Supplement to "Estimating the number of clusters using cross-validation"

Wei Fu and Patrick O. Perry
Stern School of Business, New York University

June 29, 2019

1 Clustering scree plot examples

The top row of Figure 1 displays an example where the elbow in W_k corresponds to the true number $k = 4$ of mixture components in the data-generating mechanism. The elbow approach is simple and often performs well, but it requires subjective judgment as to where the elbow is located, and, as the bottom row of Figure 1 demonstrates, the approach can easily fail.

2 Self-consistency

An important property of any estimation procedure is that in the absence of noise, the procedure correctly estimates the truth. This property is called "self-consistency" (**?**). We will now show that Gabriel cross-validation is self-consistent. That is, in the absence of noise, the Gabriel cross-validation procedure finds the optimal number of clusters.

Self-consistency by itself does not mean that a procedure performs well in practice. Self-consistency deals only with the no-noise situation. Many other methods for selecting K are likely self-consistent, but not all perform the same in practice. Self-consistency only suggests (but does not guarantee) that a procedure might perform reasonably when the noise level is low.

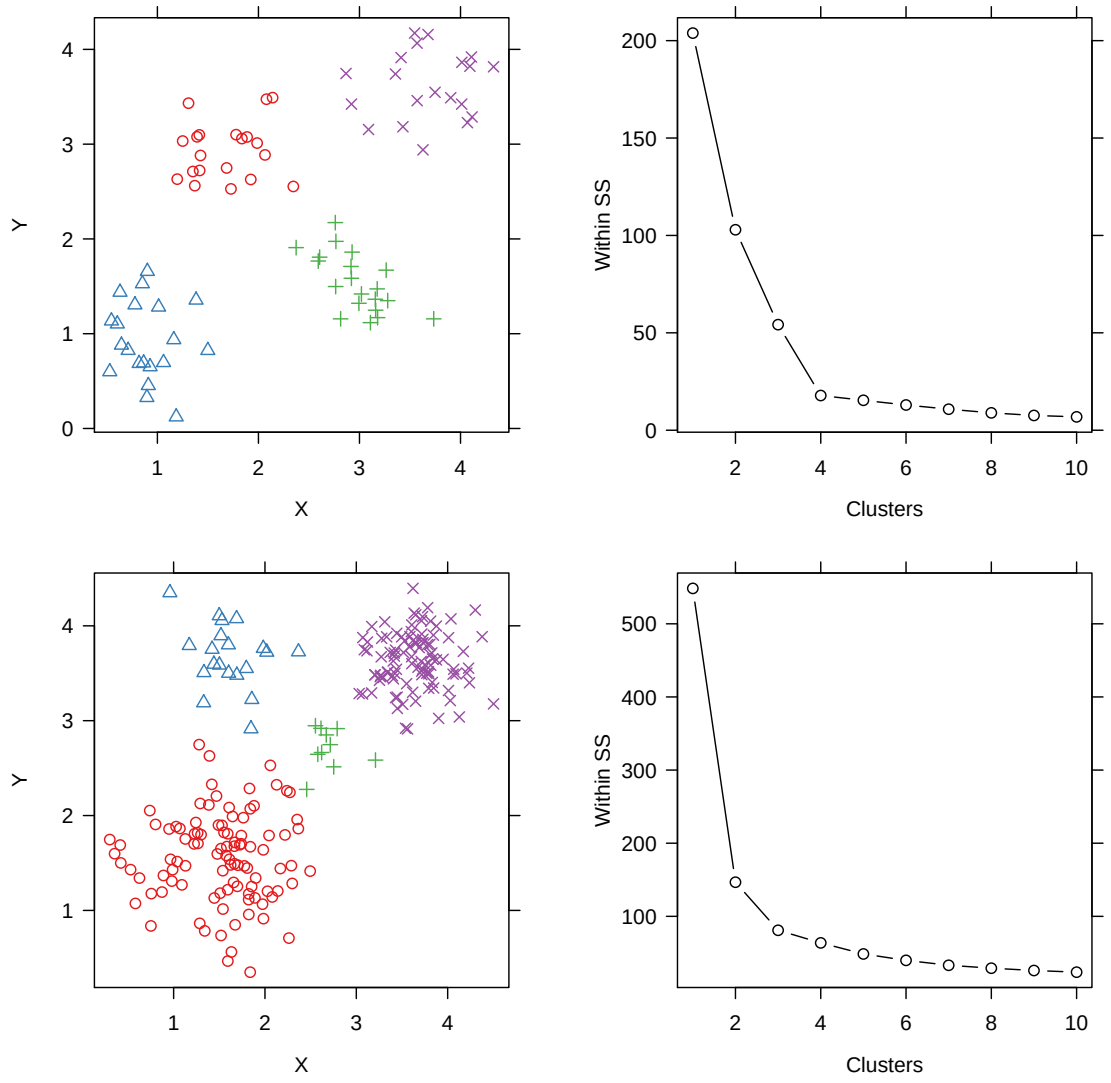

Figure 1: Scree plots for two data sets. Left panels show the sets of two-dimensional data points, generated from four clusters, with plotting symbol indicating the generated cluster. Right panels show the corresponding values of the within-cluster sum of squares W_k plotted against the number of clusters, k. The scree plot identifies the correct number of clusters in the top row, but fails in the bottom row.

It will suffice to prove self-consistency for a single fold of the cross-validation procedure. As in section **??** we assume that the P variables of the data set have been partitioned into p predictor variables represented in vector X and q response variables represented in vector Y. The N observations have been divided into two sets: n train observations and m test observations. We state the assumptions for the self-consistency result in terms of a specific split; for the result to hold in general, with high probability, these assumptions would have to hold with high probability for a random split. The following theorem gives conditions for Gabriel cross-validation to recover the true number of clusters in the absence of noise.

Proposition 1. *Let $\{(X_i, Y_i)\}_{i=1}^{n+m}$ be the data from a single fold of Gabriel cross-validation. For any k, let $\mathrm{CV}(k)$ be the cross-validation error for this fold, computed as described in Section* **??**. *We will assume that there are K "true cluster centers" $\mu(1), \ldots, \mu(K)$, with the gth cluster center partitioned as $\mu(g) = \big(\mu^X(g), \mu^Y(g)\big)$ for $g = 1, \ldots, K$. Suppose that*

