Clustering Gene Effectors

Carey E. Priebe[1], Youngser Park[1], Marta Zlatic[2]
[1]JHU AMS & [2]HHMI Janelia

May 7, 2011

1 Introduction: Optogenetics

How is sensory information processed by neural circuits and used to select specific motor outputs? How are these functions of neural circuits encoded in the genome? These are the basic questions motivating our research. We address these questions by studying an animal that is capable of complex behavior and yet simple enough to allow systematic genetic manipulation of all parts of the neural circuitry.

Drosophila larvae sense and react to a wide range of stimuli and carry out many motor behaviors. These abilities are controlled by a relatively small number of neurons (about 10,000) that can be grouped into about 300 morphologically distinct neuron classes. Using the remarkable genetic toolkit generated by the Rubin lab at Janelia, we can selectively and reproducibly label and manipulate each of these neuron classes.

Our goal is to investigate the effect of activating and inactivating single neuron classes on larval sensory processing, decision making, and motor production. For this purpose we have developed a set of automated high-throughput behavioral assays.

—excerpt from http://www.hhmi.org/research/fellows/zlatic.html

2 Data

We analyze the data received March 2011.

There are a total of 1026 gene effector folders. For each gene effector, there are (suppose to be) six feature time series files: mean curves derived from the behavior of multiple individual organisms for "area", "bias", "curve", "dir", "midline", and "speed085". For each feature time series, there are (suppose to be) 340 time series points.

Six of the 1026 gene effector folders are empty,

- `A@ch2`,
- `B@ch2`,
- `GMR_79E02AE_01@ch2`,
- `k1@ch2`,
- `k2@ch2`,
- `ppkoriginal@ch2`,

and 1 of the gene effector folders,

- `GMR_27F03_AE_o1@ch2`,

has only 314 time series points. Therefore, we proceed with a total of 1019 gene effectors.

In summary, the data set we consider herein is given by X_{ift} with $i = 1, \cdots, n = 1019$ gene effectors, $f = 1, \cdots, F = 6$ feature time series, and $t = 1, \cdots, T = 340$ time points per feature time series.

NB: The gene effectors not containing `GMR` in their file names are "important controls." There are 11 such gene effectors, and we focus our analysis largely on these 11.

NB: Some gene effector feature time series are based on significantly fewer individual organisms than the majority. Thus those feature time series (mean curves) have a larger standard error. Currently, this is not accounted for in our analysis.

3 Methodology

Given the data as described above, our multiscale clustering methodology is summarized in the following steps:

(1) Using $\{X_{ift}\}$, obtain $n \times n$ dissimilarity matrix $\Delta = [\delta_{ij}]$ given by

$$\delta(X_i, X_j) = \left(\sum_{f=1}^{6} w_f \int_{\mathcal{T}} |s(X_{if.})(t) - s(X_{jf.})(t)|^p dt \right)^{1/p}.$$

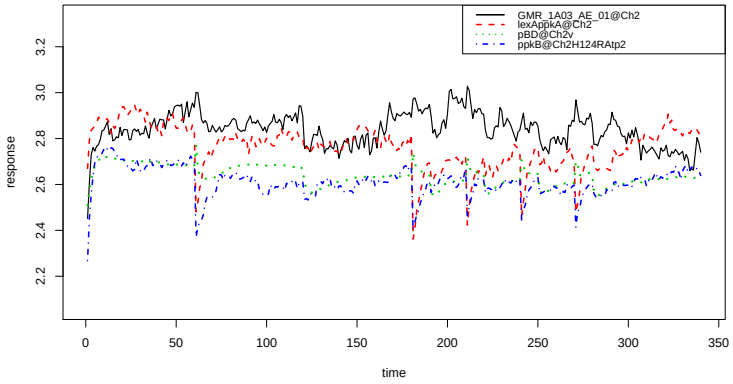
(2) Using Δ, obtain $n \times q$ Euclidean embedding Z.

(3) Using Z, jointly identify low-dimensional subspaces $D_{d_{max}}$ and clusterings $C_{K_{max}}(D)$ of the data in that subspace, for various choices of dimensionality $d_{max} \in \{1, \cdots, q\}$ and cluster complexity $K_{max} \in \{1, \cdots, n\}$.

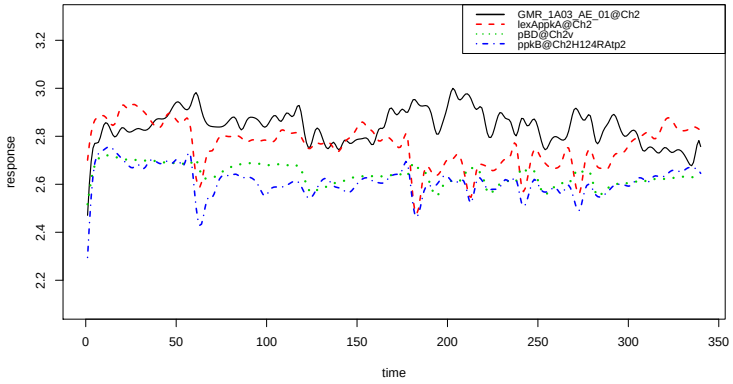
Step (1) involves generating a dissimilarity matrix from data. We proceed as in Priebe PAMI 2001 eq (15). Choosing s, $\mathcal{T}$, $\vec{w} = [w_1, \cdots, w_F]'$, and p involves understanding the scientific exploitation task and data collection protocol and subsequent exploratory data analysis. For the results presented herein, the nonparametric regression function s is a polynomial smoothing splines procedure as in Ramsey & Silverman 1997, the domain of integration is given by $\mathcal{T} = [0, 1]$, the weights $w_f = \left(\max_{ij} \int_{\mathcal{T}} |s(X_{if.})(t) - s(X_{jf.})(t)|^p dt\right)^{-1}$, and $p = 2$. Figure 1 presents illustrative smoothing results for three choices of s. We proceed with smoothing parameter chosen via cross-validation (Figure 1(b)).

Step (2) involves embedding a dissimilarity matrix into Euclidean space. We consider generalized multidimensional scaling (mds) as in Borg & Groenen 2005. For the results presented herein, we use classical multidimensional scaling into $\mathbb{R}^q$ for q large enough ($1 \ll q \leq n-1$) to capture (essentially) all of the signal. Figure 2 presents the scree plot obtained via $Z = mds(\Delta)$, suggesting that the first 40 embedding dimensions capture nearly all of the variance.

