Supplemental File to **The Mixturegram: A Visualization Tool for Assessing the Number of Components in Finite Mixture Models**

Derek S. Young,* Chenlu Ke, and Xiaoxue Zeng

August 5, 2017

1 Overview

This supplemental file provides the following:

- A discussion of the `R` function available in the `mixtools` package (Benaglia et al., 2009), including possible improvements that could be made to the mixturegram.
- Details about and additional results for all simulation settings discussed in the main text.
- Additional model selection results and mixturegrams for the quasar data and the DLBCL data analyzed in the main text.
- Construction of the mixturegram and analysis for two other well-known datasets — the Old Faithful geyser data (Azzalini and Bowman, 1990) and the Hidalgo stamp data (Izenman and Sommer, 1988).

---

*D. S. Young (derek.young@uky.edu) is an Assistant Professor and Chenlu Ke is a PhD student, Department of Statistics, University of Kentucky, Lexington, KY. Xiaoxue Zeng is a Fraud Data Analyst, Apple Inc., Austin, TX.

2 Discussion of `R` Function and Possible Improvements

A function to produce a mixturegram is available in the `R` package `mixtools` (Benaglia et al., 2009):

```
mixturegram(data, pmbs, method = c("pca", "kpca", "lda"), all.n = FALSE,
            id.con = NULL, score = 1, iter.max = 50, nstart = 25, ...)
```

Below is a description of the arguments in the above function:

- `data`: This is the user's data, which must either be a vector or a matrix. If a matrix, then the rows correspond to the observations.

- `pmbs`: This is a list of length $(K-1)$ such that each element is an $n \times k$ matrix of the posterior membership probabilities. These are obtained from each of the "best" estimated k-component mixture models, $k = 2, \ldots, K$.

- `method`: The dimension reduction method used. `"PCA"` implements principal components analysis. `"KPCA"` implements kernel principal components analysis. `"LDA"` implements reduced rank linear discriminant analysis.

- `all.n`: A logical specifying whether the mixturegram should plot the profiles of all observations (`TRUE`) or just the K-profile summaries (`FALSE`). The default is `FALSE`.

- `id.con`: This is an argument that allows one to impose some sort of (meaningful) identifiability constraint so that the mixture components are in some sort of comparable order between mixture models with different numbers of components. If `NULL`, then the components are ordered by the component means for univariate data or ordered by the first dimension of the component means for multivariate data.

- `score`: This is the value for the specified dimension reduction technique's score, which is used for constructing the mixturegram. By default, this value is `1`, which is the value that will typically be used. Larger values will result in more variability displayed on the mixturegram. Note that the largest value that can be calculated at each value of $k > 1$ on the mixturegram is $p + k - 1$, where p is the number of columns of `data`.

- `iter.max`: The maximum number of iterations allowed for the k-means clustering algorithm, which is passed to the `kmeans` function. The default is `50`.
- `nstart`: The number of random sets chosen based on k centers, which is passed to the `kmeans` function. The default is `25`.
- `...`: The ellipsis allows for additional arguments that can be passed to the underlying `plot` function.

Note that the above arguments reflect the version of the `mixturegram()` function published in `mixtools` at the time of the writing of this manuscript. We anticipate publishing future versions of this function that will afford greater flexibility for the user. We briefly discuss a couple of future improvements that we are considering.

One improvement for the `mixturegram()` function is to build it using the `ggplot2` graphics (Wickham, 2009). Development of the `mixtools` package began around 2005 and used base `R` for the graphical components. A general improvement for `mixtools` is to retrofit the graphical capabilities of the package using `ggplot2`, which will also allow for improved graphics when plotting the mixturegram.

Another consideration is to provide some intermediary functionality with the `mixturegram()` function to better understand how the different cluster means separate. This could be accomplished by checking how a group of variables in the $\mathbf{W}_k$ matrix (defined in Step 2 of the mixturegram) contribute to the separability of the cluster means.[1] An option could be included where one is able to obtain pairwise scatterplots of the variables in $\mathbf{W}_k^*$ and then color-code the points based on the clustering results, thus allowing a visual assessment of which variable(s) are possibly contributing to the mixture structure.

3 All Simulation Results

We first present the parametric forms for all of the distributions used in the simulations discussed in the main text. In the following, $\mathcal{N}()$, $\mathcal{N}_p()$, $\mathcal{G}()$, and $\mathcal{P}()$ are the univariate normal, p-variate normal,

---

[1] Recall that the clustering performed within the mixturegram is used to assign observations a particular color for plotting.

gamma, and Poisson distributions, respectively.

- Well-Separated Components:
  - *Model 1*: 4-Component Mixture of Univariate Normals

$$0.30\mathcal{N}(-12,1) + 0.30\mathcal{N}(-3,1) + 0.20\mathcal{N}(3,1) + 0.20\mathcal{N}(12,1)$$

  - *Model 2*: 3-Component Mixture of Poissons

$$0.30\mathcal{P}(3) + 0.30\mathcal{P}(40) + 0.40\mathcal{P}(90)$$

  - *Model 3*: 5-Component Mixture of 5-Variate Normals

$$\begin{aligned}
&0.20\mathcal{N}_5\left(\begin{bmatrix}-5\\-5\\-5\\-5\\-50\end{bmatrix}, \begin{bmatrix}1&0&0&0&0\\0&1&0&0&0\\0&0&1&0&0\\0&0&0&1&0\\0&0&0&0&1\end{bmatrix}\right) + 0.20\mathcal{N}_5\left(\begin{bmatrix}400\\40\\40\\40\\40\end{bmatrix}, \begin{bmatrix}27&25&20&11&5\\25&29&23&13&9\\20&23&29&20&15\\11&13&20&27&22\\5&9&15&22&28\end{bmatrix}\right)\\
&+ 0.20\mathcal{N}_5\left(\begin{bmatrix}-40\\-40\\-40\\-40\\-400\end{bmatrix}, \begin{bmatrix}7&6&6&3&0\\6&11&10&6&3\\6&10&11&7&4\\3&6&7&7&7\\0&3&4&7&11\end{bmatrix}\right) + 0.20\mathcal{N}_5\left(\begin{bmatrix}500\\100\\100\\100\\100\end{bmatrix}, \begin{bmatrix}16&12&10&7&2\\12&14&14&13&10\\10&14&20&16&13\\7&13&16&22&23\\2&10&13&23&31\end{bmatrix}\right)\\
&+ 0.20\mathcal{N}_5\left(\begin{bmatrix}125\\125\\125\\125\\-500\end{bmatrix}, \begin{bmatrix}712&538&612&291&117\\538&499&453&285&153\\612&453&711&482&192\\291&285&482&757&581\\117&153&192&581&692\end{bmatrix}\right)
\end{aligned}$$

- Moderately-Separated Components:

  - *Model 4*: 3-Component Mixture of Gammas

$$0.20\mathcal{G}(1,2) + 0.30\mathcal{G}(30,1) + 0.50\mathcal{G}(50,2)$$

  - *Model 5*: 4-Component Mixture of Poissons

$$0.20\mathcal{P}(3) + 0.20\mathcal{P}(20) + 0.30\mathcal{P}(40) + 0.30\mathcal{P}(70)$$

  - *Model 6*: 4-Component Mixture of Tetravariate Normals

$$0.30\mathcal{N}_4\left(\begin{bmatrix}-25\\-25\\-25\\-10\end{bmatrix}, \begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&1&0\\0&0&0&1\end{bmatrix}\right) + 0.20\mathcal{N}_4\left(\begin{bmatrix}-10\\-8\\-10\\-20\end{bmatrix}, \begin{bmatrix}21&19&14&6\\19&23&17&8\\14&17&22&14\\6&8&14&21\end{bmatrix}\right)$$
$$+\,0.20\mathcal{N}_4\left(\begin{bmatrix}10\\10\\10\\0\end{bmatrix}, \begin{bmatrix}8&8&5&1\\8&10&7&4\\5&7&8&6\\1&4&6&8\end{bmatrix}\right) + 0.20\mathcal{N}_4\left(\begin{bmatrix}55\\10\\55\\10\end{bmatrix}, \begin{bmatrix}1&-1&2&2\\-1&11&-2&-13\\2&-2&19&2\\2&-13&2&24\end{bmatrix}\right)$$

- Heavily-Overlapping Components:

  - *Model 7*: 3-Component Mixture of Univariate Normals

$$0.50\mathcal{N}(0,3) + 0.30\mathcal{N}(2,1) + 0.20\mathcal{N}(-2,1)$$

  - *Model 8*: 4-Component Mixture of Poissons

$$0.30\mathcal{P}(3) + 0.20\mathcal{P}(10) + 0.30\mathcal{P}(15) + 0.20\mathcal{P}(30)$$

- *Model 9*: 5-Component Mixture of Trivariate Normals

$$
\begin{aligned}
&0.20\mathcal{N}_3\left(\begin{bmatrix}0\\0\\0\end{bmatrix},\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix}\right)+0.20\mathcal{N}_3\left(\begin{bmatrix}3\\3\\3\end{bmatrix},\begin{bmatrix}16&13&8\\13&17&10\\8&10&16\end{bmatrix}\right)\\
&+0.20\mathcal{N}_3\left(\begin{bmatrix}-3\\0\\-1.5\end{bmatrix},\begin{bmatrix}1&-3&-2\\-3&11&8\\-2&8&22\end{bmatrix}\right)+0.20\mathcal{N}_3\left(\begin{bmatrix}-5\\-5\\-5\end{bmatrix},\begin{bmatrix}1&-3&-2\\-3&11&8\\-2&8&22\end{bmatrix}\right)\\
&+0.20\mathcal{N}_3\left(\begin{bmatrix}5\\5\\5\end{bmatrix},\begin{bmatrix}1&-3&5\\-3&11&-19\\5&-19&51\end{bmatrix}\right)
\end{aligned}
$$

Note that simulations involving Models 1, 6, and 7 were discussed in the main text.

We next present and discuss all of the mixturegrams for the different simulation settings. The mixturegrams are organized as follows:

- Mixturegrams for the models with well-separated components are given in Figures 1, 4, 7, 10, and 13.
- Mixturegrams for the models with moderately-separated components are given in Figures 2, 5, 8, 11, and 14.
- Mixturegrams for the models with heavily-overlapping components are given in Figures 3, 6, 9, 12, and 15.
- Mixturegrams based on the first principal component are given in Figures 1, 2, and 3.
- Mixturegrams based on the second principal component are given in Figures 4, 5, and 6.
- Mixturegrams based on the first kernel principal component are given in Figures 7, 8, and 9.
- Mixturegrams based on the second kernel principal component are given in Figures 10, 11, and 12.
- Mixturegrams based on reduced rank linear discriminant analysis are given in Figures 13, 14, and 15.

Figure 16 gives plots of the average proportion of within-cluster point scatter, $\pi^{(k)}$, versus the number of components, k, for all nine mixture models considered.