(i) Each observation i has a true cluster $G_i \in \{1, \ldots, K\}$. There is no noise, so that $X_i = \mu^X(G_i)$ and $Y_i = \mu^Y(G_i)$ for $i = 1, \ldots, n+m$.

(ii) The vectors $\mu^X(1), \ldots, \mu^X(K)$ are all distinct.

(iii) The vectors $\mu^Y(1), \ldots, \mu^Y(K)$ are all distinct.

(iv) The training set contains at least one member of each cluster: for all g in the range $1, \ldots, K$, there exists at least one i in the range $1, \ldots, n$ such that $G_i = g$.

(v) The test set contains at least one member of each cluster: for all g in the range $1, \ldots, K$, there exists at least one i in the range $n+1, \ldots, n+m$ such that $G_i = g$.

Then $\mathrm{CV}(k) > \mathrm{CV}(K)$ for $k < K$, and $\mathrm{CV}(k) = \mathrm{CV}(K)$ for $k > K$, so that Gabriel cross-validation correctly chooses $k = K$.

The proposition states that our method works well in the absence of noise, when each observation is equal to its cluster center. The assumptions here are not checkable in practice, but the proposition suggest that Gabriel cross-validation might give a reasonable answer, at least when the noise level is low.

The essential assumption here is assumption (i), which states that there is no noise. If we are willing to assume, say, that the cluster centers $\mu(g) = \big(\mu^X(g), \mu^Y(g)\big)$ for $g = 1, \ldots, K$ were randomly drawn from a distribution with a density over $\mathbb{R}^{p+q}$, then assumptions (ii) and (iii) will hold with probability one for all splits of the data. Likewise, if the clusters are not too small (relative to n and m), then assumptions (iv) and (v) will likely hold for a random split of the data into test and train.

Proposition 1 follows from Lemmas 1 and 2, which we now state and prove.

Lemma 1. *Suppose that the assumptions of Proposition 1 are in force. If $k < K$, then* $\mathrm{CV}(k) > 0$.

Proof. By definition,

$$\mathrm{CV}(k) = \sum_{i=n+1}^{n+m} \|Y_i - \bar{\mu}^Y(\hat{G}_i^X)\|^2,$$

where $\bar{\mu}^Y(g)$ is the center of cluster g returned from applying k-means to $Y_1, \ldots, Y_n$. Assumptions (i) and (v), imply that as i ranges over the test set $n+1, \ldots, n+m$, the response Y_i ranges over all distinct values in $\{\mu^Y(1), \ldots, \mu^Y(K)\}$. Assumption (iii) implies that there are exactly K such distinct values. However, there are only k distinct values of $\bar{\mu}^Y(g)$. Thus, at least one summand $\|Y_i - \bar{\mu}^Y(\hat{G}_i^X)\|^2$ is nonzero. Therefore, $\mathrm{CV}(k) > 0$. □

Lemma 2. *Suppose that the assumptions of Proposition 1 are in force. If $k \geq K$, then* $\mathrm{CV}(k) = 0$.