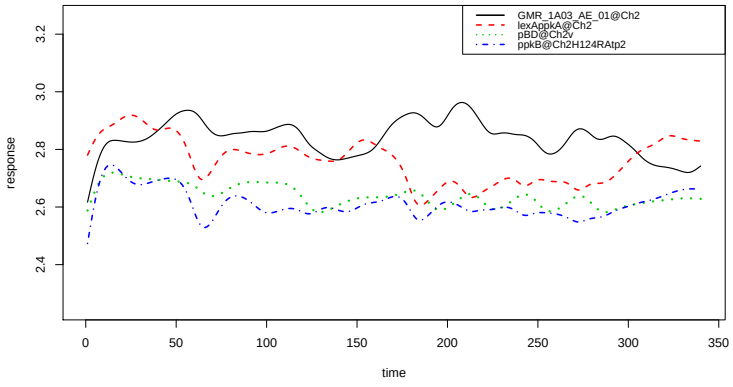
Step (3) involves joint identification of subspace and cluster structure. Given Z and d_{max}, K_{max}, the method of Raftery & Dean JASA 2006 building on the model-based clustering method of Fraley & Raftery JASA 2002 selects a canonical subspace $D \subset \{1, \cdots, d_{max}\}$ of cardinality $\widehat{d} = |D|$ and a clustering C of the data in that subspace with number of clusters $\widehat{K} \in \{1, \cdots, K_{max}\}$ so as to capture the complexity inherent in the data. Figure 3 and Table 1 present clustering results for the case $d_{max} = 2$, $K_{max} = 10$. NB: We will conclude from our analysis that $d_{max} = 2$ is too small, but this Figure 3 is illustrative and presenting higher-dimensional clustering results can be misleading. We can conclude from this Figure 3 and Table 1 that there is real structure in this data set, and that the clustering behavior of the 11 "important controls" makes scientific sense. Performing this model selection for various choices of d_{max} yields a multiscale clustering that can be analyzed for performance and consistency across scales and thereby provide confidence in claims of appropriate cluster complexity and cluster membership analysis. This multiscale clustering is pursued in the Section 4.



(a) unsmoothed



(b) cross-validated smoothing



(c) smoothing with $\lambda = 0.001$

Figure 1: Polynomial smoothing spline results for mean curves for feature times series "area" for four representative gene effectors (`GMR_1A03_AE_01@Ch2`, `lexAppkA@Ch2`, `pBD@Ch2v`, `ppkB@Ch2H124RAtp2`) with various choices of smoothing parameter λ.

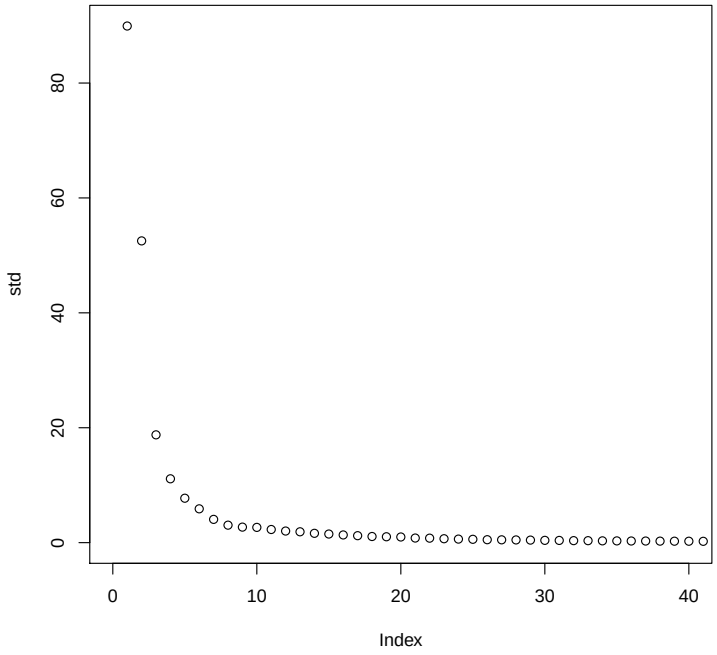
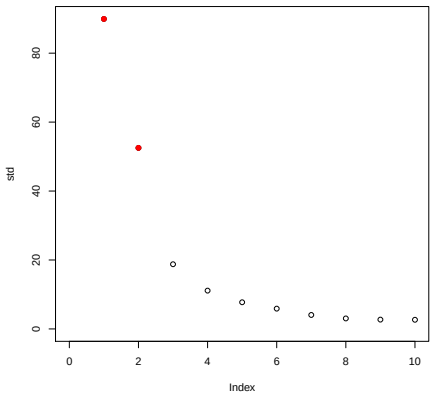


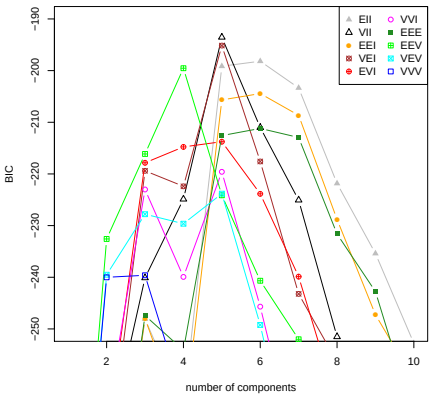
Figure 2: First 40 embedding dimensions of scree plot obtained via $Z = mds(\Delta)$. (The maximum embedding dimension allowing numerical stability is $q = 843$ for this data set.)

pBDs	
pBD@Ch2	4 (0.74)
pBD@Ch2v	4 (0.81)
non-pBDs	
lexAppkA@Ch2	1 (1.00)
lexAppkB@Ch2	1 (0.99)
ppkA@Ch2v	4 (0.95)
ppkB@Ch2BistVK5	1 (0.92)
ppkB@Ch2H124RAtp2	1,4 (0.50)
ppkB@Ch2H124RVK5	1 (0.99)
ppkB@Ch2tdTomAtp2	1 (0.89)
ppkB@Ch2tdTomVK5	3 (0.93)
ppkB@Ch2v	1 (1.00)

