Computational Statistics and Data Analysis 56 (2012) 4462–4469

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Investigating the multimodality of multivariate data with principal curves

Murat O. Ahmed, Guenther Walther *

Department of Statistics, 390 Serra Mall, Stanford University, Stanford, CA 94305, United States

ARTICLE INFO

Article history:
Received 14 July 2010
Received in revised form 1 December 2011
Accepted 4 February 2012
Available online 23 March 2012

Keywords:
Multimodality
Principal curves
Bandwidth test

ABSTRACT

We propose a simple method to assess the number of subpopulations in multivariate data by projecting the data on its principal curve and then applying Silverman's bandwidth test to the resulting univariate sample. Our results indicate that this method works well even in high-dimensional settings with relatively small sample sizes, provided that the number of subpopulations is not large compared to the number of dimensions.

1. Introduction

An important aspect of unsupervised learning concerns the inference about the number of subpopulations or groups in the data. Following the Wishart (1969) paper on Mode Analysis, there has been a considerable amount of literature concerning inference about the modes of a density, based on the premise that such modes are indicators of subpopulations. Examples for the univariate case are the penalized likelihood method of Good and Gaskins (1980), the Dip test of unimodality of Hartigan and Hartigan (1985), and the bandwidth test of Silverman Silverman (1980). The latter method has proven to be particularly popular since it is quite simple to implement. We note that it can be problematic to analyze the number of subpopulations via the number modes in the density, but this standard approach does often lead to useful results, see Walther (2003) for more details.

In the multivariate case, the problem becomes conceptually much more difficult, and the available methodology is more limited. Besides the 'curse of dimensionality', there is the computational problem of identifying a mode of a (potentially high-dimensional) surface. Hartigan (1987) and Müller and Sawitzki (1991) give arguments for basing the inference on mass concentration instead. Ray and Lindsay (2005) provide results on the topography of multivariate normals that can be used for inference on modes. Fraley and Raftery (1998) use model-based clustering for the inference about the number of subpopulations. Baudry et al. (2010) and Hennig (2010) modify model-based clustering by merging mixture components. Stuetzle and Nugent (in press) estimate the cluster tree of a density. Potential difficulties with these methods are that they are either difficult to implement computationally, or they are based on a parametric model, or they use density estimation, which may not work well in a high-dimensional setting.

We propose to compute the principal curve of the data, which is a nonlinear univariate summary of the data, see Hastie and Stuetzle (1989), and then to project the data onto the principal curve and apply Silverman's bandwidth test to the resulting one-dimensional distribution. As an example, Fig. 1 shows a sample with three subpopulations and the histogram resulting from projecting the data on the principal curve. The histogram clearly suggests a trimodal distribution, and this can be readily assessed by using a univariate method such as Silverman's bandwidth test.

* Corresponding author. Tel.: +1 650 723 3066; fax: +1 650 725 8977.
E-mail addresses: murat@stanford.edu (M.O. Ahmed), gwalther@stanford.edu, walther@stat.stanford.edu (G. Walther).

0167-9473/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2012.02.020

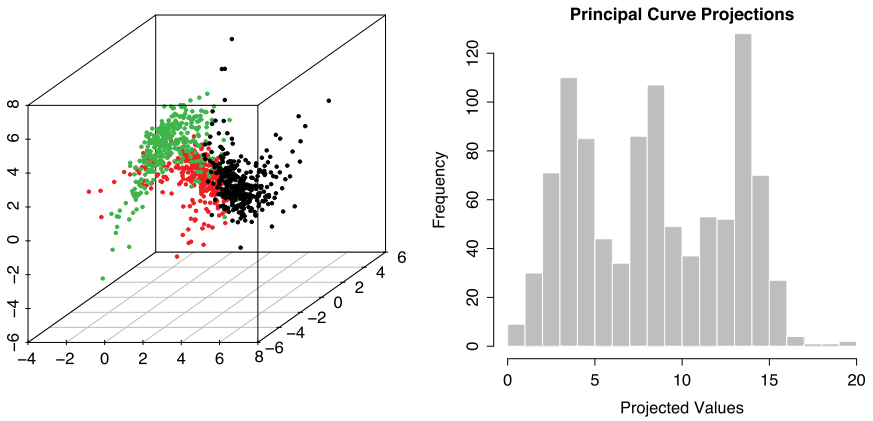

Fig. 1. A sample with three subpopulations (*left*) and the histogram of projections on the principal curve (*right*).