Proof. From assumptions (i), (iii), and (iv), we know the cluster centers gotten from applying k-means to $Y_1, \ldots, Y_n$ must include $\mu^Y(1), \ldots, \mu^Y(K)$. Without loss of generality, suppose that $\bar{\mu}^Y(g) = \mu^Y(g)$ for $g = 1, \ldots, K$. This implies that $\hat{G}_i^Y = G_i$ for $i = 1, \ldots, n$. Thus, employing assumption (i) again, we get that $\bar{\mu}^X(g) = \mu^X(g)$ for $g = 1, \ldots, K$.

Since assumption (ii) ensures that $\mu^X(1), \ldots, \mu^X(K)$ are all distinct, we must have that $\hat{G}_i^X = G_i$ for all $i = 1, \ldots, m+n$. In particular, this implies that $\bar{\mu}^Y(\hat{G}_i^X) = Y_i$ for $i = 1, \ldots, m+n$, so that $\mathrm{CV}(k) = 0$. □

3 Analysis of single cluster in more than two dimensions

Proposition 2. *Suppose that* $\{(X_i, Y_i)\}_{i=1}^{n+m}$ *is data from a single fold of Gabriel cross-validation, where each* (X, Y) *pair in* $\mathbb{R}^{p+q}$ *is an independent draw from a mean-zero multivariate normal distribution with covariance matrix* $\Sigma = \left(\begin{smallmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{smallmatrix}\right)$*, with* Σ_{YY} *has leading eigenvalue* λ_1 *and corresponding eigenvector* u_1*. In this case, the data are drawn from a single cluster; the true number of clusters is* 1*. If* $\frac{\sqrt{\lambda_1}}{2} > \frac{u_1^T \Sigma_{YX} \Sigma_{XY} u_1}{\sqrt{u_1^T \Sigma_{YX} \Sigma_{XX} \Sigma_{XY} u_1}}$*, then* $\mathrm{CV}(1) < \mathrm{CV}(2)$ *with probability tending to one as* m *and* n *increase.*

Proof. Let X and Y be jointly multivariate normal distributed with mean $\mathbf{0}$ and covariance matrix Σ, i.e.

$$(X, Y) \sim \mathcal{N}(\mathbf{0}, \Sigma)$$

where $\Sigma = \begin{bmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{bmatrix}$.

Let $\Sigma_{YY} = U \Lambda U^T$ be the eigendecomposition of Σ_{YY}, with leading eigenvalue λ_1 and corresponding eigenvector u_1. Then the centroid of k-means applying on $(y_1, .., y_n)$ is on the first principal component of Y,

$$E(u_1^T Y | u_1^T Y > 0) = \bar{\mu}_1^Y = \sqrt{2\lambda_1/\pi} u_1$$

and

$$E(u_1^T Y | u_1^T Y < 0) = \bar{\mu}_2^Y = -\sqrt{2\lambda_1/\pi} u_1$$

where $u_1^T Y \sim \mathcal{N}(0, \lambda_1)$.

To compute $\bar{\mu}_1^X = E(X | u_1^T Y > 0)$, we need to know the conditional distribution $X | u_1^T Y$. Since (X, Y) has multivariate normal distribution, $(X, u_1^T Y)$ also has a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix

$$\Sigma_{X, u_1^T Y} = \begin{bmatrix} \Sigma_{XX} & \Sigma_{XY} u_1 \\ u_1^T \Sigma_{YX} & \lambda_1 \end{bmatrix}$$

The conditional distribution $X | u_1^T Y$ is hence normal with mean

$$\mu_{X | u_1^T Y} = \Sigma_{XY} u_1 \lambda_1^{-1} u_1^T Y$$

Therefore,

$$\begin{aligned}
\bar{\mu}_1^X &= E(X \mid u_1^T Y > 0) \\
&= E\left(E[X \mid u_1^T Y] \mid u_1^T Y > 0\right) \\
&= E\left(\Sigma_{XY} u_1 \lambda_1^{-1} u_1^T Y \mid u_1^T Y > 0\right) \\
&= \lambda_1^{-1} \Sigma_{XY} u_1 E(u_1^T Y \mid u_1^T Y > 0) \\
&= \lambda_1^{-1} \Sigma_{XY} u_1 \sqrt{2\lambda_1/\pi} \\
&= \sqrt{2/(\lambda_1 \pi)} \Sigma_{XY} u_1
\end{aligned}$$