Some general observations made about these mixturegrams:

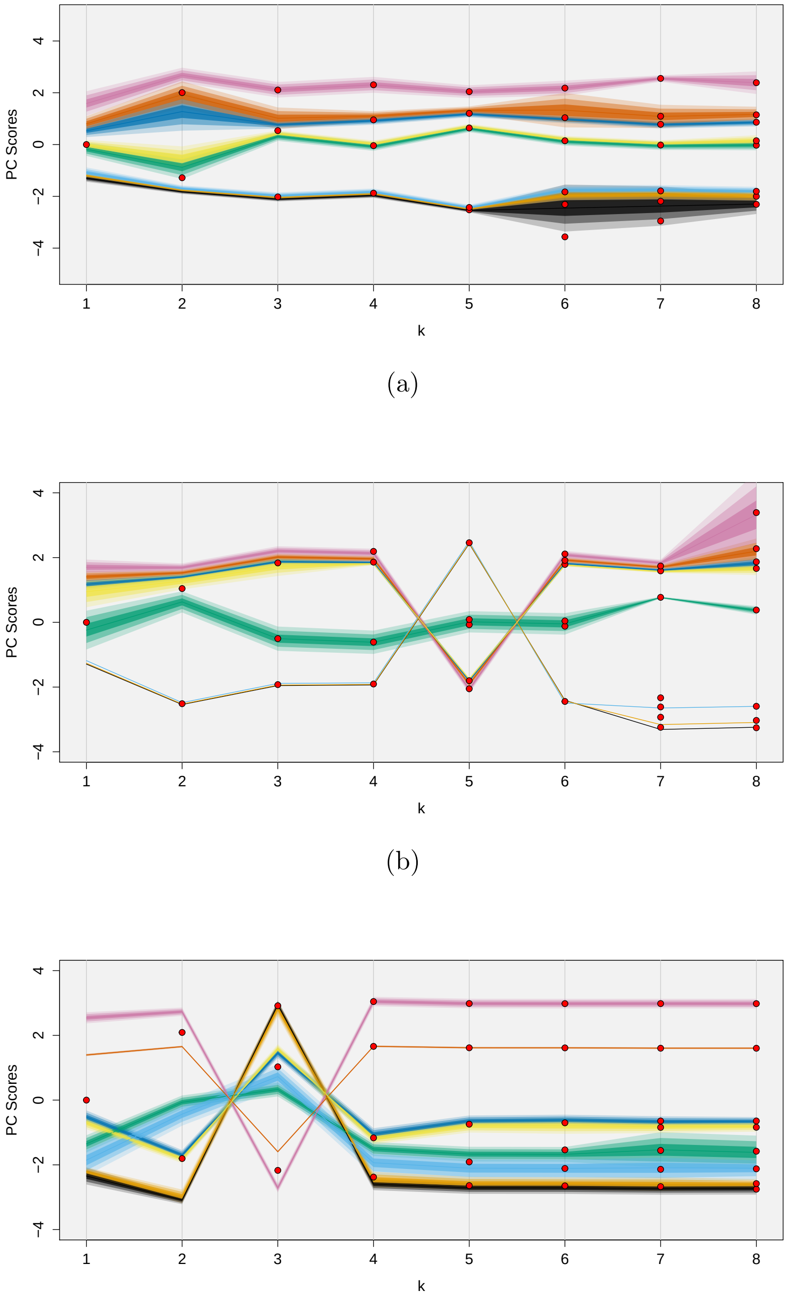
- For the well-separated components cases, principal components and reduced rank linear discriminant analysis provide mixturegrams with fairly good separation in the profiles, thus allowing one to easily determine an appropriate number of components. Kernel principal components analysis yields plots that are a bit harder to interpret. However, we are able to see in all figures values of k where there is distinct separations between the clusters and the k cluster means (i.e., red dots) have a similar type of visual separation. These values of k are what we would typically select for our number of components.

- For the moderately-separated components cases, we still have some success in interpreting the mixturegrams like with the well-separated components cases. Again, principal components analysis and reduced rank linear discriminant analysis tend to produce more visually informative mixturegrams compared to kernel principal components analysis.

- For the heavily-overlapping components cases, the mixturegrams become difficult to interpret overall. However, some distinct clusters of profiles do emerge, suggesting that there may be a mixture structure present, but providing an objective and definitive assessment of the proper number of components is challenging.

- Use of the second principal or second kernel principal component yields mixturegrams that are harder to interpret. More variability gets introduced in the figure. With that said, visual separation in the profiles is sometimes noticeable, however, again, providing a definitive assessment of the proper number of components is challenging.

- In general, reduced rank linear discriminant analysis appears to only be efficacious when the components are well-separated. In Figure 13, we can ascertain a distinct number of profiles for each of the three models when assessing k^*. In the univariate settings for the well-separated case (i.e., Figures 13(a) and 13(b)), we can easily see $k^* = 4$ and $k^* = 3$ are appropriate

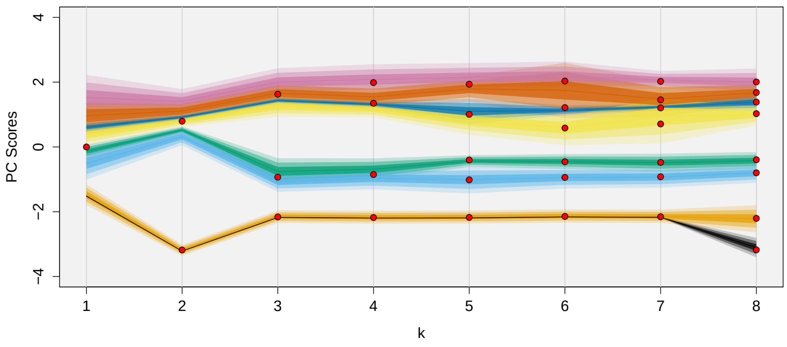


(a)

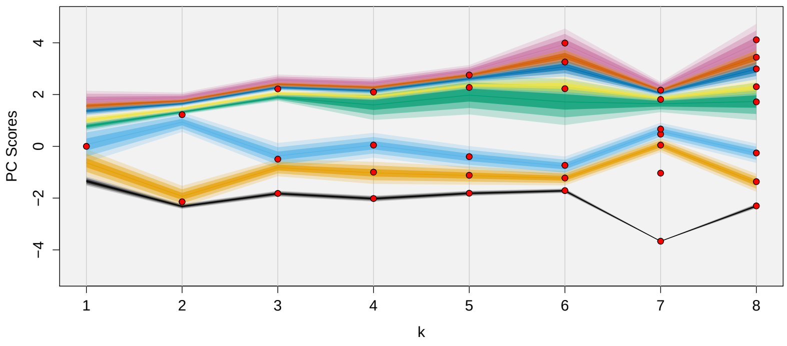
(b)

(c)

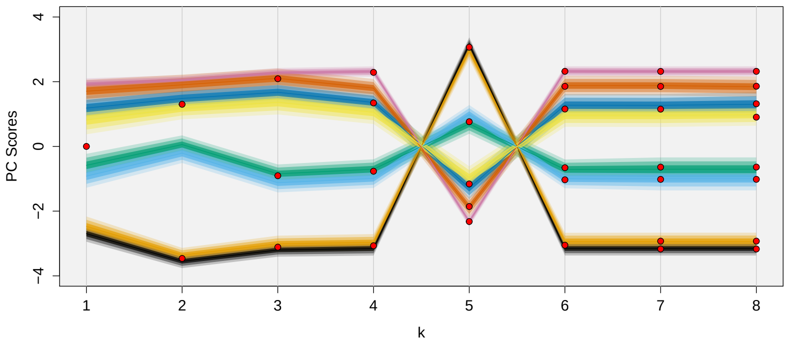
Figure 1: Mixturegrams constructed using the first principal component for (a) Model 1, (b) Model 2, and (c) Model 3. Each mixturegram shows a particular number of distinct profile groupings that correspond to the true number of components.



(a)

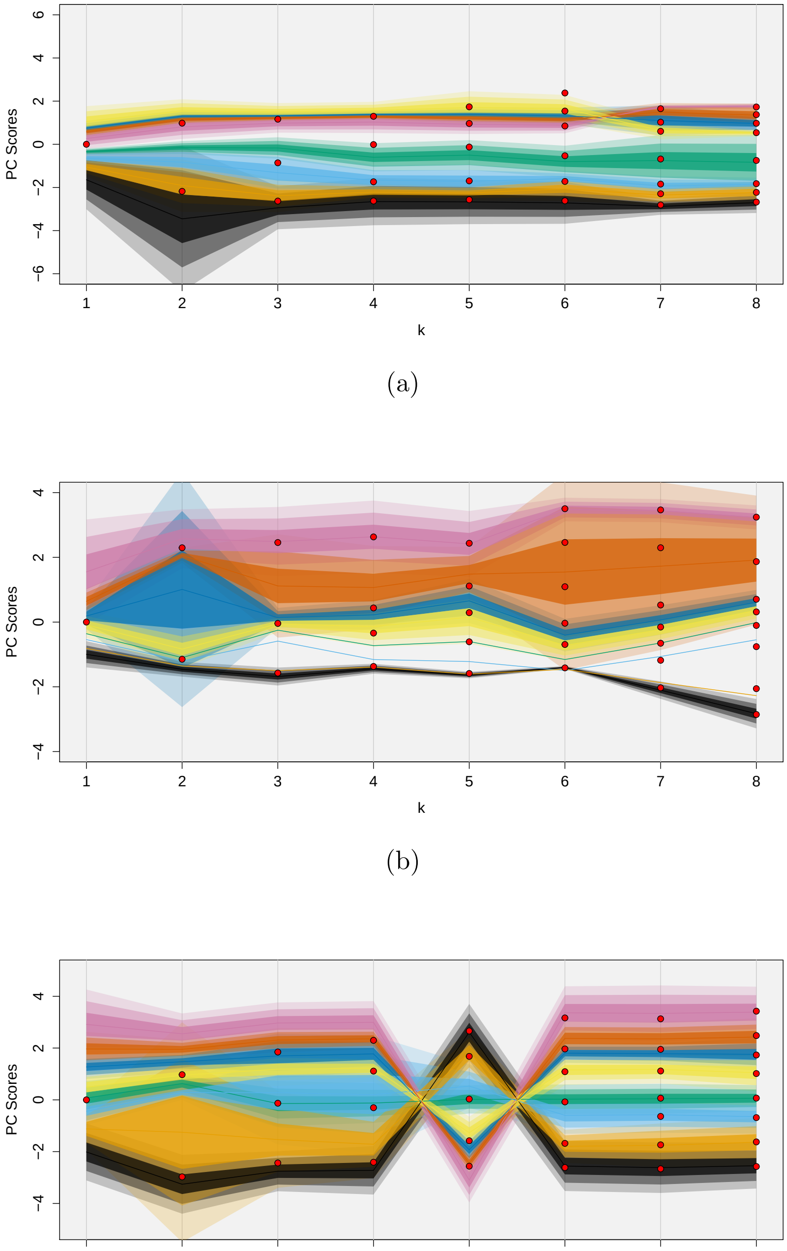


(b)



(c)

Figure 2: Mixturegrams constructed using the first principal component for (a) Model 4, (b) Model 5, and (c) Model 6. Even though the data were generated from a moderately-separated mixture component structure, each mixturegram shows a particular number of distinct profile groupings that are at or within one component of the true number of components.

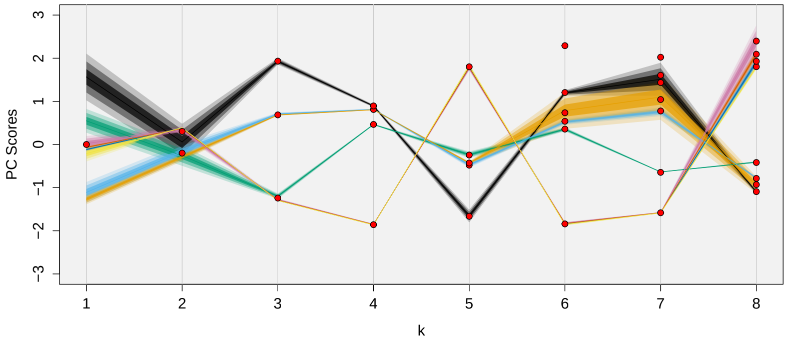


(a)

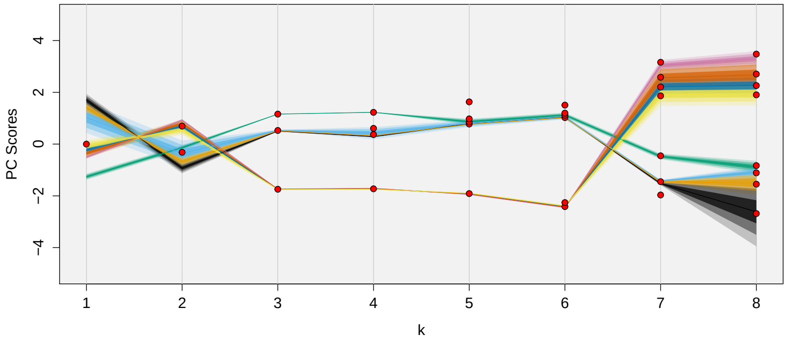
(b)

(c)

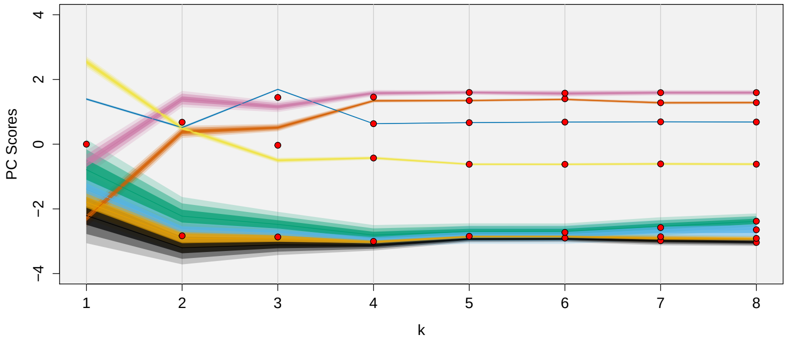
Figure 3: Mixturegrams constructed using the first principal component for (a) Model 7, (b) Model 8, and (c) Model 9. The mixturegrams give some indication of a possible mixture structure, but due to the heavily-overlapping structure of the components, it is difficult to make a definitive selection for the number of components.