Our results indicate that this method works well even in high-dimensional settings with relatively small sample sizes, provided the number of subpopulations is not large compared to the number of dimensions. A particular advantage of the methodology is that it is quite simple to implement as it draws on software, such as code for the computation of the principal curve, that is already available in many software packages.

2. Dimension reduction with principal curves

Let $\mathbf{X} \in \mathbb{R}^p$ be a random vector and let $f(\lambda)$ be a smooth curve in $\mathbb{R}^p$ that is parametrized by a real variable λ. Define the projection index $\lambda_f(\mathbf{X})$ to be the (largest) value of λ for which $f(\lambda)$ is closest to $\mathbf{X}$. Then $f(\lambda)$ is called the principal curve of $\mathbf{X}$ if it satisfies the *self-consistency* property

$$\mathbf{E}(\mathbf{X}|\lambda_f(\mathbf{X}) = \lambda) = f(\lambda).$$

This property says that $f(\lambda)$ is the mean of all points that project onto it. In this sense the principal curve gives a univariate nonlinear summary of $\mathbf{X}$. The definition is modified in an obvious way when $\mathbf{X}$ represents a sample rather than a random vector, and algorithms for the computation of the principal curve are available e.g. in R. See Hastie and Stuetzle (1989) and Tarpey and Flury (1996) for more background on principal curves.

3. Assessing multimodality with Silverman's bandwidth test

Silverman's bandwidth test is a popular method to assess whether a univariate distribution has k modes. A main attraction of this method is its conceptual and computational simplicity, which will turn out to be a useful feature when we test multiple hypotheses later.

Consider the null hypothesis that the distribution underlying the data has at most k modes. Let $\hat{f}(t; h)$ be the kernel density estimate defined by

$$\hat{f}(t; h) = (nh)^{-1} \sum_{i=1}^{n} K(h^{-1}(t - X_i))$$

where $X_1, \ldots, X_n$ are univariate observations, $K(.)$ is the Gaussian kernel function and h is the bandwidth. The relevant test statistic is the *critical* bandwidth h_k given by

$$h_k = \inf\{h : \hat{f}(.; h) \text{ has at most } k \text{ modes}\}.$$

Large values of h_k indicate rejection of the null hypothesis since a great deal of smoothing is necessary before the density estimate becomes k-modal. The Gaussian kernel ensures that $\hat{f}(.; h)$ has more than k modes if and only if $h < h_k$, see Silverman Silverman (1980).

The bootstrap is employed to assess whether the critical bandwidth is significant. The bootstrap uses as null distribution the density $\hat{f}(.; h_k)$ after rescaling to ensure that the null density has the same variance as the sample variance σ^2 of the data. The bootstrap then draws repeated samples from the null density to estimate the probability that the resulting critical bandwidth is larger than h_k, which is then used as the p-value of the test. Equivalently one can check whether the density estimate for the bootstrapped data evaluated at h_k has more than k modes. Sampling from the null density is very straightforward since $\hat{f}(.; h_k)$ is a convolution of the empirical measure with $N(0, h_k^2)$. Taking account of the above rescaling, the bootstrap sample can thus be simply generated by

$$X_j^* = (1 + h_k^2/\sigma^2)^{-1/2}(X_{i_j} + h_k\epsilon),$$

where $\epsilon \sim N(0, 1)$ and the X_{i_j} are sampled uniformly with replacement from $\mathbf{X}$.