Similar calculation yields $\bar{\mu}_2^X = -\sqrt{2/(\lambda_1\pi)}\Sigma_{XY}u_1$. The decision rule to classify any observed value of X to $\bar{\mu}_1^X$ is therefore

$$(\bar{\mu}_1^X)^T X > 0 \quad \text{or} \quad u_1^T \Sigma_{YX} X > 0$$

Since $u_1^T \Sigma_{YX} X$ is a linear combination of X, it also has normal distribution

$$\mathcal{N}\left(0, u_1^T \Sigma_{YX} \Sigma_{XX} \Sigma_{XY} u_1\right)$$

And $(Y, u_1^T \Sigma_{YX} X)$ also have multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix

$$\begin{bmatrix} \Sigma_{YY} & \Sigma_{YX}\Sigma_{XY}u_1 \\ u_1^T\Sigma_{YX}\Sigma_{XY} & u_1^T\Sigma_{YX}\Sigma_{XX}\Sigma_{XY}u_1 \end{bmatrix}$$

The conditional distribution of $Y|u_1^T\Sigma_{YX}X$ is also multivariate normal with mean

$$\mu_{Y|u_1^T\Sigma_{YX}X} = \Sigma_{YX}\Sigma_{XY}u_1(u_1^T\Sigma_{YX}\Sigma_{XX}\Sigma_{XY}u_1)^{-1}u_1^T\Sigma_{YX}X$$

The Y center for $u_1^T\Sigma_{YX}X > 0$ is

$$\begin{aligned}
\hat{\mu}_1^Y &= E(Y|u_1^T\Sigma_{YX}X > 0) \\
&= \Sigma_{YX}\Sigma_{XY}u_1(u_1^T\Sigma_{YX}\Sigma_{XX}\Sigma_{XY}u_1)^{-1}E(u_1^T\Sigma_{YX}X \mid u_1^T\Sigma_{YX}X > 0)
\end{aligned}$$

Note that $u_1^T\Sigma_{YX}X$ has normal distribution $\mathcal{N}\left(0, u_1^T\Sigma_{YX}\Sigma_{XX}\Sigma_{XY}u_1\right)$, so

$$E(u_1^T\Sigma_{YX}X \mid u_1^T\Sigma_{YX}X > 0) = \sqrt{2/\pi} \cdot \sqrt{u_1^T\Sigma_{YX}\Sigma_{XX}\Sigma_{XY}u_1}$$

Therefore, we have the Y center for $u_1^T \Sigma_{YX} X > 0$ be

$$\begin{aligned}\hat{\mu}_1^Y &= \sqrt{2/\pi} \cdot \sqrt{u_1^T \Sigma_{YX} \Sigma_{XX} \Sigma_{XY} u_1}\ \ \Sigma_{YX} \Sigma_{XY} u_1 (u_1^T \Sigma_{YX} \Sigma_{XX} \Sigma_{XY} u_1)^{-1} \\ &= \frac{\sqrt{2/\pi}}{\sqrt{u_1^T \Sigma_{YX} \Sigma_{XX} \Sigma_{XY} u_1}} \Sigma_{YX} \Sigma_{XY} u_1\end{aligned}$$

Recall that $\bar{\mu}_1^Y = \sqrt{2\lambda_1/\pi} u_1$, to judge if CV(2) > CV(1), one only need to compare the distance between $\hat{\mu}_1^Y$ and $\bar{\mu}_1^Y$ with distance between $\hat{\mu}_1^Y$ and grand mean 0. By the variance and bias decomposition of prediction MSE, when variance is the same, only bias influences the MSE.

After some linear algebra manipulation, we get $\|\hat{\mu}_1^Y - \bar{\mu}_1^Y\|^2 > \|\hat{\mu}_1^Y\|^2$ or CV(2) > CV(1) if and only if

$$\frac{\sqrt{\lambda_1}}{2} > \frac{u_1^T \Sigma_{YX} \Sigma_{XY} u_1}{\sqrt{u_1^T \Sigma_{YX} \Sigma_{XX} \Sigma_{XY} u_1}}$$

□

The condition that ensures CV(1) < CV(2) is tight but difficult to understand intuitively. We make understand this condition better by noting that it is equivalent to a condition on the correlation between the two linear combinations of X and Y variables:

$$\text{cor}(u_1^T \Sigma_{YX} X, u_1^T Y) < 1/2. \tag{1}$$

A stronger condition that is easier to understand, then, is that the canonical correlation between X and Y is below 1/2; when the latter is true the former follows and CV(1) < CV(2). A necessary but not sufficient guarantee of the stronger condition is that all correlations between variables are below 1/2.

