Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: http://www.tandfonline.com/loi/uasa20

Epistatic Clustering: A Model-Based Approach for Identifying Links Between Clusters

Jian Zhang

To cite this article: Jian Zhang (2013) Epistatic Clustering: A Model-Based Approach for Identifying Links Between Clusters, Journal of the American Statistical Association, 108:504, 1366-1384, DOI: 10.1080/01621459.2013.835661

To link to this article: https://doi.org/10.1080/01621459.2013.835661

Epistatic Clustering: A Model-Based Approach for Identifying Links Between Clusters

Jian ZHANG

Most clustering methods assume that the data can be represented by mutually exclusive clusters, although this assumption may not be the case in practice. For example, in gene expression microarray studies, investigators have often found that a gene can play multiple functions in a cell and may, therefore, belong to more than one cluster simultaneously, and that gene clusters can be linked to each other in certain pathways. This article examines the effect of the above assumption on the likelihood of finding latent clusters using theoretical calculations and simulation studies, for which the epistatic structures were known in advance, and on real data analyses. To explore potential links between clusters, we introduce an epistatic mixture model which extends the Gaussian mixture by including epistatic terms. A generalized expectation-maximization (EM) algorithm is developed to compute the related maximum likelihood estimators. The Bayesian information criterion is then used to determine the order of the proposed model. A bootstrap test is proposed for testing whether the epistatic mixture model is a significantly better fit to the data than a standard mixture model in which each data point belongs to one cluster. The asymptotic properties of the proposed estimators are also investigated when the number of analysis units is large. The results demonstrate that the epistatic links between clusters do have a serious effect on the accuracy of clustering and that our epistatic approach can substantially reduce such an effect and improve the fit.

KEY WORDS: Asymptotic property; Bootstrap test; Epistatic link; Finite epistatic mixture; Generalized EM algorithm; Model-based epistatic clustering.

1. INTRODUCTION

In this article, we propose a novel method for cluster analysis that produces epistatically linked clusters, where two clusters are defined as epistatically linked if there is another cluster whose distribution depends on these two. The epistatically linked clusters can be naturally generated from overlapping clusters. If two clusters are overlapping, then they can be partitioned into three mutually exclusive subsets, one of which is the intersection and the other two are the complements of the intersection in the two clusters. We focus on modeling the link between the intersection and the two complements. The joint behavior of multiple clusters, referred to as epistasis of clusters, is investigated. The methodology is general but was initially developed from clustering gene expression data generated by high throughput biological experiments, such as microarray experiments. In a standard microarray experiment, thousands of probes (i.e., fragments of DNA or RNA of variable lengths) are hybridized in a target DNA or RNA sample on a solid surface such as glass or a silicon chip to detect the presence of nucleotide sequences (the gene targets) that are complementary to the sequences in the probes. These experiments provide snapshots of all transcriptional activities of thousands of genes in a single DNA or RNA sample under relatively small numbers of experimental conditions and at a small number of time-points. Unlike most traditional molecular biology tools, which generally allow the study of a single gene or a small set of genes, microarray experiments facilitate the totally novel discovery of coregulated or functionally related genes (Eisen et al. 1998). For example, Gasch et al. (2000) used microarrays to analyze changes in transcript abundance in yeast cells responding to a panel of diverse environmental stresses. The resulting data matrix contains the expressions of 6152 yeast genes, where the columns of the data matrix correspond to the genes and the rows are for the different conditions, and the entries in each row are the expressions of the genes. In the more general setting, the columns stand for the analysis units (such as subjects, genes, etc.) and the rows would be the set of attributes on which one wishes to cluster the analysis units.

Although cluster analysis includes a broad suite of techniques, the existing algorithms roughly fall into two categories, namely heuristic-based methods and model-based methods. In heuristic-based settings, no probabilistic model is specified. Examples of these approaches include K-means (KM) clustering and hierarchical clustering (Kaufman and Rousseauw 1990). In model-based settings, a probabilistic model is employed. Fraley and Raftery (2002) built a Gaussian mixture (GM) model for clustering, where the model varies depending on the parameterization of the covariance matrices in each cluster.

Yeung, Medvedovic, and Bumgarner (2004) applied the GM model-based approach to both real and simulated gene expression datasets for which the underlying groupings were known in advance. They demonstrated that the GM method consistently outperforms leading heuristic-based methods, with the advantage of determining the number of clusters and suggesting an appropriate model for each cluster. Despite their success, we applied the GM method to the yeast stress dataset (Gasch et al. 2000) and found that the best GM fit can fail to capture the main data pattern as suggested by Figures 1 and 2 in Section 5. From these figures we uncovered a strongly visible inconsistency between the original dataset and an arbitrary one of 100 samples drawn from the best GM fit.

We hypothesize that the above inconsistency is mainly due to a potentially invalid model assumption that the data are

Jian Zhang is Professor of Statistics, School of Mathematics, Department of Statistics and Actuarial Science, School of Mathematics, University of Kent, Canterbury, Kent CT2 7NF, UK (E-mail: *jz79@kent.ac.uk*). We greatly appreciate the constructive and valuable comments from Co-Editor Jun Liu, an Associate Editor, two anonymous reviewers, and Professors Hal Stern and Qiwei Yao that have led to some significant improvements of the results and the presentation of the article.

Journal of the American Statistical Association
December 2013, Vol. 108, No. 504, Theory and Methods
DOI: 10.1080/01621459.2013.835661

represented by mutually exclusive clusters. Under this assumption, the profile of each analysis unit is depicted by only one of these clusters. In terms of distribution, the profile of each analysis unit follows a finite mixture distribution with the density

$$f(\mathbf{x}) = \pi_1 f_1(\mathbf{x}) + \pi_2 f_2(\mathbf{x}) + \cdots + \pi_m f_m(\mathbf{x}), \qquad (1.1)$$

where f_k, $1 \le k \le m$ are component densities and π_k, $1 \le k \le m$ are the corresponding mixing proportions which sum to one, and m is the unknown model order. However, this is not the case in microarray experiments as we expect some of genes are likely to belong to multiple clusters simultaneously because of their multiple functional roles in a cell. Singling out these genes gives rise to a new type of cluster called an epistatic cluster and also allows us to study the nature of dependence between the clusters. The epistatic clusters show the interactions between components f_k, $1 \le k \le m$. Such an epistatic structure suggests the following extension of the model (1.1),

$$f(\mathbf{x}) = \sum_{k=0}^{m} \pi_k f_k(\mathbf{x}) + \sum_{1 \le k_1 < k_2 \le m} \pi_{k_1 k_2} f_{k_1 k_2}(\mathbf{x}) + \text{high order interactions}, \qquad (1.2)$$

where the interaction terms $f_{k_1 k_2}(\mathbf{x})$, $1 \le k_1 < k_2 \le m$ are the densities for the two-way epistatic clusters with the mixing proportions $\pi_{k_1 k_2}$, $1 \le k_1 < k_2 \le m$ satisfying

$$\sum_{k=0}^{m} \pi_k + \sum_{1 \le k_1 < k_2 \le m} \pi_{k_1 k_2} + \cdots = 1.$$

These interaction terms will be modelized by a Bayesian pooling approach as described in the next section. We add the extra component $f_0(\mathbf{x})$ which is not interacting with any of other components to reflect the fact that some genes may be unrelated to any biological process under investigation. However, it can be dropped for other applications. The new model is in the same spirit as the linear epistatic regression model used in quantitative genetics (Cordell 2002).

A few questions related to the epistatic mixture model arise naturally. First, it is not clear how to construct epistatic terms. Second, it is not certain that the E-step and the M-step in the celebrated Expectation-Maximization (EM) algorithm have tractable computations for estimating the new model. Finally, it is not obvious whether a particular construction of the epistatic model will have epistatic effects on generating observations, similar to what we observed in epistatic genetics. Here, the epistatic effect is referred to the mean vector of the epistatic cluster, which accounts for the expression levels of analysis units in the epistatic cluster. Although the mean vectors of some clusters are not directly linked to each other, they have indirect links via the third parts (the epistatic clusters). In this sense, the effect of one cluster (i.e., the effects of biological processes, in which the genes in the cluster participate, in the context of gene expression) can be altered by others via epistatic clusters, and consequently the identification of these clusters may be hindered by the epistatic effects. These questions are poorly understood and need to be studied.

Our choice of epistatic terms was developed from the pioneering work of Battle, Segal, and Koller (2005), where they introduced an overlapping process model for gene expression data. In their model, the expression level of each gene is assumed to be a sum of the activities of the processes in which it participates, and the expression levels of a gene in different expression arrays (each experiment can contain a number of arrays) are assumed to be independent of each other. In terms of clustering, their assumption implies that the attribute vector of each analysis unit has a diagonal covariance structure and that the aggregation of interacting clusters is independent of their variability. The above assumption seems rather restrictive: for example, in time-course microarray experiments, the expression level of a gene may be correlated across different time-points. A gene may take part in different biological processes that have different noise levels when conditions are changing. In our model, we refine their assumption by taking into account the fact that the noises of gene activities in different processes may differ, and that the gene expression levels in different arrays can be correlated. In the more general setting of clustering, the refined assumption allows the attributes within an analysis unit to be correlated to each other. Under this refined assumption, the epistatic terms are built by aggregating the mean effects of components by the corresponding variabilities, while the correlations between different arrays are described by component covariance matrices. The links between these matrices are modeled by three types of parameterizations: (1) no links between the component covariances; (2) all the component covariances are equal; (3) the component covariances have unequal sizes but the same shapes or orientations. In general, the above methodology of modeling structural links between the clusters is innovative and represents a significant expansion of the existing Gaussian mixture models. The model estimation is carried out in two stages. In the first stage, letting the model order to be fixed for a moment, we calculate the maximum likelihood estimators using a generalized EM type algorithm. To speed up the convergence of the algorithm, we apply the so-called relative gradient approach in the M-step, exploring the non-Euclidean structure of the matrix space. In the second stage, the Bayesian Information criterion (BIC) is employed to assess the model complexity and to determine the model order. Our approach produces a miscellaneous cluster and several overlapping clusters. We term the overlapping parts of these clusters epistatic clusters and the nonoverlapped parts of these clusters primary clusters. Unlike the clusters derived from a standard mixture model, the means of the primary clusters may be linked to each other via the epistatic clusters.

We show the consistency of the BIC-based order estimator and establish some asymptotic properties of a BIC-based test statistic when the number of analysis units tends to infinity and the number of attributes is fixed. We apply the GM, KM, and epistatic methods to the yeast stress data (Gasch et al. 2000) and demonstrate that using our proposed approach can significantly improve the goodness of fit of GM to the data in terms of BIC. In the yeast stress data case, we take genes as analysis units and experiment conditions as attributes. Imitating the yeast stress study (Gasch et al. 2000), we simulate the expressions of 500 and 250 genes under eight conditions, respectively. These genes can be grouped into one miscellaneous cluster, five primary clusters, and ten epistatic clusters. Note that the simulation setting is close to the real yeast stress dataset but a slightly different number of genes, a different number of experimental conditions and a different number of clusters. We perform cluster analyses

on these synthetic data by use of the GM, the KM, and our proposed method, respectively. The results show that the GM and KM methods missed these primary clusters in most cases. In contrast, our proposed method consistently and dramatically reduced the epistatic effects on the degree of agreement between estimated clusters and underlying clusters, hence demonstrating the importance of modeling the epistatic interactions between clusters.

The article is organized as follows. The details of model specifications are provided in Section 2. The maximum likelihood estimators and the corresponding algorithms are developed in Section 3. The asymptotic properties of the proposed inference procedure are established in Section 4 when the number of analysis units tends to infinity, and the number of attributes is fixed. The simulation studies and a real data application are carried out in Section 5. The relationships with several existing models are discussed in Section 6. The additional technical details are deferred to the Appendix or to the online supplementary materials of the article.

2. MODEL SPECIFICATION

Consider a data matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ with p rows and n columns, where, for example, in genomic studies n and p, respectively, denote the number of genes and the number of microarrays, and $\mathbf{x}_i$ stands for the p-vector of the log-expression of the ith gene under the p-conditions. We first assume that the data points $\mathbf{x}_i$, $1 \le i \le n$ are independent of each other and that there are m clusters (or components) described by Gaussian densities $f_k(\mathbf{x}) = \phi(\mathbf{x}|\mu_k, \Sigma_k)$, $1 \le k \le m$ with means μ_k and covariances Σ_k, respectively. Let $\underline{\mu} = (\mu_1, \ldots, \mu_m)$ and $\underline{\Sigma} = (\Sigma_1, \ldots, \Sigma_m)$. Then the traditional mixture model defined in (1.1) can be viewed as the first-order approximation to the density of each data point by the component densities. Based on this approximation, an overall generative model is assumed, where the data are generated by first choosing one of the component densities and then sampling a data point from the chosen component. Let r_{ik} be the kth cluster-label indicator, taking the value of one or zero, according to whether $\mathbf{x}_i$ is assigned to the kth cluster. Let $|r_i|$ denote the L_1 norm of the vector $\mathbf{r}_i = (r_{i1}, \ldots, r_{im})^T$, which shows the number of clusters to which $\mathbf{x}_i$ is assigned to. In the traditional mixture model, the data point $\mathbf{x}_i$ is assigned to only one cluster and therefore $|\mathbf{r}_i| = 1$, where the conditional density of $\mathbf{x}_i$ given $\mathbf{r}_i$ can be written as

$$f(\mathbf{x}_i|\mathbf{r}_i) = \prod_{k=1}^{m} \phi(\mathbf{x}|\mu_k, \Sigma_k)^{r_{ik}}. \tag{2.1}$$

In this article, we are interested in clustering of genes wherein each gene can be assigned to more than one cluster or to a miscellaneous cluster, instead of being constrained to a single cluster. To achieve this, we remove the restriction that $|\mathbf{r}_i| = 1$, allowing $\mathbf{r}_i$ to have multiple ones. Meanwhile, we model the epistatic term by a latent Gaussian factor model $\phi(\mathbf{x}_i|\mu_{\mathbf{r}_i}, \Sigma_{\mathbf{r}_i})$, where the epistatic covariance matrix $\Sigma_{\mathbf{r}_i}$ is constructed by a weighted matrix-harmonic average of elements of $\underline{\Sigma}$ in the form of $(\sum_{k=1}^{m} \Sigma_k^{-1} r_{ik}/|\mathbf{r}_i|)^{-1}$ and the epistatic mean $\mu_{\mathbf{r}_i}$ is expressed as a matrix-weighted sum of elements of $\underline{\mu}$ in the form of $\Sigma_{\mathbf{r}_i} \sum_{k=1}^{m} r_{ik} \Sigma_k^{-1} \mu_k$. The idea behind the construction is close to Battle, Segal, and Koller (2005), where the primary clusters were assumed to be associated with m biological processes and the expression of gene i was assumed to be a sum of its expression levels in each of the processes in which it participates. Unfortunately, these authors assumed that all these biological processes have the same noise level which is not true. To extend their idea to the setting where the component covariance matrices may be different, we employ the Bayesian approach for pooling sample mean (covariance) and a prior mean (covariance)) to pooling the mean vectors (covariance matrices) of the processes in which gene i participates: We start with performing the Bayesian pooling operation on any two multivariate Gaussian variables involved in generating $\mathbf{x}_i$ by taking one as a sample variable and the other as a prior, generating a posterior variable with posterior mean and covariance. Then, we iteratively pool the posterior variable with one of the remaining multivariate Gaussian variables involved in generating $\mathbf{x}_i$ until all these variables are pooled together, giving rise to the above epistatic mean and matrix-harmonically average of $\underline{\Sigma}$ according to the weights $r_{ik}/|\mathbf{r}_i|$. We extend model (2.1) to

$$f(\mathbf{x}_i|\mathbf{r}_i) = \phi(\mathbf{x}_i|\mu_{\mathbf{r}_i}, \Sigma_{\mathbf{r}_i}). \tag{2.2}$$

A desirable property of model (2.2) is that when $\Sigma_1 = \cdots = \Sigma_m = \Sigma$, $\Sigma_{\mathbf{r}_i} = \Sigma$, and $\mu_{\mathbf{r}_i} = \sum_{k=1}^{m} r_{ik}\mu_k$ and, therefore, $f(\mathbf{x}_i|\mathbf{r}_i)$ reduces to the linear Gaussian factor model in Battle, Segal, and Koller (2005). If Σ_k, $1 \le k \le m$ can be decomposed into $\sigma_k^2 A$, $1 \le k \le m$, for some positive numbers σ_k^2 and a positive definite matrix A, then $\Sigma_{\mathbf{r}_i} = \sigma_{\mathbf{r}_i}^2 A$ and $\mu_{\mathbf{r}_i} = \sigma_{\mathbf{r}_i}^2 \sum_{k=1}^{m} r_{ik}\sigma_k^{-2}\mu_k$, where $\sigma_{\mathbf{r}_i}^2 = (\sum_{k=1}^{m} \sigma_k^{-2} r_{ik}/|\mathbf{r}_i|)^{-1}$ is the ordinary weighted-harmonic mean of σ_k^2, $1 \le k \le m$. Note that the larger the σ_k^2 is, the smaller contribution of μ_k to the epistatic mean $\mu_{\mathbf{r}_i}$. Therefore, under this model, when a gene participates in multiple biological processes, its expression is mainly determined by the processes with smaller noise levels. Based on (2.2), the marginal (i.e., overall) density of $\mathbf{x}_i$ can be written as

$$f(\mathbf{x}_i) = \sum_{d \in \Omega} w(d) f(\mathbf{x}_i|\mathbf{r}_i = d), \tag{2.3}$$

where Ω is the space for $\mathbf{r}_i$, $w(d) = P(\mathbf{r}_i = d)$ is the probability mass of $\mathbf{r}_i$ at d, and $\sum_{d\in\Omega} w(d) = 1$. The equation (2.3) is the special case of the model (1.2) with the epistatic term defined in (2.2). The above epistatic mixture model is a structured Gaussian mixture model in the sense that in the epistatic mixture model there are certain imposed structural links between the clusters while there are no such constraints in a standard mixture model. The epistatic mixture model will be more efficient than a standard mixture model if these interaction terms are correct. Note that $f(\mathbf{x}_i|\mathbf{r}_i)$ reduces to the kth component density if $|\mathbf{r}_i| = 1$ and $r_{ik} = 1$, and that for $|\mathbf{r}_i| \ge 2$, $f(\mathbf{x}_i|\mathbf{r}_i)$ defines an epistatic cluster. Consequently, the density in (2.3) gives a generic description of model (1.2). For convenience, we define a miscellaneous cluster with mean μ_0 and Σ_0 for noisy data points or outliers. Accordingly, the indicator r_{i0} of the miscellaneous cluster satisfies the equation $r_{i0} = \prod_{k=1}^{m}(1 - r_{ik})$ (i.e., $r_{i0} = 1$ when all r_{ik}, $1 \le k \le m$, are zeros). The concept of miscellaneous cluster was originally from the literature of data mining, where there is often a "rag bag" category (i.e., cluster) containing objects that cannot be merged with other objects. Such a cluster is often called miscellaneous (Han, Kamber, and Pei 2011). It could be

useful in some settings and could easily be left out of the fitting procedure by setting all $r_{i0} = 0$ when it is not applicable. We restrict the values of $\mathbf{r}_i$ to the space $\Omega_g = \{\mathbf{r} \in \{0, 1\}^m : |\mathbf{r}| \le g\}$, where the integer g, prespecified either by the users (in this article, we use a default $g = 2$) or by model selection, is used to control the maximum number of clusters to which a data point can be assigned. Only pairwise interactions between the components are considered when setting $g = 2$. Let $\{d_j, 1 \le j \le \tau(g)\}$ denote the configurations in $\Omega_g \hat{=} \Omega_{gm}$ with $d_1 = (0, \ldots, 0)^T$, $d_j = (0, \ldots, 0, 1, 0, \ldots)^T$, $2 \le j \le m + 1$, $d_{m+2} = (1, 1, \ldots)^T$, etc., where $\tau(g) = 1 + \binom{m}{1} + \cdots + \binom{m}{g}$ is the size of Ω_g. We take $w_j = w(d_j)$, $1 \le j \le \tau(g)$ as parameters to be estimated from the data. For simplicity, we define $\mathbf{r}_i/|\mathbf{r}_i| = \mathbf{0}$ (a vector of m zeros) and $\Sigma_{\mathbf{r}_i} = \Sigma_0$ when $|\mathbf{r}_i| = 0$.

