# Supplemental Materials for "A Tailored Multivariate Mixture Model for Detecting Proteins of Concordant Change Among Virulent Strains of *Clostridium Perfringens*"

Kun Chen[1]*, Neha Mishra[2], Joan Smyth[2], Haim Bar[1], Elizabeth Schifano[1], Lynn Kuo[1], Ming-Hui Chen[1]
[1]*Department of Statistics, University of Connecticut*
[2]*Department of Pathobiology and Veterinary Science, University of Connecticut*

June 14, 2017

## 1 More on Model Constraints and Penalty Forms

In some applications it may be desirable to require $\boldsymbol{\mu}_2 = -\boldsymbol{\mu}_3$, i.e., the two directions of concordant changes are exactly opposite to each other. Then $\boldsymbol{\mu}_2$ and $\boldsymbol{\mu}_3$ can be updated as follows,

$$\widehat{\boldsymbol{\mu}}_2 = -\widehat{\boldsymbol{\mu}}_3 = (\sum_{i=1}^{n}\sum_{k=2}^{3} p_{ik}^c \sum_{s\in\mathcal{S}_i} \boldsymbol{\Sigma}_k^{-1})^{-1} \left[\sum_{i=1}^{n}\{p_{i2}^c \sum_{s\in\mathcal{S}_i} \boldsymbol{\Sigma}_2^{-1}(\mathbf{y}_{i\cdot s} - \alpha_s\mathbf{1}) - p_{i3}^c \sum_{s\in\mathcal{S}_i} \boldsymbol{\Sigma}_3^{-1}(\mathbf{y}_{i\cdot s} - \alpha_s\mathbf{1})\}\right].$$

The case when all the components share the same correlation matrix can also be handled. Let $\boldsymbol{\Sigma}_1 \in \mathbb{R}^{m\times m}$ be the p.d. covariance matrix of the reference component. We parameterize $\boldsymbol{\Sigma}_2 = \mathbf{D}_2\boldsymbol{\Sigma}_1\mathbf{D}_2$ and $\boldsymbol{\Sigma}_3 = \mathbf{D}_3\boldsymbol{\Sigma}_1\mathbf{D}_3$, where $\mathbf{D}_2 = \mathrm{diag}(\mathbf{d}_2)$, $\mathbf{D}_3 = \mathrm{diag}(\mathbf{d}_3)$, and $d_{jk} > 0$ for any $j = 1,\ldots,m$ and $k = 2,3$. In the following we also write $\mathbf{D}_1 = \mathbf{I}_m$ for ease of presentation. From this parameterization, the $d_{jk}^2$'s can be regarded as variance inflation parameters. Fixing other parameters, $\boldsymbol{\Sigma}_1$ is updated as

$$\widehat{\boldsymbol{\Sigma}}_1 = \frac{1}{N}\sum_{i=1}^{n}\sum_{k=1}^{K} p_{ik}^c \sum_{s\in\mathcal{S}_i} \mathbf{D}_k^{-1}\mathbf{e}_{i\cdot s(k)}\mathbf{e}_{i\cdot s(k)}^{\mathrm{T}}\mathbf{D}_k^{-1}, \tag{1}$$

where $N = \sum_{i=1}^{n}|\mathcal{S}_i|$ is the total sample size. Denote $\mathbf{W}_k = \sum_{i=1}^{n} p_{ik}^c \sum_{s\in\mathcal{S}_i} \mathbf{e}_{i\cdot s(k)}\mathbf{e}_{i\cdot s(k)}^{\mathrm{T}}$. For updating $\mathbf{d}_2$, we need to solve

$$\max_{\mathbf{d}_2^{-1}>\mathbf{0}} \left\{ g(\mathbf{d}_2^{-1}) \equiv -(\sum_{i=1}^{n} p_{ik}^c|\mathcal{S}_i|)(\sum_{j=1}^{m}\log d_{j2}) - \frac{1}{2}\mathrm{tr}(\mathbf{D}_2^{-1}\boldsymbol{\Sigma}_1^{-1}\mathbf{D}_2^{-1}\mathbf{W}_2) \right\}. \tag{2}$$