(a)

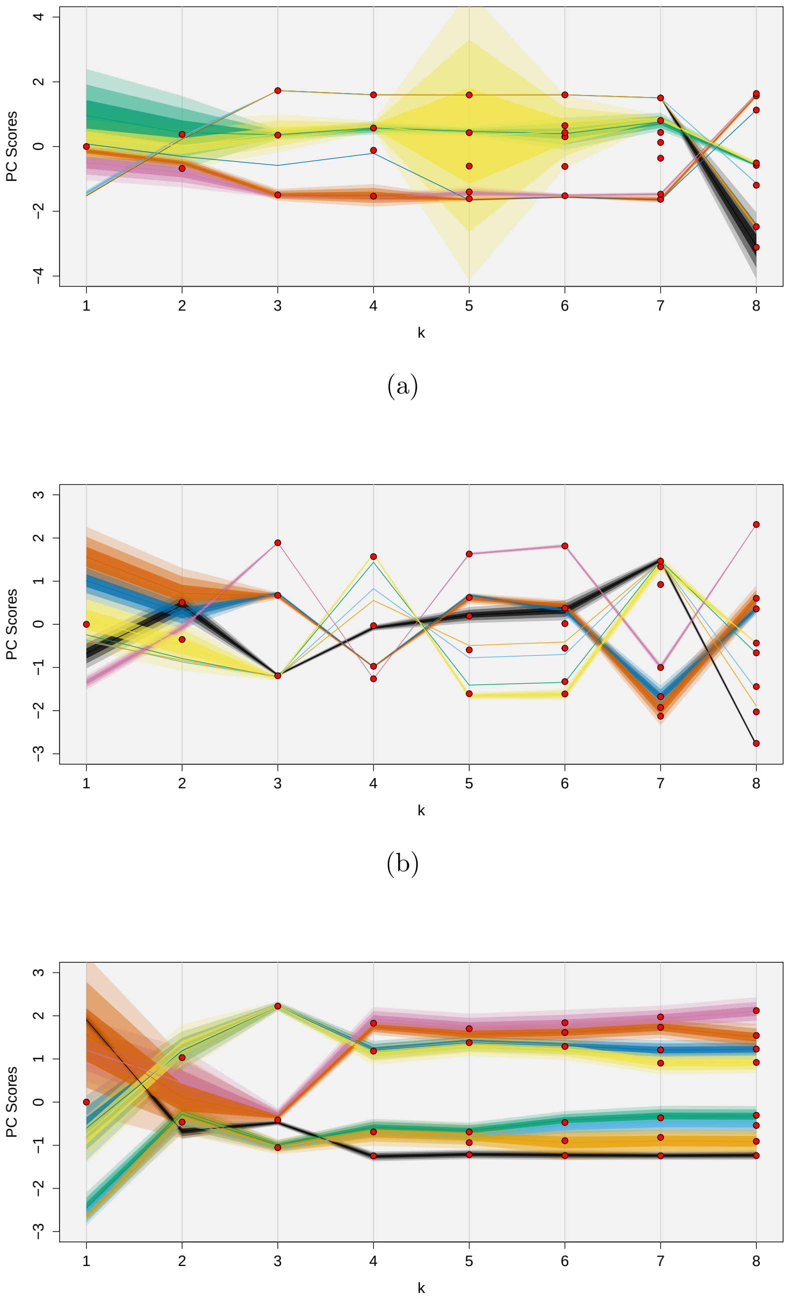


(b)



(c)

Figure 4: Mixturegrams constructed using the second principal component for (a) Model 1, (b) Model 2, and (c) Model 3. A mixture structure is clearly indicated in each dataset, but the exact number is not as clear as in the mixturegrams based on the first principal component in Figure 4.

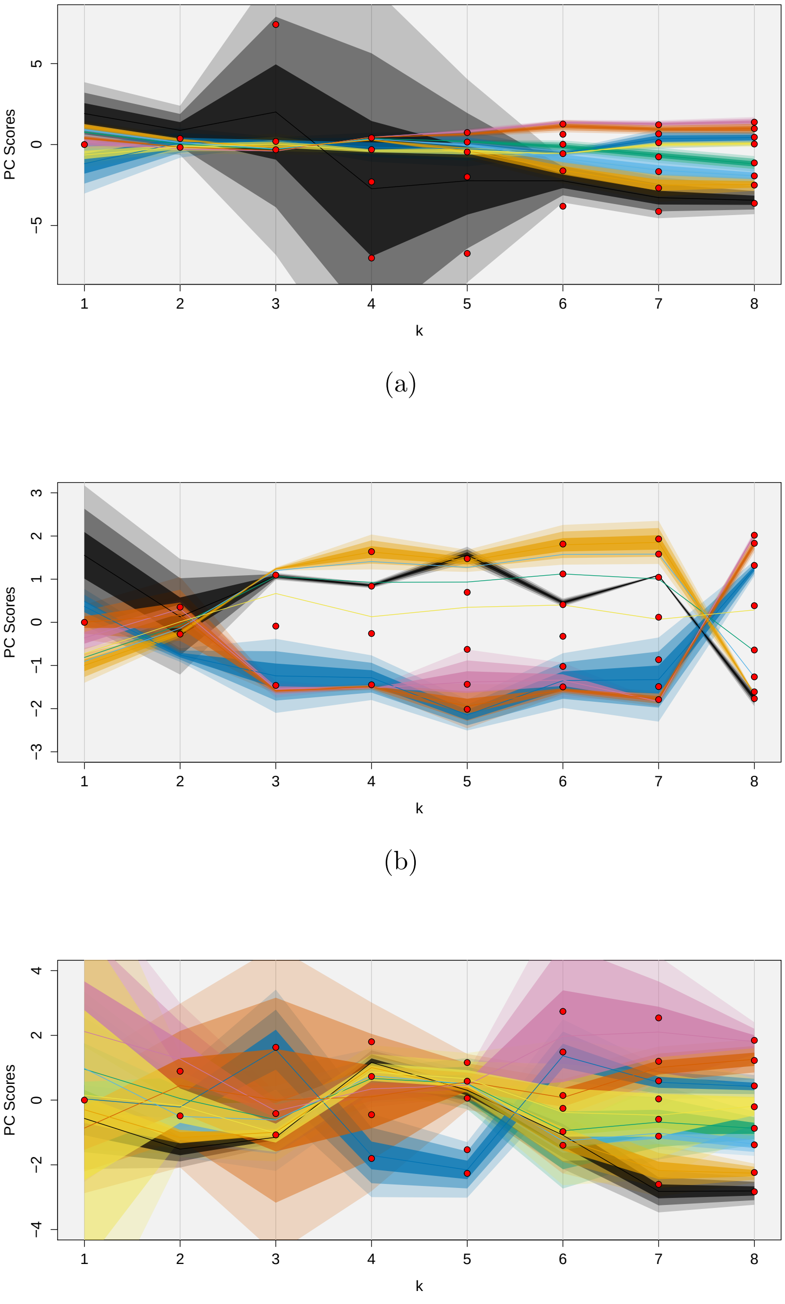


(a)

(b)

(c)

Figure 5: Mixturegrams constructed using the second principal component for (a) Model 4, (b) Model 5, and (c) Model 6. A mixture structure is clearly indicated in each dataset, but the exact number is not as clear as in the mixturegrams based on the first principal component in Figure 2.

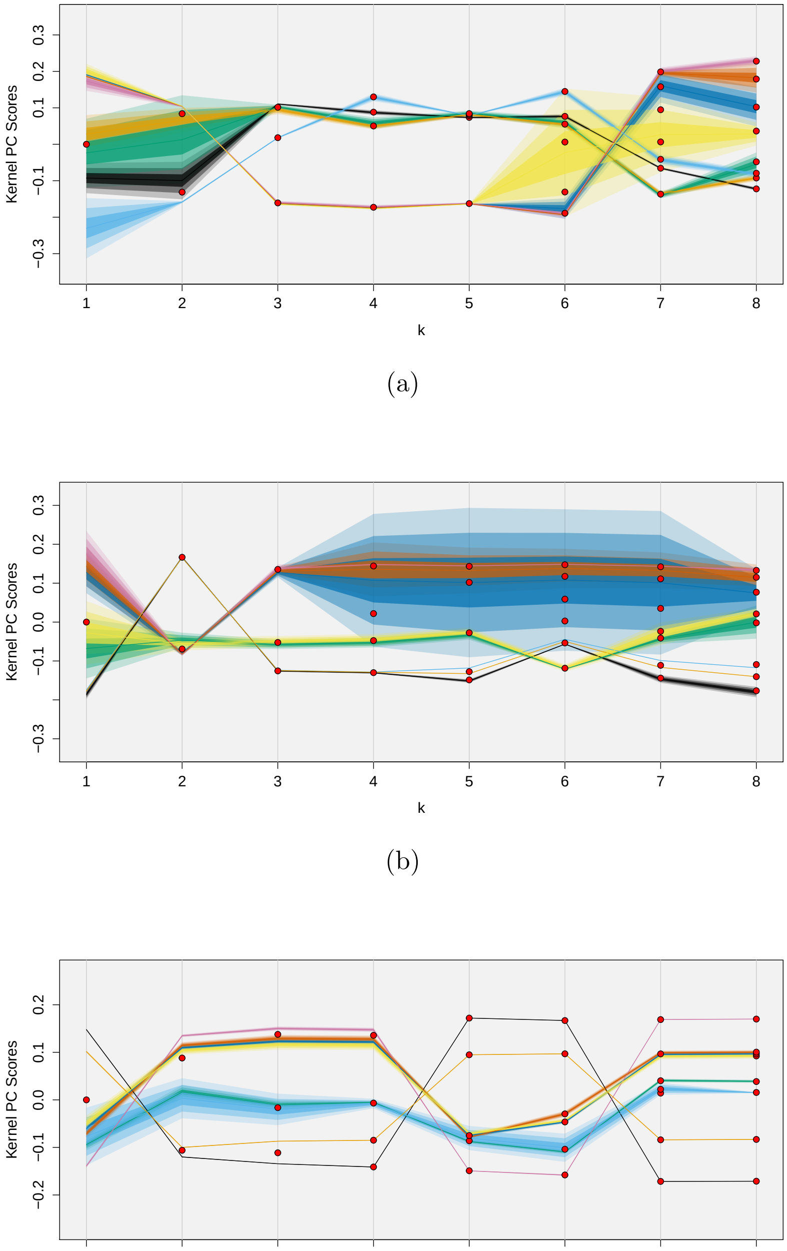


(a)

(b)

(c)

Figure 6: Mixturegrams constructed using the second principal component for (a) Model 7, (b) Model 8, and (c) Model 9. These mixturegrams do not provide any clear guidance compared to the mixturegrams based on the first principal component in Figure 3.

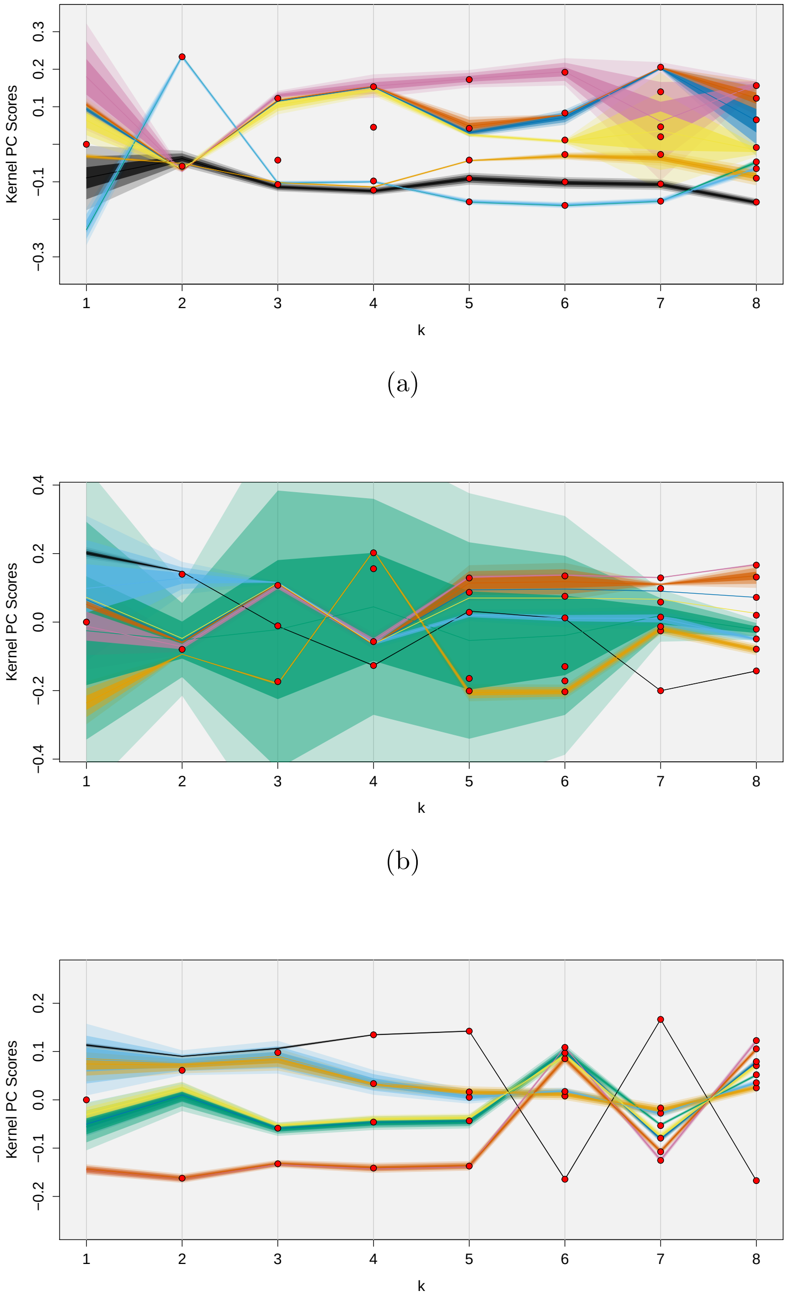


(a)

(b)

(c)

Figure 7: Mixturegrams constructed using the first kernel principal component for (a) Model 1, (b) Model 2, and (c) Model 3. A similar assessment can be made as with the mixturegrams based on the first principal component in Figure 1, however, the profiles appear to cross a bit more in these mixturegrams.