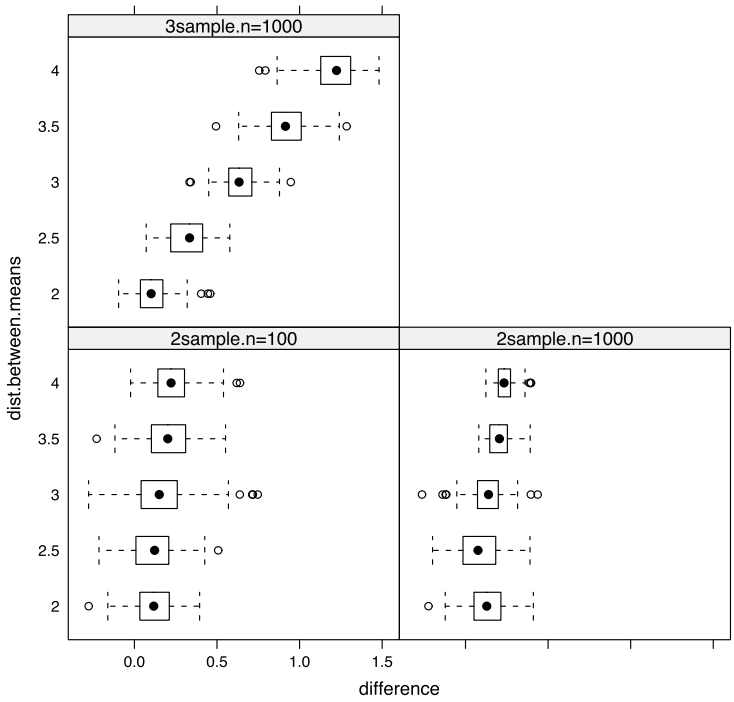

Fig. 2. Component means $(0, 0)$, $(0, d)$, $n = 100$ (*bottom left*), $n = 1000$ (*bottom right*), components means $(0, 0)$, $(0, d)$, (d, d), $n = 1000$ (*top left*) for $d \in \{2, 2.5, 3, 3.5, 4\}$.

Table 1
Power for varying distance x. Parameters in the first line are: number of components, 'p' means 1st PC and 'c' means principal curve, sample size.

x	2, p, 100	2, c, 100	2, p, 1000	2, c, 1000	3, p, 1000	3, c, 1000
2	0.02	0.01	0.03	0.11	0	0.1
2.5	0.13	0.12	0.3	0.16	0.01	0.5
3	0.24	0.19	0.93	0.91	0.04	0.99
3.5	0.68	0.62	1	1	0.02	1
4	0.93	0.93	1	1	0.04	1

4. Assessing multimodality via low-dimensional projections

Our proposal is to assess multimodality via the principal curve of the data, which is a univariate nonlinear summary of the data. Using this summary makes it possible to employ the simple univariate methods described above, such as Silverman's bandwidth test. An important question is what is gained by using the nonlinear principal curve rather than simply a linear projection. We investigated this question by simulating from two-dimensional Gaussian mixtures with two and three equally weighted components having identity covariance matrix. For each of 100 samples of size 100 and 1000 we computed the critical bandwidth associated with the projected data, both for projections on the principal curve and for projections on the first principal component, which is the most appropriate one-dimensional linear projection. Our null hypothesis was the presence of less than k modes for $k = 2, 3$. Fig. 2 displays the differences between these critical bandwidth using boxplots. The vertical axis displays the difference between the component means while the horizontal axis displays the critical bandwidth using the principal curve projections minus that using the first principal component. The mixture components have unit variance and correlation 0.5. With only two components both linear and non-linear projections produce similar results, yet principal curves typically result in larger critical values. When a third component is added the critical bandwidths for the principal curves become much larger, indicating that the projected data are less unimodal in the case of the principal curve versus the case of the principal component, and thus the former is more effective than the latter for detecting the presence of the mixture components. This result is indeed confirmed in the power analysis given in Table 1, which displays the estimated power using 100 simulations and $\alpha = 0.05$ for the level of the test. We used 1000 bootstrap samples to generate the p-values for Silverman's test of k modes, $k = 2, 3$.

The table shows that when the means are separated by at least three standard units, then in the three component case principal curves result in a power of close to 1, whereas the test based on principal components is essentially powerless. We conclude that there is a clear advantage in employing principal curves. The underlying reason is that in the case of a projection onto a one-dimensional linear space the projections of several components are often very close to each other.

Hence the density of the projected data is unimodal near this location, and thus the fact that the projection is comprised of several components is lost. In contrast, the principal curve is a much more flexible univariate summary that we expect to pass through the centers of most clusters in the data. Hence it is much less likely that different clusters are projected near the same location on the principal curve and thus the projected clusters are more likely to stay separated on the principal curve.