One might object to the condition (1) being stated in terms of a particular (X, Y) split of the data. In practice, one might want the condition to hold for all possible splits of the data. A sufficient guarantee that (1) holds for any (X, Y) split of the data is that the data covariance matrix Σ is diagonal. The ad-hoc adjustment we propose in Sec. **??** attempts to transform the data to this form.

4 Analysis of two clusters in more than two dimensions

Proposition 3. *Suppose that* $\{(X_i, Y_i)\}_{i=1}^{n+m}$ *is data from a single fold of Gabriel cross-validation, where each* (X, Y) *pair in with* $X \in \mathbb{R}^P$ *and* $Y \in \mathbb{R}^Q$ *is an independent draw from an equiprobable mixture of two multivariate normal distributions with identity covariance. Suppose that the first mixture component has mean* $\mu = (\mu^X, \mu^Y)$ *and the second has mean* $-\mu = (-\mu^X, -\mu^Y)$*, where* $\mu^X \geq 0$ *and* $\mu^Y \geq 0$*. If the cluster centers are well separated, specifically such that* $2\varphi(\mu^Y) + \mu^Y + 2\mu^Y\Phi(\mu^Y) < 4\mu^Y\Phi(\mu^X)$*, then* $\mathrm{CV}(2) < \mathrm{CV}(1)$ *with probability tending to one as* m *and* n *increase.*

Proof. There are two clusters G_1 and G_2, where observations from G_1 are distributed as

$$\mathcal{N}\left(\begin{pmatrix}\mu^X\\ \mu^Y\end{pmatrix}, \mathbf{I}\right)$$

and observations from G_2 are distributed as

$$\mathcal{N}\left(\begin{pmatrix}-\mu^X\\ -\mu^Y\end{pmatrix}, \mathbf{I}\right)$$

here μ^X and μ^Y are all vectors. Let G_i be the true cluster where observation i is generated from, by assumption

$$P(G_i = G_1) = P(G_i = G_2) = 1/2$$

To simplify the notation, let

$$\begin{pmatrix}X_1\\ Y_1\end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix}\mu^X\\ \mu^Y\end{pmatrix}, \mathbf{I}\right)$$

denote the observations from G_1 and

$$\begin{pmatrix}X_2\\ Y_2\end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix}-\mu^X\\ -\mu^Y\end{pmatrix}, \mathbf{I}\right)$$

denote the observations from G_2

Further, let's denote $\mu^X = \lambda_x e_x$ where λ_x is a scalar denotes the distance of μ^X from origin and e_x is the unit vector point at the same direction as μ^X; $\mu^Y = \lambda_y e_y$ has the same interpretation.

Apply K-mean on the Y-space, where the two clusters are $\mathcal{N}\left(\mu^Y, \mathbf{I}\right)$ and $\mathcal{N}\left(-\mu^Y, \mathbf{I}\right)$, the K-mean centroids are $\bar{\mu}_1^Y$ and $\bar{\mu}_2^Y$ with $\bar{\mu}_1^Y = -\bar{\mu}_2^Y$. Note that the boundary between the two clusters are $e_y^T Y > 0$. So

$$\begin{aligned}
\bar{\mu}_1^Y &= E(Y \mid Y > 0) && (2)\\
&= E(Y_1 \mid e_y^T Y_1 > 0) \cdot P(e_y^T Y_1 > 0) + E(Y_2 \mid e_y^T Y_2 > 0) \cdot P(e_y^T Y_2 > 0) && (3)
\end{aligned}$$

Note that $e_y^T Y_1$ projects vector Y_1 on the direction of e_y. And because the e_y is the same direct as μ^Y, it goes through the center of the sphere $\mathcal{N}\left(\mu^Y, \mathbf{I}\right)$. Because the covariance matrix is $\mathbf{I}$, the sphere is symmetric around e_y. Therefore,

$$E(Y_1 \mid e_y^T Y_1 = a) = a e_y \tag{4}$$

Also, $Y_1 \sim \mathcal{N}\left(\mu^Y, \mathbf{I}\right)$ so $e_y^T Y_1 \sim \mathcal{N}\left(\lambda_y, 1\right)$. We have

$$\begin{aligned}
E(Y_1 \mid e_y^T Y_1 > 0) &= E\left[E\left(Y_1 \mid e_y^T Y_1\right) \mid e_y^T Y_1 > 0\right] && (5)\\
from\ (3)\ above &= E(e_y^T Y_1 e_y \mid e_y^T Y_1 > 0) && (6)\\
&= e_y E(e_y^T Y_1 \mid e_y^T Y_1 > 0) && (7)
\end{aligned}$$