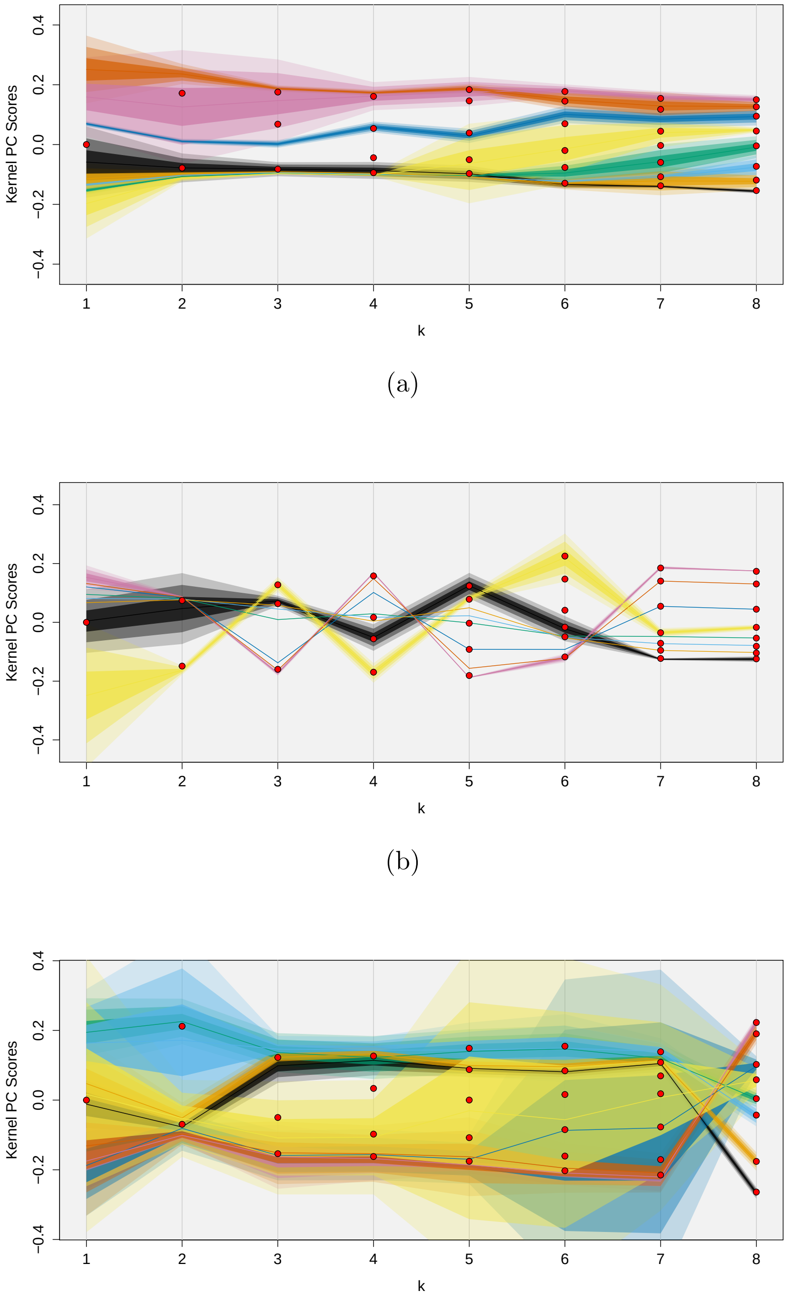


(a)

(b)

(c)

Figure 8: Mixturegrams constructed using the first kernel principal component for (a) Model 4, (b) Model 5, and (c) Model 6. A similar assessment can be made as with the mixturegrams based on the first principal component in Figure 2, however, the profiles appear to have more variable in these mixturegrams.

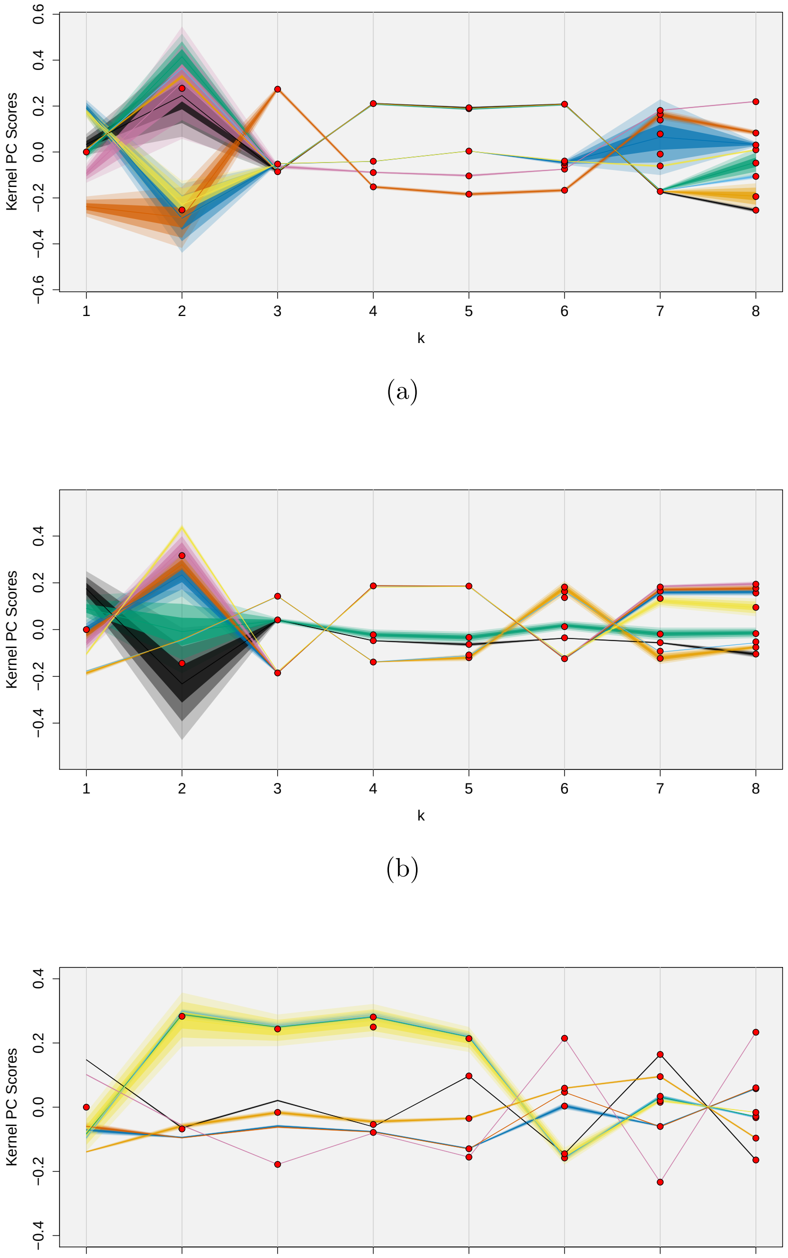


(a)

(b)

(c)

Figure 9: Mixturegrams constructed using the first kernel principal component for (a) Model 7, (b) Model 8, and (c) Model 9. A similar assessment can be made as with the mixturegrams based on the first principal component in Figure 3.

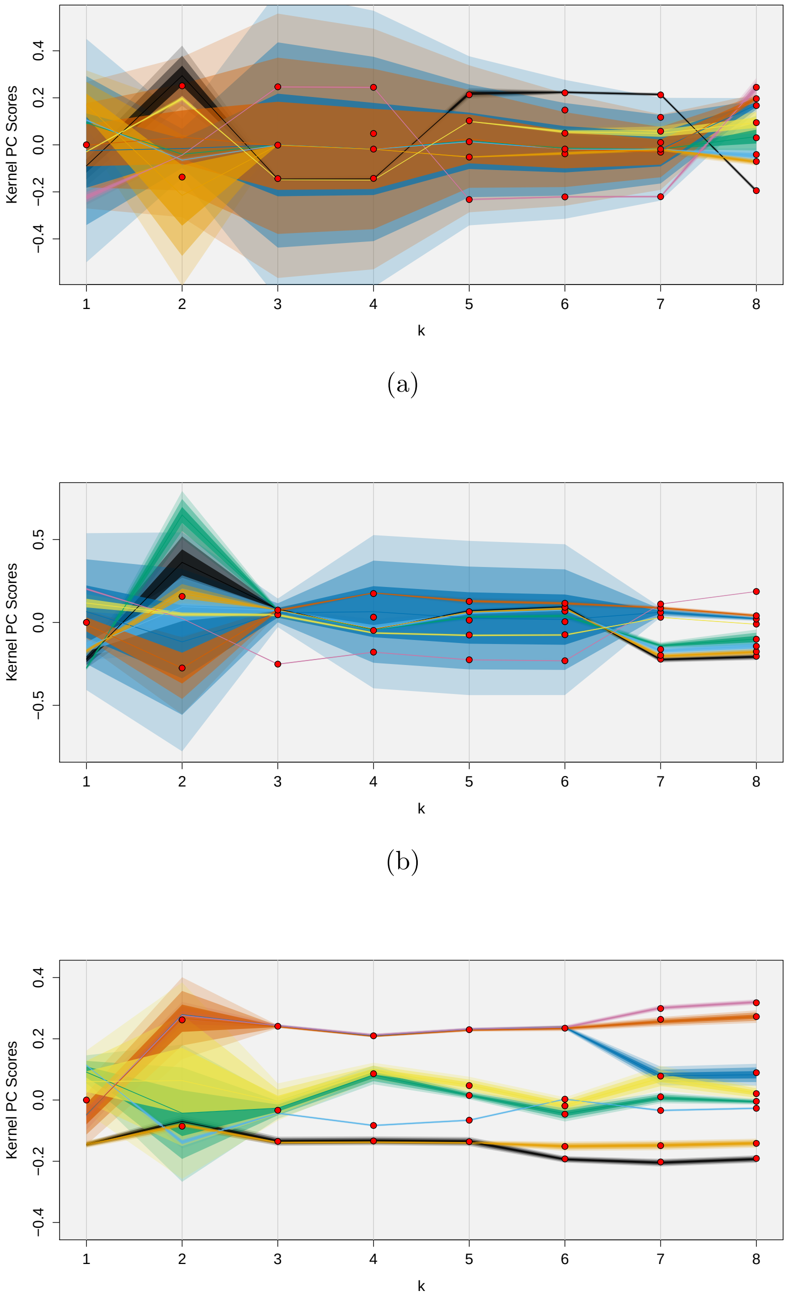


(a)

(b)

(c)

Figure 10: Mixturegrams constructed using the second kernel principal component for (a) Model 1, (b) Model 2, and (c) Model 3. A mixture structure is clearly indicated in each dataset such that there appears to be a number of distinct profiles that matches the true number of components.