Following Fraley and Raftery (2002), we consider six different parameterizations of the component covariance matrices:

- EVE (Epistatic mixture model with Varying Elliptical component covariances): Component covariances Σ_k, $0 \le k \le m$ are arbitrary and do not link each other. Here the dimension of the model is $\dim(\text{EVE}) = (m + 1)p(p + 1)/2 + p(m + 1) + \tau(g) - 1$.
- EEE (Epistatic mixture model with Equal Elliptical component covariances): All component covariances but Σ_0 are equal, that is, $\Sigma_1 = \cdots = \Sigma_m$. Here, $\dim(\text{EEE}) = p(p + m + 2) + \tau(g) - 1$.
- EED (Epistatic mixture model with Equal Diagonal component covariances): All component covariances but Σ_0 are equal with $\Sigma_1 = \cdots = \Sigma_m = D$, where D is a diagonal matrix. Here $\dim(\text{EED}) = \frac{p(p+1)}{2} + p + p(m + 1) + \tau(g) - 1$.
- EVA (Epistatic mixture model with Varying size component covariances and equal orientation matrix A): Σ_0 is arbitrary, $\Sigma_k = \sigma_k^2 A$, $1 \le k \le m$, where A is arbitrary, and $\sigma_1^2 = 1$. Here, $\dim(\text{EVA}) = m - 1 + p(p + m + 2) + \tau(g) - 1$.
- EVD (Epistatic mixture model with Varying size component covariances and equal Diagonal orientation matrix): Σ_0 is arbitrary, $\Sigma_k = \sigma_k^2 D$, $1 \le k \le m$, D is a diagonal matrix, and $\sigma_1^2 = 1$. Here, $\dim(\text{EVD}) = m - 1 + p + p(p + 1)/2 + p(m + 1) + \tau(g) - 1$.
- EVI (Epistatic mixture model with Varying size component covariances and Identity orientation matrix): Σ_0 is arbitrary and $\Sigma_k = \sigma_k^2 I_p$, $1 \le k \le m$. Here, $\dim(\text{EVI}) = m + p(p + 1)/2 + p(m + 1) + \tau(g) - 1$.

Note that in the classical factor models, all factors are assumed to have the same level of noise. The EVA, EVD, and EVI are the generalized factor models in the sense that factors may have different noise levels, while the EEE and EED are the classical linear factor models. In particular, the EED is the overlapping process model introduced by Battle, Segal, and Koller (2005). We have not considered the other parameterizations of Fraley and Raftery (2002) due to the lack of efficient algorithms for the model estimation in these cases.

3. MODEL-BASED INFERENCE

In this section, we first introduce the maximum likelihood estimators and the Bayesian Information Criterion (BIC) for estimating our epistatic mixture models, and derive the associated EM algorithms. Then, we propose an epistatic mixture-based clustering procedure and describe the existing K-mean and the Gaussian mixture-based procedures for comparison later. We introduce an adjusted RAND index for measuring the quality of clustering results and a subsampling procedure for stability analyses on the results. Finally, we develop a bootstrap procedure for testing the significance of the fitted epistatic mixture models, compared to the standard Gaussian mixture models.

3.1 Maximum Likelihood Estimators

Assume that $\mathbf{x}_i$, $1 \le i \le n$ are independent observations drawn from model (2.3) with the mixing proportions w_j, $1 \le j \le \tau(g)$ for the components $f(\mathbf{x}|\mathbf{r} = d)$, $d \in \Omega_g$. Let $\mathbf{w} = (w_1, \ldots, w_{\tau(g)})^T$, and define $f(\mathbf{x}_i|\mu_0, \Sigma_0, \underline{\mu}, \underline{\Sigma}, \mathbf{w}) = \sum_{j=1}^{\tau(g)} w_j \phi(\mathbf{x}_i|\mu_{d_j}, \Sigma_{d_j})$, where $\phi(\mathbf{x}_i|\mu_{d_j}, \Sigma_{d_j})$ is a multivariate normal with mean μ_{d_j} and covariance matrix Σ_{d_j}. We have the incomplete log-likelihood

$$l(\mu_0, \Sigma_0, \underline{\mu}, \underline{\Sigma}, \mathbf{w}|\mathbf{X}) = \sum_{i=1}^{n} \log(f(\mathbf{x}_i|\mu_0, \Sigma_0, \underline{\mu}, \underline{\Sigma}, \mathbf{w}))$$

and the complete log-likelihood

$$l_c(\mu_0, \Sigma_0, \underline{\mu}, \underline{\Sigma}, \mathbf{w}|\mathbf{X}, \mathbf{z}) = \sum_{i=1}^{n} \sum_{j=1}^{\tau(g)} z_{ij} \log(\phi(\mathbf{x}_i|\mu_{d_j}, \Sigma_{d_j}) w_j),$$

where $\mathbf{z} = (z_{ij})_{n \times \tau(g)}$, with $z_{ij} = 1$ when $\mathbf{r}_i = d_j$ and $z_{ij} = 0$ otherwise. For each m, maximizing the above log-likelihood with respect to μ_0, Σ_0, $\underline{\mu}$, $\underline{\Sigma}$, and $\mathbf{w}$, we obtain maximum likelihood (ML) estimators $\hat{\mu}_0$, $\hat{\Sigma}_0$, $\underline{\hat{\mu}}$, $\underline{\hat{\Sigma}}$, and $\hat{\mathbf{w}}$ and maximum likelihood $l_{\max}$, where their dependence on m is suppressed. For a candidate epistatic mixture model M, we define the BIC as

$$\text{BIC(M)} = -2l_{\max} + \dim(\text{M}) \log(n).$$

We select the model M and determine the latent model order m by minimizing BIC(M).

3.2 Algorithm

We develop the following generalized EM algorithm for calculating the above ML estimators, which includes two steps as follows.

E-step. Given $\mathbf{X}$ and the current values $\mu_0^{(v)}$, $\Sigma_0^{(v)}$, $\underline{\mu}^{(v)}$, $\underline{\Sigma}^{(v)}$, and $\mathbf{w}^{(v)}$, the posterior distribution of $(z_{i1}, \ldots, z_{i\tau(g)})^T$ is a multinomial distribution with probabilities

$$\begin{aligned}\tau_{ij}^{(v)} &= E\left[z_{ij}|\mathbf{X}, \mathbf{w}^{(v)}, \underline{\mu}^{(v)}, \underline{\Sigma}^{(v)}\right] \\ &= \frac{\phi\left(\mathbf{x}_i|\mu_{d_j}^{(v)}, \Sigma_{d_j}^{(v)}\right) w_j^{(v)}}{\sum_{j=1}^{\tau(g)} \phi\left(\mathbf{x}_i|\mu_{d_j}^{(v)}, \Sigma_{d_j}^{(v)}\right) w_j^{(v)}}, \quad 1 \le j \le \tau(g).\end{aligned}$$

The posterior distribution of $\underline{\mathbf{r}}$,

$$p\left(\underline{\mathbf{r}}|\mathbf{X}, \mathbf{w}^{(v)}, \underline{\mu}^{(v)}, \underline{\Sigma}^{(v)}\right) = \prod_{i=1}^{n} p\left(\mathbf{r}_i|\mathbf{X}, \mathbf{w}^{(v)}, \underline{\mu}^{(v)}, \underline{\Sigma}^{(v)}\right),$$

where $p(\mathbf{r}_i|\mathbf{X}, \mathbf{w}^{(v)}, \underline{\mu}^{(v)}, \underline{\Sigma}^{(v)})$ is equivalent to a posterior distribution of $(z_{i1}, \ldots, z_{i\tau(g)})^T$. The expectation of the complete

log-likelihood with respect to z_{ij}'s can be expressed as

$$\begin{aligned}
&\Psi(\mu_0, \Sigma_0, \underline{\mu}, \underline{\Sigma}, \mathbf{w}) \\
&\quad = E\big[l_c(\mu_0, \Sigma_0, \underline{\mu}, \underline{\Sigma}, \mathbf{w}|\mathbf{X}, \underline{\mathbf{r}})|\mathbf{X}, \mathbf{w}^{(v)}, \underline{\mu}^{(v)}, \underline{\Sigma}^{(v)}\big] \\
&\quad = \sum_{j=1}^{\tau(g)}\sum_{i=1}^{n} \tau_{ij}^{(v)} \log(w_j) - \sum_{j=1}^{\tau(g)}\sum_{i=1}^{n} \tau_{ij}^{(v)} \log(2\pi)^{p/2} \\
&\qquad + \frac{1}{2}\sum_{j=2}^{\tau(g)}\sum_{i=1}^{n} \tau_{ij}^{(v)} \log|\Sigma_{d_j}^{-1}| + \frac{1}{2}\sum_{i=1}^{n} \tau_{i1}^{(v)} \log|\Sigma_0^{-1}| \\
&\qquad - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=2}^{\tau(g)} \tau_{ij}^{(v)} (x_i - \mu_{d_j})^T \Sigma_{d_j}^{-1} (x_i - \mu_{d_j}) \\
&\qquad - \frac{1}{2}\sum_{i=1}^{n} \tau_{i1}^{(v)} (x_i - \mu_0)^T \Sigma_0^{-1} (x_i - \mu_0).
\end{aligned}$$

M-step. First, we maximize Ψ with respect to $\mathbf{w}$ subject to $\sum_{j=1}^{\tau(g)} w_j = 1$. This leads to the updated estimate of $\mathbf{w}$, $w_j^{(v+1)} = \frac{1}{n}\sum_{i=1}^{n} \tau_{ij}^{(v)}$, $j = 1, \ldots, \tau(g)$. Similarly, maximizing Ψ with respect to μ_0 and Σ_0 gives

$$\begin{aligned}
\mu_0^{(v+1)} &= \frac{\sum_{i=1}^{n} \tau_{i1}^{(v)} \mathbf{x}_i}{\sum_{i=1}^{n} \tau_{i1}^{(v)}}, \\
\Sigma_0^{(v+1)} &= \frac{\sum_{i=1}^{n} \tau_{i1}^{(v)} \big(x_i - \mu_0^{(v+1)}\big)\big(x_i - \mu_0^{(v+1)}\big)^{\tau}}{\sum_{i=1}^{n} \tau_{i1}^{(v)}}.
\end{aligned}$$

In practice, we need to regularize $\Sigma_0^{(v+1)}$ by adding a diagonal matrix $0.00001 I_p$ when the size of the miscellaneous cluster is less than p. For $\underline{\mu}$ and $\underline{\Sigma}$ under the different parameterizations of the component covariances, the derivations of their updating formulas are very involved. (See Appendix A for the details.)

Finally, for each update, we calculate the resulting value of the observed likelihood. The algorithm can be stopped when the difference between the values at two successive steps is less than a threshold. We initialize the above EM algorithms by the corresponding GM solutions or random values.

3.3 Cluster Analysis

We consider the three clustering procedures as follows.

Epistatic (EP) approach. The set of clusters, $C_j, 1 \le j \le \tau(g)$, to which the data points belong, is determined by the probability of the ith data point being in the jth cluster,

$$w_{ij} = P(z_{ij} = 1) = \frac{w_j f(\mathbf{x}_i|\mathbf{r}_i = d_j)}{\sum_{j=1}^{\tau(g)} w_j f(\mathbf{x}_i|\mathbf{r}_i = d_j)},$$

where C_1 stands for the miscellaneous cluster, $C_j, 2 \le j \le m + 1$ for the m primary clusters, and $C_j, m + 2 \le j \le \tau(g)$ for the epistatic clusters. We can reformulate the primary clusters and the epistatic clusters into m overlapping clusters $O_k, 1 \le k \le m$, where

$$\begin{aligned}
O_k = C_{k+1} \cup \{&C_j : m + 2 \le j \le \tau(g), \\
&C_j \text{ is an epistatic clusters of } C_{k+1}\}.
\end{aligned}$$

Based on the minimum BIC criterion, we obtain the estimates $\hat{m}$, $\hat{g}$, $\hat{w}_j$ and $\hat{f}(\mathbf{x}_i|\mathbf{r}_i = d_j), 1 \le j \le \tau(\hat{g})$, and the minimum BIC epistatic model $\sum_{j=1}^{\tau(\hat{g})} \hat{w}_j \hat{f}(\mathbf{x}_i|\mathbf{r}_i = d_j)$. Then, w_{ij} can be estimated by

$$\hat{w}_{ij} = \frac{\hat{w}_j \hat{f}(\mathbf{x}_i|\mathbf{r}_i = d_j)}{\sum_{j=1}^{\tau(\hat{g})} \hat{w}_j \hat{f}(\mathbf{x}_i|\mathbf{r}_i = d_j)}.$$

We assign the ith observation to the jth cluster C_j if $\hat{w}_{ij} = \max_{1 \le t \le \tau(\hat{g})} \hat{w}_{it}$. This leads to the partition $\hat{C}_j, 1 \le j \le \tau(\hat{g})$ of $\{\mathbf{x}_i\}$.

GM approach. The number of clusters is determined by minimizing the corresponding BIC. The GM method was implemented in the software Mclust by Fraley and Raftery (2002).

KM approach. In the KM, the number of clusters is determined by maximizing the so-called average silhouette width defined by Kaufman and Rousseeuw (1990). See the online supplementary materials for the details.

We use the adjusted Rand index ρ of Hubert and Arabie (1985) to assess the degree of agreement between $\hat{C}$ and another partition, say C. The larger the value of ρ is, the higher degree agreement the two partitions have. The adjusted Rand index can take value from 0 to 1. When $\rho = 1$, $\hat{C}$ perfectly matches C. On other hand, when $\rho = 0$, there are no pair of analysis units classified in the same way under two partitions. In a simulation study, where the true grouping is known, we directly calculate the adjusted Rand index to assess the quality of a clustering result. However, in a real data analysis, for the EP and the GM, we respectively calculate the minimum BIC over the corresponding model classes instead as the true groupings are unknown. Note that the minimum BIC identifies a best fit which is asymptotically closest to the underlying model in the penalized Kullback-Leibler distance. Therefore, the clustering structure defined by the best fit is expected to give a high RAND index with the true grouping. This is confirmed by simulation studies in Section 5, where we demonstrate that on average the adjusted RAND index is a linear function of the minimum BIC with an negative slope.

To assess the stability of the clustering results, we conduct the EP, the GM and the KM analyses on randomly selected subsets of the full data. We check the stability of the so-called percentage BIC reduction and the optimal number of clusters. See the online supplementary materials for the details.

3.4 Hypothesis Testing

Fraley and Raftery (2002) considered 10 multivariate Gaussian mixture models via different parameterizations of the component covariance matrices:

- EII: $\Sigma_k = \lambda I_p, 1 \le k \le m$.
- VII: $\Sigma_k = \lambda_k I_p, 1 \le k \le m$.
- EEI: $\Sigma_k = \lambda A, 1 \le k \le m$, A is a diagonal orientation matrix.
- VEI: $\Sigma_k = \lambda_k A, 1 \le k \le m$, A is a diagonal orientation matrix.
- EVI: $\Sigma_k = \lambda A_k, 1 \le k \le m$, A_k is a diagonal orientation matrix.
- VVI: $\Sigma_k = \lambda_k A_k, 1 \le k \le m$, A_k is a diagonal orientation matrix.
- EEE: $\Sigma_k = \lambda D A D^T, 1 \le k \le m$.
- EEV: $\Sigma_k = \lambda D_k A D_k^T, 1 \le k \le m$.
- VEV: $\Sigma_k = \lambda_k D_k A D_k^T, 1 \le k \le m$.
- VVV: $\Sigma_k = \lambda_k D_k A_k D_k^T, 1 \le k \le m$.

The best fit to the data can be identified by minimizing the BIC value. The main difference between the epistatic models (EVE, EEE, EED, EVA, EVD, EVI) and the Gaussian mixture models lies in whether the epistatic terms are included or not. Therefore, it is then natural to ask whether including these epistatic terms is necessary, or whether these terms are statistically significant. This leads to the problem of testing for the contrast between these models, which is equivalent to testing the competing hypotheses between two model classes $\mathcal{M}_0$ and $\mathcal{M}_1$.