Here for convenience we have equivalently expressed the problem with respect to $\mathbf{d}_2^{-1}$, for which the gradient function is

$$g'(\mathbf{d}_2^{-1}) = \frac{\partial g(\mathbf{d}_2^{-1})}{\partial \mathbf{d}_2^{-1}} = (\sum_{i=1}^{n} p_{ik}^c|\mathcal{S}_i|)\mathbf{d}_2 - \frac{1}{2}\mathcal{D}(\mathbf{W}_2^{\mathrm{T}}\mathbf{D}_2^{-1}\boldsymbol{\Sigma}_1^{-1} + \boldsymbol{\Sigma}_1^{-1}\mathbf{D}_2^{-1}\mathbf{W}_2),$$

where $\mathcal{D}(\cdot)$ denotes the vector of diagonal elements of the enclosed matrix. To solve this nonlinear optimization problem with linear inequality constraints, we can apply an adaptive barrier algorithm implemented in

*Corresponding author;kun.chen@uconn.edu

the free software environment R. Although in principle the optimal solution can occur on the boundary of the feasible region, e.g., $d_{j2} = 0$ for some $j$, we do not observe such occurrence in our numerical experiments. The problem of updating $\mathbf{d}_3$ has exactly the same form as that of updating $\mathbf{d}_2$; we thus omit the details.

Beside the group $\ell_0$ penalty used in the paper, we may also consider the group lasso penalty (Yuan and Lin, 2006)

$$\rho(\boldsymbol{\gamma}_i; \lambda) = \lambda \|\boldsymbol{\gamma}_i\|_2. \tag{3}$$

In this case, (15) is a group lasso regression problem with a single group of size $m$ (Yuan and Lin, 2006), which, however, does not admit an explicit solution in general. There are numerous existing methods for efficiently solving group lasso (Huang *et al.*, 2012); however, in our problem, naively applying a general group lasso solver $n$ many times in each M-step can be extremely inefficient. We thus explore further the properties of the solution of (15), denoted as $\widehat{\boldsymbol{\gamma}}_i$. Based on the Karush-Kuhn-Tucker (KKT) conditions, $\widehat{\boldsymbol{\gamma}}_i$ satisfies

$$-\widetilde{\mathbf{X}}^{\mathrm{T}}(\widetilde{\mathbf{y}}_i - \widetilde{\mathbf{X}}\widehat{\boldsymbol{\gamma}}_i) + \lambda_i \mathbf{s}_i = \mathbf{0},$$

where $\lambda_i = \lambda/(p_{i1}|\mathcal{S}_i|)$, and $\mathbf{s}_i$ is a subgradient vector of $\|\boldsymbol{\gamma}_i\|_2$ at $\widehat{\boldsymbol{\gamma}}_i$, i.e., $\mathbf{s}_i = \widehat{\boldsymbol{\gamma}}_i / \|\widehat{\boldsymbol{\gamma}}_i\|_2$ if $\widehat{\boldsymbol{\gamma}}_i \neq \mathbf{0}$, and $\mathbf{s}_i$ is a vector with $\|\mathbf{s}_i\|_2 < 1$ if $\widehat{\boldsymbol{\gamma}}_i = \mathbf{0}$. It follows that $\widehat{\boldsymbol{\gamma}}_i = \mathbf{0}$ whenever $\|\widetilde{\mathbf{X}}^{\mathrm{T}}\widetilde{\mathbf{y}}_i\|_2 < \lambda_i$, because the above KKT condition can be satisfied with $\widehat{\boldsymbol{\gamma}}_i = \mathbf{0}$ and $\mathbf{s}_i = \widetilde{\mathbf{X}}^{\mathrm{T}}\widetilde{\mathbf{y}}_i/\lambda_i$. When $\|\widetilde{\mathbf{X}}^{\mathrm{T}}\widetilde{\mathbf{y}}_i\| \geq \lambda_i$, the solution satisfies $\widehat{\boldsymbol{\gamma}}_i = (\widetilde{\mathbf{X}}^{\mathrm{T}}\widetilde{\mathbf{X}} + (\lambda_i/\|\widehat{\boldsymbol{\gamma}}_i\|)\mathbf{I})^{-1}\widetilde{\mathbf{X}}^{\mathrm{T}}\mathbf{y}_i$. Therefore, in our implementation, we directly set $\widehat{\boldsymbol{\gamma}}_i = \mathbf{0}$ for any $\|\widetilde{\mathbf{X}}^{\mathrm{T}}\widetilde{\mathbf{y}}_i\|_2 < \lambda_i$, and only otherwise a general group lasso solver is used. Since most of $\boldsymbol{\gamma}_i$'s are expected to be zero vectors, this approach greatly improves computational efficiency.