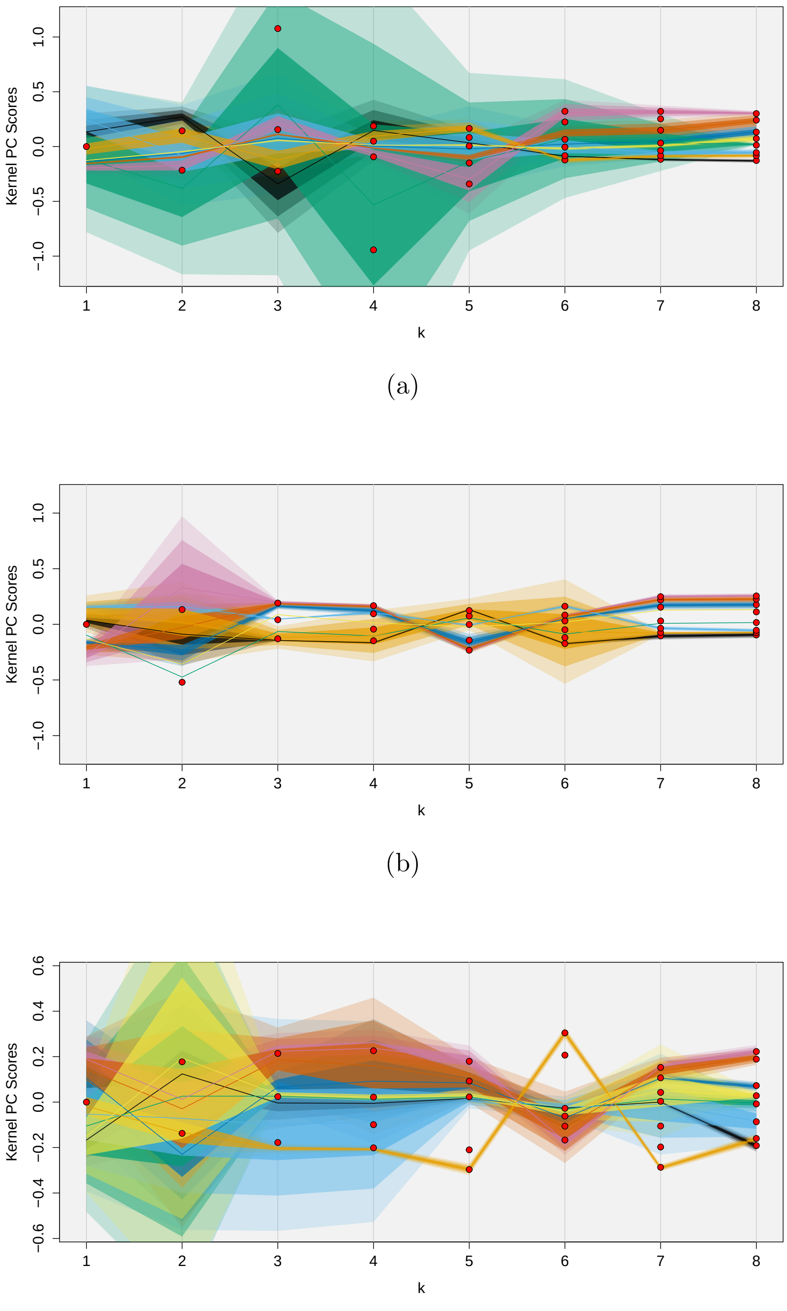


(a)

(b)

(c)

Figure 11: Mixturegrams constructed using the second kernel principal component for (a) Model 4, (b) Model 5, and (c) Model 6. A mixture structure is clearly indicated in each dataset, but the exact number is not as clear as in the mixturegrams based on the first kernel principal component in Figure 8.

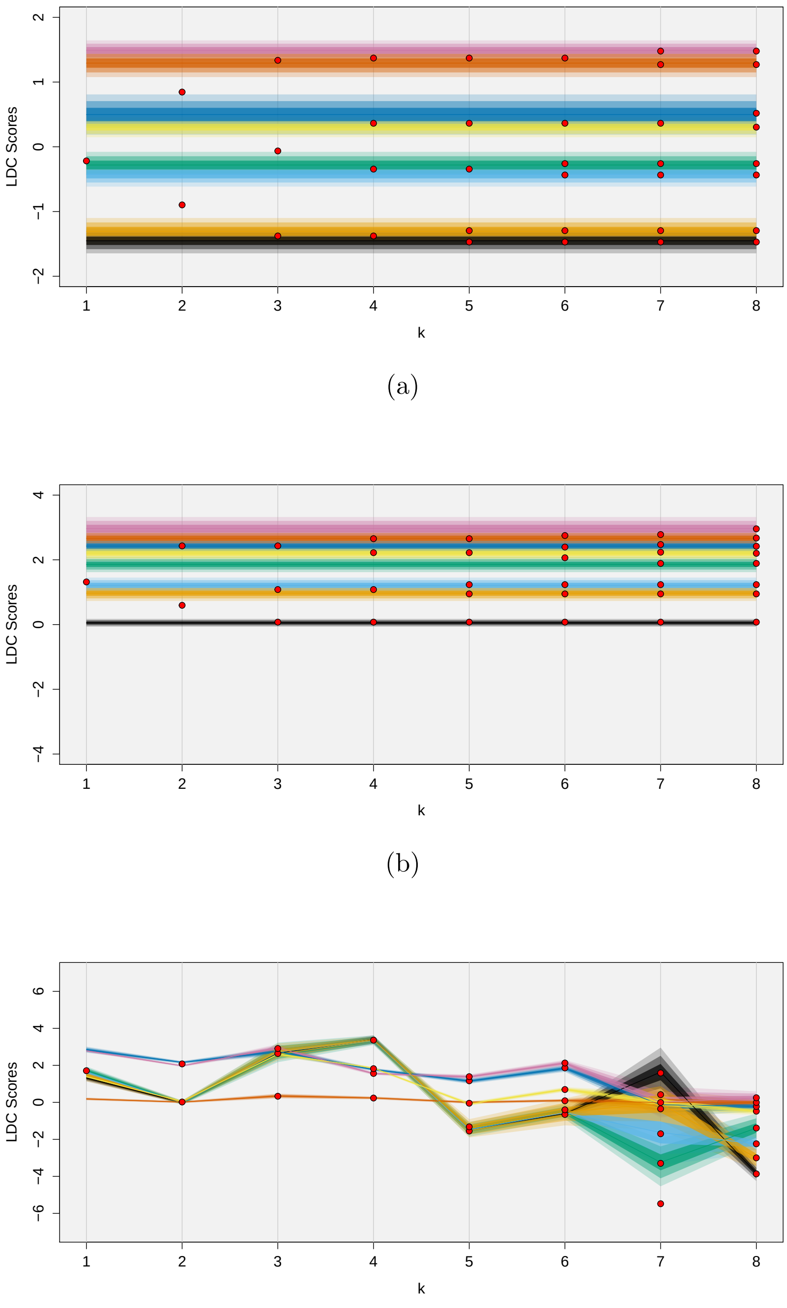


(a)

(b)

(c)

Figure 12: Mixturegrams constructed using the second kernel principal component for (a) Model 7, (b) Model 8, and (c) Model 9. Due to the heavily-overlapping structure of the mixture components, it is difficult to provide any clear guidance on the appropriate number of components.

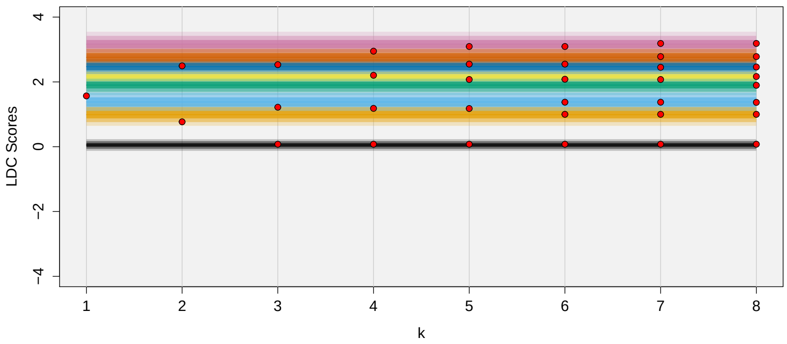


(a)

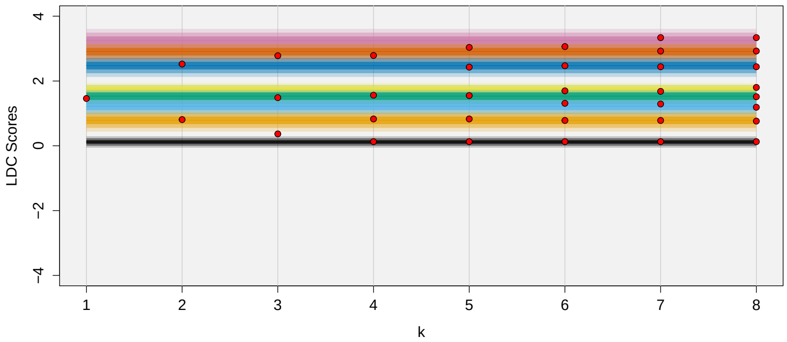
(b)

(c)

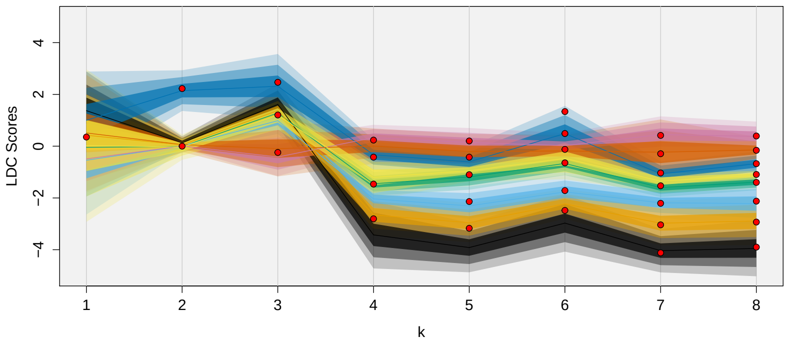
Figure 13: Mixturegrams constructed using reduced rank linear discriminant anlaysis for (a) Model 1, (b) Model 2, and (c) Model 3. The number of distinct profiles (which are horiztonal lines) for Models 1 and 2 match the true number of components, however, the number of profiles for Model 3 seems to indicate less than the true number of components.



(a)

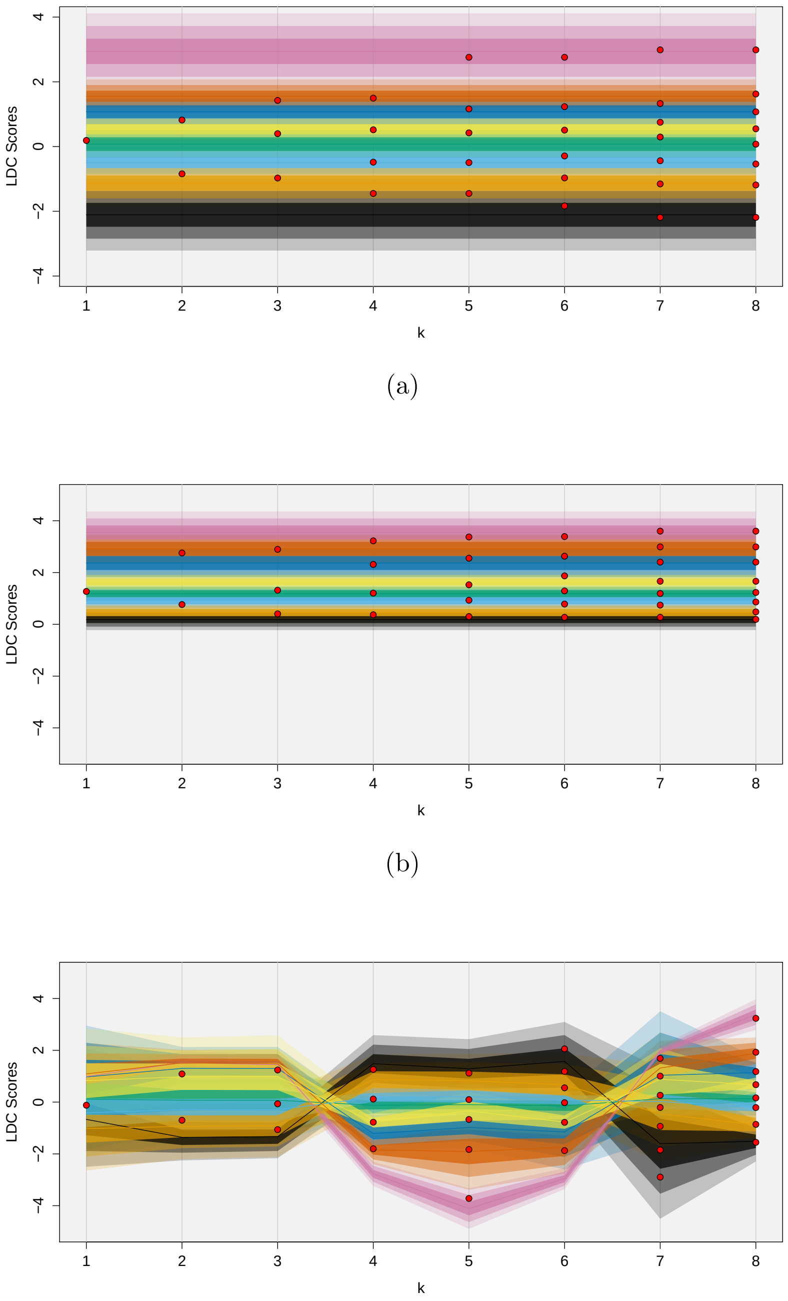


(b)



(c)

Figure 14: Mixturegrams constructed using reduced rank linear discriminant anlaysis for (a) Model 4, (b) Model 5, and (c) Model 6. There appears to be some distinct groupings of profiles shown in each mixturegram, but there is no clear guidance that can be provided from these plots.