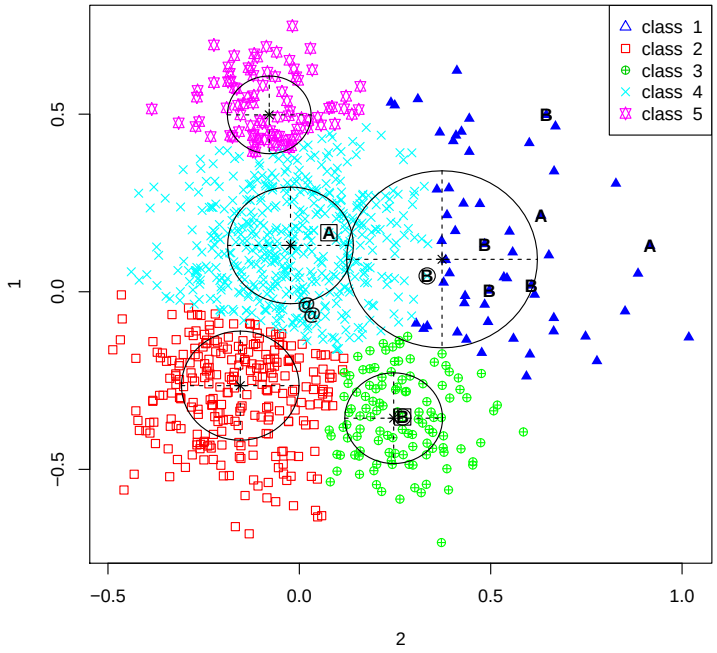
Table 1: Clustering results for the 11 controls with $d_{max} = 2$ and $K_{max} = 10$ yielding $\widehat{d} = 2$ and $\widehat{K} = 5$ clusters (see Figure 3). The integer represents cluster membership, and the probability (in parentheses) represents the posterior probability of class membership in that cluster. 10 of these eleven posteriors are convincingly high; the one ambiguously-clustered observation is highlighted in red. The clustering behavior of these 11 "important controls" makes scientific sense.



(a) subspace selection



(b) cluster complexity selection



(c) clustering

Figure 3: Clustering results with $d_{max} = 2$ and $K_{max} = 10$. Panel (a) presents the zoomed scree plot; both candidate canonical embedding dimensions 1 & 2 are selected ($\widehat{d} = 2$). Panel (b) presents the model selection plot for model-based clustering in these two dimensions ($\widehat{K} = 5$ clusters are selected with spherical unequal volume (VII) cluster structure and cluster cardinalities 56,266,116,496,85). Panel (c) presents the scatter plot for the clustering. The color symbols represent class membership and the letter symbols represent 11 control gene effectors by the fourth letter in their names (see Table 1). The two `pBDs` fall into cluster 4 (cyan) along with one `ppkA` (denoted with a square), one `ppkB` (denoted with a square) falls into cluster 3 (green), and six of the remaining seven non-`pBDs` fall into cluster 1 (blue). One `ppkB` (denoted with a circle) is displayed in cyan but its cluster membership is ambiguous between cluster 4 (cyan) and cluster 1 (blue). Clusters 2 (red) and 5 (pink) are not represented among the controls.

4 Multiscale Clustering Results

Multiscale clustering results for a suite of seven choices of the pair (d_{max}, K_{max}) given by

$$(d_{max}, K_{max}) \in \{(2,10), (3,10), (4,10), (10,20), (20,20), (30,20), (40,20)\}$$

are presented in Table 2 and Figures 3-9. The results demonstrate that the subspace identification model selection operation continues to choose more dimensions until $d_{max} = 40$, at which point it selects a canonical subspace consisting of $\widehat{d} = |D| = 18$ dimensions and selection of new dimensions from amongst the final candidates becomes sufficiently rare; from this we conclude that $d_{max} = 40$ captures the signal subspace adequately. Furthermore, the cluster structure model selection operation settles in at $\widehat{K}$ in the teens and the analysis of the cluster behavior for the 11 controls (Table 2) shows that for large d_{max} the two `pBDs` continue to cluster together while none of the non-`pBDs` cluster with the two `pBDs`, as desired; from this we conclude that $\widehat{K} = 13$ captures the cluster complexity adequately.

d_{max}	2	3	4	10	20	30	40
K_{max}	10	10	10	20	20	20	20
$\widehat{d}$	2	3	4	9	15	18	18
$\widehat{K}$	5	5	4	4	13	18	13
pBDs							
`pBD@Ch2`	4 (0.74)	1 (0.81)	3 (0.85)	3 (0.86)	8 (0.98)	10 (0.65)	6 (0.77)
`pBD@Ch2v`	4 (0.81)	1 (0.82)	3 (0.88)	3 (0.88)	8 (0.98)	10 (0.80)	6 (0.70)
non-pBDs							
`lexAppkA@Ch2`	1 (1.00)	5 (1.00)	1 (1.00)	4 (1.00)	1 (0.97)	11 (0.98)	13 (1.00)
`lexAppkB@Ch2`	1 (0.99)	5 (0.98)	1 (1.00)	4 (1.00)	1 (1.00)	18 (0.99)	13 (0.99)
`ppkA@Ch2v`	4 (0.95)	1 (0.88)	3 (0.97)	3 (1.00)	8 (0.97)	6 (0.98)	1 (0.94)
`ppkB@Ch2BistVK5`	1 (0.92)	5 (0.53)	1 (0.98)	4 (1.00)	1 (1.00)	15 (0.91)	4 (0.81)
`ppkB@Ch2H124RAtp2`	1,4 (0.50)	1 (0.97)	1 (0.97)	4 (1.00)	1 (1.00)	18 (0.81)	2 (0.76)
`ppkB@Ch2H124RVK5`	1 (0.99)	5 (1.00)	1 (1.00)	4 (1.00)	1 (1.00)	18 (1.00)	13 (0.91)
`ppkB@Ch2tdTomAtp2`	1 (0.89)	1 (0.94)	1 (0.99)	4 (1.00)	1 (1.00)	1 (1.00)	9 (0.95)
`ppkB@Ch2tdTomVK5`	3 (0.93)	3 (0.99)	4 (0.97)	4 (1.00)	6 (1.00)	1 (0.96)	9 (1.00)
`ppkB@Ch2v`	1 (1.00)	5 (1.00)	1 (1.00)	4 (1.00)	11 (1.00)	11 (1.00)	13 (1.00)

Table 2: Multiscale clustering results for the 11 controls. The integers represent cluster membership, and the probabilities (in parentheses) represent the posterior probability of class membership in that cluster. Posterior probabilities less than 0.6 are highlighted in red. (The leftmost column is the 2-dimensional example presented in Section 3.)

The final clustering depicted in Figure 9 ($d_{max} = 40$ and $K_{max} = 20$ yields $\widehat{d} = 18$ and $\widehat{K} = 13$) produces clusters with cardinalities $c_1 = 138, c_2 = 42, c_3 = 56, c_4 = 150, c_5 = 53, c_6 = 140, c_7 = 68, c_8 = 66, c_9 = 131, c_{10} = 6, c_{11} = 68, c_{12} = 83, c_{13} = 18$. Only cluster #10, with $c_{10} = 6$, is worrisomely small. No single cluster accounts for more than 15% of the 1019 total observations. The two `pBD` controls fall into cluster #6 with a total of $c_6 = 140$ observations. Cluster #13 contains a total of $c_{13} = 18$ observations, including 4 of the 9 non-`pBD` controls. Cluster #9 contains a total of $c_9 = 131$ observations, including both of the `ppkB@Ch2tdTom*` controls. Figures 10 and 11 show that this clustering solution does in fact capture the general structure inherent in our data set. Figure 12 investigates cluster #13 in detail.

