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 Analysis of single cluster in more than two dimensions

Proposition 1. *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

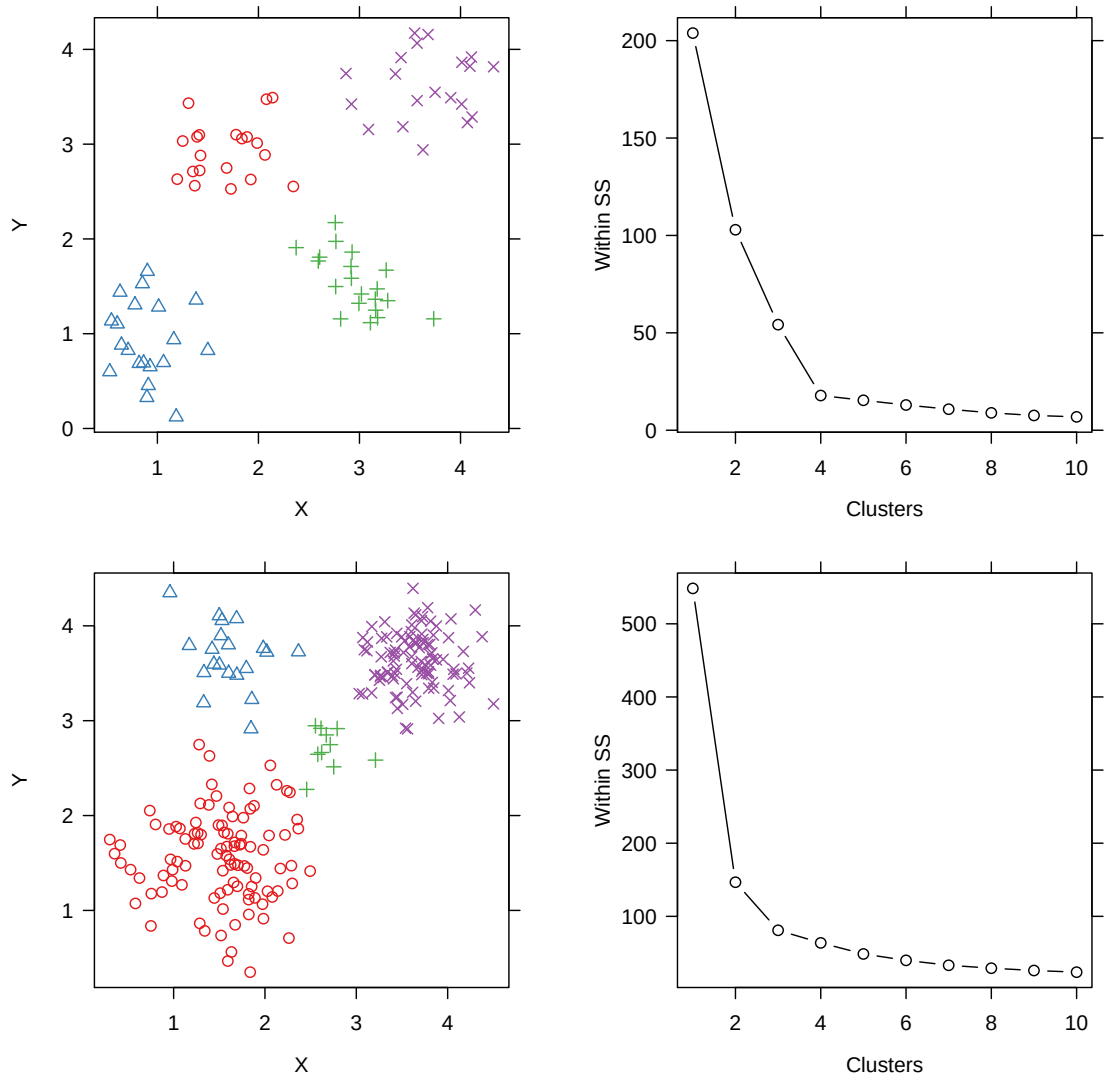

Figure 1: Scree plots for two data sets. Left panels show the sets of two-dimensional data points, generated from four clusters, with plotting symbol indicating the generated cluster. Right panels show the corresponding values of the within-cluster sum of squares W_k plotted against the number of clusters,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

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.

3 Analysis of two clusters 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 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* CV(2) < 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 $e_y^T Y_1 \sim \mathcal{N}(\lambda_y, 1)$. We have

$$E(Y_1 \mid e_y^T Y_1 > 0) = E\left[E\left(Y_1 \mid e_y^T Y_1\right) \mid e_y^T Y_1 > 0\right] \tag{5}$$

$$from\ (3)\ above = E(e_y^T Y_1 e_y \mid e_y^T Y_1 > 0) \tag{6}$$

$$= e_y E(e_y^T Y_1 \mid e_y^T Y_1 > 0) \tag{7}$$

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

$$E(e_y^T Y_1 \mid e_y^T Y_1 > 0) = E(\lambda_y + Z \mid Z > -\lambda_y) \tag{8}$$

$$= \lambda_y + E(Z \mid Z > -\lambda_y) \tag{9}$$

$$= \lambda_y + \frac{\varphi(\lambda_y)}{\Phi(\lambda_y)} \tag{10}$$

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

$$E(Y_2 \mid e_y^T Y_2 > 0) = -E(Y_1 \mid e_y^T Y_1 < 0) \tag{12}$$

$$= -e_y E(e_y^T Y_1 \mid e_y^T Y_1 < 0) \tag{13}$$

$$= \left[\frac{\varphi(\lambda_y)}{1 - \Phi(\lambda_y)} - \lambda_y\right] e_y \tag{14}$$

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

$$P(e_y^T Y_1 > 0) = P(Z > -\lambda_y) \tag{15}$$

$$= \Phi(\lambda_y) \tag{16}$$

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

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

$$\begin{aligned}
\text{X independent of Y} &= E(Y_1) \cdot P(e_x^T X_1 > 0) + E(Y_2) \cdot P(e_x^T X_2 > 0) && (22)\\
&= \mu^Y \cdot P(e_x^T X_1 > 0) - \mu^Y \cdot P(e_x^T X_2 > 0) && (23)\\
&= \mu^Y \left(P(e_x^T X_1 > 0) - P(e_x^T X_2 > 0)\right) && (24)\\
&= \mu^Y \left[\Phi(\lambda_x) - (1 - \Phi(\lambda_x))\right] && (25)\\
&= (2\Phi(\lambda_x) - 1)\mu^Y && (26)\\
&= (2\Phi(\lambda_x) - 1)\lambda_y e_y && (27)
\end{aligned}$$

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 1, 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 $\lambda_y = \mu^Y$. □

4 Technical Lemmas

Lemma 1. *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 1 has some important special cases:

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

5 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, $G_1, G_2, \ldots, G_N$.

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 $\mu(G_i)$.