H_0: The data are sampled from the model in $\mathcal{M}_0$,

against

H_1: The data are sampled from the model in $\mathcal{M}_1$,

where

$$\mathcal{M}_0 = \{\text{GM models EII, VII, EEI, VEI, EVI,VVI, EEE, EEV, VEV, VVV}\}$$

and

$$\mathcal{M}_1 = \{\text{EP models EVE, EEE, EED, EVA, EVD, EVI with some positive epistatic mixing proportions}\}.$$

Let $\text{BIC}(\mathcal{M}_1)$ and $\text{BIC}(\mathcal{M}_0)$ be the minimum BIC values over all models in $\mathcal{M}_0$ and $\mathcal{M}_1$, respectively. Then, the BIC difference $\text{BIC}(\mathcal{M}_1) - \text{BIC}(\mathcal{M}_0)$ is a natural statistic for testing the above hypotheses. However, its null-distribution is hard to calculate. In practice, we employ a parametric bootstrap approach to calculate it instead. In the bootstrap procedure, we first conduct cluster analyses by fitting the GM and the proposed epistatic models to the data to find the best GM and epistatic fits. Suppose that these best fits have the minimum BIC values of BIC_{OG} and BIC_{OE}, respectively. We then repeatedly draw samples from the best GM fit, obtaining N bootstrap samples, say $\mathbf{X}^{(t)}$, $1 \le t \le N$. We fit the GM and epistatic models to these bootstrap samples, respectively, obtaining N bootstrapped pairs of the minimum BIC values, $\text{BIC}_{bG}^{(t)}$ and $\text{BIC}_{bE}^{(t)}$, and the BIC differences $\text{BIC}_{bE}^{(t)} - \text{BIC}_{bG}^{(t)}$, $1 \le t \le N$. Finally, the bootstrap p-value is defined as the proportion of these bootstrap BIC differences which are less than the observed.

4. THEORETICAL PROPERTIES

In microarray studies, the number of genes n can be very large but the number of microarrays p is limited. In this section, we investigate the asymptotic behavior of our procedure when n tends to infinity and p is fixed. We first extend the result on the consistency of the estimated order of the Gaussian mixtures (Gassiat 2002) to the epistatic mixtures in Theorem 1 and then develop an asymptotic theory for the maximum likelihood ratio test and the BIC boostrap test. In particular, we show that under some regulrality conditions, the BIC bootstrap p-value converges to the target p-value in probability.

4.1 Order Consistency

To describe the result of order consistency, we first introduce the concept of identifiability in the context of epistatic mixture models as follows. Let f_0 denote the underlying epistatic mixture density of order m_0. Let $\mathcal{Q}_p$ be the set of all $p \times p$ positive definite matrices with eigenvalues being bounded from below and above by two positive constants $\delta_1 < \delta_2$ and $\mathbf{R}^p$ be the p-dimensional Euclidean space. Let $\Omega_{g^*}^*$ denote $\Omega_{g^* m^*}$. Finite epistatic mixtures from the Gaussian family $\{\phi(\cdot|\mu, \Sigma) : \mu \in R^p, \Sigma \in \mathcal{Q}_p\}$ are called identifiable if a relation for all x of the form,

$$\sum_{k=0}^{m} \pi_k \phi(x|\mu_k, \Sigma_k) + \sum_{|d|\ge 2, d\in\Omega_g} w(d)\phi(x|\mu_d, \Sigma_d)$$
$$= \sum_{k=0}^{m^*} \pi_k^* \phi(x|\mu_k^*, \Sigma_k^*) + \sum_{|d|\ge 2, d\in\Omega_{g^*}^*} w^*(d)\phi(x|\mu_d^*, \Sigma_d^*),$$

where m and m^* are positive integers, $\sum_{k=0}^{m} \pi_k + \sum_{|d|\ge 2, d\in\Omega_g} w(d) = \sum_{k=0}^{m^*} \pi_k^* + \sum_{|d|\ge 2, d\in\Omega_g^*} w^*(d) = 1$ and $\pi_k > 0, 1 \le k \le m, \pi_k^* > 0, 1 \le k \le m^*$, implies that $m = m^*$, $g = g^*$ and that there exists a permutation ν on $1, 2, \ldots, m$ such that $(\pi_k^*, \mu_k^*, \Sigma_k^*) = (\pi_{\nu(k)}, \mu_{\nu(k)}, \Sigma_{\nu(k)})$ and that the mixing proportions of the associated epistatic terms in the second expression are the same as those in the first expression after this permutation. The above definition is the standard notation of weak identifiability of the mixtures but restricted to the family of the epistatic mixture models. See McLachlan and Peel (2000). It follows from the theory of finite Gaussian mixtures that the Gaussian epistatic mixtures are identifiable under certain conditions, for instance, when $\mu_d, d \in \Omega_g$ are different from each other.

Let ψ_m denote $(\mu_0, \Sigma_0, \underline{\mu}, \underline{\Sigma}, w(d), d \in \Omega_g)$. For some positive constants δ_i, $i = 1, 2$ and b, let

$$\Psi_m = \left\{ \psi_m : w(d) \ge 0, \sum_{d\in\Omega(g)} w(d) = 1, ||\mu_k|| \le b, \right.$$
$$\left. \Sigma_k \in \mathcal{Q}_p, 1 \le k \le m \right\}, \quad \mathcal{G}_m = \{f_\psi : \psi \in \Psi_m\}.$$

For the theory that follows, we refer to the identifiability of an epistatic mixture model of order m_0 and the underlying density f_0 belongs to $\mathcal{G}_{m_0}$ as Condition (C1). Note that for $m_1 < m_2$ and any $f_{\psi_{m_1}} \in \mathcal{G}_{m_1}$, we can rewrite $f_{\psi_{m_1}}$ as the form $f_{\psi_{m_2}}$ with $\psi_{m_2} \in \Psi_{m_2}$ by simply adding some mixture components with zero mixing proportions to $f_{\psi_{m_1}}$. For each m, we consider the constrained estimator $\hat{\psi}_m$ over $\psi_m \in \Psi_m$. Accordingly, we define the minimum BIC estimator $\hat{m}$. Let $H(f, f_0) = \sqrt{\int (\sqrt{f} - \sqrt{f_0})^2 dx}$ be the Hellinger distance between f and f_0. Then, under Condition (C1), we obtain that $\hat{m}$ is asymptotically not less than m_0 in the following proposition.

Proposition 4.1. Suppose that Condition (C1) holds. Then, as n tends to infinity and p is fixed, $P(\hat{m} \ge m_0) \to 1$ and for $m \ge m_0$, $H(f_{\hat{\psi}_m}, f_0) = o(1)$ almost surely.

Remark 4.1. For $m \ge m_0$, we rewrite f_0 in the form $f_{\psi_m^*}$ with $\psi_m^* \in \Psi_m$ by adding the mixture components with zero proportions. Then, $H(f_{\hat{\psi}_m}, f_{\psi_m^*}) = H(f_{\hat{\psi}_m}, f_0) = o(1)$ almost surely. This implies that $\hat{\psi}_m \to \psi_m^*$ almost surely, up to a permutation on the order of the components in ψ_m^*. Finally, we note that the similar results hold for the Gaussian mixture models.

We now show that the probability of $\hat{m}$ being larger than m_0 tends to zero. For this purpose, we reparameterize $f \in \mathcal{G}_m, m \geq m_0$ as follows. For $m \geq m_0$, define $\mathcal{G}_m = \{f : f \text{ is of order } m, \psi_m \in \Psi_m\}$. Let q be the dimension of $(\gamma^{(0)}, v, \eta)$, where the dependence of q on m is suppressed, $(\gamma^{(0)}, v, \eta)$ takes values in a compact subset Θ_q of $\mathbf{R}^q$, and $f_0 = f_{\gamma^{(0)},0,\eta}$. All these f's form the set $\{f_{\gamma^{(0)},v,\eta} : (\gamma^{(0)}, v, \eta) \in \Theta_q\}$, which is still denoted by $\mathcal{G}_m$. See Appendix B for more details. For any density f of order $m \geq m_0$, define $s_f = (\sqrt{f/f_0} - 1)/H(f, f_0)$. Let $\mathcal{S}_m = \{s_f : f \in \mathcal{G}_m\}$, and $\mathcal{S} = \cup_{m=m_0}^{M} \mathcal{S}_m$. Then the above reparametrization allows us to calculate the limit points of s_f as $H(f, f_0)$ tends to zero as follows. If we denote by $R(\gamma^{(0)}, v, \eta)$ the partial derivative of $\sqrt{f/f_0} - 1$ with respect to v, then we have

$$s_f = \frac{v^T \int_0^1 R\big(\gamma^{(0)}, tv, \eta\big)dt}{\big|\big|v^T \int_0^1 R\big(\gamma^{(0)}, tv, \eta\big)dt\big|\big|_2} = \frac{v^{*T} \int_0^1 R\big(\gamma^{(0)}, tv, \eta\big)dt}{\big|\big|v^{*T} \int_0^1 R\big(\gamma^{(0)}, tv, \eta\big)dt\big|\big|_2},$$

where $v^* = v/||v||$, $||v^{*T} \int_0^1 R(\gamma^{(0)}, tv, \eta)dt||_2 = \sqrt{E_{f_0}[v^{*T} \int_0^1 R(\gamma^{(0)}, tv, \eta)dt]^2}$, E_{f_0} denotes the expectation operator with respect to the density f_0 and the dependence of $R(\gamma^{(0)}, tv, \eta)$ on $\mathbf{x}$ is suppressed. Letting $||v|| \to 0$, we are able to obtain the limit points of s_f.

To estimate the so-called covering entropy of set $\mathcal{G}_m$, we let

$$C_1(\mathbf{x}) = \sup_{(\gamma^{(0)},v,\eta)\in\Theta_q} \left| \int_0^1 R(\gamma, tv, \eta)dt \right|,$$

$$C_2(\mathbf{x}) = \inf_{(\gamma^{(0)},v,\eta)\in\Theta_q} \min \left\{ \left|\left| \int_0^1 \frac{\partial R(\gamma, tv, \eta)}{\partial \gamma^{(0)}} dt \right|\right|, \left|\left| \int_0^1 \frac{\partial R(\gamma, tv, \eta)}{\partial v^T} dt \right|\right|, \left|\left| \int_0^1 \frac{\partial R(\gamma, tv, \eta)}{\partial \eta^T} dt \right|\right| \right\},$$

and

$$C(\mathbf{x}) = \max\{C_1(\mathbf{x}), C_2(\mathbf{x})\}.$$

We impose the following regularity condition on the likelihood, of which the first part implies that the coordinates in the function vector $\int_0^1 R(\gamma^{(0)}, tv, \eta)dt$ are linearly independent while the second part is a uniformly integrable condition for the first two derivatives of the likelihood:

(C2) The eigenvalues of the matrix $E_{f_0}[\int_0^1 R(\gamma^{(0)}, tv, \eta)dt \int_0^1 R^T(\gamma^{(0)}, tv, \eta)dt]$ are uniformly bounded away from zero over $(\gamma^{(0)}, v, \eta) \in \Theta_q$ for a large constant M and $m_0 \leq m \leq M$. And $E_{f_0}[C(\mathbf{x})^2] < \infty$.

The above condition would hold in most practical situations. If we reparameterize f_{ψ_t} by $f_{\gamma^{(0)},vt,\eta}$, then by the definition and the mean value theorem, there exists $0 \leq t^* = t^*(\mathbf{x}) \leq 1$ such that

$$\begin{aligned} &E_{f_0}\left[\int_0^1 R\big(\gamma^{(0)}, tv, \eta\big)dt \int_0^1 R^T\big(\gamma^{(0)}, tv, \eta\big)dt\right] \\ &= \frac{1}{4}\int \left[\int_0^1 \frac{\partial \log(f_{\psi_t}(\mathbf{x}))}{\partial v}\sqrt{f_{\psi_t}(\mathbf{x})}dt \right. \\ &\quad \left. \times \int_0^1 \frac{\partial \log(f_{\psi_t}(\mathbf{x}))}{\partial v^T}\sqrt{f_{\psi_t}(\mathbf{x})}dt\right] d\mathbf{x} \\ &= \frac{1}{4}\int \left[\frac{\partial \log(f_{\psi_{t^*}}(\mathbf{x}))}{\partial v}\frac{\partial \log(f_{\psi_{t^*}}(\mathbf{x}))}{\partial v^T}\right] f_{\psi_{t^*}}(\mathbf{x})d\mathbf{x}, \end{aligned}$$

which is a generalized Fisher-type information matrix. So Condition (C2) requires the generalized Fisher-type information matrix to be nondegenerate over a compact parameter space. Note that, for any $k > 0$ and the standard normal density $\phi(x)$, $\sup_x |x^k \phi(x)| < +\infty$ holds. By using this property, we can show that $E_{f_0}[C(\mathbf{x})^2] < \infty$ holds for a number of cases.

The following theorem extends the result of the order consistency of the Gaussian mixtures (Gassiat 2002) to the epistatic mixtures.

Theorem 1. Suppose that Conditions (C1) and (C2) hold. Then, as n tends to infinity and p is fixed, the BIC order estimator $\hat{m}$ is consistent with the true value m_0 in probability. Moreover, $P(\hat{m} = m_0) \to 1$ and the constrained maximum likelihood estimator $\hat{\psi}_{m_0}$ converges to ψ_{m_0} in probability, up to a permutation on the order of the m_0 mixture components.

4.2 Asymptotic Distribution of the Likelihood Ratio Test

In this section, we develop an asymptotic theory for testing the competing hypotheses between the ten Gaussian mixture models (VVV, EEE, EEI, VEI, VII, EII, EVI, VVI, EEV, and VEV) and the six epistatic mixture models (EVE, EEE, EED, EVD, EVI, and EVA).

We begin with the pairwise BIC test. Let $\mathrm{M}_0 \in \mathcal{M}_0$, one of the Gaussian mixture models and $\mathrm{M}_1 \in \mathcal{M}_1$, one of the epistatic mixture models, where notations $\mathcal{M}_0$ and $\mathcal{M}_1$ are defined in Section 3. Let $\mathcal{F}_0$ and $\mathcal{F}_1$ be the parametric density families defined by the models M_0 and M_1 respectively. We can employ the pairwise BIC statistic $T_n = \mathrm{BIC}(\mathrm{M}_1) - \mathrm{BIC}(\mathrm{M}_0)$ to test the null hypothesis H_0: $f \in \mathcal{F}_0$ versus the alternative hypothesis H_1: $f \in \mathcal{F}_1 \setminus \mathcal{F}_0$. The p-value for the observed value T_{n0} is defined as $P(T_n < T_{n0})$ under the null hypothesis H_0. The p-value can be calculated via the distribution of the maximum log-likelihood ratio statistic $\lambda_n = 2(\sup_{f\in\mathcal{F}_1} l_n(f) - \sup_{f\in\mathcal{F}_0} l_n(f))$. Let m_0 be the order of the underlying density f_0 and ψ_{m_0} be its parameter before the reparameterization. Let λ_{n0} be the observed value of λ_n and $H_{n\psi_{m_0}}(\cdot)$ be the cumulative distribution of λ_n. Note that the BIC and the likelihood ratio statistics differ by the same constant in the general and observed cases when the model order is consistently estimated. Then, the p-value admits

$$\begin{aligned} &P(T_n < T_{n0}|H_0, T_{n0}) \\ &\quad = P(\lambda_{n_0} < \lambda_n|H_0, T_{n0}) = 1 - H_{n\psi_{m_0}}(\lambda_{n0}). \end{aligned}$$

Therefore, the above p-value can be estimated by replacing $H_{n\psi_{m_0}}(\cdot)$ by its parametric bootstrap estimator $H_{n\hat{\psi}_{m_0}}(\cdot)$, where $\hat{\psi}_{m_0}$ is the maximum likelihood estimator under the null model.

In the following, we investigate the asymptotic behavior of $H_{n\psi_{m_0}}(\cdot)$ and $H_{n\hat{\psi}_{m_0}}(\cdot)$. We first consider the case where $\mathcal{F}_0$ is not a subset of $\mathcal{F}_1$, $f_0 \in \mathcal{F}_0$ and $\sup_{f\in\mathcal{F}_1} E_{f_0}[\log(f/f_0)] < 0$ (i.e., $f_0 \notin \mathcal{F}_1$). We have the following proposition.

Proposition 4.2. Suppose that Condition (C1) holds and that $f_0 \in \mathcal{F}_0$ and $f_0 \notin \mathcal{F}_1$. Then, under the null hypothesis H_0, for any $\lambda > -\infty$, both $H_{n\psi_{m_0}}(\lambda)$ and $H_{n\hat{\psi}_{m_0}}(\lambda)$ tend to 1 as n tends to infinity and p is fixed.

Remark 4.2. Similar to Proposition 4.1, we can show that when the alternative hypothesis is true, λ_n (thus the observed λ_{n0}) is bounded from $-\infty$ below in probability. Therefore, the p-value $1 - H_{n\psi_{m_0}}(\lambda_{n0})$ will tend to 1 in probability as n tends to infinity and p is fixed.

We now turn to the case where $\mathcal{F}_0$ is a subset of $\mathcal{F}_1$ and $f_0 \in \mathcal{F}_0$. It follows from Theorem 1 that under $\mathcal{M}_0$ or $\mathcal{M}_1$, $P(\hat{m} = m_0) \approx 1$ when n is large. So, $P(\lambda_n < \lambda) \approx P(\lambda_n < \lambda, \hat{m} = m_0)$, indicating that the asymptotic behavior of λ_n depends only on the densities in $\mathcal{G}_{m_0}$. Therefore, for simplicity, we assume that $\mathcal{F}_1 = \mathcal{G}_{m_0}$ and $\mathcal{F}_0 = \{f \in \mathcal{F}_1 : f$ has zero mixing proportions for epistatic terms$\}$. Then the above testing problem reduces to testing whether the epistatic terms are significant or not.