The multiscale clustering results taken holistically provide important information concerning co-clustering behavior – persistence of co-clustering across scales is strong evidence of commonality. Multiscale clustering results for all n=1019 gene effectors (cluster membership and posterior probabilities for all seven clusterings) are provided as Supplementary Material file "`subj-class-posterior.txt`". Tables 3 and 4 present illustrative results therefrom.

d_{max}	2	3	4	10	20	30	40
K_{max}	10	10	10	20	20	20	20
$\widehat{d}$	2	3	4	9	15	18	18
$\widehat{K}$	5	5	4	4	13	18	13
pBDs							
pBD@Ch2	4 (0.74)	1 (0.81)	3 (0.85)	3 (0.86)	8 (0.98)	10 (0.65)	6 (0.77)
pBD@Ch2v	4 (0.81)	1 (0.82)	3 (0.88)	3 (0.88)	8 (0.98)	10 (0.80)	6 (0.70)
clustered with the two pBDs							
GMR_36H01_AE_01@Ch2	4 (0.87)	1 (0.80)	3 (0.90)	3 (0.69)	8 (0.89)	10 (0.84)	6 (0.79)
GMR_69F06_AE_01@Ch2	4 (0.95)	1 (0.61)	3 (0.90)	3 (0.90)	8 (0.97)	10 (0.66)	6 (0.55)
GMR_88H03_AE_01@Ch2	4 (0.79)	1 (0.78)	3 (0.86)	3 (0.84)	8 (0.98)	10 (0.84)	6 (0.93)

Table 3: Multiscale clustering results for those gene effectors which are clustered with the two pBDs across all clustering scales. The integers represent cluster membership, and the probabilities (in parentheses) represent the posterior probability of class membership in that cluster.

d_{max}	2	3	4	10	20	30	40
K_{max}	10	10	10	20	20	20	20
$\widehat{d}$	2	3	4	9	15	18	18
$\widehat{K}$	5	5	4	4	13	18	13
one non-pBD							
ppkA@Ch2v	4 (0.95)	1 (0.88)	3 (0.97)	3 (1.00)	8 (0.97)	6 (0.98)	1 (0.94)
clustered with this one non-pBD							
GMR_10F10_AE_01@Ch2	4 (0.92)	1 (0.95)	3 (0.96)	3 (1.00)	8 (0.96)	6 (0.99)	1 (0.79)
GMR_22C07_AE_01@Ch2	4 (0.90)	1 (0.97)	3 (0.97)	3 (1.00)	8 (0.95)	6 (0.89)	1 (0.79)
GMR_25A08_AE_01@Ch2	4 (0.91)	1 (0.58)	3 (0.98)	3 (0.98)	8 (0.98)	6 (0.97)	1 (0.79)
GMR_43D09_AE_01@Ch2r	4 (0.88)	1 (0.98)	3 (0.95)	3 (0.89)	8 (0.98)	6 (0.98)	1 (0.79)
GMR_44D10_AE_01@Ch2r	4 (0.93)	1 (0.91)	3 (0.93)	3 (0.99)	8 (0.99)	6 (0.96)	1 (0.79)
GMR_50A11_AE_01@Ch2r	4 (0.88)	1 (0.85)	3 (0.89)	3 (1.00)	8 (0.89)	6 (0.99)	1 (0.79)
GMR_65H03_AE_01@Ch2	4 (0.95)	1 (0.98)	3 (0.98)	3 (0.99)	8 (0.54)	6 (0.98)	1 (0.79)

Table 4: Multiscale clustering results for those gene effectors which are clustered with the non-pBD ppkA@Ch2v across all clustering scales. The integers represent cluster membership, and the probabilities (in parentheses) represent the posterior probability of class membership in that cluster.

5 References

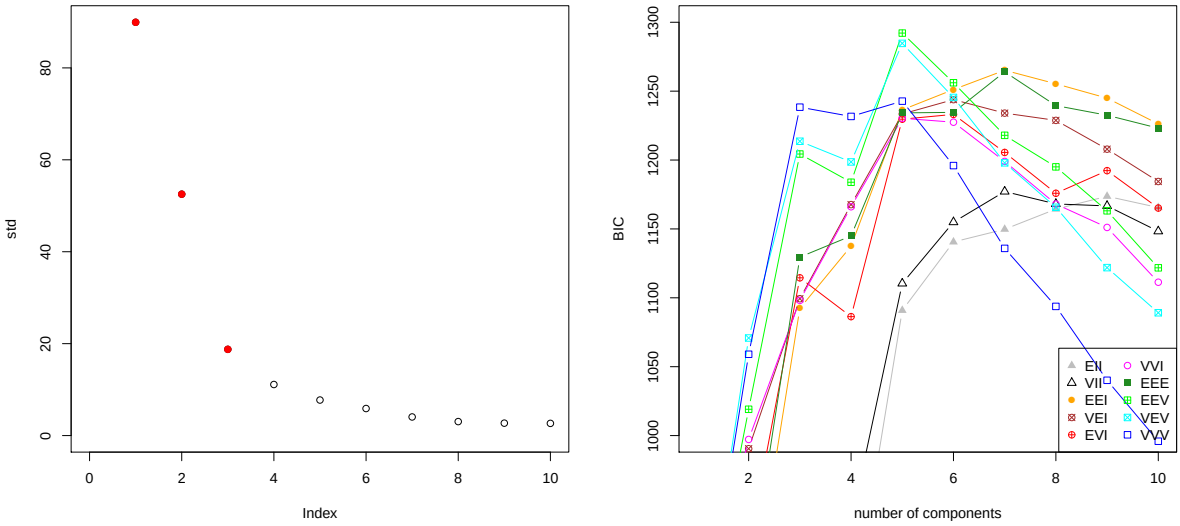
[1] I. Borg and P.J.F. Groenen, *Modern Multidimensional Scaling*, 2nd edition. New York, Springer, 2005.

[2] C. Fraley and A.E. Raftery, "Model-based Clustering, Discriminant Analysis, and Density Estimation," *Journal of the American Statistical Association* Vol. 97, pp. 611-631, 2002.

[3] A.E. Raftery and N. Dean, "Variable Selection for Model-Based Clustering," *Journal of the American Statistical Association*, Vol. 101, pp. 168-178, 2006.

