Supplement to "Model-based Clustering for Gene Expression Data"
Data Transformations and the Gaussian mixture assumption

Ka Yee Yeung, Chris Fraley, Alejandro Murua, Adrian E. Rafter, Walter L. Ruzzo

June 30, 2001

Before applying model-based clustering to gene expression data, we assessed the extent to which the Gaussian mixture assumption holds. Since we do not expect raw expression data to satisfy the Gaussian mixture assumption, we explored the degree of normality of each class after applying different data transformations. In particular, we studied two types of data transformations: the Box-Cox transformations [Box and Cox, 1964], and the standardization of each gene (or clone) to have mean 0 and standard deviation 1.

The Box-Cox transformation [Box and Cox, 1964] is a parametric family of transformations from y to $y^{(\lambda)}$ with parameter λ:

$$y^{(\lambda)} = \begin{cases} \frac{y^{\lambda}-1}{\lambda} & \text{if } \lambda \neq 0 \\ \log y & \text{if } \lambda = 0. \end{cases} \tag{1}$$

The Box-Cox transformation subsumes many commonly used transformations, including the log transformation which is very popular for microarray data (for example, [Speed, 2000]).

Standardizing each gene (or clone) to have mean 0 and standard deviation 1 is another very popular data transformation step before clustering, for example, [Tamayo *et al.*, 1999] and [Tavazoie *et al.*, 1999]. Note that this standardization of subtracting the mean and dividing by the standard deviation makes correlation and Euclidean distance equivalent in the transformed data set.

1 Methodology to test Gaussian mixture assumption

In order to test the Gaussian mixture assumption, gene expression data sets with external criteria in Section 3.1 in [Yeung *et al.*, 2001] were used. We tested the multivariate normality of *each class* in each data set. There are large collections of tests for multivariate normality. We used three different approaches: goodness of fit tests based on the empirical distribution function, e.g. [Aitchison, 1986], skewness and kurtosis tests, e.g. [Jobson, 1991], and maximum likelihood estimation of the transformation parameters, e.g. [Andrews *et al.*, 1973].

Aitchison tests: [Aitchison, 1986] tested three aspects of the data for multivariate normality: the marginal univariate distribution, the bivariate angle distribution and the radius distribution. Suppose a gene expression data set has n genes and p experiments. Since we are interested in clustering the genes, the p experiments are our variables. There are a total of p tests for each of the marginal distribution, a total of $p(p-1)/2$ bivariate angle tests, and one radius test.

Let x_{ij} be the expression level of gene i under experiment j. Suppose the data set has G classes, and class g has n_g genes ($\sum_{g=1}^{G} n_g = n$). Let $\hat{\mu}^{\mathbf{g}} = [\hat{\mu}^{\mathbf{g}}_{\mathbf{j}}]$ and $\hat{\Sigma}^g = [\hat{\sigma}^g_{kj}]$ (where $k, j = 1, \ldots, p$) be the sample mean vector and covariance matrix for class g:

$$\hat{\mu}^g_j = \sum_{i=1}^{n_g} x_{ij}/n_g, \tag{2}$$

$$\hat{\sigma}^g_{kj} = \sum_{i=1}^{n_g} (x_{ik} - \hat{\mu}^g_k)(x_{ij} - \hat{\mu}^g_j)/(n_g - 1). \tag{3}$$

In the **marginal test**, the normality of the marginal distribution of each experiment j is evaluated. Let $\Phi(\cdot)$ denote the standard normal distribution function, and let $z_i^g = \Phi\{(x_{ij} - \hat{\mu}_j^g)/\sqrt{\hat{\sigma}_{jj}^g}\}$ (where $i = 1, \ldots, n_g$). If the x_{ij}'s are normally distributed in class g under experiment j, the sorted values of z_i^g in ascending order should approximate the order statistics of a uniform distribution over the interval (0,1).

Three different forms of empirical distribution functions (Anderson-Darling, Cramer-won Mises, and Watson) were used to measure departures of the sorted z_i^g values from the order statistics of the uniform distribution. Assuming that z_i^g are the sorted values from class g, The Anderson-Darling statistic is defined as $Q_A = -\{\sum_{i=1}^{n_g}(2i-1)\{\log z_i^g + \log(1 - z_{n_g+1-i}^g)\} - n_g\}/n_g$. The Cramer-von Mises statistic is defined as $Q_C = \sum_{i=1}^{n_g}\{z_i^g - (2i-1)/(2n_g)\}^2 + 1/(12n_g)$. The Watson statistic is defined as $Q_W = Q_C - n_g(\bar{z} - \frac{1}{2})^2$ where $\bar{z} = \sum_{i=1}^{n_g} z_i^g/n_g$. Critical values of the empirical distribution function test statistics are given in [Aitchison, 1986]. We used the critical values corresponding to the 1% significance level. For each class, we computed the empirical distribution function test statistics for each of the Anderson-Darling, Cramer-won Mises, and Watson forms using the z_i^g's. If a given test statistic for experiment j is greater than the critical value, we say that the marginal distribution of experiment j shows departure from normality.