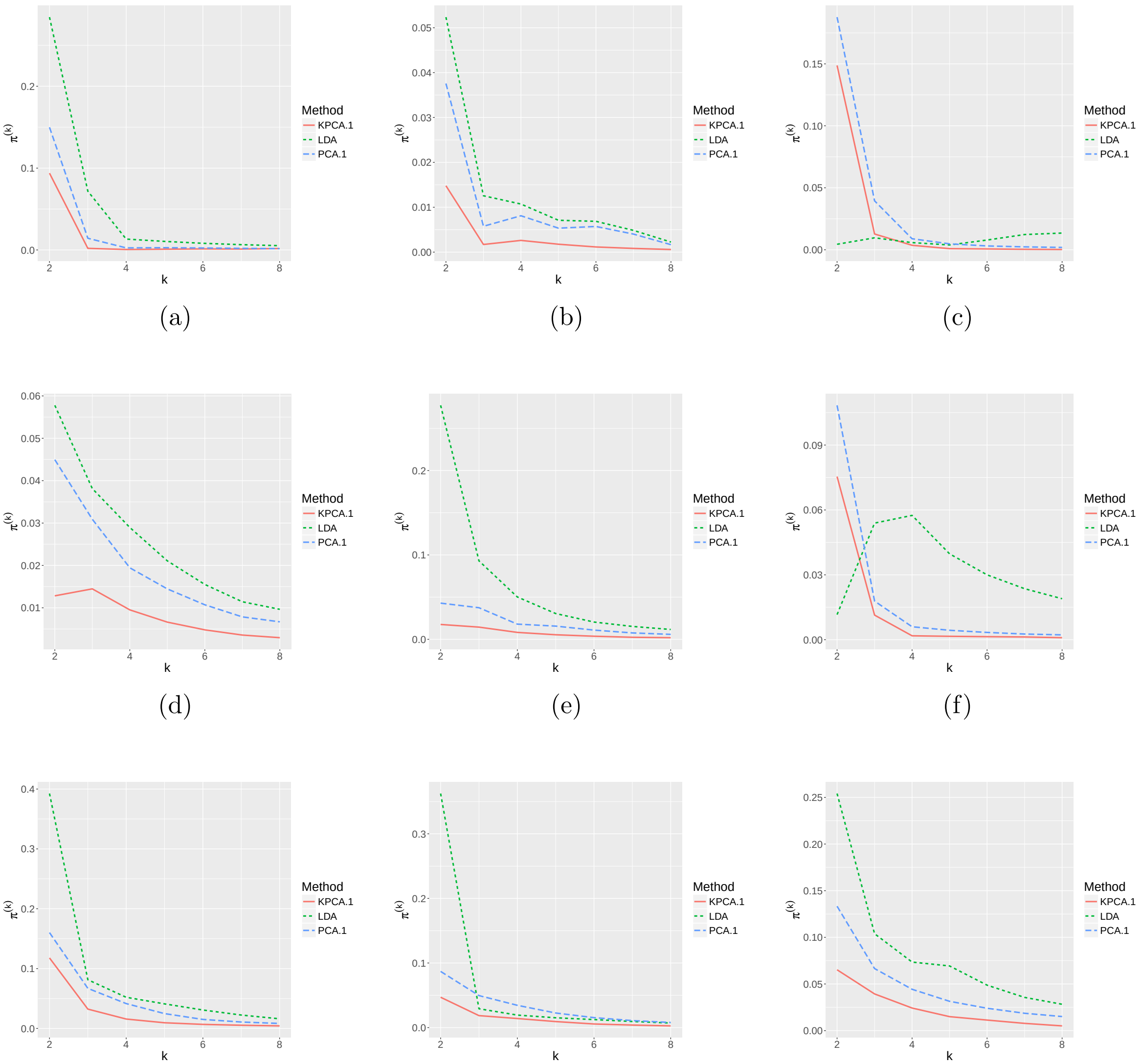


(a)

(b)

(c)

Figure 15: Mixturegrams constructed using reduced rank linear discriminant anlaysis for (a) Model 7, (b) Model 8, and (c) Model 9. All of these mixturegrams indicate that a mixture structure is not present.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 16: Plots of the average proportion of within-cluster point scatter, $\pi^{(k)}$, versus the number of components, k, for (a) Model 1, (b) Model 2, (c) Model 3, (d) Model 4, (e) Model 5, (f) Model 6, (g) Model 7, (h) Model 8, and (i) Model 9. Wherever the average values for each $\pi^{(k)}$ tend to level-off or $\pi^{(k+1)} > \pi^{(k)}$ is indicative of an appropriate choice for the value of k.

choices. However, Figure 13(c) indicates that at least $k^* = 4$ is appropriate, but it is much more subjective about the interpretation at $k^* = 5$, which is the correct number of components for this mixture simulation. Thus, using reduced rank linear discriminant analysis as the dimension reduction technique can yield more subjective results – compared to principal components and kernel principal components – even in a well-separated setting.

- In Figure 16, we see locations where the change in the average trend becomes considerably smaller. These locations on the figures are where we suspect reasonable candidates for the number of components. For the well-separated components cases, the results obtained from these plots generally agree with the true number of components, regardless of the dimension reduction technique employed. However, for some of the moderately-separated components cases and the heavily-overlapping components cases, these results are difficult to interpret and one might select a large number of components if the profile does not have a clear elbow in its shape. These figures also emphasize the *ad hoc* nature of this criterion. The results appear to be consistent with our interpretations of the mixturegrams for the well-separated cases, but less informative for some of the moderately-separated and heavily-overlapping cases. However, the interpretations that we make using the $\pi^{(k)}$ values for the data analyses tend to agree with what we select based on our assessments of the corresponding mixturegrams as well as the calculated information criteria. Thus, this *ad hoc* criterion can serve a confirmatory role regarding the final selected value k^*.

In Table 1, we report the percentage of times each model selection criterion selected the appropriate number of k for our $B = 100$ simulated datasets. Clearly the more separation we have between the components, the better all of the criteria perform at selecting the correct number of components. In practice, if one uses a traditional quantitative metric for assessing the number of components – like model selection criteria or other approaches discussed in Section 2 of the main text – we advocate also displaying the mixturegram as an effective visualization component for such an assessment.

Model	k	AIC	BIC	ICL	CAIC
Well-Separated Components					
Model 1	4	93%	100%	100%	100%
Model 2	3	96%	100%	100%	100%
Model 3	5	31%	100%	100%	100%
Moderately-Separated Components					
Model 4	3	93%	100%	100%	100%
Model 5	4	92%	100%	100%	100%
Model 6	4	46%	100%	100%	100%
Heavily-Overlapping Components					
Model 7	3	17%	7%	8%	5%
Model 8	4	51%	11%	13%	5%
Model 9	5	12%	6%	6%	1%

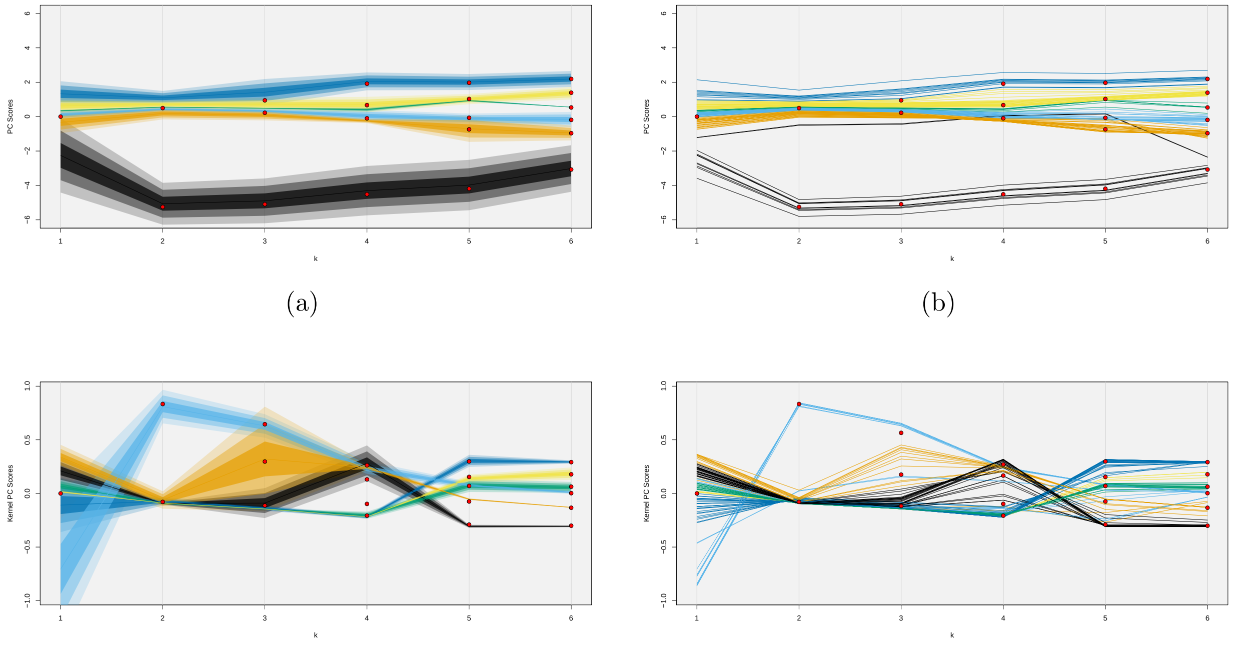
Table 1: Percentage of times each model selection criterion selected the correct number of components for the three simulation conditions. The results for the models having well-separated or moderately-separated components are in good agreement, wherease those for the heavily-overlapping components are not able to accurately select the true number of components.

4 Data Analysis

4.1 Additional Results for Quasar Data

We constructed mixturegrams for the quasar data based on PCA with the K-profile summaries (Figure 17(a)) and with all observations plotted (Figure 17(d)) as well as based on KPCA with the K-profile summaries (Figure 17(c)) and with all observations plotted (Figure 17(d)). There are clearly two distinct groupings of profiles in the PCA-based mixturegrams. The KPCA-based mixturegrams also demonstrate two groupings of profiles, but the profiles cross and do not demonstrate the same visual separation as with the PCA-based mixturegrams. Regardless, these mixturegrams indicate that a 2-component mixture of normals appears appropriate for these data.

All of the model selection results and values of $\pi^{(k)}$ for the quasar data are reported in Table 2. All of the model selection criteria would select $k^* = 2$ as an appropriate number of components. The values of $\pi^{(k)}$ also indicate that $k^* = 2$ is appropriate, mainly because going from $k = 1$ to $k = 2$ has a substantial relative decrease, whereas going from $k = 2$ to $k = 3$ is a much smaller relative decrease. Thus, the mixturegram interpretation, the model selection criteria, and the criterion based on $\pi^{(k)}$ are all in agreement.



(a) (b)

(c) (d)

Figure 17: Mixturegrams of the quasar data based on (a) PCA with the K-profile summaries, (b) PCA with all observations plotted, (c) KPCA with the K-profile summaries, and (d) KPCA with all observations plotted. There are clearly two distinct groupings of profiles in the PCA-based mixturegrams, while the KPCA-based mixturegrams demonstrate a clear separation at $k = 2$. Thus, a 2-component mixture of normals appears appropriate for these data.