Because $e_y^T Y_1 \sim \mathcal{N}\left(\lambda_y, 1\right) = \lambda_y + Z$, where Z is standard normal, by Lemma 3 from Appendix C we have

$$\begin{aligned}
E(e_y^T Y_1 \mid e_y^T Y_1 > 0) &= E(\lambda_y + Z \mid Z > -\lambda_y) && (8)\\
&= \lambda_y + E(Z \mid Z > -\lambda_y) && (9)\\
&= \lambda_y + \frac{\varphi(\lambda_y)}{\Phi(\lambda_y)} && (10)
\end{aligned}$$

where $\varphi()$ and $\Phi()$ are the standard normal probability and cumulative distribution function respectively. So, by (6) we have

$$E(Y_1 \mid e_y^T Y_1 > 0) = \left[\lambda_y + \frac{\varphi(\lambda_y)}{\Phi(\lambda_y)}\right] e_y \tag{11}$$

Similarly, we can have

$$\begin{aligned} E(Y_2 \mid e_y^T Y_2 > 0) &= -E(Y_1 \mid e_y^T Y_1 < 0) && (12)\\ &= -e_y E(e_y^T Y_1 \mid e_y^T Y_1 < 0) && (13)\\ &= \left[\frac{\varphi(\lambda_y)}{1-\Phi(\lambda_y)} - \lambda_y\right] e_y && (14) \end{aligned}$$

Because $e_y^T Y_1 \sim \mathcal{N}(\lambda_y, 1) = \lambda_y + Z$, it's easy to get

$$\begin{aligned} P(e_y^T Y_1 > 0) &= P(Z > -\lambda_y) && (15)\\ &= \Phi(\lambda_y) && (16) \end{aligned}$$

By symmetry, we can get

$$P(e_y^T Y_2 > 0) = 1 - \Phi(\lambda_y) \tag{17}$$

Put everything together, we have

$$\begin{aligned} \bar{\mu}_1^Y &= E(Y_1 \mid e_y^T Y_1 > 0) \cdot P(e_y^T Y_1 > 0) + E(Y_2 \mid e_y^T Y_2 > 0) \cdot P(e_y^T Y_2 > 0) && (18)\\ &= \left[\lambda_y + \frac{\varphi(\lambda_y)}{\Phi(\lambda_y)}\right] \cdot \Phi(\lambda_y) e_y + \left[\frac{\varphi(\lambda_y)}{1-\Phi(\lambda_y)} - \lambda_y\right] \cdot (1-\Phi(\lambda_y))\, e_y && (19)\\ &= [2\lambda_y \Phi(\lambda_y) + 2\varphi(\lambda_y) - \lambda_y] e_y && (20) \end{aligned}$$

After training the classifier, because of the identity covariance matrix, the classification boundary is $e_x^T X > 0$. So the Y center for observation with $e_x^T X > 0$ is

$$\hat{\mu}_1^Y = E(Y_1 \mid e_x^T X_1 > 0) \cdot P(e_x^T X_1 > 0) + E(Y_2 \mid e_x^T X_2 > 0) \cdot P(e_x^T X_2 > 0) \tag{21}$$

$$\text{X independent of Y} = E(Y_1) \cdot P(e_x^T X_1 > 0) + E(Y_2) \cdot P(e_x^T X_2 > 0) \tag{22}$$

$$= \mu^Y \cdot P(e_x^T X_1 > 0) - \mu^Y \cdot P(e_x^T X_2 > 0) \tag{23}$$

$$= \mu^Y \left(P(e_x^T X_1 > 0) - P(e_x^T X_2 > 0) \right) \tag{24}$$

$$= \mu^Y \left[\Phi(\lambda_x) - (1 - \Phi(\lambda_x)) \right] \tag{25}$$

$$= (2\Phi(\lambda_x) - 1)\mu^Y \tag{26}$$

$$= (2\Phi(\lambda_x) - 1)\lambda_y e_y \tag{27}$$

Because of symmetry and $P(G_i = G_1) = P(G_i = G_2) = 1/2$, it's sufficient to show that for observations with $e_x^T X > 0$, if $CV(2) < CV(1)$ then the Gabriel CV method correctly picks $k = 2$ over $k = 1$.