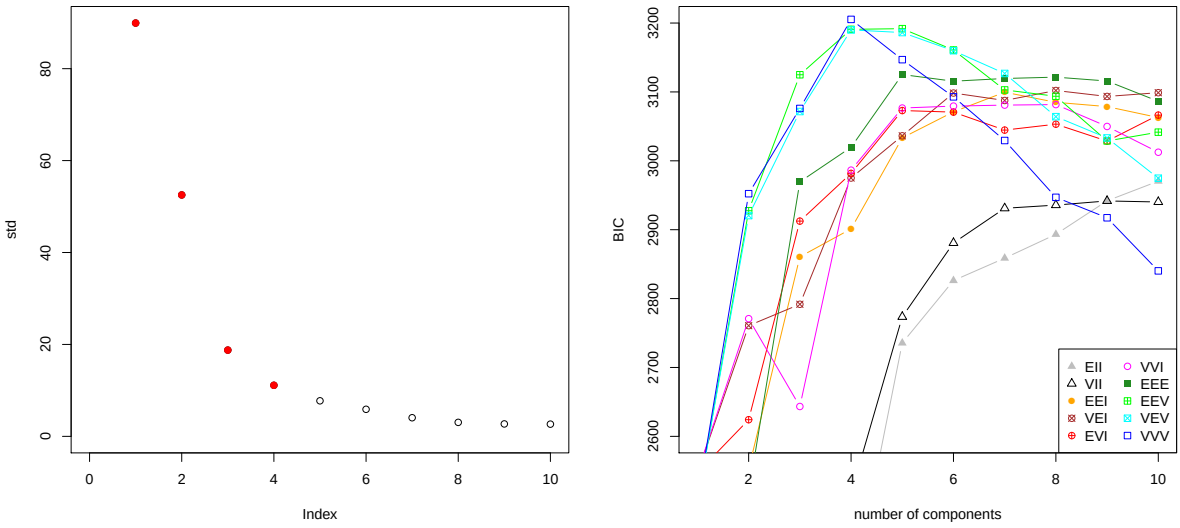
[4] C.E. Priebe, "Olfactory Classification via Interpoint Distance Analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 23, pp. 404-413, 2001.

[5] J.O. Ramsay and B.W. Silverman, *Functional Data Analysis*. New York, Springer, 1997.



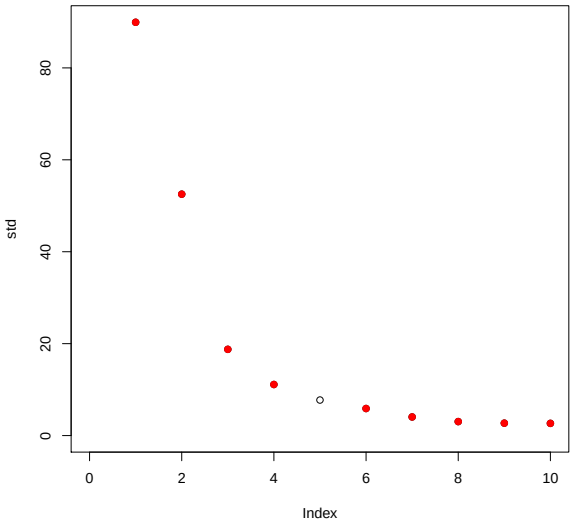
(a) $D = \{2, 3, 1\}$ (b) $\widehat{d} = 3$, $\widehat{K} = 5$ (EEV)

Figure 4: $d_{max} = 3$ and $K_{max} = 10$ yields $\widehat{d} = 3$ and $\widehat{K} = 5$ with ellipsoidal equal volume and equal shape (EEV) cluster structure and cluster cardinalities 364,228,186,214,27.

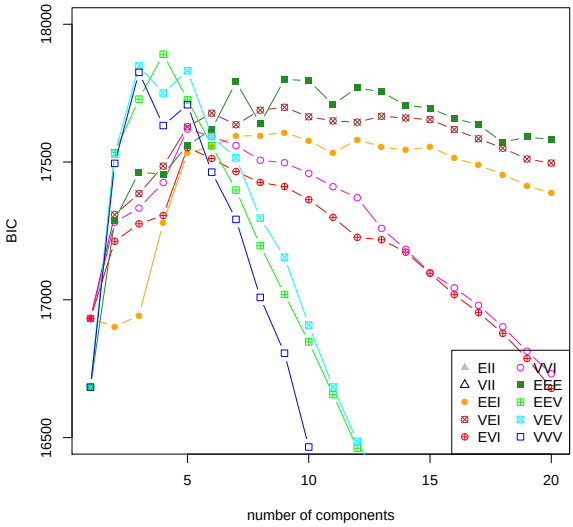


(a) $D = \{2, 3, 4, 1\}$ (b) $\widehat{d} = 4$, $\widehat{K} = 4$ (VVV)

Figure 5: $d_{max} = 4$ and $K_{max} = 10$ yields $\widehat{d} = 4$ and $\widehat{K} = 4$ with ellipsoidal varying volume varying shape varying orientation (VVV) cluster structure and cluster cardinalities 60,296,507,156.

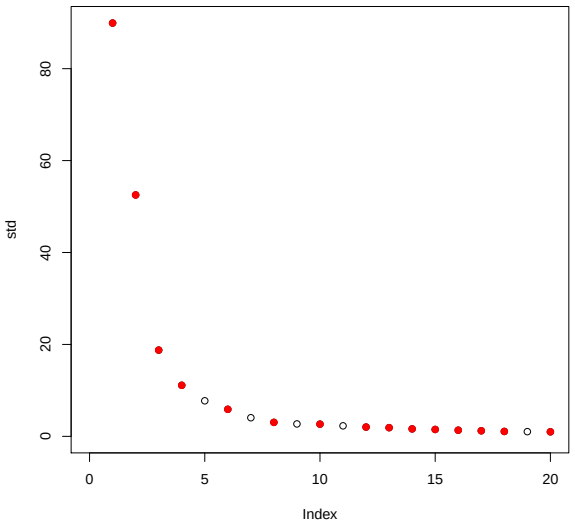


(a) $D = \{2, 3, 4, 1, 6, 10, 9, 7, 8\}$

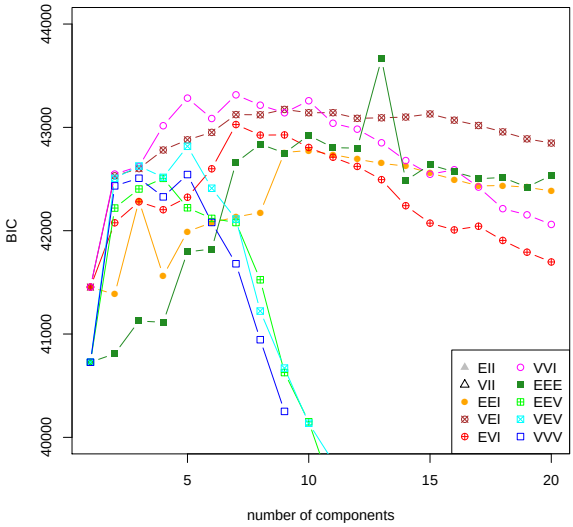


(b) $\widehat{d} = 9$, $\widehat{K} = 4$ (EEV)

Figure 6: $d_{max} = 10$ and $K_{max} = 20$ yields $\widehat{d} = 9$ and $\widehat{K} = 4$ with ellipsoidal equal volume equal shape (EEV) cluster structure and cluster cardinalities 296,187,460,77.