4.2 Additional Results for DLBCL Data

Lee and McLachlan (2013) discuss model-based clustering with non-normal multivariate mixture distributions, such as the (unrestricted) multivariate skew t distribution. $\mathbf{X} \in \mathbb{R}^p$ is said to follow an (unrestricted) multivariate skew t distribution with location vector $\boldsymbol{\mu} \in \mathbb{R}^p$, scale matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$, skewness vector $\boldsymbol{\delta} \in \mathbb{R}^p$, and (scalar) degrees of freedom ν, if it has probability density function

$$g(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\delta}, \nu) = 2^p t_{p,\nu}\left(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) T_{p,\nu}\left(\boldsymbol{\Delta}\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\sqrt{\frac{\nu + p}{\nu + \|\boldsymbol{\Omega}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})\|^2}}; \mathbf{0}, \boldsymbol{\Lambda}, \nu + p\right),$$

where $\boldsymbol{\Delta} = \operatorname{diag}(\boldsymbol{\delta})$, $\boldsymbol{\Omega} = \boldsymbol{\Sigma} + \boldsymbol{\Delta}^2$, $\boldsymbol{\Lambda} = \mathbf{I}_{p \times p} - \|\boldsymbol{\Omega}^{-1/2}\boldsymbol{\Delta}\|^2$, and $t_{p;\nu}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $T_{p;\nu}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ are the probability density function and cumulative distribution function, respectively, of the p-dimensional t distribution with location vector $\boldsymbol{\mu}$ and scale matrix $\boldsymbol{\Sigma}$.

We constructed mixturegrams for the DLBCL data based on PCA with the K-profile summaries

Criterion	Number of Components (k)					
	1	2	3	4	5	6
AIC	55.487	**76.818**	74.516	73.035	71.043	70.040
BIC	51.784	**67.559**	59.701	52.664	45.116	38.558
ICL	51.784	**67.855**	60.585	53.841	46.575	40.196
CAIC	50.784	**65.059**	55.701	47.164	38.116	30.058
$\pi^{(k)}$ (PCA)	1.000	**0.041**	0.029	0.019	0.015	0.019
$\pi^{(k)}$ (KPCA)	1.000	**0.009**	0.057	0.018	0.015	0.006

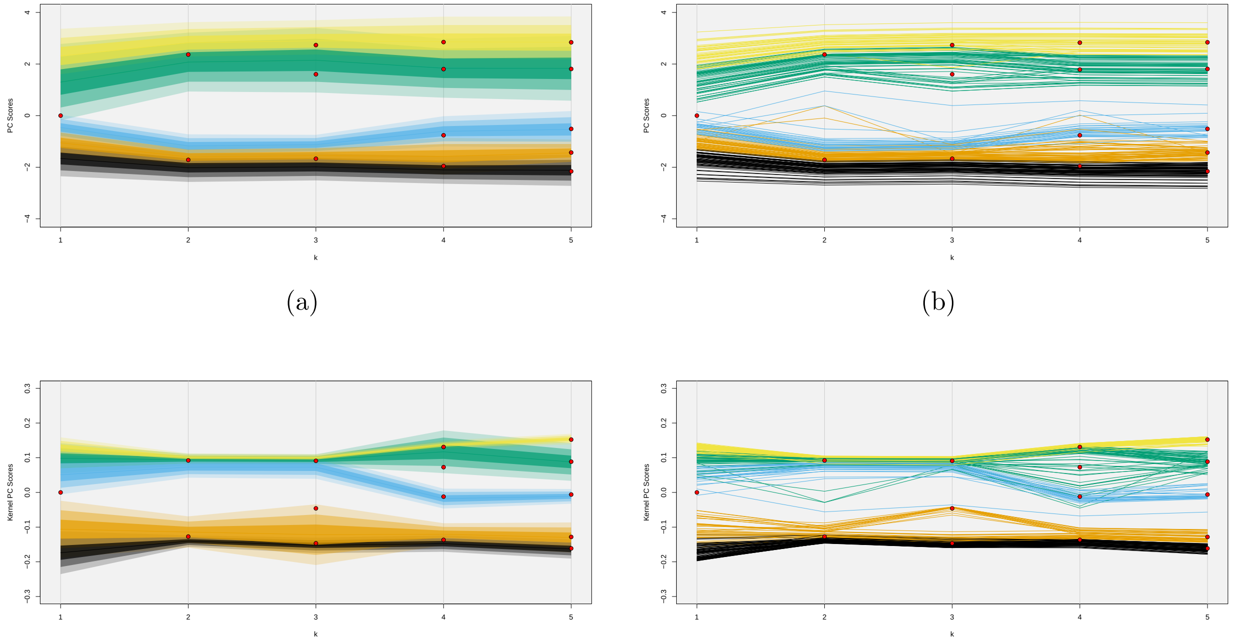
Table 2: Model selection and $\pi^{(k)}$ results for the quasar data. Quantities in boldface pertain to the respective number of components k that are selected for the mixture model. There is agreement across all measures to select $k^* = 2$.

(Figure 18(a)) and with all observations plotted (Figure 18(d)) as well as based on KPCA with the K-profile summaries (Figure 18(c)) and with all observations plotted (Figure 18(d)). There are clearly two distinct groupings of profiles in all of these mixturegrams. Thus, a 2-component mixture of trivariate skew t distributions appears appropriate for these data.

All of the model selection results and values of $\pi^{(k)}$ for the DLBCL data are reported in Table 3. With the exception of AIC, the other selection criteria all select $k^* = 2$ as an appropriate number of components. Moreover, the values of $\pi^{(k)}$ indicate that the value of $k^* = 2$ is appropriate since there is only a moderate decrease in the value of $\pi^{(3)}$ relative to $\pi^{(2)}$. Thus, we would select $k^* = 2$ as the number of components to use for this mixture problem.

Criterion	Number of Components (k)				
	1	2	3	4	5
AIC	-4537.789	-4370.414	**-4359.522**	-4360.523	-4365.228
BIC	-4560.678	**-4417.953**	-4431.712	-4457.363	-4486.718
ICL	-4560.678	**-4418.294**	-4432.138	-4458.001	-4487.496
CAIC	-4558.928	**-4414.203**	-4425.962	-4449.613	-4476.968
$\pi^{(k)}$ (PCA)	1.000	**0.054**	0.035	0.027	0.019
$\pi^{(k)}$ (KPCA)	1.000	**0.025**	0.010	0.012	0.008

Table 3: Model selection and $\pi^{(k)}$ results for the DLBCL data. Quantities in boldface pertain to the respective number of components k that are selected for the mixture model. There is majority agreement across the measures to select $k^* = 2$.



(a)

(b)

(c)

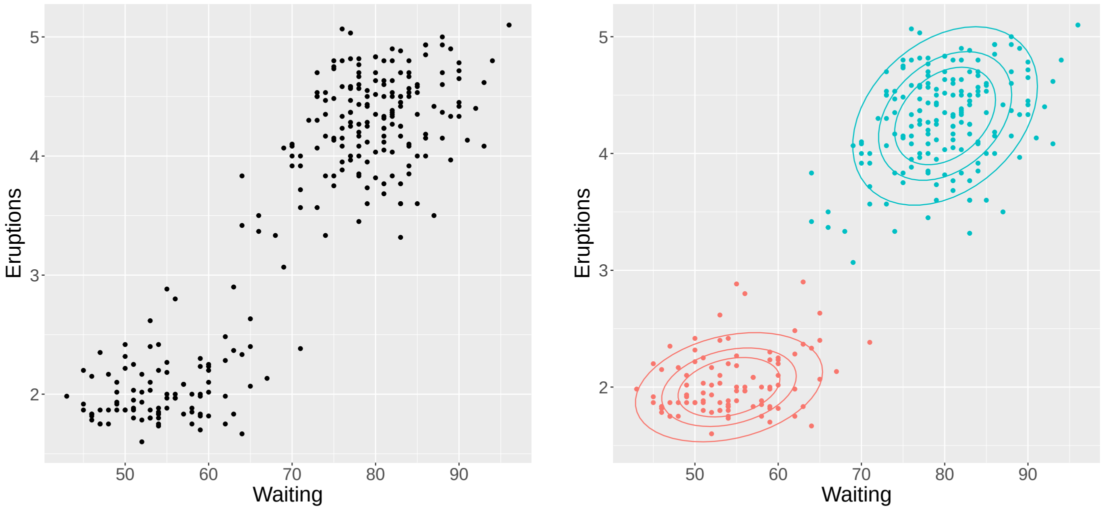
(d)

Figure 18: Mixturegrams of the DLBCL data based on (a) PCA with the K-profile summaries, (b) PCA with all observations plotted, (c) KPCA with the K-profile summaries, and (d) KPCA with all observations plotted. There are clearly two distinct groupings of profiles in all of the mixturegrams, thus indicating that a 2-component mixture of trivariate skew t distributions is appropriate for these data.

4.3 Old Faithful Data

The Old Faithful geyser dataset consists of the waiting time between eruptions and the duration of the eruptions (both recorded in minutes) for the Old Faithful Geyser in Yellowstone National Park. This bivariate dataset consists of $n = 272$ records and are depicted in Figure 19(a). The data were first reported in Azzalini and Bowman (1990) and have been analyzed using various statistical approaches, such as multivariate density estimation (Scott, 2004), cluster analysis (Hennig, 2003), and mixture models (Benaglia et al., 2009).

We consider mixture models with bivariate normal component distributions and $k = 1, 2, 3, 4$ components. We used 20 random starts for the EM algorithm and for each $k > 1$, we retained the estimates that had the largest log-likelihood. Figure 20 gives the mixturegrams that we constructed for the Old Faithful geyser data. We again construct mixturegrams based on PCA with the K-profile summaries (Figure 20(a)) and with all observations plotted (Figure 20(d)) as well as based on KPCA

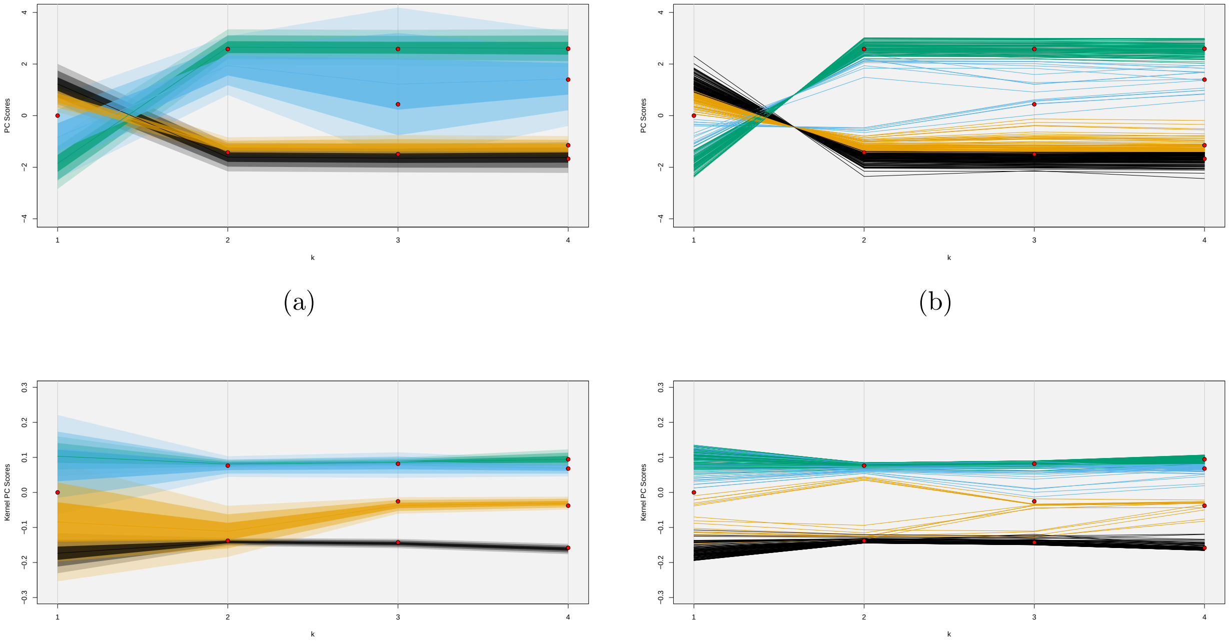


(a) (b)

Figure 19: (a) Scatterplot of all $n = 272$ recorded measurements of the waiting time between eruptions (y-axis) and duration of the eruptions (x-axis) for the Old Faithful geyser data. (b) The same scatterplot with the means and 50%, 70%, and 90% bivariate normal contours for the estimated 2-component mixture of bivariate normals.

with the K-profile summaries (Figure 20(c)) and with all observations plotted (Figure 20(d)). There are clearly two distinct groupings of profiles in the PCA-based mixturegrams, while the KPCA-based mixturegram based on the K-profile summaries seems to indicate 3-components. However, a plot of all the observations based on KPCA indicates more overlap of the profiles. Specifically, this indicate there is a stronger representation of two groupings of profiles, a feature that is masked by how we construct the K-profile summaries. This highlights that occasionally a mixturegram based on all of the observations may provide further insight beyond what is plotted on the mixturegram using the K-profile summaries. Regardless, a 2-component mixture of bivariate normals appears appropriate for these data.