Similar as in the proof of Proposition 2, by the variance and bias decomposition of MSE, the variance is the same, so only the bias influences the result. Note the predicted center is grand 0 for $CV(1)$, so to see if $CV(2) < CV(1)$ one only need to see if $||\bar{\mu}_1^Y - \hat{\mu}_1^Y||^2 < ||\hat{\mu}_1^Y - 0||^2$, which is true if

$$2\Phi(\lambda_y) + 2\frac{\varphi(\lambda_y)}{\lambda_y} < 4\Phi(\lambda_x) - 1$$

this result reduces to the original result of Proposition **??** if one sets $\lambda_x = \mu^X$ and $\lambda_y = \mu^Y$. □

5 Technical Lemmas

Lemma 3. *If Z is a standard normal random variable, then*

$$\mathrm{E}(Z \mid a < Z < b) = -\frac{\varphi(b) - \varphi(a)}{\Phi(b) - \Phi(a)}$$

and

$$\mathrm{E}\{(Z - \delta)^2 \mid a < Z < b\} = \delta^2 + 1 - \frac{(b - 2\delta)\varphi(b) - (a - 2\delta)\varphi(a)}{\Phi(b) - \Phi(a)}$$

for all constants a, b, and δ, where $\varphi(z)$ and $\Phi(z)$ are the standard normal probability density and cumulative distribution functions. These expressions are valid for $a = -\infty$ or $b = \infty$ by taking limits.

Proof. We will derive the expression for the second moment. Integrate to get

$$\begin{aligned}\mathrm{E}[(Z-\delta)^2 1\{Z<b\}] &= \int_{-\infty}^{b} (z-\delta)^2 \varphi(z)\, dz \\ &= (\delta^2+1)\Phi(b) - (b-2\delta)\varphi(b).\end{aligned}$$

Now,

$$\mathrm{E}\{(Z-\delta)^2 \mid a<Z<b\} = \frac{\mathrm{E}[(Z-\delta)^2 1\{Z<b\}] - \mathrm{E}[(Z-\delta)^2 1\{Z<a\}]}{\Phi(b)-\Phi(a)}.$$

□

Lemma 3 has some important special cases:

$$\begin{aligned}\mathrm{E}\{Z \mid Z>0\} &= 2\varphi(0) = \sqrt{2/\pi}, \\ \mathrm{E}\{(Z-\delta)^2 \mid Z>0\} &= \delta^2 + 1 - 4\delta\varphi(0), \\ \mathrm{E}\{(Z-\delta)^2 \mid Z<0\} &= \delta^2 + 1 + 4\delta\varphi(0).\end{aligned}$$

6 Wold cross-validation

In Wold cross-validation, we perform "speckled" hold-outs in each fold, leaving out a random subset of the entries of the data matrix $\mathfrak{X} \in \mathbb{R}^{N\times P}$. For each value of k and each fold, we perform the following set of actions to get an estimate of cross-validation error, $\mathrm{CV}(k)$, which we average over all folds.

1. Randomly partition the set of indices $\{1,2,\ldots,N\} \times \{1,2,\ldots,P\}$ into a train set S_{train} and a test set S_{test}.

2. Apply a k-means fitting procedure that can handle missing data to the training data $\{\mathfrak{X}_{i,j} : (i,j) \in S_{\text{train}}\}$. This gives a set of cluster means $\mu(1),\ldots,\mu(k) \in \mathbb{R}^P$ and cluster labels for the rows, $G_1, G_2, \ldots, G_N$.

Table 1: Biological process enrichment within gene clusters

Cluster	Cluster Size	Process Category (In Cluster/Total Genes)	p-value
1	550	response to oxidative stress (24/55)	1.5×10^{-5}
		response to chemical (64/213)	2.2×10^{-5}
2	590	mitochondrion organization (79/159)	1.1×10^{-16}
		mitochondrial translation (28/51)	2.9×10^{-8}
		generation of precursor metabolites and energy (37/80)	7.3×10^{-8}
3	654	transcription from RNA polymerase II promoter (75/214)	5.5×10^{-6}
		mRNA processing (30/67)	2.7×10^{-5}
		mitotic cell cycle (63/183)	6.2×10^{-5}
4	634	cytoplasmic translation (105/134)	3.3×10^{-47}
		ribosomal subunit biogenesis (73/138)	7.7×10^{-17}
		rRNA processing (61/131)	5.7×10^{-11}
		ribosome assembly (21/36)	1.5×10^{-6}
5	517	chromosome segregation (53/106)	6.0×10^{-15}
		cellular response to DNA damage stimulus (71/172)	3.6×10^{-14}
		DNA repair (64/147)	3.7×10^{-14}
		DNA replication (42/78)	1.8×10^{-13}
		mitotic cell cycle (70/183)	4.8×10^{-12}

3. Compute the cross-validation error as

$$\mathrm{CV}(k) = \sum_{(i,j) \in S_{\text{test}}} \{\mathfrak{X}_{i,j} - \mu_j(G_i)\}^2,$$

where $\mu_j(G_i)$ denotes the jth component of $\mu(G_i)$.