It is clear that the main difference between the group lasso and the group $\ell_0$ penalization methods is in their ways of adjusting for the outlying effects; while the former induces shrinkage estimation so the outlying effects of the proteins may be partially adjusted in a continuous/smooth fashion as $\lambda$ varies, the latter acts in a discrete way, i.e., either fully adjusting for the outlying effects using the least squares solutions or doing nothing at all. We note that the AICc and the formula for the degrees of freedom given in Section 3.4 still apply.

# 2 Connections to Trimmed Likelihood

To better understand the identifiability and robustness of our proposed regularized estimation approach, we explore its connections with other robust clustering approaches. To focus on the main idea, we consider a generic setup of the problem, i.e., the task of clustering observations $\mathbf{y}_i \in \mathbb{R}^m$, $i = 1, \ldots, n$, using the group $\ell_0$ penalization approach under the eigenvalue-ratio condition A0. Our results can be readily applied to handle replicated data and additional structural constraints.

Consider

$$\max_{\boldsymbol{\Theta} \in \Omega, \boldsymbol{\Gamma}} \left[ \sum_{i=1}^{n} \log \left\{ \sum_{k=1}^{K} \pi_k \phi(\mathbf{y}_i; \boldsymbol{\mu}_k + \boldsymbol{\gamma}_{i(k)}, \boldsymbol{\Sigma}_k) \right\} - \frac{\lambda^2}{2} \sum_{i=1}^{n} I(\|\boldsymbol{\Gamma}_i\|_F \neq 0) \right], \tag{4}$$

where

$$\Omega = \{(\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), k = 1, \ldots, K; 0 \leq \pi_k \leq 1, \sum_{k=1}^{K} \pi_k = 1, \delta_{\max}/\delta_{\min} = c\},$$

$\boldsymbol{\Gamma}_i = (\boldsymbol{\gamma}_{i(1)}, \ldots, \boldsymbol{\gamma}_{i(K)}) \in \mathbb{R}^{m \times K}$, $\boldsymbol{\Gamma} = \{\boldsymbol{\Gamma}_i; i = 1, \ldots, n\}$, and $\|\cdot\|_F$ stands for the Frobenius norm. (The tailored approach we use in the protein application corresponds to solving (4) with the extra structural constraints in A1–A3 and with setting $\boldsymbol{\gamma}_{i(k)} = \mathbf{0}$ for $k \neq 1$.)

**Lemma 2.1.** *Suppose* $(\widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Gamma}})$ *is a solution from solving (4). Let* $\widehat{\mathcal{H}} = \{i; \|\widehat{\boldsymbol{\Gamma}}_i\|_F \neq 0, i = 1, \ldots, n\}$ *and* $h = h(\lambda) = |\widehat{\mathcal{H}}|$*. Then solving (4) is equivalent to*

$$(\widehat{\boldsymbol{\Theta}}, \widehat{\mathcal{H}}) = \arg \max_{\boldsymbol{\Theta} \in \Omega, \mathcal{H}: |\mathcal{H}| = h} \left[ \sum_{i \notin \mathcal{H}} \log \left\{ \sum_{k=1}^{K} \pi_k \phi(\mathbf{y}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\} + \sum_{i \in \mathcal{H}} \log \left\{ \sum_{k=1}^{K} \pi_k \phi(\mathbf{y}_i; \mathbf{y}_i, \boldsymbol{\Sigma}_k) \right\} \right]. \tag{5}$$