5. Comparison with other methods

One recent suggestion for determining the number of modes is to fit the data with a Gaussian mixture model and then merge some of the components. The idea is that even if the data are far from normal we can approximate the distribution by merging several Gaussian distributions. For comparison purposes we adopt the method proposed by Hennig (2010) that makes use of the topography of multivariate normals as introduced in Ray and Lindsay (2005). This method performed favorably in a comparison with other methods given in Hennig (2010). Let

$$g(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \phi(\mathbf{x};\, \boldsymbol{\mu}_k,\, \Sigma_k), \quad \mathbf{x} \in \mathbb{R}^p$$

where $\pi_k \in [0, 1]$, $\sum_{k=1}^{K} \pi_k = 1$, be a Gaussian mixture density, $\phi(.)$ denoting a multivariate normal density. For $K = 2$ components Ray and Lindsay (2005) define the *ridgeline* function as the one-dimensional curve

$$\mathbf{x}^*(\alpha) = ((1-\alpha)\Sigma_1^{-1} + \alpha\Sigma_2^{-1})^{-1}((1-\alpha)\Sigma_1^{-1}\boldsymbol{\mu}_1 + \alpha\Sigma_2^{-1}\boldsymbol{\mu}_2)$$

where $\alpha \in [0, 1]$. They show that all the modes of a multivariate normal density occur along the *ridgeline* and in the two component case the modes can be found by solving

$$\frac{1}{\pi_1} = 1 + \frac{\alpha\phi_1(\mathbf{x}^*(\alpha))}{(1-\alpha)\phi_2(\mathbf{x}^*(\alpha))}$$

for α. If the density is unimodal there is only a single solution to the above equation.

Hennig (2010) suggests the *ridgeline unimodal method* for determining which components to merge in the Gaussian mixture. The algorithm is as follows:

1. Start with components estimated via a Gaussian mixture model using BIC to select the number of components.
2. Compute whether pairwise the mixture of any two components would be unimodal using the criterion described above and π as determined by hard classification using the posterior probabilities.
3. If none are unimodal use the current clustering as the final one. Otherwise:
   (a) Merge pairs that represent unimodal submixtures.
   (b) Only merge multiple components if all pairs of components represent unimodal submixtures.
   (c) If some pair in (b) does not represent a unimodal submixture, then merge the pair with the closest mean vectors.

This method is similar to *complete linkage* hierarchical clustering in the sense that a component is merged or added to a group if all pairwise comparisons within a current group lead to unimodal mixtures. It should be noted that Baudry et al. (2010) suggest merging the two clusters that yield the greatest increase in entropy, where the stopping criterion is an entropy penalized BIC criterion. There is no guarantee that this approach will result in unimodal components.

We compare the above method with our method in a simulation study. We simulated data from a mixture with equal mixture proportions and K components that are generated as follows: Let $X^k \sim MVN(\mu^k, \Sigma), k = 1, \ldots, K, \mu^k \in \mathbb{R}^p$, where $\Sigma_{ii} = 1$, $\Sigma_{ij} = .4$, $i \neq j$ and $\mu^k = aU^k$ with $U^k \sim \mathcal{U}(0, 1)^p$

Now the larger the distance between component means, the easier it is to determine the correct number of modes. However, for the simulation study the elements of the component means are uniformly distributed, and so the expected Euclidean distance between component means increases with dimension. To keep the expected distance constant across dimensions, we introduce the multiplier a, which we set to $a = d\sqrt{2}/.92$,[1] where d is the desired Euclidean distance between component means. This way the value chosen for d results in performance comparable to the results in Fig. 2.

The jth coordinate Y_j of the kth component is then generated as follows: $Y_1 = X_1^k$ and for $j \geq 2$:

$$Y_j = X_j^k + c_j(X_1^k - \mu_1^k)^2$$

with $c_j \sim \mathcal{U}(-1, 1)$. A plot of data sampled from the above scheme is given in Fig. 1 using three components in three dimensions.