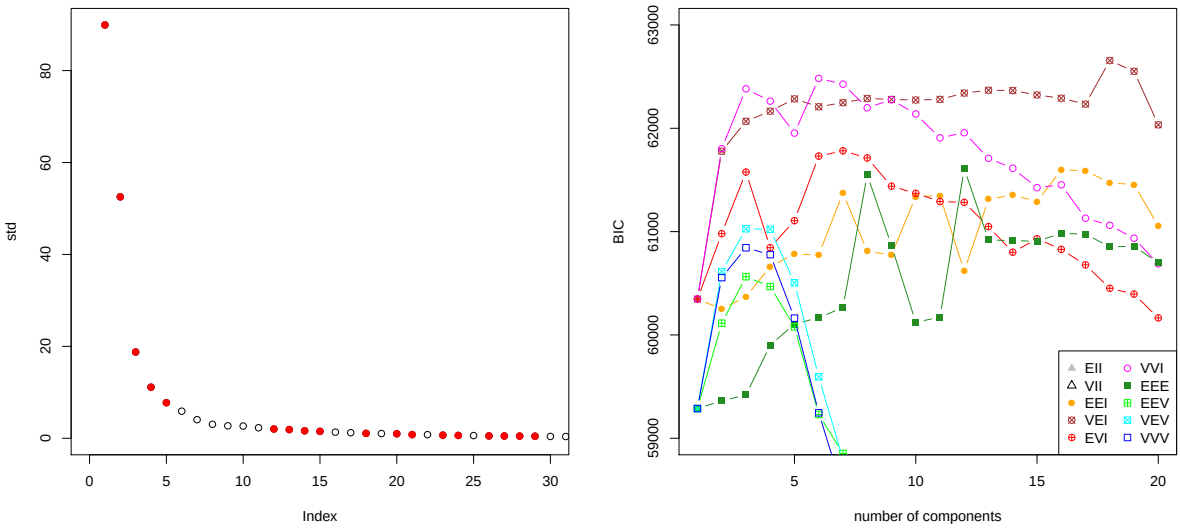


(a) $D = \{12, 14, 15, 13, 20, 2, 3, 18, 1, 17, 4, 6, 8, 10, 16\}$



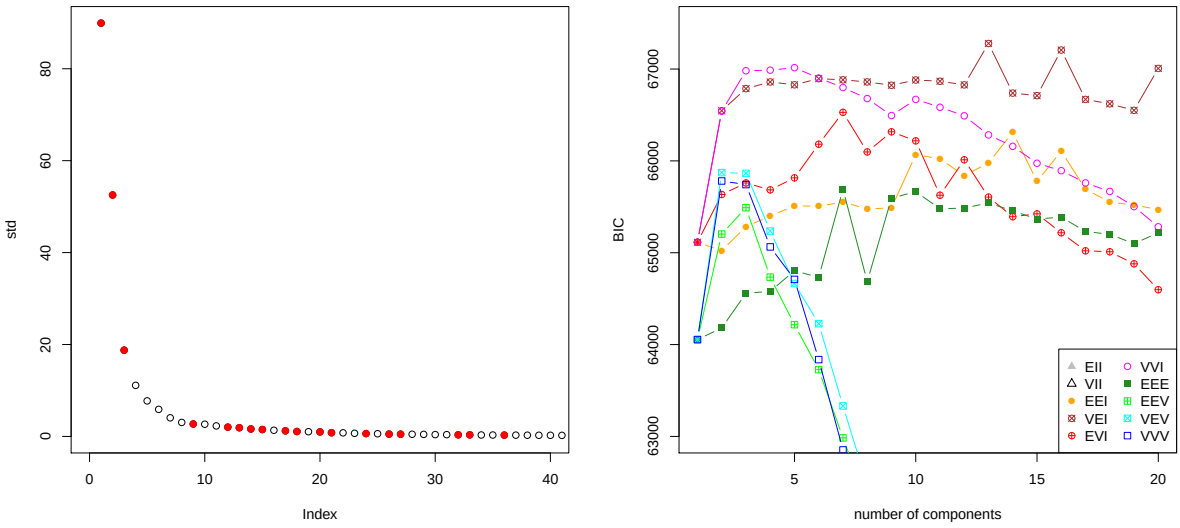
(b) $\widehat{d} = 15$, $\widehat{K} = 13$ (EEE)

Figure 7: $d_{max} = 20$ and $K_{max} = 20$ yields $\widehat{d} = 15$ and $\widehat{K} = 13$ with ellipsoidal equal volume equal shape equal orientation (EEE) cluster structure and cluster cardinalities 22,48,24,93,105,178,4,392,5,26,10,46,66.



(a) $D = \{12, 14, 15, 13, 20, 2, 28, 3, 21, 26, 18, 1, 5, 24, 4, 29, 27, 23\}$ (b) $\widehat{d} = 18$, $\widehat{K} = 18$ (VEI)

Figure 8: $d_{max} = 30$ and $K_{max} = 20$ yields $\widehat{d} = 18$ and $\widehat{K} = 18$ with diagonal varying volume equal shape (VEI) cluster structure and cluster cardinalities 129,6,23,54,91,86,48,93,46,71,11,21,108,26,49,64,75,18.



(a) $D = \{12, 14, 15, 13, 20, 2, 3, 21, 26, 18, 1, 36, 24, 33, 17, 32, 27, 9\}$ (b) $\widehat{d} = 18$, $\widehat{K} = 13$ (VEI)

Figure 9: $d_{max} = 40$ and $K_{max} = 20$ yields $\widehat{d} = 18$ and $\widehat{K} = 13$ with diagonal varying volume equal shape (VEI) cluster structure and cluster cardinalities 138,42,56,150,53,140,68,66,131,6,68,83,18. See also Figures 10 and 11.