Consider first the special case of equal and known covariances. The second term in (5) becomes a constant $h \log\{\phi(\mathbf{0}; \mathbf{0}, \boldsymbol{\Sigma})\}$. Consequently, it is easy to see that our method becomes exactly the same as the trimmed likelihood approach, in which $h$ observations are completely discarded to achieve the largest likelihood value possible with the remaining $(n-h)$ observations. More generally, when the covariance matrices are unknown, our method still searches for $h$ "outlying" observations, each of which is then modeled in a case-specific way, i.e., $\mathbf{y}_i \sim \sum_{k=1}^{K} \pi_k \phi(\cdot; \mathbf{y}_i, \boldsymbol{\Sigma}_k)$, for each $i \in \mathcal{H}$. Alternatively, since $\sum_{i \in \mathcal{H}} \log\{\sum_{k=1}^{K} \pi_k \phi(\mathbf{y}_i; \mathbf{y}_i, \boldsymbol{\Sigma}_k)\} = h \log\{\sum_{k=1}^{K} \pi_k \phi(\mathbf{0}; \mathbf{0}, \boldsymbol{\Sigma}_k)\}$, it can also be viewed that the $h$ original observations are replaced by $h$ many $\mathbf{0}$'s from the scale mixture $\sum_{k=1}^{K} \pi_k \phi(\cdot; \mathbf{0}, \boldsymbol{\Sigma}_k)$ ($\mathbf{0}$ is its mode). As such, the second term in (5) essentially becomes a penalty term on the scales of the mixture components. Intuitively, the $h$ observations picked up by the method tend to be the isolated points in the low-density areas of the mixture data clouds, in order to achieve a compact mixture structure to best fit the rest of the data. In the protein application, we have restricted $\boldsymbol{\gamma}_{i(k)} = \mathbf{0}$ for all $k \neq 1$, so that the $h$ observations are modeled as $\mathbf{y}_i \sim \pi_1 \phi(\cdot; \mathbf{y}_i, \boldsymbol{\Sigma}_k) + \sum_{k=2}^{K} \pi_k \phi(\cdot; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_k)$; as such, the method aims to identify the discordant proteins around the reference component.

Our method can be further reformulated through using an assignment function $z(; \boldsymbol{\Theta})$, which assigns each observation to either the regular mixture model $f(\mathbf{y}; \boldsymbol{\Theta}) = \sum_{k=1}^{K} \pi_k \phi(\mathbf{y}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ or its case-specific scale mixture model $\sum_{k=1}^{K} \pi_k \phi(\mathbf{y}; \mathbf{y}, \boldsymbol{\Sigma}_k)$ (the same as $\sum_{k=1}^{K} \pi_k \phi(\mathbf{0}; \mathbf{0}, \boldsymbol{\Sigma}_k)$). Let $\alpha = h/n$ be the proportion of identified "outliers". Let $P_n$ be the empirical measure $P_n(\cdot) = \sum_{i=1}^{n} \delta_{\mathbf{y}_i}(\cdot)/n$ where $\delta_{\mathbf{y}}$ is the Dirac measure. Define the distribution function of $f$ as $G(u; \boldsymbol{\Theta}, P_n) = P_n(f(\cdot; \boldsymbol{\Theta}) \leq u)$ and the quantile function of $f$ as $R_\alpha(\boldsymbol{\Theta}; P_n) = \inf_u \{G(u; \boldsymbol{\Theta}, P_n) \geq \alpha\}$.

**Lemma 2.2.** *Suppose* $(\widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Gamma}})$ *is a solution from solving (4). Let* $\widehat{\mathcal{H}} = \{i; \|\widehat{\boldsymbol{\Gamma}}_i\|_F \neq 0, i = 1, \ldots, n\}$ *and* $h = h(\lambda) = |\widehat{\mathcal{H}}|$ *and* $\alpha = h/n$*. Then solving (4) is equivalent to*

$$\widehat{\boldsymbol{\Theta}} = \arg \max_{\boldsymbol{\Theta} \in \Omega} L(\boldsymbol{\Theta}, P_n) \equiv \mathbb{E}_{P_n} \left[ z(\cdot; \boldsymbol{\Theta}) \log\{\sum_{k=1}^{K} \pi_k \phi(\cdot; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\} + \alpha \log\{\sum_{k=1}^{K} \pi_k \phi(\mathbf{0}; \mathbf{0}, \boldsymbol{\Sigma}_k)\} \right], \tag{6}$$

*where* $z(\cdot; \boldsymbol{\Theta}) = I\{f(\cdot; \boldsymbol{\Theta}) \geq R_\alpha(\boldsymbol{\Theta}; P_n)\}$.