As pointed out in the proof of Theorem 1, the distributional behavior of λ_n asymptotically depends only on the limits of s_f's when $H(f, f_0) \to 0$. To study the entropy property of these limits, we can reparameterize f_0 and $f \in \mathcal{F}_1$ as $f_{\gamma^{(0)},0}$ and $f_{\gamma^{(0)},v}$ for a given partition $(J_k)_{1\le k\le m_0}$ of $\{1, 2, \ldots, m_0\}$, where $v^* = v/||v||$ and the total increase in the mixing proportions of f over f_0 defined as sum(v^*) is zero. The details are deferred to Appendix B. Under Condition $(C1)$, we can choose partition $(J_k)_{1\le k\le m_0}$ such that the maximum likelihood estimator of v tends to zero. Let $R(\gamma^{(0)}, v)$ denote the partial derivative of $\sqrt{f/f_0} - 1$ with respect to v. Then, we are able to identify the following sets of the limits of s_fs for $f \in \mathcal{F}_1$ and $f \in \mathcal{F}_0$, respectively:

$$\mathcal{S}_{0+}^{(1)} = \left\{ \frac{v^{*T} R\left(\gamma^{(0)}, 0\right)}{\left|\left|v^{*T} R\left(\gamma^{(0)}, 0\right)\right|\right|_2} : ||v^*|| = 1, \text{sum}(v^*) = 0 \right\},$$

$$\mathcal{S}_{0+}^{(0)} = \left\{ s \in \mathcal{S}_{0+}^{(1)} : \alpha_{k_1 j_1 k_2 j_2} = 0, j_1 \in J_{k_1}, j_2 \in J_{k_2}, k_1 < k_2 \le m_0 \right\},$$

where $\alpha_{k_1 j_1 k_2 j_2}$ is defined in Appendix B. Let q' be the dimension of $(\gamma^{(0)}, v)$ and $\Theta_{q'}$ a compact subset of $R^{q'}$. Accordingly, we modify the definitions of $C_1(\mathbf{x})$, $C_2(\mathbf{x})$ and $C(\mathbf{x})$ by

$$C_1(\mathbf{x}) = \sup_{(\gamma^{(0)},v)\in\Theta_{q'}} \left| \int_0^1 R\left(\gamma^{(0)}, tv\right) dt \right|,$$

$$C_2(\mathbf{x}) = \inf_{(\gamma^{(0)},v)\in\Theta_{q'}} \min \left\{ \left|\left| \int_0^1 \frac{\partial R\left(\gamma^{(0)}, tv\right)}{\partial \gamma^{(0)}} dt \right|\right|, \left|\left| \int_0^1 \frac{\partial R\left(\gamma^{(0)}, tv\right)}{\partial v^T} dt \right|\right| \right\},$$

and

$$C(\mathbf{x}) = \max\{C_1(\mathbf{x}), C_2(\mathbf{x})\}.$$

Assume the following regularity condition—

(C2′): The eigenvalues of the matrix $E_{f_0}[\int_0^1 R(\gamma^{(0)}, tv)dt \int_0^1 R^T(\gamma^{(0)}, tv)dt]$ are uniformly bounded away from zero over $v \in \Theta_{q'}$. And $E_{f_0}[C(\mathbf{x})^2] < \infty$.

Under Conditions (C1) and (C2′), the following theorem gives an asymptotic distribution of λ_n.

Theorem 2. If Conditions (C1) and (C2′) hold, then under the null hypothesis H_0, as n tends to infinity and p is fixed, λ_n converges weakly to $\sup_{s\in\mathcal{S}_{0+}^{(1)}} \max\{W(s), 0\} - \sup_{s\in\mathcal{S}_{0+}^{(0)}} \max\{W(s), 0\}$, where $W(.)$ is a Gaussian process defined on $\mathcal{S}_{0+}^{(1)}$ with covariance $\Sigma(s_1, s_2) = E_{f_0}[s_1 s_2], s_1, s_2 \in \mathcal{S}_{0+}^{(1)}$.

The above theorem says that under the weak identifiability of the epistatic mixture model and the boundedness constraints on the generalized Fisher-information matrix, λ_n converges weakly to a functional of a Gaussian process. Since the above asymptotic distribution is not easy to calculate, we approximate the distribution of λ_n, $H_{n\psi_{m_0}}(\cdot)$ by the parametric bootstrap estimator $H_{n\hat{\psi}_{m_0}}(\cdot)$ suggested in Section 3.4. A basic question required to address is whether $||H_{n\hat{\psi}_{m_0}} - H_{n\psi_{m_0}}|| \hat{=} \sup_{z\in\mathbf{R}^1} |H_{n\hat{\psi}_{m_0}}(z) - H_{n\psi_{m_0}}(z)| \to 0$ in probability under the null hypothesis. The following theorem gives a positive answer.

Theorem 3. If Conditions (C1) and (C2′) hold, then under the null hypothesis H_0, as n tends to infinity and p is fixed, $||H_{n\hat{\psi}_{m_0}} - H_{n\psi_{m_0}}|| \to 0$ in probability.

Finally, we present a theory on testing a group of Gaussian models, namely H_0: Gaussian mixture models against H_1: epistatic mixture models. Let M_{0t}, $1 \le t \le 10$ denote Gaussian mixture models VVV, EEE, EEI, VEI, VII,EII, EVI, VVI, EEV, and VEV, respectively. Similarly, let M_{1t}, $1 \le t \le 6$ stand for epistatic mixture models EVE, EEE, EED, EVD, EVI, and EVA. We consider the following minimum BIC test statistic

$$\mathrm{BIC}_n = \min_{1\le t\le 6} \mathrm{BIC}(\mathrm{M}_{1t}) - \min_{1\le t\le 10} \mathrm{BIC}(\mathrm{M}_{0t}).$$

Note that for $1 \le t \le 5$, M_{0t} is a submodel of M_{1t}. Using this fact, we show that the problem of the group model testing is asymptotically equivalent to a problem of pairwise model testing as follows.

Proposition 4.3. Suppose that Conditions (C1) and (C2′) hold. Then,

$$\mathrm{BIC}_n = \mathrm{BIC}(\mathrm{M}_{1t}) - \mathrm{BIC}(\mathrm{M}_{0t}) + o_p(1),$$

if model M_{0t} with $1 \le t \le 5$ is true. Furthermore, if model M_{0t} with $6 \le t \le 10$ is true, then BIC_n converges to $-\infty$ in probability as n tends to infinity and p is fixed. The similar result holds for the corresponding bootstrap estimator.

Note that due to computational feasibility, we haven't considered the epistatic models associated with GM models EII, EVI, VVI, EEV, and VEV. However, the above proposition holds for any two subsets of GM and EP model classes.

5. NUMERICAL RESULTS

In this section, we tested the proposed method (named EP), the GM and the KM on a real dataset and a number of simulated datasets.

5.1 Real Data Analysis

We assessed our proposed method on a publicly available dataset arising from genetic studies. Using the dataset, we showed that the proposed EP had significant better fit than the GM by the BIC test.

Yeast stress dataset. Gasch et al. (2000) used DNA microarrays to explore genome-wide expression patterns in the yeast *Saccharomyces cerevisiae* in response to diverse environmental changes. This dataset contains log-expression levels of 6152 yeast genes. For simplicity of presentation, we considered the expressions of these genes under heat shock from 25C to 37C (15 time points). We carried out our analysis in two stages. Note that the usual purpose of large-scale microarray studies is to reduce a vast set of possibilities to a much smaller set of differentially expressed genes. It is not wise to fit an epistatic mixture

model to the entire dataset because the potential accumulation of noise can deteriorate the quality of the fit and because the computation is very expensive. So, in the first stage we prescreened the data using the following empirical procedure, based on gene variability. Here we removed noisy analysis units rather than noisy attributes from the clustering model. First, to screen the genes, we calculated a sample variance over the conditions (or time-points) for each gene. We selected these variances within three-folds of the minimum sample variance. We calculated the average of these selected variances, $\hat{\sigma}_0^2$, and used it to estimate the background noise level σ_0^2. Note that it is reasonable to benchmark the background noise variance off the minimum sample variance because the expression levels have already been rescaled by microarray normalization. For each gene, we calculated the sample variability index which is the ratio between the sample variance and the estimated background noise level. In the online supplementary materials, we showed that under certain conditions (attributes are iid or satisfy some mixing conditions), $\hat{\sigma}_0^2/\sigma_0^2 \in (1/3, 3)$ and $\hat{\sigma}_i^2/\sigma_i^2 \in (1/3, 3)$ for the ith gene with high probabilities. This suggests that for most of the genes, the sample variability index being larger than 9 will imply that the underlying signal-noise-ratio $\sigma_i^2/\sigma_0^2 > 1$. Therefore, to obtain genes of higher variability, we thresholded the gene sample variability indices by 9, leading to 496 genes left. The multiple pairwise plots of the resulting dataset displayed in Figure 1 reveal a visible pattern of the pairwise column correlations, which corresponding to the relationships among the 15 experimental conditions for the selected genes: Conditions 1–15 can be divided into five blocks, composed of Conditions 1–8, 9–11, 12, 13–14, and 15, respectively, where big increases in heat happened at time points 1 and 13 giving shocks. Within block 1, the conditions are positively correlated each other; block 1 is negatively correlated to block 2, weakly correlated to block 3, positively correlated to block 4, and negatively correlated to block 5. Within block 2, the conditions are positively correlated each other; block 2 is weakly correlated to block 3, negatively correlated to block 4, and positively correlated to block 5. Within block 4 the conditions are positively correlated each other; block 4 is negatively correlated to block 5. These facts suggest a general sine pattern of the cell reaction to the shock, where the gene expressions are increasing under block 1, decreasing under block 2, increasing under block 4, and decreasing again under block 5. This was consistent with what had happened with the yeast cell: When the heat shock was introduced, the cell increasingly reacted from the low level to the peak to change its expression programs as demonstrated in the first time block. Once the cell got used to the current level of shock, the reaction would decrease as demonstrated in the second time block, followed by a short stable period in the third time block. However, when the shock reached another level, the cell would react again to adopt the significant change of the shock, followed a reaction decreasing period.

In the second stage we grouped these selected genes by applying the GM, the KM, and the EP (with $g = 1, 2, 3$) to the above data. The best GM fit gives three clusters of sizes 190, 106, and 200, respectively, while the KM fit yields two clusters of sizes 263 and 233, respectively. The best GM fit is a VEV model with a BIC value of 13160.99. However, the best EP fit produces a miscellaneous cluster, four primary clusters, and three epistatic clusters (generated by interactions between the primary clusters 1 and 2, 1 and 3, and 2 and 4), which are of sizes 8, 149, 63, 73, 84, 95, 21, and 3, respectively. Among six epistatic models, the best EP fit is an EVA model with the BIC value of 12,534.79, improving the best GM fit by a reduction of 626.2 and the best EED (i.e, the overlapping process model, Battle, Segal, and Koller 2005) fit by a reduction of 3055.47. To test the significance of this reduction, from the best GM fit we generated 200 (parametric) bootstrap samples of size 496 each. We fitted the EVA and VEV models to each bootstrap sample and calculated the BIC difference between two methods. We then calculated the proportion of these values that were less than what we observed (i.e., -626.2), obtaining a bootstrap p-value of 0 which is against the GM model VEV. This is consistent with what you would get asymptotically if the model you selected was the true one (i.e., the asymptotic p-value is 0 if the EVA is true as the EVA and VEV models are different from each other). See Section 4.2. Regarding the KM, if we view the KM as a particular sort of Gaussian mixture model, it should be the GM (EEI) model. For the yeast stress data, the best GM fit (i.e., VEV) was found much better than the GM(EEI) in terms of the BIC. As the best EP fit is better than the best GM fit, the former is thus better than the GM(EEI) fit in terms of the BIC. We also arbitrarily resampled 100 datasets of size 496 from the best GM fit and the best EP fit, respectively, and made the multiple pair plots of these samples. The typical patterns are displayed in Figures 2 and 3, respectively. Interestingly, Figures 1–3 reveal that the EP sample has recovered the visible pattern of the original sample, while the GM sample has not. This again supports that the EP model is more appropriate than the GM in fitting the yeast stress data. We also notice that the results derived from the GM, KM, and EP methods are largely different in general, which can be seen from the low adjusted Rand indices between them: 0.39 between the GM clusters and the EP clusters; 0.26 between the KM and the EP; 0.57 between the GM and the KM. To assess the stability of the above results, we randomly drew 21 subsets of the above 496 genes and conducted the EP, the GM and the KM analyses on these subsets. We found that unlike the GM, the optimal numbers of clusters derived from the EP and the KM fits are much more stable than the GM and that the EP outperformed the GM in all these subsamples. In the KM setting, there is only one model so the misspecification doesn't cause things to move around as much as in the GM setting. See the online supplementary materials for the further details.

To evaluate whether these results are biologically significant, we tested whether the genes associated with each cluster show any enrichment for known annotations, using the yeast Gene Ontology (GO) database which includes 7167 background genes. For each cluster and each annotation in GO, we counted the number of genes in this cluster with that annotation. If a cluster indeed corresponds to known biological processes and functions, then we would expect the cluster under investigation to contain a high fraction of the genes with the corresponding annotation compared to the background gene distribution. The statistical significance (p-value) was calculated by use of the hypergeometric distribution. See Boyle et al. (2004) for the details on how it was done. We applied this evaluation procedure to the above clusters via the GO term finder at *http://www.yeastgenome.org/* on 25/04/2011. The results are summarized as follows.

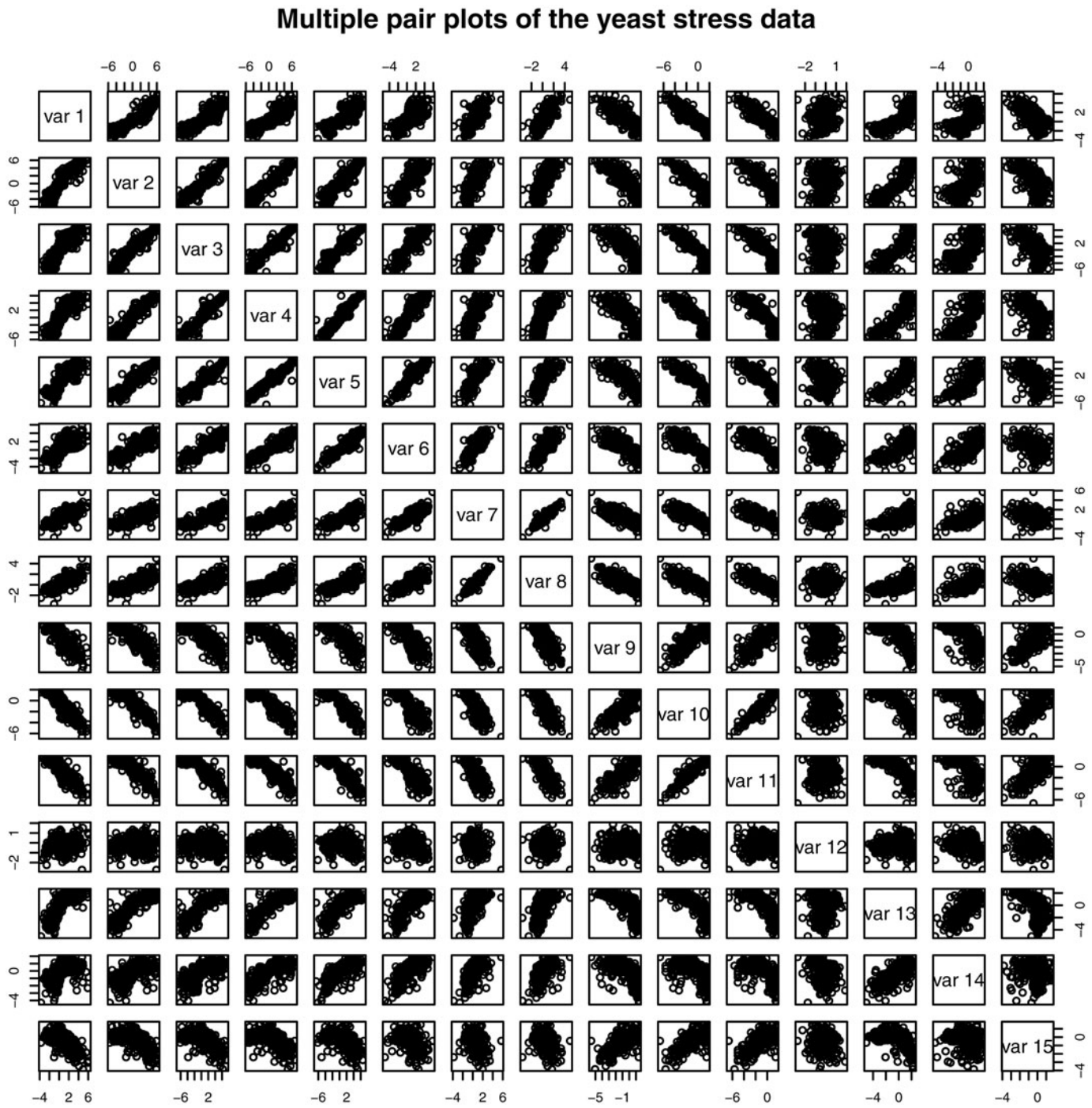

Figure 1. Multiple pair plots of the expression data of the 496 selected yeast genes under 15 heat shock conditions, revealing a visible pattern that the pairwise correlations of these genes are positive under conditions 1–8, negative under conditions 9–11, positive again under conditions 13–14, and negative under condition 15.