We varied the dimension over $p = 2, 10, 100$ and the number of components over $k = 2, 5, 10$ and simulated $n = 2500$ observations in each case. Fig. 3 displays boxplots of the p-values from Silverman's test on the principal curve projections based on 100 Monte Carlo samples. In the case of two modes the method using principal curve projections works very well. Nearly every simulation resulted in significant p-values for rejecting the null that there are fewer than k modes for $k = 1$ and

[1] Note that for $X, Y \sim \mathcal{U}(0, 1)^p$, $\mathbf{E}\|X - Y\| \approx .92\sqrt{p/6}$. Allowing 30% of the components to be a distance d apart and the rest noise, i.e. distance 0 apart, results in the above choice for a. For the present analysis $d = 4$ was used.

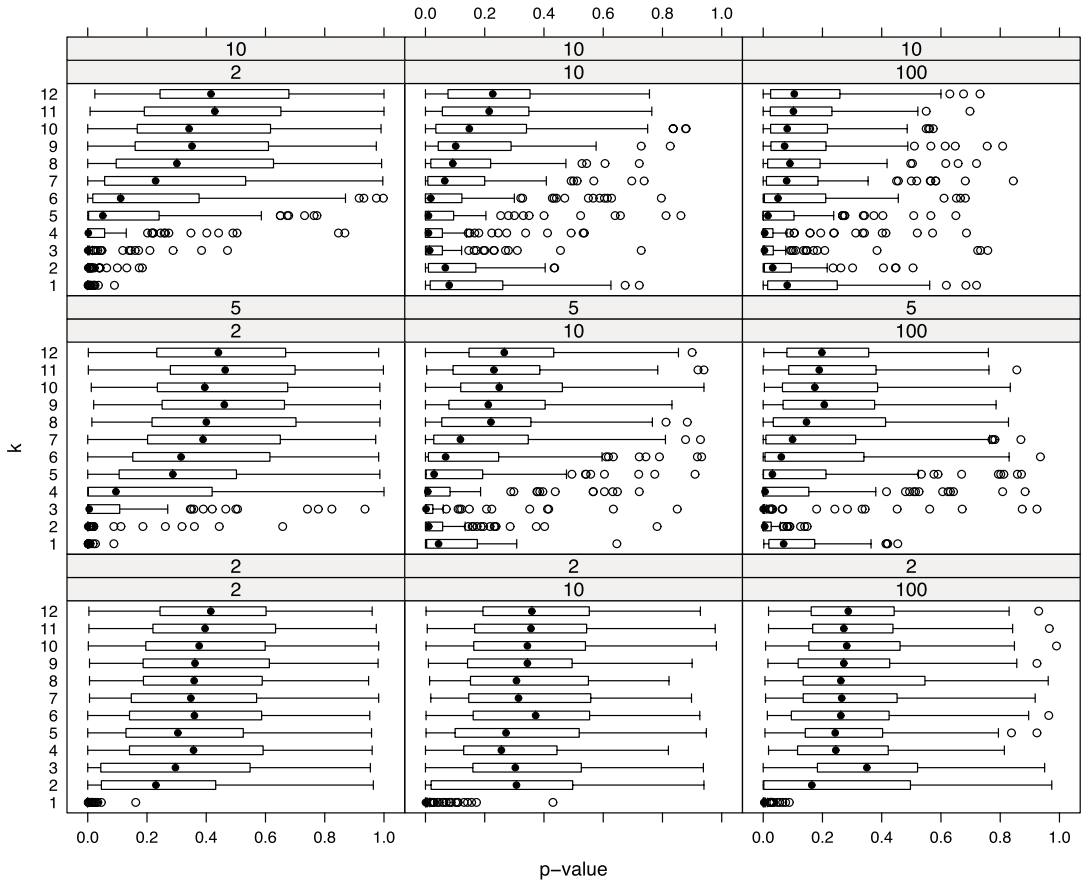

Fig. 3. Results for the method using principal curves: the vertical axis contains the number of modes k for the null hypothesis, the horizontal axis displays the p-value. For each plot the top number refers to the number of modes while the bottom one shows the dimension.