The above result is from the formulation established in Lemma 2.1 together with the simple fact that $\sum_{k=1}^{K} \pi_k \phi(\mathbf{y}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \leq \sum_{k=1}^{K} \pi_k \phi(\mathbf{0}; \mathbf{0}, \boldsymbol{\Sigma}_k)$ for any $\mathbf{y}$. Therefore, in order to achieve the best partition, it must be true that the points with the smallest regular mixture likelihood values are assigned as outliers; we emphasize here that the assignment function $z(\cdot; \boldsymbol{\Theta})$ itself is a function of the model parameters. Our method thus closely relates to the mixture model approach proposed by Fraley and Raftery (1998), in which an additional mixture component is introduced to account for the noise, and the trimmed clustering approach proposed by García-Escudero *et al.* (2008), in which the "worst" points are trimmed. From Lemma 2.2, the main difference is that in our method, the outlying observations are still modeled in a case-specific and data-adaptive way. Nevertheless, our method provides a new perspective of conducting the robust clustering through the celebrated regularized estimation, while the determination of the trimming proportion then naturally translates to the problem of tuning parameter selection. Moreover, our formulation provides more flexibility on controlling for the extreme observations based on application needs; in the protein application, we are able to restrict the anomaly detection only around the reference component, in order to account for the discordant proteins.

*Proof of Lemma 2.1.* Recall $(\widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Gamma}})$ is the maximizer of (4). We can write

$$\widehat{\boldsymbol{\Gamma}} = \arg \max_{\boldsymbol{\Gamma}} \left[ \sum_{i=1}^{n} \log\{\sum_{k=1}^{K} \widehat{\pi}_k \phi(\mathbf{y}_i; \widehat{\boldsymbol{\mu}}_k + \boldsymbol{\gamma}_{i(k)}, \widehat{\boldsymbol{\Sigma}}_k)\} - \frac{\lambda^2}{2} \sum_{i=1}^{n} I(\|\boldsymbol{\Gamma}_i\|_F \neq 0) \right].$$

The above problem is separable in each $\mathbf{\Gamma}_i$, i.e.,

$$\widehat{\mathbf{\Gamma}}_i = \arg\max_{\mathbf{\Gamma}_i} \left[ \log\{\sum_{k=1}^{K} \widehat{\pi}_k \phi(\mathbf{y}_i; \widehat{\boldsymbol{\mu}}_k + \boldsymbol{\gamma}_{i(k)}, \widehat{\mathbf{\Sigma}}_k)\} - \frac{\lambda^2}{2} I(\|\mathbf{\Gamma}_i\|_F \neq 0) \right]. \tag{7}$$

If $\widehat{\mathbf{\Gamma}}_i = \mathbf{0}$, (7) becomes $\log\{\sum_{k=1}^{K} \widehat{\pi}_k \phi(\mathbf{y}_i; \widehat{\boldsymbol{\mu}}_k, \widehat{\mathbf{\Sigma}}_k)\}$, and if $\widehat{\mathbf{\Gamma}}_i \neq \mathbf{0}$, it must be true that $\widehat{\boldsymbol{\gamma}}_{i(k)} = \mathbf{y}_i - \widehat{\boldsymbol{\mu}}_k$, $k = 1, \ldots, K$, and (7) then becomes $\log\{\sum_{k=1}^{K} \widehat{\pi}_k \phi(\mathbf{y}_i; \mathbf{y}_i, \widehat{\mathbf{\Sigma}}_k)\} - \lambda^2/2$. Let $\widehat{\mathcal{H}} = \{i; \|\widehat{\mathbf{\Gamma}}_i\|_F \neq 0, i = 1, \ldots, n\}$ and $h = h(\lambda) = |\widehat{\mathcal{H}}|$. It then follows that the maximum value of the objective function in (4) is

$$\sum_{i \notin \widehat{\mathcal{H}}} \log\{\sum_{k=1}^{K} \pi_k \phi(\mathbf{y}_i; \boldsymbol{\mu}_k, \mathbf{\Sigma}_k)\} + \sum_{i \in \widehat{\mathcal{H}}} \log\{\sum_{k=1}^{K} \pi_k \phi(\mathbf{y}_i; \mathbf{y}_i, \mathbf{\Sigma}_k)\} - \frac{\lambda^2}{2}(n-h).$$

For any given $\lambda$, the number of non-zero $\widehat{\mathbf{\Gamma}}_i$ is determined and hence the third term becomes a constant. Therefore, the original problem in (4) is the same as searching for an index set of size $h$ such that the criterion in (5) is maximized. This proves the results. □