EP clusters: As expected, the miscellaneous cluster is not enriched for any GO terms. Primary cluster 1 is enriched for the following GO terms: small molecule metabolic process (p-value: 1.3×10^{-8}), carboxylic acid metabolic process (p-value: 8.64×10^{-6}), oxidation–reduction process (p-value: 1.2×10^{-4}), and alcohol metabolic process (p-value: 1.1×10^{-3}). Primary cluster 2 is enriched for the GO terms: energy reserve metabolic process (p-value: 2.2×10^{-6}) and glucan metabolic process (p-value: 3.7×10^{-4}), which are subprocesses of carbohydrate metabolic process. The epistatic cluster between clusters 1 and 2 is enriched for the GO terms: carbohydrate metabolic process (p-value: 2.1×10^{-6}), alcohol metabolic process (p-value: 6.8×10^{-5}), and glucose metabolic process (p-value: 2.3×10^{-4}). Primary cluster 3 is enriched for the GO terms: translation (p-value: 3.28×10^{-40}), protein metabolic process (p-value: 9.7×10^{-23}), macromolecule metabolic process (p-value: 3.3×10^{-11}), gene expression (p-value: 7.7×10^{-21}), and ribosome biogenesis (p-value: 1.1×10^{-9}). The epistatic

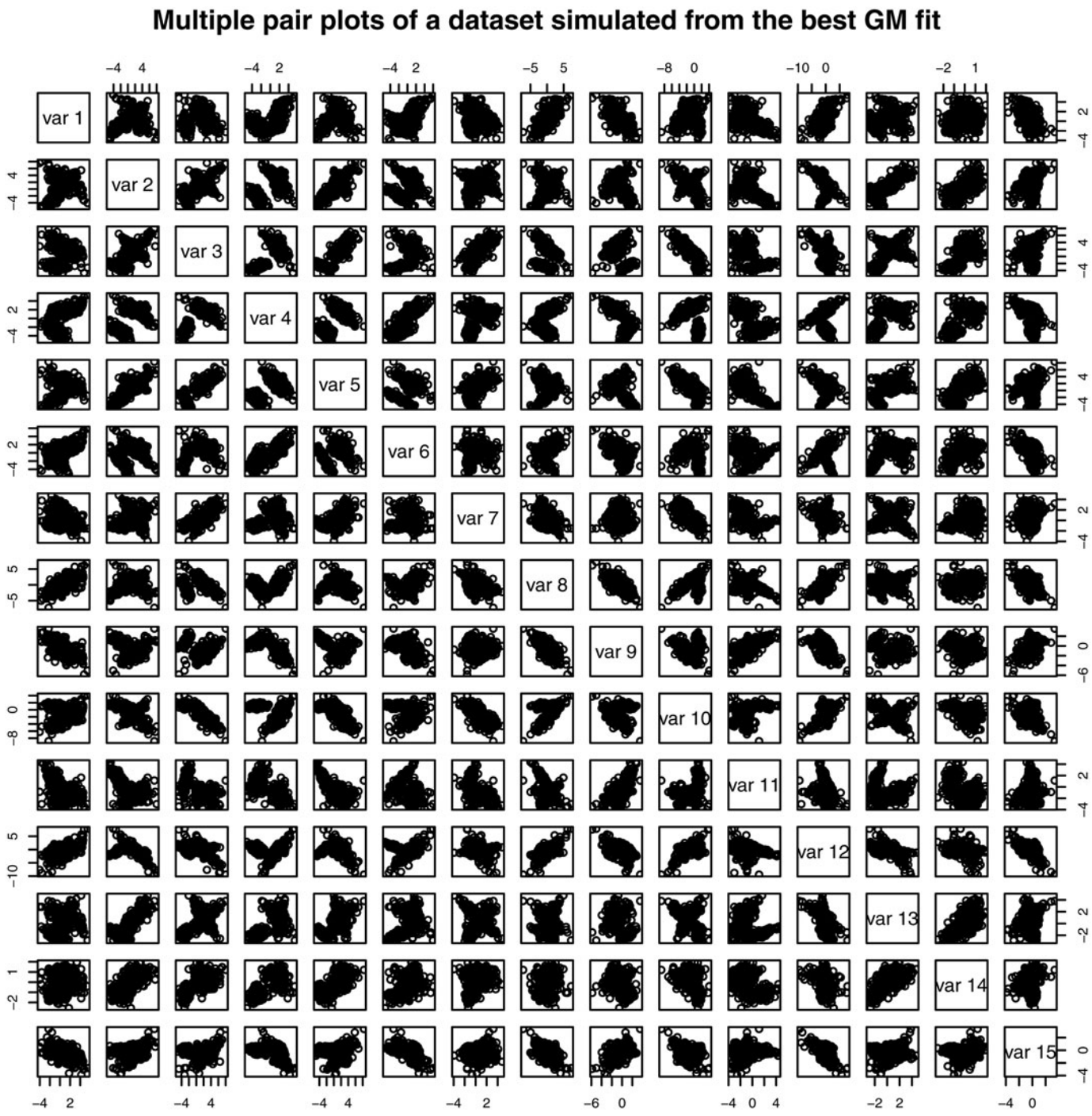

Figure 2. Multiple pair plots of a bootstrap sample from the best GM fit of the expression data of the 496 selected yeast genes under 15 heat shock conditions shows a completely different pattern from that in the original yeast stress data.

cluster between clusters 1 and 3 is enriched for the GO terms: macromolecule metabolic process (p-value: 9.2×10^{-4}), cellular protein metabolic process (p-value: 4.9×10^{-7}), and translation (p-value: 1.8×10^{-12}). Primary cluster 4 is enriched for the GO terms: ribosome biogenesis (p-value: 3.7×10^{-43}), RNA metabolic process (p-value: 1.7×10^{-13}) and gene expression (p-value: 2.3×10^{-9}). The epistatic cluster between clusters 2 and 4 is enriched for the GO term: ribosomal small subunit biogenesis (p-value: 7×10^{-3}).

GM clusters: Cluster 1 is enriched for the GO terms: small molecule metabolic process (p-value: 1.2×10^{-9}) and carboxylic acid metabolic process (p-value: 6.6×10^{-6}). Cluster 2 is enriched for the GO terms: oxidation-reduction process (p-value: 9.6×10^{-5}) and carbohydrate metabolic process (p-value: 5.6×10^{-4}). Cluster 3 is enriched for the GO terms: ribosome biogenesis (p-value: 1.5×10^{-66}), rRNA metabolic process (p-value: 6.8×10^{-39}), and gene expression (p-value: 2.7×10^{-36}).

KM clusters: Cluster 1 is enriched for the GO terms: ribosome biogenesis (p-value: 1.3×10^{-69}), rRNA metabolic process (p-value: 2.8×10^{-40}), gene expression (p-value: 3.2×10^{-30}), and translation (p-value: 5.8×10^{-22}).

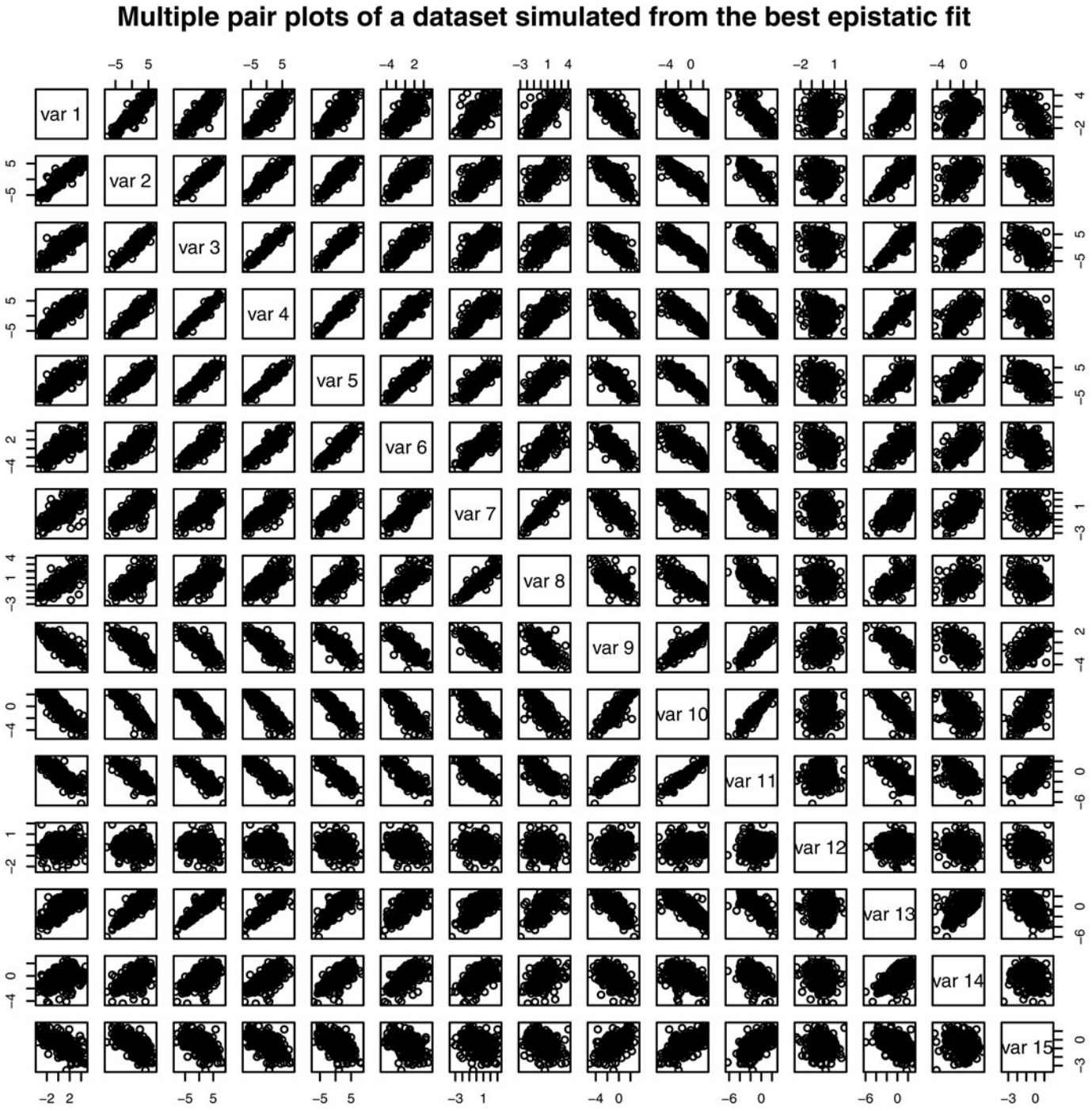

Figure 3. Multiple pair plots of a bootstrap sample from the best epistatic fit of the expression data of the 496 selected yeast genes under 15 heat shock conditions. It shows that the pattern, which has been seen in the original yeast stress data, is recovered.

Cluster 2 is enriched for the GO terms: oxidation-reduction process (p-value: 1.9×10^{-13}), carbohydrate metabolic process (p-value: 4.5×10^{-10}), and generation of precursor metabolites and energy (p-value: 5.5×10^{-11}).

By using the GO term finder, we found that all these clusters show a significant enrichment for known annotations even after multiple testing corrections in terms of false discovery rate (FDR). In fact, all the estimated FDR values are equal to zero. See Boyle et al. (2004) for the details on correction for multiple hypotheses through FDR. We also found that many genes have participated multiple biological processes. Compared to the GM and the KM, the EP has taken advantage of multiple functional roles of some genes. In fact, the GO term finder has indicated that among the 496 selected genes, there are a number of them that have played multiple roles in the cell. We should assign these genes to multiple clusters that are enriched for the corresponding GO terms. The EP achieved this objective by assigning these genes to the epistatic clusters. For example, gene PGM2/YMR105C is known to play an indirect catalyzing role in both carbohydrate metabolic process and alcohol metabolic process. The carbohydrate metabolic process denotes the various biochemical reactions and pathways responsible for the

formation, breakdown, and interconversion of carbohydrates in living organisms, while the alcohol metabolic process denotes the chemical reactions and pathways for metabolizing alcohol. The two processes are interacting as the latter has a detrimental effect on the former, for instance, in maintaining healthy blood sugar levels. The EP assigned PGM2/YMR105C to the epistatic cluster between primary clusters 1 and 2, suggesting that this gene is linked to both primary clusters 1 and 2. In contrast, the GM and the KM were unable to do it, because the GM- (KM-) clusters are mutually exclusive: the GM put PGM2/YMR105C only in the GM-cluster 1 (alcohol metabolic process) and not in the GM-cluster 2 (carbohydrate metabolic process). The KM put PGM2/YMR105C only in the KM-cluster 2 (carbohydrate metabolic process). To conclude, the above gene enrichment analysis confirms that the EP can produce better clusters of biologically meaningful explanation than both the GM and the KM.

5.2 Simulation Studies

By using simulations, we answer the following questions: What is the impact of epistatic interactions on the performances of the GM, KM, and EP methods in terms of the RAND index? When will the EP outperform the other two methods in the presence of epistatic effects between clusters? How is the minimum BIC related to the RAND index?

We simulated a number of datasets that mimic different levels of epistatic interactions, pretending that we did not know their true grouping structures, where there were one miscellaneous and several overlapping clusters. When the epistatic interactions appear in the data, the GM model is misspecified. This makes the BIC-based estimation of the model order biased. The silhouette width-based determination of the number of clusters can also be biased in the presence of the overlapping clusters in the data. To assess the extent of such effects, we conducted the GM and KM analyses and calculated their oracle model order estimates at which their RAND indices with the true clusters attain the maximum values. For each method, the estimated error rate is defined as the average of the absolute differences between the estimated numbers of (overlapping) clusters and the number of underlying overlapping clusters over the replicates.

Synthetic yeast stress data. To reduce the computing burden, we considered only a subset of the yeast stress data, the log-expressions of the above 496 selected yeast genes under Conditions 1–8. The multiple pair plots in Figure 1 have already shown that the expression levels under these eight conditions are positively correlated with each other. Applying EP to this subset of the data, we obtained one miscellaneous cluster, and five primary clusters as well as their epistatic products. These primary clusters together with their epistatic products can be reformulated as $m_0 = 5$ overlapping clusters. The corresponding best EP fit has the estimated parameters: $\hat{\mathbf{w}} = (\hat{w}_j)_{1\le j\le 16} = (0.01, 0.047, 0.226, 0.202, 0.129, \quad 0.122, 0.011, 10^{-8}, 0.002, 0.008, 0.004, 0.114, \ 0.044, 0.016, 10^{-13}, 0.064)$, $\hat{\underline{\mu}}$, and $\hat{\underline{\Sigma}}$. Here $\hat{w}_j$ for $7 \le j \le 16$ are the mixing proportions of the epistatic components, of which two have nearly zero mixing proportions. We multiplied each of these 10 epistatic mixing proportions by a constant q and then renormalized the entire set of 16 mixing proportions so that they sum to one. This leads to a new set of mixing proportions $\hat{w}_j(q)$, $1 \le j \le 16$. When $q = 1$, these new mixing proportions reduce to the original ones. When $q = 0$, the generating distribution comes from a GM (VEV) model. Values of q greater than 1 would put more weights on the epistatic components. When q is increasing, epistatic effects are increasing. The identifiability is weakening as the sizes of the primary clusters are decreasing, where the EP is expected to be superior to the other two methods. For each of $q = 0, 1, 2$, we drew 30 datasets from the above modified EP model with the parameters: $\hat{w}_j(q)$, $1 \le j \le 16$, $\hat{\underline{\mu}}$, and $\hat{\underline{\Sigma}}$. Each dataset contains the log-expressions of 500 genes under Conditions 1–8.

We performed the EP-, the GM-, and the KM-based cluster analyses on these datasets, respectively. The degrees of agreement between the groupings derived and the underlying ones were then compared with each other in terms of the Rand index and also in terms of error rate of the model order estimation. These quantities, summarized in Table 1, ranked EP at the top, followed by KM and GM. In particular, the results indicate that

Table 1. Comparison of the EP, GM, and KM methods via ERRm and ave(ρ) over the 30 synthetic yeast stress datasets

	$q = 0$		$q = 1$		$q = 2$	
	ERRm	ave(ρ)	ERRm	ave(ρ)	ERRm	ave(ρ)
			$n = 250$			
EP	0.10 (0.056)	0.89 (0.001)	0.43 (0.092)	0.70 (0.002)	0.53 (0.09)	0.62 (0.003)
GM	2.57 (0.218)	0.62 (0.007)	4 (0)	0.14 (0.001)	4.03 (0.033)	0.11 (0.001)
Oracle GM	0.83 (0.097)	0.82 (0.002)	1.13 (0.178)	0.58 (0.004)	1.33 (0.194)	0.51 (0.004)
KM	4 (0)	0.45 (0.001)	4 (0)	0.26 (0.001)	4 (0)	0.20 (0.001)
Oracle KM	0.9 (0.161)	0.61 (0.002)	0.7 (0.085)	0.50 (0.002)	0.8 (0.074)	0.50 (0.001)
			$n = 500$			
EP	0.07 (0.046)	0.90 (0.001)	0.03 (0.033)	0.78 (0.001)	0.1 (0.056)	0.70 (0.001)
GM	1.03 (0.033)	0.89 (0.001)	3.57 (0.124)	0.23 (0.004)	3.93 (0.046)	0.12 (0.002)
Oracle GM	0.87 (0.063)	0.90 (0.001)	1.03 (0.155)	0.65 (0.001)	1.63 (0.182)	0.57 (0.001)
KM	4 (0)	0.45 (0.001)	4 (0)	0.26 (0.0004)	4 (0)	0.20 (0.0004)
Oracle KM	1.03 (0.101)	0.62 (0.001)	0.83 (0.136)	0.53 (0.002)	0.87 (0.063)	0.49 (0.001)

Here, ave(ρ) is the average of the adjusted Rand indices over the 30 synthetic datasets. And ERRm stands for the average of the error rates in estimating the numbers of clusters over the 30 synthetic datasets, which is defined in Section 5.2. In the table, the number in the parentheses is the standard error.