7 Enrichment analysis

To further validate our clusters, we follow **?**, performing an enrichment analysis to discover which functional gene groups are significantly over-represented in each cluster. In the Saccharomyces Genome Database, each gene is mapped to a set of Gene Ontology categories. We focus on the 103 biological process categories.

For each category and each cluster, we compute a p-value for the null hypothesis that

genes from the category are distributed across all clusters without any bias towards the particular cluster in question. Under the null hypothesis, the number of genes from the category that end up in the cluster is distributed as a hypergeometric random variable. For each cluster, we compute p-values for all 103 biological process categories, and we report those that are significantly enrigched in Table 1. Using a Bonferroni correction to control the family-wise error rate at level 5%, we only report p-values that are less than $0.05/103 = 4.8 \times 10^{-4}$.

From Table 1, we can see that Cluster 1 is enriched with genes that somatize cell stress, such as oxidative heat-induce proteins. Cluster 2 contains genes that govern mitochondrial translation and mitochondrion organization. Cluster 3, the first period cluster, contains cell cycle genes related to budding and cell polarity, along with genes that govern RNA processing and transcription. Cluster 4 contains genes related to cytoplasmic translation and genes encoding ribosomes. Cluster 5, the second periodic cluster, contains genes that participate cell-cycle processes, along with DNA replication and DNA repair.

8 Comparison with Tavazoie clusters

In the **?** analysis, those authors performed k-means clustering with $k = 30$; they found 23 of the clusters to be uninterpretable, and they found 7 clusters to be meaningful. To compare our clusters with the Tavazoie et al. clusters, we prepared a confusion matrix comparing our clusters with the 7 interpretable Tavazoie clusters in Table 2. Entry (i, j) of the confusion matrix gives the number of genes in Tavazoie's Cluster i and our Cluster j.

Figure **??** provides a more in-depth comparison with the Tavazoie clusters, using a graphical confusion matrix. The plot in cell (i, j) of the upper left part of this figure gives the mean expression level for genes in the intersection of Tavazoie's Cluster i and our Cluster j; the plots in the margins give the mean expression levels for Tavazoie's clusters (top right) and our clusters (bottom left). In Figure **??**, we only include a plot for cell (i, j) if the number of genes in that cell is greater than 20.

Our Cluster 1 mainly consists of genes that Tavazoie et al. found to be in uninterpretable clusters. Our Cluster 2 contains high concentrations of Tavazoie's Clusters 4 and 8. Our first periodic cluster, Cluster 3, contains high concentrations of Tavazoie's Clulsters 3, 7,

Table 2: Confusion matrix comparing the 5 clusters found by Gabriel cross-validation to the 7 interpretable clusters found by the **?** analysis

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Total
Cluster 1	0	0	1	161	2	164
Cluster 2	1	0	0	0	185	186
Cluster 3	0	0	91	11	2	104
Cluster 4	0	102	2	66	0	170
Cluster 7	1	10	83	7	0	101
Cluster 8	3	145	0	0	0	148
Cluster 14	0	1	29	6	38	74
Other	545	332	448	383	290	1998
Total	550	590	654	634	517	2945

and 14; this is notable, because Tavazoie et al. highlighted their Clusters 7 and 14 as being periodic. Our Cluster 4 contains almost all of Tavazoie's Cluster 1, along with part of Tavazoie's Cluster 4. Finally, our second periodic cluster, Cluster 5, contains almost all of Tavazoie's Cluster 2, along with part of Tavazoie's Cluster 14; this, again, is notable, because Tavazoie et al. highlited these clusters as being periodic.

For the clusters that Tavazoie et al. were able to characterize, our analysis broadly agrees with the earlier clustering. The major difference between our analysis and that of **?** is that we are able to identify meaningful groups of genes with a much smaller value of k ($k = 5$ instead of $k = 30$), and we are able to interpret all of the clusters found by our analysis.