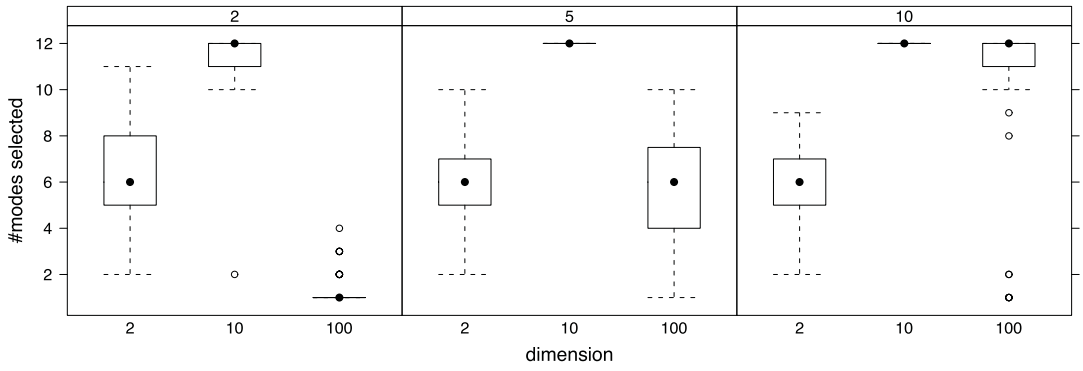

Fig. 4. Results for the method using Gaussian components merging. Vertical axis contains the number of modes selected by the method. Horizontal axis displays the data dimension. For each plot the top number indicates the true number of modes.

typically gave large p-values for $k \geq 2$. In the case of five and 10 modes the results were reasonable for higher dimensional data but poor for the two dimensional data. Note that the median p-value is minimized at the correct value of $k = 4$ for the five mode data in 10 and 100 dimensions.

One explanation for the poor performance when the number of modes is large relative to the dimension is that the principal curve algorithm converges before interpolating each mode. We discuss this issue further in Section 7, where we also give a rule of thumb that describes when our method works well.

The performance of merging Gaussian components is much worse, as evidenced by Fig. 4. In the two component case principal curve projections work very well across all dimensions. On the other hand the selected number of components from Gaussian merging has a wide range in two dimensions, almost never merges in 10 dimensions and practically merges all components in 100 dimensions. In fact, regardless of the number of components the 10 dimensional data results in

almost no merging of components. In addition, for the 100 dimensional data practically no components are merged in the 10 component case.[2] But one needs to keep in mind that due to computational constraints we limited the maximum number of fitted components to 12. Hence in the 10 component case there are not many components that need to be merged.

6. Size analysis of projections

An important question is whether our procedure produces p-values that are at least approximately valid. The main reason for this concern is that, especially in a high dimensional situation, data from one component may produce several spurious clusters, which might then be picked up by the principal curve due to its flexibility. If that concern proved true, then our procedure would be anti-conservative, i.e. indicate more components than are really there. In addition, there is the general question whether the bootstrap procedure inherent in Silverman's test results in valid inference, but since Silverman's test has found widespread application in the literature, this issue seems unlikely to result in a major problem.

To investigate the approximate validity of the p-values produced by our procedure, we estimated its size by simulating 1000 samples under the null hypothesis of a single Gaussian component. Fig. 5 displays the simulated probability of the p-value being less than x for $x \in [0, 1]$, when testing the null hypothesis of a single mode. Standard multivariate normal data were generated across dimensions 3, 10, 100 and 250. It is apparent that the test is slightly conservative when $x < 0.20$ and its size seems quite independent of dimension. These results show the above concern appears to be unfounded, and that the p-values produced by the procedure are even conservative.

7. Dimension dependence of performance

The difficulty of determining the number of modes when the number of clusters is large relative to dimension is suggested by Fig. 3. One explanation for the poor performance when the number of modes is large relative to the dimension is that the principal curve algorithm converges before interpolating each mode. The span used to smooth the data for computing conditional expectations is selected via cross-validation, see Hastie and Stuetzle (1989). Controlling the smoothness of the principal curve by cross-validation is reasonable with no prior knowledge of the problem. However, when the support of the data is "crowded" by many modes, then the principal curve converges before interpolating each one. We investigate this phenomenon by simulating MVN components $X^i \in \mathbb{R}^p$, $i = 1, \ldots, k$ for $k, p = \{2, 3, \ldots, 20\}$ as described above and perform Silverman's test with $k - 1$ modes as the null.[3] We see from Fig. 6 that as long as the number of modes is not much larger than the dimension, then the method works well. This is a promising feature for high-dimensional problems. From the results of Fig. 6 we can derive as a rule of thumb that our method works well if

the number of components does not exceed the dimension

with a particularly good performance if

the number of components does not exceed $2\sqrt{\text{dimension}}$.