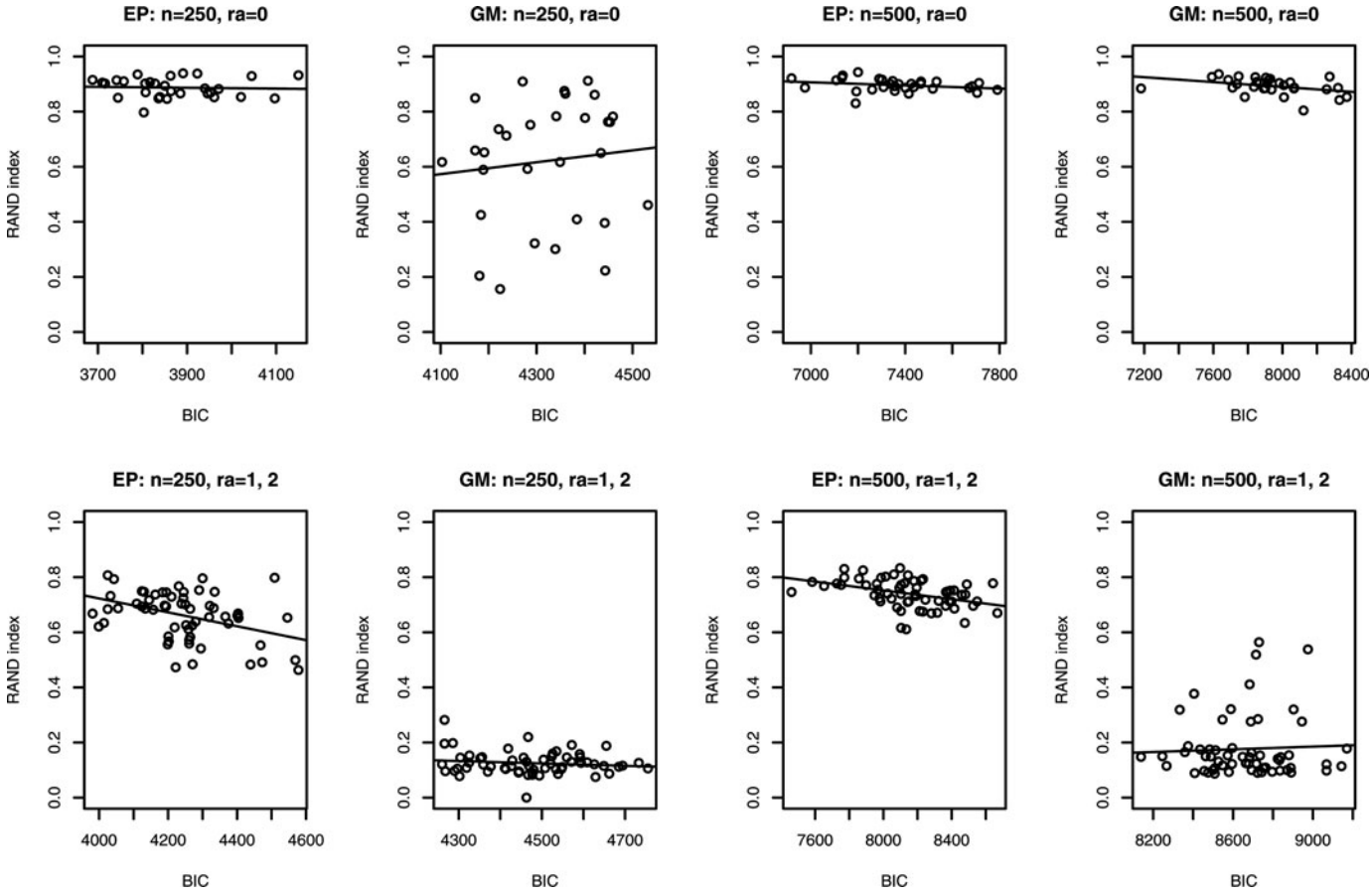

Figure 4. The scatterplots of the values of the RAND index against the minimum BIC values and their regression line fits for the synthetic yeast stress datasets. The subplots in the top rows are for the cases, where $n = 250$, $q = 0$, and $n = 500$, $q = 0$, respectively. The subplots in the bottom rows are for the cases, where $n = 250$, $q = 1, 2$ and $n = 500$, $q = 1, 2$, respectively.

compared to the EP and their Oracle versions, the GM and the KM performed very badly in the presence of epistatic effects. Even with the oracle order determination, these two methods were still inferior to the EP. In the unpublished simulation studies, instead of choosing the model order by the BIC or by the average silhouette width, we chose the orders of the GM and the KM via maximizing their RAND indices with the best EP fit. The resulting clusters turned out to have the RAND indices close to their oracle versions.

The best average values of the RAND index that the EP can attain in Table 1 are 0.90, 0.78, and 0.70, respectively, when $q = 0, 1$, and 2. Compared to the performance of the GM, it is not surprising, since Yeung et al. (2001) found that in many practical scenarios the best values of the RAND index that the GM can attain were between 0.70 and 0.85. The above fact has an implication to the real data analysis that there potentially were 10%, 22%, and 30% of mismatch between the estimated and the underlying true clusters, respectively, when $q = 0, 1$, and 2. Therefore, it is important to check the potential biological meaning of the resulting assignments of the genes by using the enrichment analysis.

In Figure 4, the values of the RAND index were plotted against the corresponding minimum BIC values for the best EP fits and the best GM fits over the 30 synthetic yeast stress data, respectively. It suggests that unlike the case of the GM, for the EP, the smaller BIC gives a bigger RAND index on average. This is not surprising because for the GM, the BIC can be biased when epistatic clusters are presented in the data. Figure 4 also suggests that the EP works well even when n is not too large. Note that the EP procedure was initialized by some GM solutions. We have not tried other strategies of initializations that can potentially further improve the performance of the EP method.

In the online supplementary materials, some more simulation results can be found on the relationship between the percentage of BIC reduction of EP over GM and the corresponding percentage of RAND index increases.

6. DISCUSSION

Epistatsis is defined generally as the interaction between different genes. In the past decade, scientists have made great progress in using this concept to construct gene networks (van Drissche et al. 2005). In the same spirit we have developed an epistatic approach for clustering gene expression data, exploring the hypothesis that gene clusters can be epistatically linked in forming subsets of biological modules. The new approach is based on our epistatic mixture model, where the epistatic mean effects are defined by the weighted summation of the activities of the involved processes and the noise covariances of the epistatic terms are determined by weighted matrix-harmonic mean of

the covariances of the involved processes. When all component covariances are the same, our model reduces to a linear factor model. We have provided a BIC test for the significance of the epistatic terms and developed the asymptotic theory for model comparison when the number of genes tends to infinity and the number of microarrays is fixed. The performance of the proposed method has been assessed by simulations and a real data analysis. We have also evaluated the biological significance of our results by use of a software of gene enrichment analysis. We have argued that if there are epistatic structures shared by some clusters, then fitting them via our models would be more efficient and more informative. It should be noted that the key idea behind our model is rather general. For example, we can apply it to modeling latent node-positions for complex networks in society, which have been shown to have the overlapping structures (Palla et al. 2005). Note also that the KM is very close to the Laplace-distribution based mixture model. Therefore, to fully compare the EP and the KM, it would be very interesting to extend the EP to the non-Gaussian mixture setting in future studies (Zhang and Liang 2010).

Conceptually, the present work is most closely related to the overlapping process model (Battle, Segal, and Koller 2005), which aims at describing gene activities by multiple biological processes. We have extended their work by removing the unrealistic assumptions that the effects of different biological processes are of the same noise level within an array, and that gene expression levels are independent of each other across arrays. Our model also bears interesting relationships to the partial membership model (Heller, Williamson, and Ghahramani 2008). In their model, these authors relaxed the constraint $r_{ik} \in \{0, 1\}$ to allow r_{ik} to take any continuous value in the range $[0, 1]$. They defined the epistatic term $f(\mathbf{x}_i|\mathbf{r}_i)$ as the normalized product of all component densities involved. In the Gaussian case, $f(\mathbf{x}_i) = \int \phi(\mathbf{x}_i|\mu_{\mathbf{r}_i}/|\mathbf{r}_i|, \Sigma_{\mathbf{r}_i}/|\mathbf{r}_i|)dP(\mathbf{r}_i)$, where $P(\mathbf{r}_i)$ is a distribution of $\mathbf{r}_i$, and the epistatic mean $\mu_{\mathbf{r}_i}/|\mathbf{r}_i| = \sum_{k=1}^{m} \Sigma_{\mathbf{r}_i}\Sigma_k^{-1}\mu_k r_{ik}/|\mathbf{r}_i|$ and the epistatic covariance $\Sigma_{\mathbf{r}_i}/|\mathbf{r}_i| = (\sum_{k=1}^{m} \Sigma_k^{-1} r_{ik})^{-1}$. They are the matrix-weighted average of the component means and the matrix-harmonic sum of the component covariances, respectively. In contrast, our model defines the epistatic term by a Gaussian factor model with mean $\mu_{\mathbf{r}_i}$ and covariance $\Sigma_{\mathbf{r}_i}$. We have assessed the performances of both procedures by applying them to the yeast stress data. We found that our epistatic approach may be more appropriate than theirs in modeling overlapping processes of gene regulation (the numerical details are omitted). Our model is different from the mixed-membership models such as the Latent Dirichlet Allocation (LDA) model and the admixture model (Pritchard, Stephens, and Donnelly 2000; Blei, Ng, and Jordan 2003) arising in the areas of information retrieval and population genetics. These models usually aim at capturing admixed population structures, whereas our model intends to detect interactions between biological processes of gene regularization. Unlike our model, these mixed-membership models are built on the assumption that attributes x_{ik}'s are independent given membership scores. This assumption may not be true for gene expression data, because the expressions of a gene under different experimental conditions can be correlated.

We have fitted the proposed model to the yeast stress data and shown it is a significant improvement over GM and KM. But in the chemostat-based transcriptome data of yeast, only a weak improvement has been found (see the online supplementary materials). To improve the fitness, we may consider more flexible models such as non-Gaussian mixture-based epistatic models. Exponential power mixtures may be useful for defining such models (e.g., Zhang and Liang 2010). It would be very interesting to compare the performance of Gaussian epistatic models with those of non-Gaussian epistatic models. It would be also interesting to investigate both the sensitivity/specificity of the GM (EP) in correctly clustering genes that participate in multiple processes.

There is a computational issue left for further study as follows. For moderate m or small g, we can estimate mixing proportions $w(d), d \in \Omega$. However, for large m or g, it can be very time-consuming to do that directly, because it involves estimating $\tau(g) = O(m^g)$ mixing proportions. To reduce this burden, we set $g = 2$ in our simulation studies. For a general g, we may approximate these proportions by a discrete random field over the space Ω_g, such that the probability

$$w_j = p(d_j|\mathbf{a}, \mathbf{b}) = \frac{\exp\left(\sum_{k=1}^{m} d_{jk}\left(a_k + b_k \sum_{k_1 \neq k} d_{jk_1}\right)\right)}{C(\mathbf{a}, \mathbf{b})}, \tag{6.1}$$

where $C(\mathbf{a}, \mathbf{b}) = \sum_{d \in \Omega_g} \exp(d^T\mathbf{a} + (|d| - 1)d^T\mathbf{b})$ with parameters $\mathbf{a} = (a_1, \ldots, a_m)^\tau \in (-\infty, \infty)^m$ and $\mathbf{b} = (b_1, \ldots, b_m)^T \in (-\infty, 0]^m$. For $2 \leq j \leq m + 1$, $w_j = \exp(a_{j-1})/C(\mathbf{a}, \mathbf{b})$ are the mixing proportions for the primary clusters. For $j \geq m + 2$, w_j depends on the mixing proportions of the primary clusters and is penalized by using the number of clusters that a data point will be assigned to. By this approximation, the number of mixing parameters reduces from $O(m^g)$ to $2m$. Under the above assumptions on $\mathbf{x}_i$ and $\mathbf{r}_i$, the marginal density of $\mathbf{x}_i$ can be formulated as another mixture model with components $f(\mathbf{x}_i|\mathbf{r}_i = d_j)$, $1 \leq j \leq \tau(g)$ and the mixing weights $w_j, 1 \leq j \leq \tau(g)$. The proposed random field model is flexible since it allows for various degrees of epistatic interactions between the clusters. For example, letting $\mathbf{b}$ in (6.1) tend to $-\infty$, we see that the limit of $p(\mathbf{r}_i|\mathbf{a}, \mathbf{b})$ only has positive masses on $m + 1$ configurations, that is,

$$\pi_0 = \frac{1}{1 + \sum_{j=1}^{m} \exp(a_j)},$$
$$\pi_k = p(e_k|\mathbf{a}, \mathbf{b}) = \frac{\exp(a_k)}{1 + \sum_{j=1}^{m} \exp(a_j)},$$

where e_k is a unit vector whose kth component is one. Therefore, any interaction between the clusters is excluded. On the other hand, if we let $\mathbf{b}$ tend to zero, then $p(\mathbf{r}_i|\mathbf{a}, \mathbf{b})$ may place positive mass on all configurations in $\Omega_{\tau(g)}$. When w_j's are approximated by a random field defined in (6.1), we need to replace w_j by $p(d_j|\mathbf{a}, \mathbf{b})$ in the expectation of the complete log-likelihood. To estimate unknown parameters $\mathbf{a}$ and $\mathbf{b}$, in the M-step, we need to maximize Ψ with respect to $\mathbf{a}$ and $\mathbf{b}$ rather than w_j's. We can also adaopt the hierarchical strategy by assigning a multivariate Normal and a log-Normal to $\mathbf{a}$ and $-\mathbf{b}$, respectively. The further study along this direction is beyond the scope of this article.

APPENDIX A: TECHNICAL DETAILS OF THE M-STEP

EVE case:

Letting $\Sigma_j = \Sigma_j^{(v)}$ be held fixed, we maximize Ψ with respect to $\underline{\mu}$ by solving the simultaneous equations $\frac{\partial \Psi}{\partial \underline{\mu}} = 0$. This gives rise to

$$\operatorname{vec}(\underline{\mu}^{(v+1)}) = \operatorname{diag}(\Sigma_1, \ldots, \Sigma_m) \left(\sum_{j=1}^{\tau(g)} \left(\sum_{i=1}^{n} \tau_{ij}^{(v)} \right) d_j d_j^T \otimes \Sigma_{d_j} \right)^{-1} \times (T^{(v)} D^T \otimes I_p)^T \operatorname{vec}(\mathbf{X}),$$

where $d_1 = \mathbf{0}$, $D = (d_1, \ldots, d_{\tau(g)})$, $T^{(v)} = (\tau_{ij}^{(v)})_{n \times \tau(g)}$, vec is an operator which stacks the column vectors of a matrix, I_p is a $p \times p$ unit matrix and $\otimes$ denotes the Kronecker product operator. Letting $\underline{\mu} = \underline{\mu}^{(v)}$ be held fixed, we calculate the partial derivative of Ψ with respect to Σ_t^{-1},

$$\frac{\partial \Psi}{\partial \Sigma_t^{-1}} = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=2}^{\tau(g)} \tau_{ij}^{(v)} \frac{d_{jt}}{|d_j|} \{ \Sigma_{d_j} - (x_i - |d_j|\mu_t)(x_i - |d_j|\mu_t)^T + |d_j|^2 (\mu_t - \mu_{d_j}/|d_j|)(\mu_t - \mu_{d_j}/|d_j|)^T \}.$$

Let $\Sigma_{jt}^{(v+1)} = \sum_{i=1}^{n} \tau_{ij}^{(v)} (x_i - |d_j|\mu_t)(x_i - |d_j|\mu_t)^T / (n w_j^{(v+1)})$, $2 \le j \le \tau(g)$. Observe that d_{t+1}, $1 \le t \le m$ all have the L_1 norms of 1, among which for each $1 \le t \le m$, only d_{t+1} has the value of one in its tth coordinate. Therefore, the above partial derivatives can be written as

$$\frac{\partial \Psi}{\partial \Sigma_t^{-1}} = w_{t+1}^{(v+1)} \left\{ \Sigma_t - \Sigma_{(t+1)t}^{(v+1)} + \sum_{j=m+2}^{\tau(g)} \frac{w_j^{(v+1)} d_{jt}}{w_{t+1}^{(v+1)} |d_j|} \times \left(\Sigma_{d_j} - \Sigma_{jt}^{(v+1)} + |d_j|^2 (\mu_t - \mu_{d_j}/|d_j|)(\mu_t - \mu_{d_j}/|d_j|)^T \right) \right\}.$$

The parametric space of covariances is not Euclidean but has a Riemanian metric structure, where the ordinary gradient does not give the steepest direction of increment for the target function Ψ; the steepest direction may be given by the so-called relative gradient (Cardoso and Laheld 1996). This motivates us to consider the direction of increment, $\hat{\Sigma}_t^{-\beta} U \hat{\Sigma}_t^{-\beta}$, $0 \le \beta \le 1/2$, where U is a positive definite matrix determined by optimizing the target function Ψ. This leads to

$$\Sigma_t^{(v+1)} = \hat{\Sigma}_t - \alpha_v \hat{\Sigma}_t^{1-2\beta} \frac{\partial \Psi}{\partial \Sigma_t^{-1}} \hat{\Sigma}_t^{1-2\beta}.$$

We set $\beta = 1/4$ which provides the best numerical performance according to our experience. The learning rate α is often set to be proportional to $1/(v+1)$ (Cardoso and Laheld 1996).

EEE case:

When all component covariances but Σ_0 are equal to Σ, the above updating formula for $\underline{\mu}$ still holds. Setting

$$\frac{\partial \Psi}{\partial \Sigma^{-1}} = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{\tau(g)} \tau_{ij}^{(v)} [\Sigma - (\mathbf{x}_i - \mu_{d_j})(\mathbf{x}_i - \mu_{d_j})^T] = 0,$$

and replacing $\underline{\mu}$ by $\underline{\mu}^{(v+1)}$, we have the updating formula for Σ

$$\Sigma^{(v+1)} = \frac{1}{\sum_{i=1}^{n} \sum_{j=2}^{\tau(g)} \tau_{ij}^{(v)}} \sum_{i=1}^{n} \sum_{j=2}^{\tau(g)} \tau_{ij}^{(v)} \left(\mathbf{x}_i - \mu_{d_j}^{(v+1)} \right) \left(\mathbf{x}_i - \mu_{d_j}^{(v+1)} \right)^T.$$

EED case:

In this case, the updating formulas but $\Sigma^{(v+1)}$ in the EEE still hold. The updating formula for $\Sigma = D$ is

$$\Sigma^{(v+1)} = \frac{1}{\sum_{i=1}^{n} \sum_{j=2}^{\tau(g)} \tau_{ij}^{(v)}} \sum_{i=1}^{n} \sum_{j=2}^{\tau(g)} \tau_{ij}^{(v)} \operatorname{diag} \left((\mathbf{x}_i - \mu_{d_j}^{(v+1)})^2 \right),$$

where $(\mathbf{x}_i - \mu_{d_j}^{(v+1)})^2$ denotes the component-wise square of $\mathbf{x}_i - \mu_{d_j}^{(v+1)}$.