In the **bivariate angle test**, the bivariate normality of each pair of experiments (k, j) is evaluated. The idea is that if a pair of variables (u_1, u_2) is circular normal, then the radian angle between the vector from the origin (0,0) to (u_1, u_2) and the u_1-axis is approximately uniform in the interval $[0, 2\pi]$. Since any bivariate normal distribution can be reduced to a circular normal distribution by a suitable transformation, we applied the transformation to each pair of experiments (k, j) and tested the resulting angle for the uniform property. Again, the empirical distribution function test statistics are used to measure the departure from the uniform distribution.

In the **radius test**, the radius of each gene i in class g is defined as $u_i = (\mathbf{x_i} - \hat{\mu}^\mathbf{g})^\mathbf{T}(\hat{\mathbf{\Sigma}}^\mathbf{g})^{\mathbf{-1}}(\mathbf{x_i} - \hat{\mu}^\mathbf{g})$, where $\mathbf{x_i}$ is the vector of expression levels of gene i under all p experiments. Under the multivariate normal assumption of $\mathbf{x_i}$'s, the radii u_i's are approximately distributed as $\chi^2(p)$. If we define z_i as the sorted values of $F(u_i)$, where F is the distribution function of $\chi^2(p)$, we can again use the empirical distribution function test statistics to measure deviation from the uniform distribution.

Skewness and Kurtosis: Skewness measures the amount of asymmetry in a distribution. For a normal distribution, the skewness is 0. Kurtosis measures the extent to which the data are peaked or flat relative to the normal distribution. For the standard normal distribution, the kurtosis is 3. We computed the skewness and kurtosis of each class g in the data. Let $m_{ir} = (\mathbf{x_i} - \hat{\mu}^\mathbf{g})^\mathbf{T}(\hat{\mathbf{\Sigma}}^\mathbf{g})^{\mathbf{-1}}(\mathbf{x_r} - \hat{\mu}^\mathbf{g})$, where $i, r = 1, \ldots, n_g$. Multivariate skewness and kurtosis are defined by $\sum_{i=1}^{n_g}\sum_{r=1}^{n_g} m_{ir}^3/n_g^2$ and $\sum_{i=1}^{n_g} u_i^2/n_g$, and there are distributions for both the multivariate skewness and kurtosis [Mardia, 1970]. A small p-value suggests the multivariate normal assumption to be questionable.

Maximum likelihood estimation of the transformation parameters: The parameter λ in the Box-Cox transformation in Equation 1 is estimated by maximum likelihood using the observations [Andrews *et al.*, 1973]. The estimated value of λ suggests both the scale on which the data are closest to normality, and also the extent to which the data on other scales deviate from normality.

2 Results of testing the Gaussian mixture assumption

We focused on the popular array data transformations: the logarithmic and square root transformations and the standardization to mean 0 and standard deviation 1. We applied the Aitchison tests and the skewness and kurtosis tests to each class in the transformed ovary data and the transformed yeast cell cycle data. Due to the large number of test statistics from the Aitchison tests ($(p + p(p-1)/2 + 1) * 3$) for each class on any data, only a summary of the Aitchison tests is presented in this technical report. In addition, we found the maximum likelihood estimates of the transformation parameter for each class.

Geometrically, the standardization of subtracting the mean and dividing by the standard deviation of each observation puts the data points on the (p-2) dimensional surface of a (p-1)-dimensional sphere. Moreover, the covariance matrices of the standardized data sets are singular. Hence, the skewness and kurtosis tests and the radius test (which involve the inverse of the covariance matrix) are not applicable to the standardized data.