As discussed above, this limitation on the performance seems to be due in large part to shortcomings in the standard algorithms for computing the principal curve, which select the span used to smooth the data by cross-validation. Typically, smaller spans should be beneficial for detecting many clusters whereas larger ones would be more robust to noisy data. Other methods for making principal curves more robust can be found in Banfield and Raftery (1992). These authors suggest a two-dimensional extension of a *twicing* procedure for bias-correcting the estimated principal curve. These and other possible improvements for computing the principal curve may lead to improvements in our method for detecting modes. We leave this as an open problem for future research.

8. An example: the olive oil data

The Olive Oil data consists of measurements of eight chemical components on 572 samples of olive oil, see Zupan et al. (1994). The samples come from three different regions of Italy. We applied our methodology by projecting the data on its principal curve and then applying Silverman's bandwidth test with 10,000 bootstrap samples. We obtained the following sequence of p-values for testing whether the projected data possess at most $k = 1, \ldots, 15$ modes:

0.002 0.000 0.014 0.078 0.061 0.469 0.160 0.445 0.345 0.165 0.330 0.396 0.259 0.514 0.347

Thus we conclude at the 1% significance level that there are at least three modes in the data. Interestingly, at $k = 7$ (eight modes) the p-value dips to 0.165 which is not significant but noticeably lower than the p-values for neighboring values of k. Indeed, the olive oils from the three regions can be further partitioned into originating from nine areas, see Zupan et al. (1994).

---

[2] The covariance matrices are modeled as $\Sigma_k = \lambda I$ when fitting the Gaussian mixture models as this resulted in the best performance.

[3] We let d = 5 to emphasize the method's dependence on dimension and not its power; sample size is n =10 000. Each pixel is the mean p-value across 10 simulations.

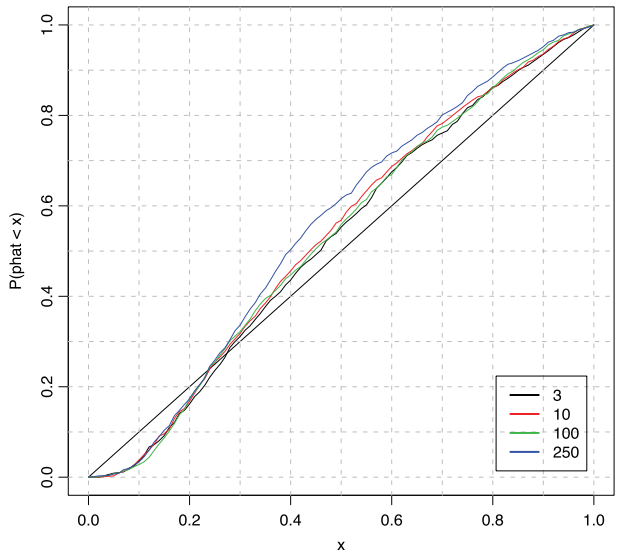

Fig. 5. Size analysis with a single component standard Gaussian null. $\mathbf{P}(\hat{p} < x)$ (*vertical axis*) for $x \in [0, 1]$ across 3, 10, 100 and 250 dimensional data.