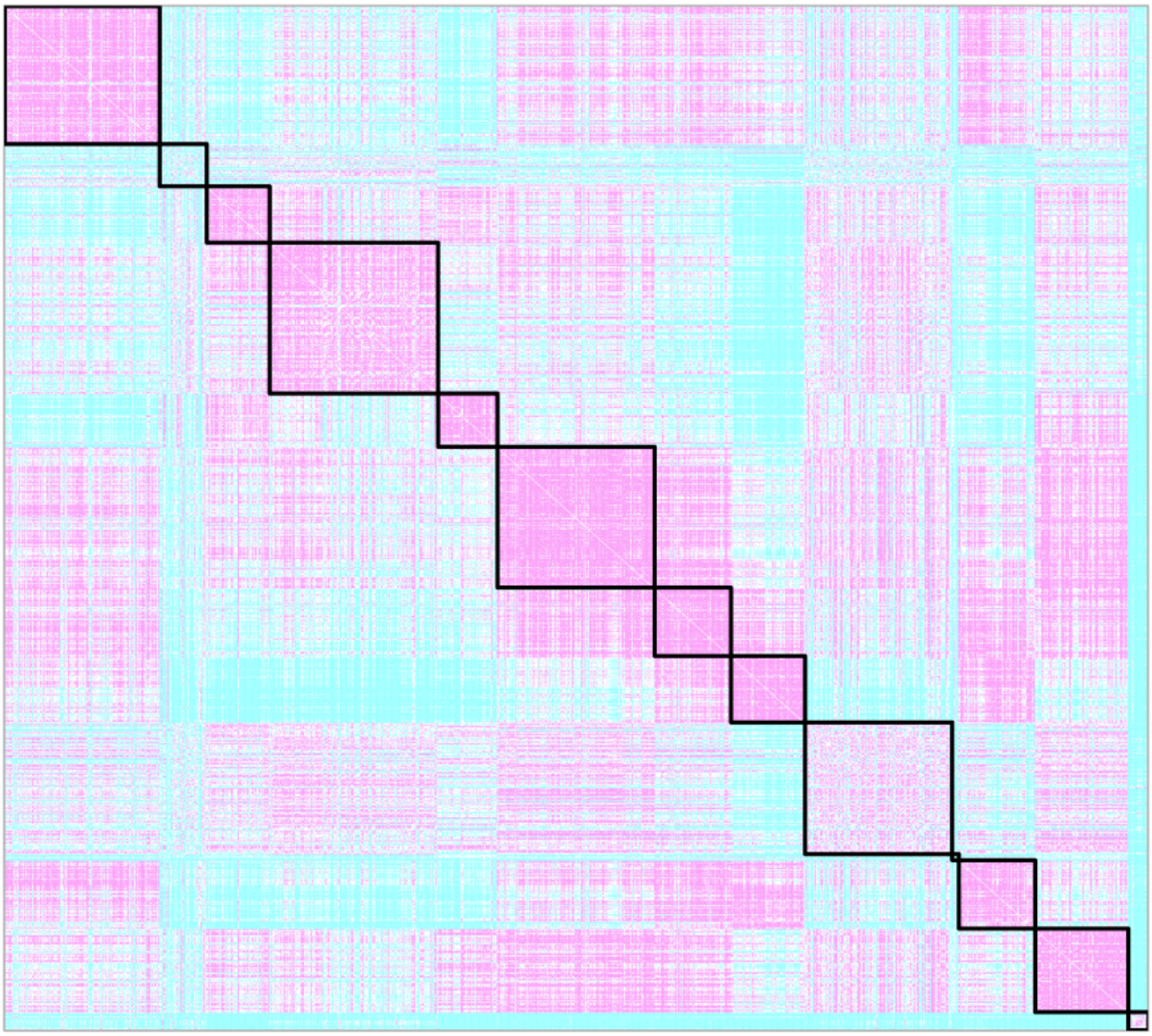


Figure 10: Depicted is the colormap of the 1019 x 1019 interpoint distances between the 1019 gene effectors in our final ($\widehat{d} = 18$)-dimensional space wherein we obtain $\widehat{K} = 13$ clusters (see also Figures 9 and 11). Pink represents small distances and cyan represents larger distances. This plot demonstrates that this clustering solution does in fact capture the general structure inherent in our data set: the structure captured by these 13 clusters is real, and most of the structure is captured by these 13 clusters.