*Proof of Lemma 2.2.* From Lemma 2.1 and using the fact that $\phi(\mathbf{y}_i; \mathbf{y}_i, \mathbf{\Sigma}_k) = \phi(\mathbf{0}; \mathbf{0}, \mathbf{\Sigma}_k)$, we can write

$$\widehat{\mathcal{H}} = \arg \max_{\mathcal{H}:|\mathcal{H}|=h} \left[ \frac{1}{n} \sum_{i \notin \mathcal{H}} \log\{\sum_{k=1}^{K} \widehat{\pi}_k \phi(\mathbf{y}_i; \widehat{\boldsymbol{\mu}}_k, \widehat{\mathbf{\Sigma}}_k)\} + \frac{h}{n} \log\{\sum_{k=1}^{K} \widehat{\pi}_k \phi(\mathbf{0}; \mathbf{0}, \widehat{\mathbf{\Sigma}}_k)\} \right].$$

For any $\mathbf{\Theta}$, it always holds that $f(\mathbf{y}; \mathbf{\Theta}) = \sum_{k=1}^{K} \pi_k \phi(\mathbf{y}; \boldsymbol{\mu}_k, \mathbf{\Sigma}_k) \leq \sum_{k=1}^{K} \pi_k \phi(\mathbf{0}; \mathbf{0}, \mathbf{\Sigma}_k)$, for any $\mathbf{y}$. Therefore, at the point of solution when $\mathbf{\Theta} = \widehat{\mathbf{\Theta}}$, the estimated index set $\widehat{\mathcal{H}}$ must be corresponding to the observations at the lower $\alpha$-quartile of the values $f(\mathbf{y}_i; \widehat{\mathbf{\Theta}}) = \sum_{k=1}^{K} \widehat{\pi}_k \phi(\mathbf{y}_i; \widehat{\boldsymbol{\mu}}_k, \widehat{\mathbf{\Sigma}}_k)$. This completes the proof. □

# 3 Problem of Unbounded Likelihood

We have also shown that the problem of unbounded likelihood in our model setup is resolved under mild conditions similar to García-Escudero *et al.* (2008), so the solution of the proposed method is well defined.

**Theorem 3.1.** *Suppose A0 holds and that any $(n-h)$ points of the data $\{\mathbf{y}_i \in \mathbb{R}^m; i = 1, \ldots, n\}$ are not concentrated on less than or equal to $K$ points. Then there exists some $\mathbf{\Theta} \in \Omega$ such that the maximum of (4) is achieved.*

*Proof of Theorem 3.1.* It suffices to consider the problem in (6). We acknowledge that the proof is similar to that in García-Escudero *et al.* (2008) for the trimmed likelihood method, so we shall only sketch the main steps here. We first show that $L(\mathbf{\Theta}, P_n)$ is upper bounded by a simpler criterion. Let

$$z_k(\mathbf{y}, \mathbf{\Theta}) = I\{\max_j \pi_j \phi(\mathbf{y}; \boldsymbol{\mu}_j, \mathbf{\Sigma}_j) = \pi_k \phi(\mathbf{y}; \boldsymbol{\mu}_k, \mathbf{\Sigma}_k)\}, \text{ and } \widetilde{z}_k(\mathbf{\Theta}) = I\{\max_j \pi_j \phi(\mathbf{0}; \mathbf{0}, \mathbf{\Sigma}_j) = \pi_k \phi(\mathbf{0}; \mathbf{0}, \mathbf{\Sigma}_k)\}.$$

Then we have

$$\begin{aligned}
& z(\mathbf{y}_i; \mathbf{\Theta}) \log\{\sum_{k=1}^{K} \pi_k \phi(\mathbf{y}_i; \boldsymbol{\mu}_k, \mathbf{\Sigma}_k)\} + \alpha \log\{\sum_{k=1}^{K} \pi_k \phi(\mathbf{0}; \mathbf{0}, \mathbf{\Sigma}_k)\}, \\
\leq & z(\mathbf{y}_i; \mathbf{\Theta}) \log\{K \max_j \pi_j \phi(\mathbf{y}; \boldsymbol{\mu}_j, \mathbf{\Sigma}_j))\} + \alpha \log\{K \max_j \pi_j \phi(\mathbf{0}; \mathbf{0}, \mathbf{\Sigma}_j)\} \\
\leq & \log K + \sum_{k=1}^{K} z(\mathbf{y}_i; \mathbf{\Theta}) z_k(\mathbf{y}_i; \mathbf{\Theta}) \log(\pi_k \phi(\mathbf{y}_i; \boldsymbol{\mu}_k, \mathbf{\Sigma}_k)) + \alpha \sum_{k=1}^{K} \widetilde{z}_k(\mathbf{\Theta}) \log(\pi_k \phi(\mathbf{0}; \mathbf{0}, \mathbf{\Sigma}_k)).
\end{aligned}$$