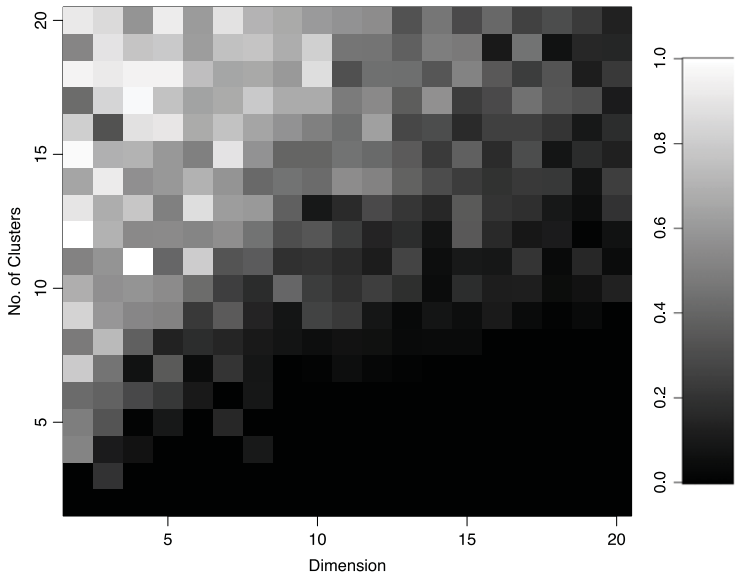

Fig. 6. p-values from Silverman's test with $k - 1$ modes as null; $k =$ number of clusters.

9. Technical details

We computed the principal curves with the `princurve v1.1-9` package for R. The curve is initialized as the first linear principal component and the conditional expectation step is computed using smoothing splines via the default settings for `smooth.spline()`. The Gaussian mixtures were fit with the `mclust` package when performing the merging analysis.

10. Summary and conclusions

Projecting multivariate data on the principal curve and applying Silverman's univariate bandwidth test gives a simple, easily implementable and effective way to assess the number of subpopulations in the data. Simulations show that this method is conservative at relevant significance levels, more powerful than other recent techniques, and works well even in high-dimensional settings with relatively small sample sizes, provided that the number of subpopulations is not large compared to the number of dimensions.

Acknowledgments

This work was supported by NSF grant DMS-1007722 and NIH grant 1R21AI069980.

References

Banfield, J.D., Raftery, A.E., 1992. Ice floe identification in satellite images using mathematical morphology and clustering about principal curves. J. Amer. Statist. Assoc. 87 (417), 7–16.
Baudry, J.P., Raftery, A., Celeux, G., Lo, K., Gottardo, R., 2010. Combining mixture components for clustering. J. Comput. Graph. Stat. 9, 332–353.
Fraley, C., Raftery, A., 1998. How many clusters? Answers via model-based cluster analysis. Comput. J. 4, 578–588.
Good, I., Gaskins, R., 1980. Density estimation and bump- hunting by the penalized likelihood method exemplified by scattering and meteorite data. J. Amer. Statist. Assoc. 75, 42–56.
Hartigan, J.A., 1987. Estimation of a convex density contour in two dimensions. J. Amer. Statist. Assoc. 82, 267–270.
Hartigan, J., Hartigan, P., 1985. The dip test of unimodality. Ann. Stat. 13, 80–84.
Hastie, T., Stuetzle, W., 1989. Principal curves. J. Amer. Statist. Assoc. 84, 502–516.
Hennig, C., 2010. Methods for merging gaussian mixture components. Adv. Data Anal. Classif. 4, 3–34.
Müller, D.W., Sawitzki, G., 1991. Excess mass estimates and tests for multimodality. J. Amer. Statist. Assoc. 86, 738–746.
Ray, S., Lindsay, B., 2005. The Topography of multivariate normal mixtures. Ann. Stat. 33, 2042–2065.
Silverman Silverman, B., 1980. Using kernel density estimates to investigate multimodality. J. R. Statisc. Soc. Ser. B 43, 97–99.
Stuetzle, W., Nugent, R., 2009. A generalized single linkage method for estimating the cluster tree of a density. J. Computational Graphical Stat. (in press).
Tarpey, T., Flury, B., 1996. Self-consistency: a fundamental concept in statistics. Stat. Sci. 11, 229–243.
Walther, G., 2003. Bikernel mixture analysis. In: Misra, J.C. (Ed.), industrial mathematics and statistics. Narosa, pp. 586–604.
Wishart, D., 1969. Mode analysis: A generalization of nearest neighbor which reduces chaining effects. In: Cole, A.J. (Ed.), Numerical Taxonomy. Academic Press, pp. 282–311.
Zupan, J., Novic, M., Li, X., Gasteiger, J., 1994. Classification of multicomponent analytical data of olive oils using different neural networks. Anal. Chim. Acta 292, 219–234.