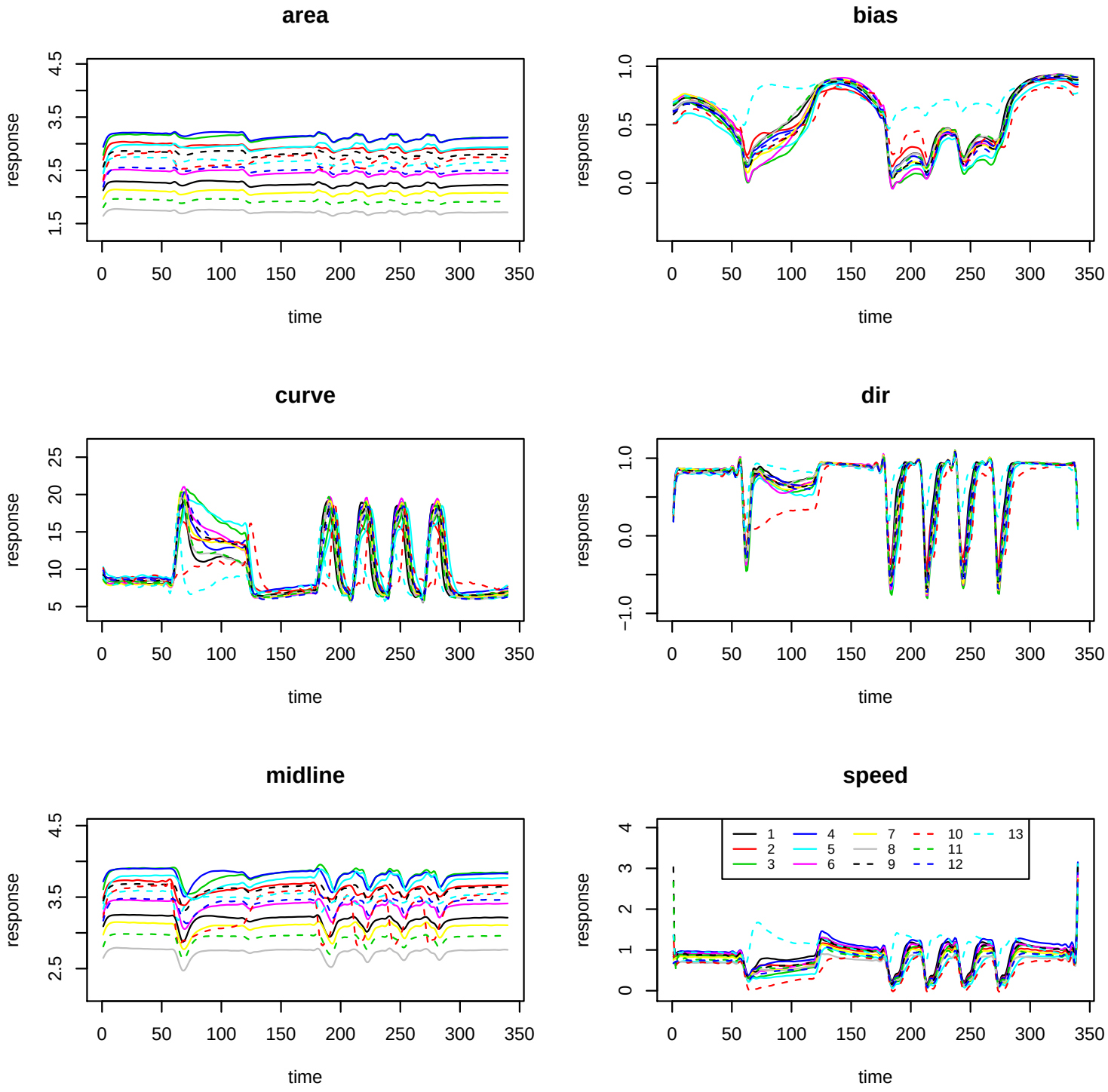


Figure 11: Depicted are the mean curves for the $\widehat{K} = 13$ clusters (see also Figures 9 and 10) for each of the six time series features.

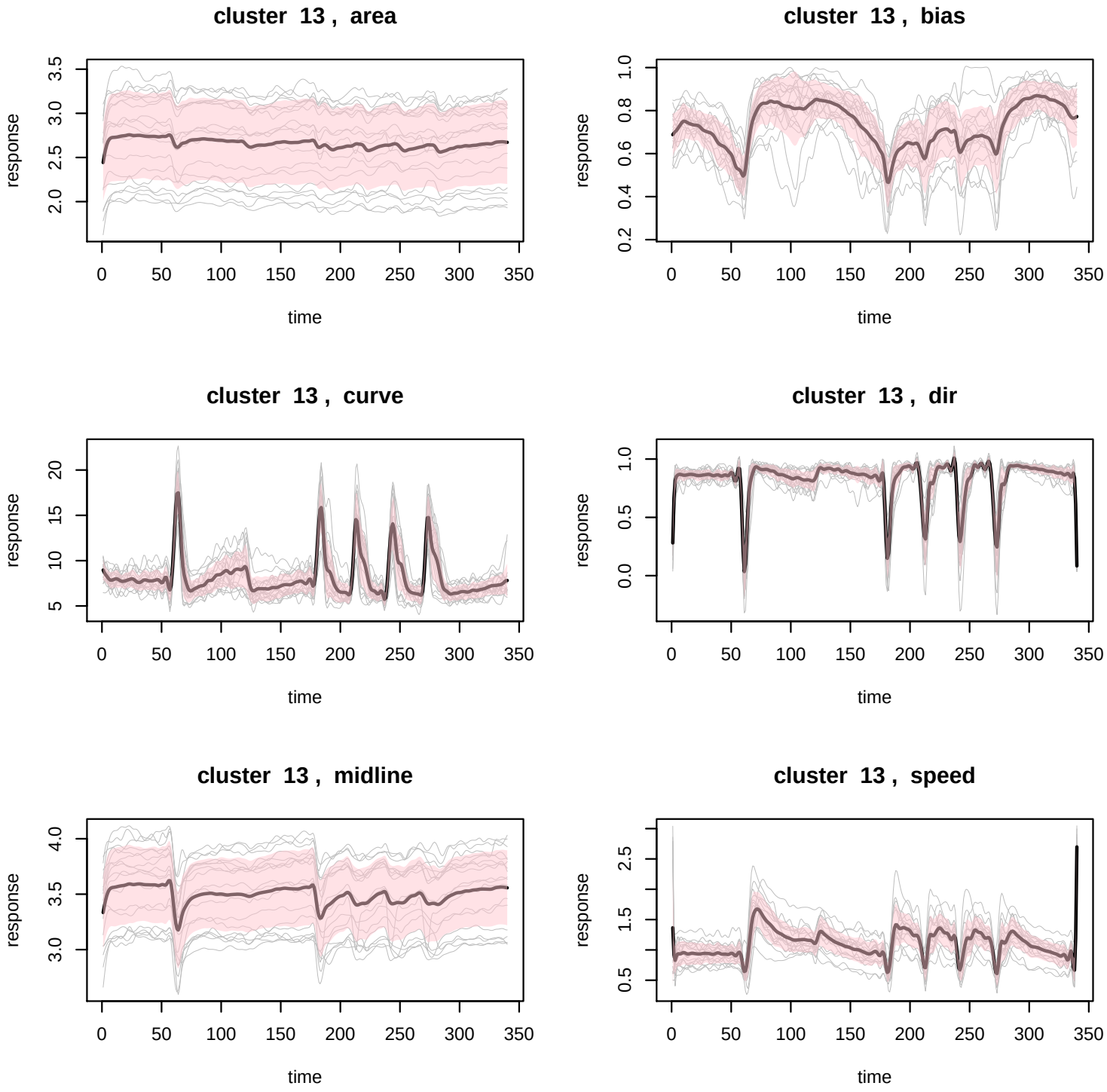


Figure 12: Depicted is a cluster analysis plot for cluster #13 from our $\widehat{K} = 13$ cluster solution depicted in Figures 9, 10, and 11. The six time series features are plotted for all $c_{13} = 18$ observations in that cluster; the mean time series for that cluster is represented in solid black, and the shaded region is $\pm$ one standard deviation. Notice, for instance, that cluster #13 seems to have two clear outliers in the "bias" time series at around time 50. An interactive version of this plot is available at http://www.cis.jhu.edu/∼parky/Data/HHMI/Excel/Z2-smoothed-d40-K20-c13.xlsx, in which the individual outlier observations can be identified. Analogous figures for all 13 clusters are available at http://www.cis.jhu.edu/∼parky/Data/HHMI/TSplots/, and interactive versions at http://www.cis.jhu.edu/∼parky/Data/HHMI/Excel/.