Ovary data: Table 1 shows the results of the Aitchison tests on each of the four classes in the ovary data. In the marginal test, if the test statistics of an experiment j from all three empirical distribution functions are greater than their corresponding critical values at 1 % significance level, we adopt the shorthand convention of saying that experiment j *violates* the normality assumption. The column **m** in Table 1 shows the number of violations from the 24 marginal tests on each class in the ovary data. Similarly, the column **b** in Table 1 shows the number of violations from $\binom{24}{2} = 276$ bivariate angle tests on each class in the ovary data. The column **r** has an entry 1 if the test statistics from all three empirical distribution functions are greater than their corresponding critical values at 1 % significance level in the radius test. Otherwise, the column **r** has an entry 0. The results from Table 1 suggest that the square root transformation is closer to multivariate normal than the log transformation. On the square root transformed data, the marginal test shows that only one experiment (out of 24) deviates from normality in class 1. Similarly, class 2 has 6 experiments, class 3 has 4 experiments and class 4 has 3 experiments that deviate from marginal normality. None of the classes in the square root transformed data shows any deviation in the bivariate angle or radius tests. On the standardized data, the radius tests are not applicable, so the **r** columns for the standardized data are marked "NA" in Table 1.

	class 1			*class 2*			*class 3*			*class 4*		
	m	b	r	m	b	r	m	b	r	m	b	r
raw	0	0	0	5	0	0	18	12	0	4	1	0
log	9	0	0	14	12	0	2	0	0	4	0	0
sqrt	1	0	0	6	0	0	4	0	0	3	0	0
standardized	3	0	NA	7	13	NA	6	0	NA	5	2	NA

Table 1: Results of Aitchison tests on the ovary data.

		class 1	*class 2*	*class 3*	*class 4*
raw	skewness	0.844	0	0	1
raw	kurtosis	0.999	0.001	0.31	1
log	skewness	0.002	0	0.854	1
log	kurtosis	0.826	0	0.999	1
sqrt	skewness	0.768	0	0.559	1
sqrt	kurtosis	0.999	0.057	0.998	1

Table 2: p-values of skewness and kurtosis on the ovary data.

Table 2 shows the p-values of skewness and kurtosis for each class on the raw, log and square root transformed ovary data. A small p-value indicates deviations from the skewness and kurtosis criteria. From Table 2, class 2 deviates from the skewness and kurtosis criteria in the raw, log and square root transformed data. On the other hand, class 4 does not violate the skewness or kurtosis criteria. Both the square root and log transformations improve skewness in the raw data, but the log transformation makes class 1 skewed. To summarize, the skewness and kurtosis tests show the same overall picture as the Aitchison tests: the square root transformation is relatively close to multivariate normal.

Table 3 shows the results of the maximum likelihood estimation of the transformation parameters on each of the four classes on the raw ovary data. $\mathcal{L}_{max}(0.5)$ and $\mathcal{L}_{max}(0)$ are the maximum likelihood of the square root and log transformations respectively From Table 2, the optimal parameters for the Box-Cox transformation ($\hat{\lambda}$) lie between 0.40 and 0.73 for the four classes in the ovary data. Comparing the maximum likelihood values of the square root transformation to those of the log transformation shows that the square root transformation is closer to the multivariate normal distribution in all four classes.

class	$\hat{\lambda}$	$\mathcal{L}_{max}(\hat{\lambda})$	$\mathcal{L}_{max}(0.5)$	$\mathcal{L}_{max}(0)$
1	0.728	750	744	678
2	0.658	1195	1188	1060
3	0.405	1221	1219	1179
4	0.590	725	724	689

Table 3: Estimates of the transformation parameter for the ovary data.

Yeast cell cycle data with the 5-phase criterion: Table 4 shows the results of the Aitchison tests on the yeast cell cycle data with the 5-phase criterion. The results from Table 4 show that the log transformed yeast cell cycle data is relatively close to the multivariate normal distribution than the square root transformation. With the log transformation, classes 1, 3, and 4 show no deviation from any of the marginal, bivariate angle and radius tests. The only deviations from normality in this data set are: class 2 shows deviation from the radius test, and one experiment (out of 17) in class 5 shows deviation from marginal normality. The Aitchison tests show that the log transformation greatly enhances normality in all of the 5 classes: the raw data shows significant deviations from the marginal, bivariate angle and radius tests in all of the 5 classes. The standardized yeast cell cycle data is also much more Gaussian than the raw data, but not as much as the log transformed data.

	class 1			*class 2*			*class 3*			*class 4*			*class 5*		
	m	b	r	m	b	r	m	b	r	m	b	r	m	b	r
raw	17	49	1	17	136	1	17	94	1	17	0	1	17	33	1
log	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0
sqrt	8	0	1	17	1	1	15	0	1	0	0	0	7	0	0
standardized	5	0	NA	4	5	NA	1	0	NA	1	0	NA	2	0	NA

Table 4: Results of Aitchison tests on the yeast cell cycle data with the 5-phase criterion.

		class 1	*class 2*	*class 3*	*class 4*	*class 5*
raw	skewness	0	0	0	0	0
raw	kurtosis	0	0	0	0	0
log	skewness	0.051	0	0	0.046	0
log	kurtosis	0.735	0	0	0.678	0.001
sqrt	skewness	0	0	0	0	0
sqrt	kurtosis	0	0	0	0.003	0.001

Table 5: p-values of skewness and kurtosis on the yeast cell cycle data with the 5-phase criterion.

