) and reverse this process.

To generate a random matrix Q_{sim} via inverse transforms as in eq (2), simulate from the Gaussian random generator matrix $Z \sim N_{k\times n}(0, I)$. For given $\alpha > 0$, put $\beta = 1/\alpha$. Define Q_{sim} by

$$Q_{sim} = (\Sigma_\alpha^{1/2}Z)W_\beta, \tag{16}$$

where Σ_α denotes the covariance matrix of Q_α, see eq(12), which is used now to *un-sphere* Z, and W_β is a diagonal $n \times n$ matrix which applies an inverse power transforms to $\Sigma_\alpha^{1/2}Z$. The non-zero elements of W_β are defined by $w_j^\beta = \left\| \underset{\rightarrow}{\widetilde{z}}_j \right\|^{\beta-1}$ – see also eq (10) – where $\underset{\rightarrow}{\widetilde{z}}_j$denotes the j^{th} column vector of $\Sigma_\alpha^{1/2}Z$. The matrix Q_{sim} is a $k \times n$ matrix which describes the distributional properties of the sphered kidney data Q more closely than Z. One can simulate the non-Gaussian random variable matrix X from Q_{sim} as follows

$$X = U_{\cdot,1:k}\, S_{1:k,1:k}Q_{sim} \tag{17}$$

with Q_{sim} generated from Gaussian random matrix $Z \sim N_{k\times n}(0, I)$ as described in eq (16). Such simulated data X will have a covariance matrix which depends on Σ_α. The process of multiplying Z by $\Sigma_\alpha^{1/2}$ can be thought of as a type of 'un-sphering' the Gaussian Z, and transforming Z into a non-Gaussian matrix with covariance matrix Σ_α.

8 Acknowledgement

Blinded.

References

[1] Blinded

[2] Cleveland, W. S. (1993) *Visualizing Data*, Hobart Press, Summit, New Jersey, U.S.A.

[3] Friedman, J. H. and Tukey, J. W. (1974) A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers*, 23, 881-890.

[4] Friedman, J. H. (1987) Exploratory projection pursuit, *Journal of the American Statistical Association*, 82, 249–266..

[5] F. Hernández-Campos, J. S. Marron, G. Samorodnitsky and F. D. Smith (2002) Variable Heavy Tail Duration in Internet Traffic, internet available at http://www.cs.unc.edu/Research/dirt/proj/marron/VarHeavyTails/. Part 1 appeared in the proceedings of IEEE MASCOTS'02, October, 2002. Part 2 appeared in the Proceedings of the 40th Allerton Conference in Communications, Control and Computing, October, 2002.

[6] Hyvärinen, A and Oja, E. (1999) *Independent Component Analysis: A Tutorial*, internet available at http://www.cis.hut.fi/projects/ica

[7] Hyvärinen, A., Karhunen, J. and Oja, E. (2001) *Independent Component Analysis*, Wiley, New York.

[8] Inselberg, A. (1985) The plane with parallel coordinates, *The Visual Computer*, 1, 69-91.

[9] Inselberg, A. and Dimsdale, B. (1987) Parallel coordinates for visualizing multi-dimensional geometry, *Computer Graphics 1987*, Ed: T. L. Kunii, 25–44. Springer-Verlag, Berlin.

[10] Jones, M.C. and Sibson, R (1987) What is Projection Pursuit? *The Journal of the Royal Statisitcal Society Series A*, 150, pp1-36

[11] Wand, M. P. and Jones, M. C. (1995) *Kernel Smoothing*, Chapman and Hall, New York.