EVA case:

Under this situation, letting $\sigma_{d_j}^2 = \left(\sum_{k=1}^{m} \sigma_k^{-2} d_{jk} / |d_j| \right)^{-1}$, we have

$$\Sigma_{d_j} = \sigma^2(d_j) A, \quad |\Sigma_{d_j}^{-1}| = \sigma_{d_j}^{-2p} |A^{-1}|, \quad \mu_{d_j} = \sigma^2(d_j) \sum_{k=1}^{m} \sigma_k^{-2} \mu_k d_{jk}.$$

Given $\sigma_{d_j}^2$ and A, by solving the equations $\frac{\partial \Psi}{\partial \underline{\mu}} = 0$, we have

$$\operatorname{vec}(\underline{\mu}) = \left(\operatorname{diag}(\sigma_1^2, \ldots, \sigma_m^2) \otimes I_p \right) \left(\sum_{j=2}^{\tau(g)} \left(\sum_{i=1}^{n} \tau_{ij}^{(v)} \right) d_j d_j^T \otimes \left(\sigma_{d_j}^2 I_p \right) \right)^{-1} (T^{(v)} D^T \otimes I_p)^T \operatorname{vec}(\mathbf{X}).$$

Letting $\underline{\mu}$ be held fixed, we alternatively update A and σ_k, $1 \le k \le m$. First, holding σ_k^2, $1 \le k \le m$ at the current values, we update A by solving the equations $\frac{\partial \Psi}{\partial A^{-1}} = 0$, which leads to

$$A^{(v+1)} = \frac{\sum_{j=2}^{\tau(g)} \sigma_{d_j}^{-2} \sum_{i=1}^{n} \tau_{ij}^{(v)} (\mathbf{x}_i - \mu_{d_j})(\mathbf{x}_i - \mu_{d_j})^T}{\sum_{j=2}^{\tau(g)} \sum_{i=1}^{n} \tau_{ij}^{(v)}}.$$

Then, letting A be held fixed and dropping the requirement of $\sigma_1 = 1$, we calculate the first and second partial derivatives of Ψ:

$$\frac{\partial \Psi}{\partial \underline{\sigma}^{-2}} = \frac{np}{2} \sum_{j=2}^{\tau(g)} w_j^{(v)} \frac{d_j}{|d_j|} \odot \left(\sigma_{d_j}^2 \mathbf{1}_m - H \right),$$

where $\odot$ is the coordinate-wise product operator, $\underline{\sigma}^{-2} = (\sigma_1^{-2}, \ldots, \sigma_m^{-2})^T$ and $H = (h_1, \ldots, h_m)^T$ with

$$h_k = \frac{\sum_{i=1}^{n} \tau_{ij}^{(v)} (\mathbf{x}_i - |d_j|\mu_k)^T A^{-1} (\mathbf{x}_i - |d_j|\mu_k)}{n w_j^{(v)} p} - \frac{|d_j|^2 (\mu_k - \mu_{d_j}/|d_j|)^T A^{-1} (\mu_k - \mu_{d_j}/|d_j|)}{p}.$$

And

$$\frac{\partial^2 \Psi}{\partial \underline{\sigma}^{-2} \partial \underline{\sigma}^{-2T}} = -\frac{np}{2} \sum_{j=2}^{\tau(g)} w_j^{(v)} \sigma_{d_j}^2 \frac{d_j d_j^T}{|d_j|^2} \odot \Delta_2$$

$$\Delta_2 = \sigma_{d_j}^2 \mathbf{1}_m \mathbf{1}_m^T + 2 \frac{|d_j|^2}{p} \left(\underline{\mu} - \mu_{d_j} \mathbf{1}_m^T / |d_j| \right)^T \times A^{-1} \left(\underline{\mu} - \mu_{d_j} \mathbf{1}_m^T / |d_j| \right),$$

where $\mathbf{1}_m$ is the m-vector of 1s. Making a log-transformation on σ_k^2, $\beta_k = \log(\sigma_k^{-2})$, so that we can apply the Newton-Raphson algorithm to β_k, $1 \le k \le m$. This gives the following updating formula

$$\underline{\sigma}^{(v+1)} = \underline{\sigma}^{2(v)} \odot \exp \left(-\operatorname{diag} \left(\underline{\sigma}^{2(v)} \right) G^{-1} F \right),$$

where $\operatorname{diag}(\underline{\sigma}^{2(v)})$ denotes the diagonal matrix based on $\underline{\sigma}^{2(v)}$, and

$$G = \sum_{j=2}^{\tau(g)} w_j^{(v)} \sigma_{d_j}^2 \frac{d_j d_j^T}{|d_j|^2} \odot \Delta_3$$

$$\Delta_3 = \sigma_{d_j}^2 \mathbf{1}_m \underline{1}_m^T + 2 \frac{|d_j|^2}{p} \left(\underline{\mu} - \mu_{d_j} \mathbf{1}_m^T / |d_j| \right)^T A^{-1} \left(\underline{\mu} - \mu_{d_j} \mathbf{1}_m^T / |d_j| \right)$$

$$F = \sum_{j=2}^{\tau(g)} w_j^{(v)} \frac{d_j}{|d_j|} \odot \left(\sigma_{d_j}^2 \mathbf{1}_m - H \right)$$

with $\underline{\mu}, \sigma_k^2$, $1 \le k \le m$ and A inside being replaced by $\underline{\mu}^{(v)}, (\sigma_k^2)^{(v)}$, $1 \le k \le m$ and $A^{(v+1)}$, respectively. Finally, we adjust $\underline{\sigma}^{2(v+1)}$ and $A^{(v+1)}$ by $\underline{\sigma}^{2(v+1)} / \sigma_1^{2(v+1)}$ and $\sigma_1^{2(v+1)} A^{(v+1)}$ to match the requirement of $\sigma_1^2 = 1$.

EVD case:

In this situation the updating formulas are similar to the EVA case, that is, these formulas still hold if we estimate the matrix A by the following diagonal matrix $\mathrm{diag}(1/\hat{a}_1, \ldots, 1/\hat{a}_p)$, where

$$\hat{a}_k = \frac{\sum_{j=2}^{\tau(g)} \sum_{i=1}^{n} \tau_{ij}^{(v)} (x_{ik} - \mu_{d_j}[k])^2 / \sigma_{d_j}^2}{\sum_{j=2}^{\tau(g)} \sum_{i=1}^{n} \tau_{ij}^{(v)}}, 1 \le k \le p,$$

and $\mu_{d_j}[k]$ is the k-th coordinate of the vector μ_{d_j}.

EVI case:

In this situation the updating formulas are similar to the EVA case, that is, these formulas still hold if we replace the matrix A there by the identity matrix I_p.

APPENDIX B: REPARAMETERIZATION

The reparameterization for $f \in \mathcal{G}_m$, $m \ge m_0$. For ease of presentation, we focus on the EVA models with $g = 2$. We write f_0 and f respectively as follows:

$$\begin{aligned} f_{\psi_{m_0}}(.) &= \sum_{k=0}^{m_0} w_k^{(0)} \phi\big(.|\mu_k^{(0)}, \sigma_j^{(0)2} A^{(0)}\big) \\ &\quad + \sum_{1 \le j_1 < j_2 \le m_0} w_{j_1 j_2}^{(0)} \phi\left(.|2\left(\frac{\mu_{j_1}^{(0)}/\sigma_{j_1}^{(0)2} + \mu_{j_2}^{(0)}/\sigma_{j_2}^{(0)2}}{1/\sigma_{j_1}^{(0)2} + 1/\sigma_{j_2}^{(0)2}}, \right.\right. \\ &\quad \left.\left. \frac{2A^{(0)}}{1/\sigma_{j_1}^{(0)2} + 1/\sigma_{j_2}^{(0)2}}\right)\right). \\ f_{\psi_m}(.) &= \sum_{k=0}^{m} w_k \phi\left(.|\mu_k, {\sigma_j}^2 A\right) \\ &\quad + \sum_{1 \le j_1 < j_2 \le m} w_{j_1 j_2} \phi\left(.|2\frac{\mu_{j_1}/\sigma_{j_1}^2 + \mu_{j_2}/\sigma_{j_2}^2}{1/\sigma_{j_1}^2 + 1/\sigma_{j_2}^2}, \frac{2A}{1/\sigma_{j_1}^2 + 1/\sigma_{j_2}^2}\right). \end{aligned}$$

Following the road map of Azaïs, Gassiat, and Mercadier (2009), given $(J_k)_{1 \le k \le m_0+1}$, a partition of $\{1, 2, \cdots, m\}$, we can reparameterize f as $f_{\gamma^{(0)}, v, \eta}$ in terms of new parameters $(\gamma^{(0)}, v, \eta)$, where we denote

$$\begin{aligned} u &= \big(\beta_0, \mathrm{vec}(\Delta\Sigma), (\beta_{kj})_{j \in J_k, 1 \le k \le m_0}, \big(\sigma_{kj}^2\big)_{j \in J_k, 1 \le k \le m_0}, \mathrm{vec}(\Delta A)\big), \\ \gamma^{(1)} &= ((w_j)_{j \in J_{m_0+1}}, (w_{j_1 j_2})_{j_1 < j_2 \in J_k, 1 \le k \le m_0}, (w_{j_1 j_2})_{j_1 \in J_k, j_2 \in J_{m_0+1}, 1 \le k \le m_0}, \\ &\quad (w_{j_1 j_2})_{j_1 < j_2 \in J_{m_0+1}}), \\ v &= \big(\alpha_0, (\alpha_{kj})_{j \in J_k, 1 \le k \le m_0}, (\alpha_{k_1 j_1 k_2 j_2})_{j_1 \in J_{k_1}, j_2 \in J_{k_2}, k_1 < k_2 \le m_0}, \gamma^{(1)}, u\big), \\ \eta &= \big((\mu_j)_{j \in J_{m_0+1}}, (\sigma_j^2)_{j \in J_{m_0+1}}\big) \\ \gamma^{(0)} &= \big(w_0^{(0)}, (\alpha_{kj}^{(0)})_{j \in J_k, 1 \le k \le m_0}, (\alpha_{k_1 j_1 k_2 j_2}^{(0)})_{j_1 \in J_{k_1}, j_2 \in J_{k_2}, k_1 < k_2}\big). \end{aligned}$$

Here v is determined by the decompositions $w_0 = w_0^{(0)} + \alpha_0$, $\mu_0 = \mu_0^{(0)} + \beta_0$, $\Sigma_0 = \Sigma_0^{(0)} + \Delta\Sigma, A = A^{(0)} + \Delta A$; $w_{kj} = \alpha_{kj}^{(0)} + \alpha_{kj}$, $\mu_j = \mu_k^{(0)} + \beta_{kj}$, $\sigma_j^2 = \sigma_k^{(0)2} + \sigma_{kj}^2 (j \in J_k, 1 \le k \le m_0)$; $w_{j_1 j_2} (j_1 < j_2 \in J_k, 1 \le k \le m_0)$; $w_{j_1 j_2} = \alpha_{k_1 j_1 k_2 j_2}^{(0)} + \alpha_{k_1 j_1 k_2 j_2}$ $(j_1 \in J_{k_1}, j_2 \in J_{k_2}, k_1 < k_2 \le m_0)$. And

$$\begin{aligned} \sum_{j \in J_k} \alpha_{kj}^{(0)} &= w_k^{(0)}, \quad \sum_{j_1 \in J_{k_1}, j_2 \in J_{k_2}} \alpha_{k_1 j_1 k_2 j_2}^{(0)} = w_{k_1 k_2}^{(0)}, \\ 0 &= \alpha_0 + \sum_{k=1}^{m_0} \sum_{j \in J_k} \alpha_{kj} + \sum_{j \in J_{m_0+1}} w_j + \sum_{k=1}^{m_0} \sum_{j_1 < j_2 \in J_k} w_{j_1 j_2} \\ &\quad + \sum_{k_1 < k_2 \le m_0} \sum_{j_1 \in J_{k_1}, j_2 \in J_{k_2}} \alpha_{k_1 j_1 k_2 j_2} + \sum_{k=1}^{m_0} \sum_{j_1 \in J_k, j_2 \in J_{m_0+1}} w_{j_1 j_2} \\ &\quad + \sum_{j_1 < j_2 \in J_{m_0+1}} w_{j_1 j_2}. \end{aligned}$$

The reparameterization may not be unique. However, we can permutate the mixture components so that $\|v\| + \|\eta\|$ attains the minimum, since the value of the density f is not dependent of the order of its mixture components. The idea behind this reparameterization is that the asymptotic behavior of the model order estimator depends on the convergence property of an empirical process over a Hellinger neighborhood of f_0. Under the identifiability Condition (C1), for any sequence $f_i \in \mathcal{F}_m$, when $H(f_i, f_0) \to 0$, we can find a permutation on $\{1, \ldots, m\}$, such that the new parameter $\|v_i\| \to 0$.

The reparameterization for $f \in \mathcal{G}_{m_0}$. We reparameterize f into the form $f_{\gamma^{(0)}, v}$. Let $v^* = v/\|v\|$. We take the EVA models with $g = 2$ as examples, where $\gamma^{(0)}$ and v are defined as follows.

$$\begin{aligned} u &= \big(\beta_0, \mathrm{vec}(\Delta\Sigma), (\beta_{kj})_{j \in J_k, 1 \le k \le m_0}, (\sigma_{kj}^2)_{j \in J_k, 1 \le k \le m_0}, \mathrm{vec}(\Delta A)\big), \\ v &= \Big(\alpha_0, (\alpha_{kj})_{j \in J_k, 1 \le k \le m_0}, (\alpha_{k_1 j_1 k_2 j_2})_{j_1 \in J_{k_1}, j_2 \in J_{k_2}, k_1 < k_2 \le m_0}, u\Big), \\ \gamma^{(0)} &= \Big(w_0^{(0)}, (\alpha_{kj}^{(0)})_{j \in J_k, 1 \le k \le m_0}\Big), \\ 0 &= \alpha_0 + \sum_{k=1}^{m_0} \sum_{j \in J_k} \alpha_{kj} + \sum_{k_1 < k_2} \sum_{j_1 \in J_{k_1}, j_2 \in J_{k_2}} \alpha_{k_1 j_1 k_2 j_2} \hat{=} \mathrm{sum}(v^*), \\ w_k^{(0)} &= \alpha_{kj}^{(0)}, j \in J_k, 1 \le k \le m_0. \end{aligned}$$

APPENDIX C: LEMMAS AND THE PROOFS

Proof of Proposition 4.1. See the online supplementary materials for the details. □

To prove Theorem 1, we need the following lemma to provide an upper bound to the bracketing number of the function space $\mathcal{S}$, $N_{[]}(\delta, \mathcal{S}, \|.\|_2)$, which is defined as the minimum number of brackets of size δ needed to cover $\mathcal{S}$. See van der Vaart and Wellner (1996).

Lemma 6.1. Under Condition (C2), there exists a constant c_0 such that for any functions $f_1 = f_{\gamma_1^{(0)}, v_1, \eta_1}$ and $f_2 = f_{\gamma_2^{(0)}, v_2, \eta_2}$ in $\mathcal{G}_m$ with $m \ge m_0$ and for all $\mathbf{x}$,

$$\begin{aligned} |s_{f_1}(\mathbf{x}) - s_{f_2}(\mathbf{x})| \le c_0 C(\mathbf{x}) (&\|v_1^* - v_2^*\| + \|\|v_2\| - \|v_1\|\| \\ &+ \|\gamma_1^{(0)} - \gamma_2^{(0)}\| + \|\eta_1 - \eta_2\|), \end{aligned}$$

where $C(\mathbf{x})$ is defined in Section 4.1. Moreover, $N_{[]}(\delta \|C(X)\|_2, \mathcal{S}, \|.\|_2) \le \mathrm{const.} \delta^q$, where q depends on $m_0 \le m \le M$ and p.

Proof of Lemma 6.1. See the online supplementary materials for the details. □

Proof of Theorem 1. Let $l_n(f) = \sum_{i=1}^n \log(f(x_i))$ denote the log-likelihood based on the density f. Then, by using Lemma 6.1 we can show that under Conditions (C1) and (C2), $\mathcal{G}_m$ is Donsker (van der Vaart and Wellner 1996) with a square integrable envelope and that $\inf_{f \in \mathcal{G}_m} E_{f_0}[(s_f)_-^2] \ne 0$. Denote $y_- = -\min\{y, 0\}$. Applying the likelihood ratio inequality of Gassiat (2002), we have

$$\begin{aligned} l_n(f) - l_n(f_0) &\le 2 \sum_{i=1}^n s_f(x_i) H(f, f_0) - s_f(x_i)_-^2 H^2(f, f_0), \\ \sup_{f \in \mathcal{G}_m} l_n(f) &\ge l_n(f_0), \quad E_{f_0}[s_f] = -\frac{H(f, f_0)}{2}, \end{aligned}$$

which imply $\sup_{f \in \mathcal{G}_m} |E_{f_0}[s_f]| = O_p(1/\sqrt{n})$ and

$$H(f, f_0) \le \frac{2 \sum_{i=1}^n (s_f(x_i) - E_{f_0} s_f)}{\sum_{i=1}^n (s_f(x_i))_-^2} = O_p(1/\sqrt{n}).$$

□

Consequently,

$$\begin{aligned}
&P(\hat{m} > m_0)\\
&\le \sum_{m=m_0+1}^{M} P\left(\mathrm{BIC}(\mathcal{G}_m) < \mathrm{BIC}(\mathcal{G}_{m_0})\right)\\
&\le \sum_{m=m_0+1}^{M} P\left(\sup_{s_f\in\mathcal{S}_m} \frac{(\sum_{i=1}^n s_f(x_i))^2}{\sum_{i=1}^n (s_f(x_i)_-)^2} \ge \frac{\log(n)}{2}\right.\\
&\qquad \left.\times\ (\dim(\mathcal{G}_m) - \dim(\mathcal{G}_{m_0}))\right)\\
&= \sum_{m=m_0+1}^{M} P\left(\dim(\mathcal{G}_m) - \dim(\mathcal{G}_{m_0}) \le O_p(1)/\log(n)\right) = o(1)
\end{aligned}$$

when $n \to \infty$. This together with Proposition 4.1 yields that $P(\hat{m} = m_0) \to 1$. It follows from the fact that $\{\log(f/f_0) : f \in \mathcal{G}_{m_0}\}$ is a Glivenko–Cantelli class; therefore,

$$\frac{1}{n}[l_n(f) - l_n(f_0)] = E_{f_0}\log(f/f_0) + o_p(1) \tag{C.2}$$

uniformly in $f \in \mathcal{G}_{m_0}$. Note that under Condition (C1), the divergence $d(f, f_0) \hat{=} E_{f_0}\log(f/f_0) \le 0$ and attains zero if and only if f and f_0 have the same mixture components, up to a permutation on the order of these components. Let $\hat{\psi}_{m_0}$ denote the constrained maximum likelihood estimator. Then, Equation (C.2) implies that

$$0 \le \frac{1}{n}\left[l_n(f_{\hat{\psi}_{m_0}}) - l_n(f_0)\right] = d\left(f_{\hat{\psi}_{m_0}}, f_0\right) + o_p(1),$$

which yields $d(f_{\hat{\psi}}, f_0) = o_p(1)$. Therefore, $\hat{\psi}_{m_0}$ converges to ψ_{m_0} in probability, up to a permutation on the order of the mixture components. The proof is completed.