Table 5 portrays a different picture than the Aitchison tests: the raw, square root and log transformed data all show deviations from the skewness and kurtosis criteria. However, the log transformation seems to show relatively less deviation.

Table 6 supports the conclusions from the other approaches: the optimal transformation is closer (in terms of difference between Box-Cox power parameter) to the log transformation than to than the square root transformation. The estimates $\hat{\lambda}$ are between 0.14 and 0.22 for all 5 classes.

Yeast cell cycle data with the MIPS criterion: In general, the Aitchison tests, the skewness and kurtosis tests, and the maximum likelihood estimation all show similar patterns to the 5-phase criterion: the log transform is relatively more Gaussian than the square root transformation (see Tables 7, 8 and 9). However, class 4 (ribosomal

class	$\hat{\lambda}$	$\mathcal{L}_{max}(\hat{\lambda})$	$\mathcal{L}_{max}(0.5)$	$\mathcal{L}_{max}(0)$
1	0.136	-4833	-4910	-4844
2	0.140	-9398	-9591	-9429
3	0.202	-4920	-4975	-4945
4	0.153	-3422	-3468	-3431
5	0.219	-3676	-3713	-3701

Table 6: Estimates of the transformation parameter for the yeast cell cycle data with the 5-phase criterion.

proteins) shows significantly more deviations from normality with very low p-values for both the skewness and kurtosis tests using the log and square root transformations.

	class 1			*class 2*			*class 3*			*class 4*		
	m	b	r	m	b	r	m	b	r	m	b	r
raw	17	3	1	17	48	0	17	2	1	9	0	1
log	0	0	0	0	0	0	4	0	1	17	67	1
sqrt	8	0	0	15	0	0	12	0	1	14	1	1
standardized	6	1	NA	2	0	NA	3	0	NA	15	28	NA

Table 7: Results of Aitchison tests on the yeast cell cycle data with the MIPS criterion.

		class 1	*class 2*	*class 3*	*class 4*
raw	skewness	0	0	1	0
raw	kurtosis	0	0.046	1	0
log	skewness	0.136	0.999	1	0
log	kurtosis	0.896	0.999	1	0
sqrt	skewness	0	0.747	1	0
sqrt	kurtosis	0.014	0.996	1	0

Table 8: p-values of skewness and kurtosis on the yeast cell cycle data with the MIPS criterion.

References

[Aitchison, 1986] Aitchison, J. (1986) *The Statistical Analysis of Compositional Data*. Chapman and Hall.

[Andrews *et al.*, 1973] Andrews, D. F., Gnanadesikan, R. and Warner, J. L. (1973) Methods for assessing multivariate normality. In Krishnaiah, P. R. (ed.), *Multivariate analysis III*, New York: Academic Press, 95–116.

[Box and Cox, 1964] Box, G. E. P. and Cox, D. R. (1964) An analysis of transformations. *J. R. Statist. Soc. B*, **26**, 211–252.

[Jobson, 1991] Jobson, J. D. (1991) *Applied multivariate data analysis*. New York: Springer-Verlag.

[Mardia, 1970] Mardia, K. V. (1970) Measures of multivariate skewness and kurtosis with applications. *Biometrika*, **57**, 519–530.

[Speed, 2000] Speed, T. P. (2000) Speed group microarray page: Hints and prejudices. Http://stat-www.berkeley.edu/users/terry/zarray/Html/hintsindex.html.

class	$\hat{\lambda}$	$\mathcal{L}_{max}(\hat{\lambda})$	$\mathcal{L}_{max}(0.5)$	$\mathcal{L}_{max}(0)$
1	0.175	-3448	-3483	-3459
2	0.096	-1912	-1951	-1915
3	0.088	-808	-998	-969
4	0.308	-13188	-13234	-13323

Table 9: Estimates of the transformation parameter for the yeast cell cycle data with the MIPS criterion.

[Tamayo *et al.*, 1999] Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S. and Golub, T. R. (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Science USA*, **96**, 2907–2912.

[Tavazoie *et al.*, 1999] Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J. and Church, G. M. (1999) Systematic determination of genetic network architecture. *Nature Genetics*, **22**, 281–285.

[Yeung *et al.*, 2001] Yeung, K. Y., Fraley, C., Murua, A., Raftery, A. E. and Ruzzo, W. L. (2001) Model-based clustering and data transformations for gene expression data. *To appear in Bioinformatics*.