We also report the model selection results and values of $\pi^{(k)}$ in Table 4. With the exception of AIC, the other selection criteria all select $k^* = 2$ as an appropriate number of components. Moreover, the values of $\pi^{(k)}$ indicate that a value of $k^* = 2$ is appropriate, especially since the value of $\pi^{(3)}$ is larger than $\pi^{(2)}$. Thus, we would select $k^* = 2$ as the number of components to use for this mixture problem. A scatterplot with the estimated component means and selected bivariate normal contours based on this 2-component fit is given in Figure 19(b).



(a) (b)

(c) (d)

Figure 20: Mixturegrams of the Old Faithful geyser data based on (a) PCA with the K-profile summaries, (b) PCA with all observations plotted, (c) KPCA with the K-profile summaries, and (d) KPCA with all observations plotted. There are clearly two distinct groupings of profiles in the PCA-based mixturegrams, while the KPCA-based mixturegram based on the K-profile summaries seems to indicate 3-components. However, a plot of all the observations in this case indicates more overlap of the profiles such that there is a stronger representation of two groupings of profiles. Thus, a 2-component mixture of bivariate normals appears appropriate for these data.

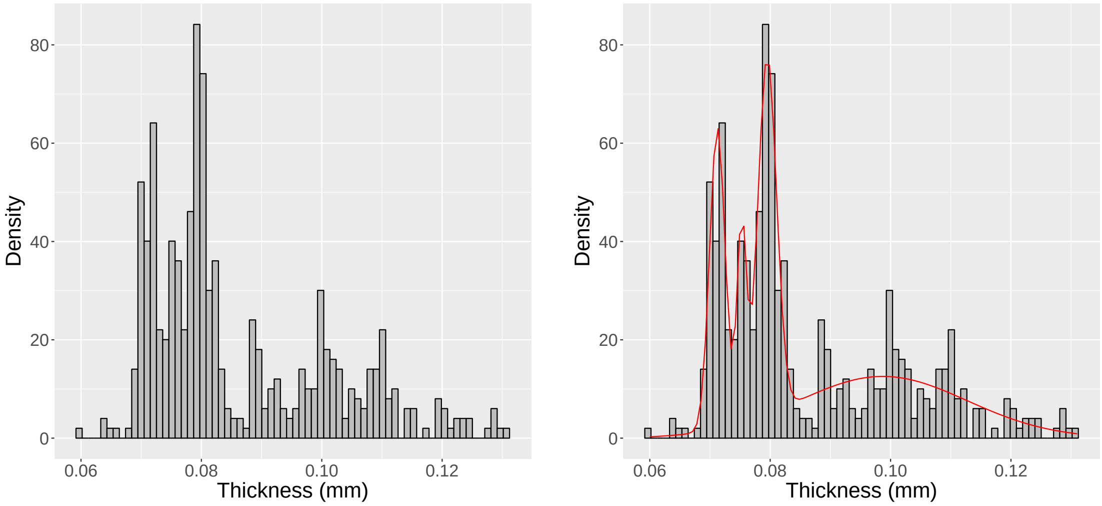
4.4 Hidalgo Stamp Data

The Hidalgo stamp dataset contains $n = 485$ records of the thickness of stamps having images of Miguel Hidalgo y Costilla (a famous leader of the Mexican War of Independence) that were issued by Mexico in 1872 and circulated until 1874. Due to poor quality control at that time, the thicknesses of the stamps varied considerably, which are depicted in Figure 21(a). This specific example of a philatelic mixture was presented and extensively analyzed in Izenman and Sommer (1988) using both a nonparametric approach and a mixture-of-normals approach in order to identify the different components.

When fitting a mixture model to these data, it is more challenging to assess the number of components since the component densities are clearly not well-separated. These data have been analyzed many times in the literature with different results depending on the strategy used for determining the

Criterion	Number of Components (k)			
	1	2	3	4
AIC	-1291.799	-1141.264	**-1136.300**	-1138.919
BIC	-1296.013	**-1164.444**	-1172.124	-1187.387
ICL	-1296.013	**-1163.793**	-1171.244	-1186.106
CAIC	-1297.013	**-1169.944**	-1180.624	-1198.887
$\pi^{(k)}$ (PCA)	1.000	**0.026**	0.027	0.016
$\pi^{(k)}$ (KPCA)	1.000	**0.008**	0.009	0.008

Table 4: Model selection and $\pi^{(k)}$ results for the Old Faithful data. There is majority agreement across the measures to select $k^* = 2$.

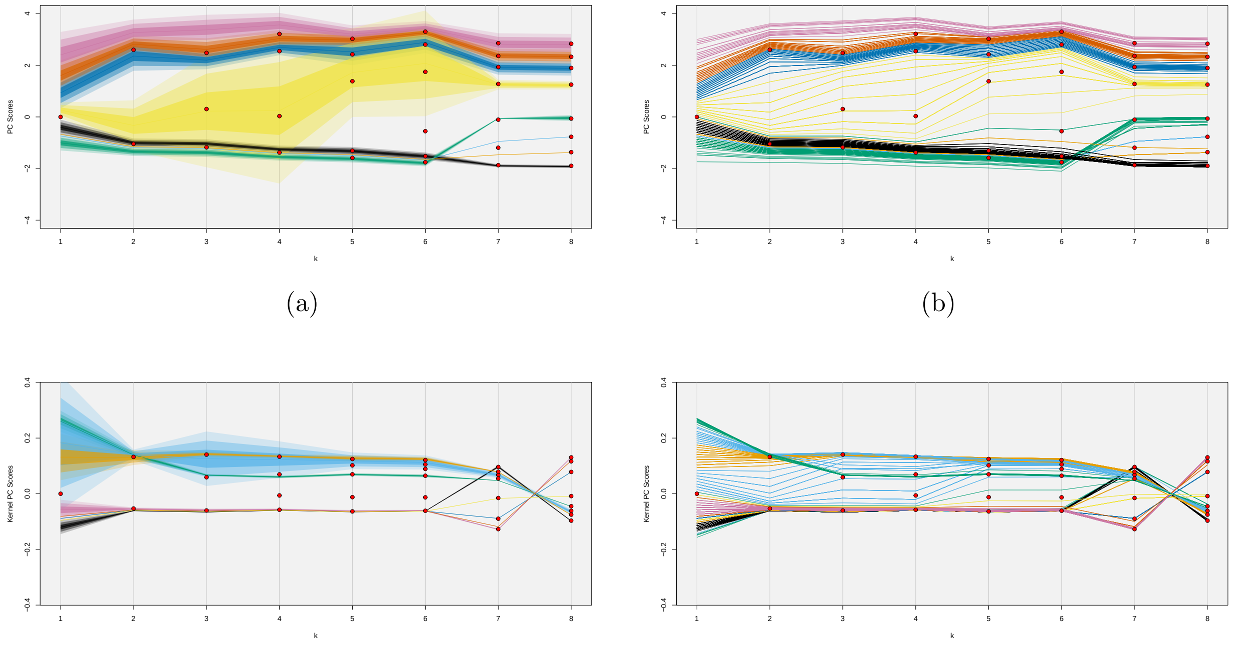


(a)

(b)

Figure 21: (a) Histogram of all $n = 485$ thickness measurements for the Hidalgo stamp data. (b) The same histogram with the estimated 4-component mixture of normals density curve plotted.

number of components. The number of components that have been selected include $k^* = 3$ (Izenman and Sommer, 1988), $k^* = 4$ (McLachlan and Peel, 2000), and $k^* = 7$ (Basford et al., 1997). We consider mixture models with univariate normal component distributions and $k = 1, \ldots, 8$ components. We, again, used 20 random starts for the EM algorithm and for each $k > 1$, we retained the estimates that had the largest log-likelihood. Figure 22 are the mixturegrams we constructed for the Hidalgo stamp data. We again construct mixturegrams based on PCA with the K-profile summaries (Figure 22(a)) and with all observations plotted (Figure 22(d)) as well as based on KPCA with the K-profile summaries (Figure 22(c)) and with all observations plotted (Figure 22(d)). The results of these mixturegrams are much more subjective. In the top-half of the mixturegrams based on PCA, the profiles appear to be similar, but then separate at $k = 4$. In particular, some of the profiles start to have an overall increasing trend while some of the profiles continue a decreasing trend. This indicates at least two components. In the bottom-half of the mixturegram, there appears to be a profile grouping that stays fairly constant as k increases. This indicates one more component. Finally, in the bottom-half there appears to be a small subset of four observations where the profiles noticeably increase from $k = 2$ to $k = 3$ and then decrease starting at $k = 5$. This indicates one final component. For the KPCA-based mixturegrams, however, there appears to be two or three groupings of profiles. Regardless, all of the mixturegrams for this heavily-overlapping setting have a greater degree of subjectivity in their interpretation. But combining the minimum number of components we discern from these mixturegrams with the results in Table 5 (which we discuss in the next paragraph), our assessment is that $k^* = 4$ components is an appropriate choice.



(a)

(b)

(c)

(d)

Figure 22: Mixturegrams of the Hidalgo stamp data based on (a) PCA with the K-profile summaries, (b) PCA with all observations plotted, (c) KPCA with the K-profile summaries, and (d) KPCA with all observations plotted. There appears to be at least three distinct groupings of profiles in the PCA-based mixturegrams, while the KPCA-based mixturegrams appear to suggest at least two distinct grouping. The mixturegrams based on the K-profile summaries are better for making such an assessment compared to what is observed with the mixturegrams where all observations are plotted. Based on the lack of separability, a conservative assessment is to suggest a mixture of normals model where $k^* \geq 3$.

Criterion	Number of Components (k)							
	1	2	3	4	5	6	7	8
AIC	1348.338	1437.625	1467.687	1464.687	1473.457	1470.457	1492.947	**1496.341**
BIC	1344.634	1428.365	1452.871	1444.316	1447.531	1438.975	**1455.909**	1453.747
ICL	1344.634	1428.966	1453.646	1445.484	1448.920	1440.545	**1457.729**	1455.531
CAIC	1343.634	1425.865	**1448.871**	1438.816	1440.531	1430.475	1445.909	1442.247
$\pi^{(k)}$ (PCA)	1.000	0.057	0.037	**0.015**	0.007	0.004	0.005	0.002
$\pi^{(k)}$ (KPCA)	1.000	0.032	0.018	**0.006**	0.002	0.001	0.001	0.001

Table 5: Model selection and $\pi^{(k)}$ results for the Hidalgo stamp data. There is no clear agreement across the different measures.

We also report the model selection results and values of $\pi^{(k)}$ in Table 5. As we can see, there is no unanimous agreement between the methods. AIC selects at least $k^* = 8$, BIC and ICL both select $k^* = 7$, CAIC selects $k^* = 3$, and the values of $\pi^{(k)}$ seem to indicate that a value of $k^* = 4$ is appropriate since the decrease in $\pi^{(5)}$ from $\pi^{(4)}$ is relatively small. Thus, there is a lot of variability in what one might decide based on the model selection criteria or $\pi^{(k)}$. Regardless, we proceed to select $k^* = 4$ components based on the assessment of the mixturegram. A histogram with the estimated 4-component mixture of normals density curve overlaid is given in Figure 21(b).

References

Azzalini, A. and Bowman, A. W. (1990). A look at some data on the Old Faithful geyser. *Applied Statistics*, 39(3):357–365.

Basford, K. E., McLachlan, G. J., and York, M. G. (1997). Modelling the distribution of stamp paper thickness via finite normal mixtures: The 1872 stamp issue of Mexico revisited. *Journal of Applied Statistics*, 24(2):169–180.

Benaglia, T., Chauveau, D., Hunter, D. R., and Young, D. S. (2009). mixtools: An R package for analyzing finite mixture models. *Journal of Statistical Software*, 32(6):1–29.

Hennig, C. (2003). Clusters, outliers, and regression: Fixed point clusters. *Journal of Multivariate Analysis*, 86(1):183–212.

Izenman, A. J. and Sommer, C. J. (1988). Philatelic mixtures and multimodal distributions. *Journal of the American Statistical Association*, 83(404):941–953.

Lee, S. X. and McLachlan, G. J. (2013). Model-based clustering and classification with non-normal mixture distributions. *Statistical Methods and Applications*, 22(4):427–454.

McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*. Wiley, New York.

Scott, D. W. (2004). Multivariate density estimation and visualization. In Gentle, J. E., Härdle, W., and Mori, Y., editors, *Handbook of Computational Statistics: Concepts and Methods*, pages 549–570. Springer.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer, New York.