Proof of Proposition 4.2. See the online supplementary materials for the details. □

Proof of Theorem 2. See the online supplementary materials for the details. □

Proof of Theorem 3. It follows from Theorem 1 that $\hat{\psi}_{m_0}$ converges to the true value ψ_{m_0} in probability, up to a permutation on the order of the mixture components. As $\|H_{n\hat{\psi}_{m_0}} - H_{n\psi_{m_0}}\|$ is not dependent on the order of the mixture component, without loss of generality we assume that $\hat{\psi}_{m_0} \to \psi_{m_0}$ in probability. Therefore, invoking Template A in Beran (2003), we only need to prove that for any $\psi_n \to \psi_{m_0}$, there exists a distribution $H_{\psi_{m_0}}(\cdot)$ such that $\|H_{n\psi_n} - H_{\psi_{m_0}}\| \to 0$ and that $H_{\psi_{m_0}}(\cdot)$ is continuous. □

In doing so, we first show that Condition (C2′) holds uniformly in a neighborhood of ψ_{m_0}, $B(\psi_{m_0}, \delta)$ below. Let f_ψ denote the epistatic mixture density of order m_0 with parameter ψ. For any epistatic mixture density f, we reparameterization f around f_ψ so that $f_\psi = f_{\gamma_\psi,0}$ accordingly define $\mathcal{F}_\psi$ and $\mathcal{S}_\psi$. Let $R_\psi(\gamma_\psi, v)$ denote the partial derivative of $\sqrt{f/f_\psi} - 1$ with respect to v and

$$\begin{aligned}
s_{ff_\psi} &= \frac{v^{*T}\int_0^1 R_\psi(\gamma_\psi, tv)dt}{\|v^{*T}\int_0^1 R_\psi(\gamma_\psi, tv)dt\|_2},\\
C_{1\psi}(\mathbf{x}) &= \sup_{(\gamma_\psi, v)\in\Theta_{q'}} \left|\int_0^1 R_\psi(\gamma_\psi, tv)dt\right|,\\
C_{2\psi}(\mathbf{x}) &= \inf_{(\gamma_\psi, v)\in\Theta_{q'}} \min\left\{\left\|\int_0^1 \frac{\partial R_\psi}{\partial \gamma_\psi}dt\right\|, \left\|\int_0^1 \frac{\partial R_\psi}{\partial v^T}dt\right\|\right\},\\
C_\psi(\mathbf{x}) &= \max\{C_{1\psi}(\mathbf{x}), C_{2\psi}(\mathbf{x})\}.
\end{aligned}$$

Let $\mathcal{S}_\psi = \{s_{ff_\psi} : f \in \mathcal{G}_{m_0}\}$ and

$$\begin{aligned}
\mathcal{S}^{(1)}_{\psi 0+} &= \left\{\frac{v^{*T}R_\psi(\gamma^{(0)}_\psi, 0)}{\|v^{*T}R_\psi(\gamma^{(0)}_\psi, 0)\|_2} : \|v^*\| = 1, \mathrm{sum}(v^*) = 0\right\},\\
\mathcal{S}^{(0)}_{\psi 0+} &= \left\{s \in \mathcal{S}^{(1)}_{\psi 0+} : \alpha_{k_1 j_1 k_2 j_2} = 0, j_1 \in J_{k_1}, j_2 \in J_{k_2}, k_1 < k_2 \le m_0\right\},
\end{aligned}$$

where $\mathcal{S}^{(1)}_{\psi 0+}$ tends to $\mathcal{S}^{(1)}_{0+}$ as $\psi \to \psi_{m_0}$. For any small $\delta > 0$, let $B(\psi_{m_0}, \delta)$ be a δ-neighborhood of ψ_{m_0}. Let $\mathcal{S}_\delta = \cup_{\psi\in B(\psi_{m_0},\delta)}\mathcal{S}^{(1)}_{\psi 0+}$ and $W(.)$ be a Gaussian process over $\mathcal{S}_\delta$ with covariance $\Sigma(s_1, s_2) = E_{f_0}[s_1(\mathbf{x})s_2(\mathbf{x})] - E_{f_0}[s_1(\mathbf{x})]E_{f_0}[s_2(\mathbf{x})]$. Let $H_{\psi_{m_0}}(\cdot)$ be the distribution of $\sup_{s\in\mathcal{S}^{(1)}_{0+}}\max\{W(s), 0\} - \sup_{s\in\mathcal{S}^{(0)}_{0+}}\max\{W(s), 0\}$. Then, under Condition (C2′), by the continuity of $R^T_\psi(\gamma_\psi, tv)$ in ψ, we can find a small $\delta > 0$ such that the eigenvalues of the matrix $E_{f_\psi}[\int_0^1 R_\psi(\gamma^{(0)}_\psi, tv)dt \int_0^1 R^T_\psi(\gamma^{(0)}_\psi, tv)dt]$ are uniformly bounded away from zero over $v \in \Theta_{q'}$ and $\psi \in B(\psi_{m_0}, \delta)$, and that $\sup_{\psi\in B(\psi_{m_0},\delta)} E_{f_\psi}[C_\psi(\mathbf{x})^2] < \infty$. By the Taylor expansion, we can show that for any small $\epsilon > 0$,

$$\sup_{\psi\in B(\psi_{m_0},\delta)} N_{[]}(\epsilon\|C_\psi(\mathbf{x})\|_2, \mathcal{S}^{(1)}_{\psi 0+}, \|\cdot\|_{\psi 2}) \le \text{ const. } \epsilon^{q'}, \tag{C.3}$$

where $\|\cdot\|_{\psi 2}$ is the L_2 norm under the expectation under f_ψ.

For any $\psi_n \to \psi_{m_0}$, the inequality (C.3) remains true if the $\|\cdot\|_{\psi 2}$ is replaced by the L_2 norm under the expectation f_{ψ_n}. Therefore, if n is large enough, then by Theorem 2.8.3 in van der Vaart and Wellner (1996), $\mathcal{S}_\delta$ is Donsker uniformly in ψ_n. Furthermore, Theorem 1 in Azaïs, Gassiat, and Mercadier (2009) can be modified and applied to the current setting. We obtain

$$\begin{aligned}
&2\sup_{f\in\mathcal{F}_1}(l_n(f) - l_n(f_\psi))\\
&= \sup_{s_{ff_\psi}\in\mathcal{S}^{(1)}_{\psi 0+}}\left[\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n s_{ff_\psi}(\mathbf{x}_i)\right)^2 I\left(\sum_{i=1}^n s_{ff_\psi}(\mathbf{x}_i) \ge 0\right)\right] + o_p(1)
\end{aligned}$$

uniformly in $\psi \in B(\psi_{m_0}, \delta)$, where $I(\cdot)$ is an indicator function and the underlying probability density of $\mathbf{x}_i$ is f_ψ. In particular, we have

$$\begin{aligned}
&2\sup_{f\in\mathcal{F}_1}(l_n(f) - l_n(f_{\psi_n}))\\
&= \sup_{s_{ff_{\psi_n}}\in\mathcal{S}^{(1)}_{\psi_n 0+}}\left[\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n s_{ff_{\psi_n}}(\mathbf{x}_i)\right)^2 I\left(\sum_{i=1}^n s_{ff_{\psi_n}}(\mathbf{x}_i) \ge 0\right)\right] + o_p(1)
\end{aligned} \tag{C.4}$$

when the underlying probability density of $\mathbf{x}_i$ is f_{ψ_n}. We consider the following empirical process

$$W_n(s) = \frac{1}{\sqrt{n}}\sum_{i=1}^n (s(\mathbf{x}_i) - E_{f_{\psi_n}}[s(\mathbf{x}_i)]), \quad s \in \mathcal{S}_\delta.$$

Note that as ψ_n tends to ψ_{m_0},

$$\begin{aligned}
&\sup_{s_1, s_2\in\mathcal{S}_\delta}\left|\sqrt{E_{f_{\psi_n}}(s_1(\mathbf{x}) - s_2(\mathbf{x}))^2} - \sqrt{E_{f_{\psi_{m_0}}}(s_1(\mathbf{x}) - s_2(\mathbf{x}))^2}\right| \to 0,\\
&E_{f_{\psi_n}}[C_1(\mathbf{x})^2 I(C_1(\mathbf{x}) \ge \epsilon\sqrt{n})] \to 0, \text{ for every } \epsilon > 0,
\end{aligned}$$

where $C_1(\mathbf{x})$ is defined in Section 4.1. Combining these facts with the inequality (C.3) and with Theorem 2.8.10 in van der Vaart and Wellner (1996) yields that the empirical process $\{W_n(s) : s \in \mathcal{S}_\delta\}$ converges weakly to $\{W(s) : s \in \mathcal{S}_\delta\}$, where W is uniformly continuous with respect to the norm $\sqrt{E_{f_0}[s^2(\mathbf{x})]}$. Note that $E_{f_{\psi_n}}[s_{ff_{\psi_n}}(\mathbf{x})] = 0$ and $\mathcal{S}^{(1)}_{\psi_n 0+} \subset \mathcal{S}_\delta$ for large n. Thus, by the so-called continuous mapping

theorem [Theorem 1.9.5 in van der Vaart and Wellner (1996)], we have

$$\sup_{s_{ff_{\psi_n}}\in\mathcal{S}^{(1)}_{\psi_n 0+}}\left[\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}s_{ff_{\psi_n}}(\mathbf{x}_i)\right)^2 I\left(\sum_{i=1}^{n}s_{ff_{\psi_n}}(\mathbf{x}_i)\geq 0\right)\right]$$
$$\leq \sup_{s\in\mathcal{S}_\delta}\left[\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(s(\mathbf{x}_i)-E[s(\mathbf{x})])\right)^2 I\left(\sum_{i=1}^{n}(s(\mathbf{x}_i)-E[s(\mathbf{x})])\geq 0\right)\right]$$
$$\xrightarrow{\text{weakly}} \sup_{s\in\mathcal{S}_\delta} W(s)^2 I(W(s)\geq 0).$$

Due to the continuity of W, the last term above tends to $\sup_{s\in\mathcal{S}^{(1)}_{0+}} W(s)^2 I(W(s)\geq 0)$ as $\delta\to 0$. On the other hand, for any small $\epsilon>0$, by the finiteness of the covering number in (C.3) and by the uniform continuity of W, we can find $\{s_k, 1\leq k\leq k_0\}\subset\mathcal{S}^{(1)}_{0+}$ such that $\sup_{g\in\mathcal{S}^{(1)}_{0+}} W(s)^2 I(W(s)\geq 0) < \sup_{1\leq k\leq k_0} W(s_k) I(W(s_k)\geq 0)+\epsilon$. We can also find $\{s_{kn}, 1\leq k\leq k_0\}\subset\mathcal{S}^{(1)}_{\psi_n+}$ such that $E_{f_0}[s_{kn}(\mathbf{x})-s_k(\mathbf{x})]^2\to 0$, $1\leq k\leq k_0$ as $n\to\infty$. Note that

$$\sup_{s_{ff_{\psi_n}}\in\mathcal{S}^{(1)}_{\psi_n 0+}}\left[\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}s_{ff_{\psi_n}}(\mathbf{x}_i)\right)^2 I\left(\sum_{i=1}^{n}s_{ff_{\psi_n}}(\mathbf{x}_i)\geq 0\right)\right]$$
$$\geq \sup_{1\leq k\leq k_0}\left[\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}s_{kn}(\mathbf{x}_i)\right)^2 I\left(\sum_{i=1}^{n}s_{kn}(\mathbf{x}_i)\geq 0\right)\right]$$
$$\xrightarrow{\text{weakly}} \sup_{1\leq k\leq k_0} W(s_k) I(W(s_k)\geq 0),$$

where the last term follows from the Lindeberg central limit theorem and the continuous mapping theorem mentioned before. Consequently,

$$\sup_{s_{ff_{\psi_n}}\in\mathcal{S}^{(1)}_{\psi_n 0+}}\left[\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}s_{ff_{\psi_n}}(\mathbf{x}_i)\right)^2 I\left(\sum_{i=1}^{n}s_{ff_{\psi_n}}(\mathbf{x}_i)\geq 0\right)\right]$$
$$\xrightarrow{\text{weakly}} \sup_{s\in\mathcal{S}^{(1)}_{0+}} W(s)^2 I(W(s)\geq 0). \tag{C.5}$$

Analogously, we can prove that

$$2\sup_{f\in\mathcal{F}_0}(l_n(f)-l_n(f_{\psi_n}))$$
$$= \sup_{s_{ff_{\psi_n}}\in\mathcal{S}^{(0)}_{\psi_n 0+}}\left[\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}s_{ff_{\psi_n}}(\mathbf{x}_i)\right)^2 I\left(\sum_{i=1}^{n}s_{ff_{\psi_n}}(\mathbf{x}_i)\geq 0\right)\right]+o_p(1),$$
$$\sup_{s_{ff_{\psi_n}}\in\mathcal{S}^{(0)}_{\psi_n 0+}}\left[\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}s_{ff_{\psi_n}}(\mathbf{x}_i)\right)^2 I\left(\sum_{i=1}^{n}s_{ff_{\psi_n}}(\mathbf{x}_i)\geq 0\right)\right]$$
$$\xrightarrow{\text{weakly}} \sup_{s\in\mathcal{S}^{(0)}_{0+}} W(s)^2 I(W(s)\geq 0).$$

Combining these equations with (C.4) and (C.5), we complete the proof.

Proof of Proposition 4.3. It is similar to the proofs of Theorem 1 and Proposition 4.1. The details are omitted. □

SUPPLEMENTARY MATERIALS

The supplementary materials contain additional simulations, examples, and proofs from the main article.

[Received May 2011. Revised September 2012.]

REFERENCES

Azaïs, J., Gassiat, E., and Mercadier, C. (2009), "The Likelihood Ratio Test for General Mixture Models With or Without Structural Parameter," *ESAIM: Probability and Statistics*, 13, 301–327. [1382,1383]

Battle, A., Segal, E., and Koller, D. (2005), "Probabilistic Discovery of Overlapping Cellular Processes and Their Regulation," *Journal of Computational Biology*, 12, 907–927. [1367,1368,1369,1374,1380]

Beran, R. (2003), "The Impact of the Bootstrap on Statistical Algorithms and Theory," *Statistical Science*, 18, 175–184. [1383]

Blei, D., Ng, A., and Jordan, M. (2003), "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, 3, 993–1022. [1380]

Boyle, E., Wang, S., Gollub, J., Jin, H., Botstein, D., Cherry, J. M., and Sherlock, G. (2004), "GO: TermFinder-Open Source Software for Accessing Gene Ontology Information and Finding Significantly Enriched Gene Ontology Terms Associated With a List of Genes," *Bioinformatics*, 20, 3710–3715. [1374,1377]

Cardoso, J. F., and Laheld, B. H. (1996), "Equivariant Adaptive Source Separation," *IEEE Transactions on Signal Processing*, 44, 3017–3030. [1381]

Cordell, H. (2002), "Epistasis: What It Means, What It Doesn't Mean, and Statistical Methods to Detect It in Humans," *Human Molecular Genetics*, 11, 2463–2468. [1367]

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998), "Cluster Analysis and Display of Genome-wide Expression Patterns," *Proceedings of the National Academy of Sciences*, 95, 14863–14868. [1366]

Fraley, C., and Raftery, A. E. (2002), "Model-Based Clustering, Discriminant Analysis, and Density Estimation," *Journal of the American Statistical Association*, 97, 611–631. [1366,1369,1370]

Gasch, A. P., Spellman, P. T., Kao, C. M., Carmel-Harel, O., Eisen, M. B., Storz, G., Botstein, D., and Brown, P. O. (2000), "Genomic Expression Program in the Response of Yeast Cells to Environmental Changes," *Molecular Biology of the Cell*, 11, 4241–4257. [1366,1367,1373]

Gassiat, E. (2002), "Likelihood Ratio Inequalities With Applications to Various Mixtures," *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, 6, 897–906. [1371,1372,1382]

Han, J., Kamber, M., and Pei, J., (2011), *Data Mining: Concepts and Techniques* (3rd ed.), Waltham: Morgan Kaufmann. [1368]

Heller, K., Williamson, S., and Ghahramani, Z. (2008), "Statistical Models for Partial Membership," in *Proceedings of the 25th International Conference on Machine Learning, Helsinki, Finland*, pp. 392–398. [1380]

Hubert, L. J., and Arabie, P. (1985), "Comparing Partitions," *Journal of Classification*, 2, 193–218. [1370]

Kaufman, L., and Rousseauw, P. J. (1990), *Finding Groups in Data: An Introduction to Cluster Analysis*, New York: Wiley. [1366,1370]

McLachlan, G., and Peel, D. (2000), *Finite Mixture Models*, New York: Wiley. [1371]

Palla, G., Derenyi, I., Farkas, I., and Vicsek, T. (2005), "Uncovering the Overlapping Community Structure of Complex Networks in Nature and Society," *Nature*, 435, 814–818. [1380]

Pritchard, J., Stephens, M., and Donnelly, P. (2000), "Inference of Population Structure Using Multilocus Genotype Data," *Genetics*, 155, 945–959. [1380]

van Driessche, N., Demsar, J., Booth, E., et al. (2005), "Epistasis Analysis With Global Transcriptional Phenotypes," *Nature Genetics*, 37, 471–477. [1379]

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," *Bioinformatics*, 17, 977–987. [1379]

Yeung, K. Y., Medvedovic, M., and Bumgarner, R. E. (2004), "From Co-expression to Co-regulation: How Many Microarray Experiments Do We Need?" *Genome Biology*, 5, Article R48. [1366]

van der Vaart, A., and Wellner, J. (1996), *Weak Convergence and Empirical Processes With Applications to Statistics*, New York: Springer. [1382,1383]

Zhang, J., and Liang, F. (2010), "Robust Clustering Using Exponential Power Mixtures," *Biometrics*, 66, 1078–1086. [1380]