Define $z^*(\mathbf{y}_i;\boldsymbol{\Theta}) = z(\mathbf{y}_i;\boldsymbol{\Theta})z_k(\mathbf{y}_i;\boldsymbol{\Theta}) = I[\{\mathbf{f}(\mathbf{y}_i;\boldsymbol{\Theta}) \geq R_\alpha(\boldsymbol{\Theta};P_n)\}\cap\{\max_j \pi_j\phi(\mathbf{y};\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j) = \pi_k\phi(\mathbf{y};\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\}]$. Then we have

$$L(\boldsymbol{\Theta},P_n) \leq \log K + \mathbb{E}_{P_n}\left\{\sum_{k=1}^{K} z^*(\cdot;\boldsymbol{\Theta})\log(\pi_k\phi(\cdot;\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)) + \alpha\sum_{k=1}^{K}\widetilde{z}_k(\boldsymbol{\Theta})\log(\pi_k\phi(\mathbf{0};\mathbf{0},\boldsymbol{\Sigma}_k))\right\}. \tag{8}$$

Consider a sequence $\{\boldsymbol{\Theta}^t\}_{t=1}^{\infty}$ such that

$$\lim_{t\to\infty} L(\boldsymbol{\Theta}^t,P_n) = \sup_{\boldsymbol{\Theta}\in\Omega} L(\boldsymbol{\Theta},P_n) = M > -\infty. \tag{9}$$

It can be easily verified that $L(\boldsymbol{\Theta},P_n)$ is bounded from below. Since $[0,1]^K$ is compact, we can find a subsequence of $\{\boldsymbol{\Theta}^t\}_{t=1}^{\infty}$ (denoted as the original one) such that $\pi_k^t \to \pi_k \in [0,1]$, $1\leq k\leq K$, and satisfying for some $g\in 0,1,\ldots,K$ that $\boldsymbol{\mu}_k^t \to \boldsymbol{\mu}_k$, $0\leq k\leq g$, and $\min_{k>g}\|\boldsymbol{\mu}_k^t\| \to \infty$. We then consider a further subsequence (denoted as the original one) admitting one of the following: (I) $\boldsymbol{\Sigma}_k^t \to \boldsymbol{\Sigma}_k$, $1\leq k\leq K$, (II) $\delta_{\max}^t \to \infty$, and (III) $\delta_{\min}^t \to 0$. We show that only (I) is possible. From (8), we obtain

$$\begin{aligned} L(\boldsymbol{\Theta}^t,P_n) \leq& \log K + \mathbb{E}_{P_n}\{\sum_{k=1}^{K} z_k^*(\cdot;\boldsymbol{\Theta}^t)(\log\pi_k^t - \frac{m}{2}\log 2\pi - \frac{m}{2}\log\delta_{\min}^t - \frac{1}{2}(\delta_{\max}^t)^{-1}\|\cdot-\boldsymbol{\mu}_k^t\|^2) \\ &+ \alpha\sum_{k=1}^{K}\widetilde{z}_k(\boldsymbol{\Theta}^t)(\log\pi_k^t - \frac{m}{2}\log 2\pi - \frac{m}{2}\log\delta_{\min}^t)\}. \end{aligned}$$

If $\delta_{\max}^t \to \infty$, it must be true that $\delta_{\min}^t \to \infty$ due to the eigenvalue-ratio condition, and consequently $L(\boldsymbol{\Theta}^t,P_n) \to -\infty$, leading to contradiction with (9). Now consider the case $\delta_{\min}^t \to 0$. Since

$$\begin{aligned} L(\boldsymbol{\Theta}^t,P_n) \leq& \log K + (1-\alpha)(-\frac{m}{2}\log 2\pi - \frac{m}{2}\log\delta_{\min}^t) - \frac{1}{2}(\delta_{\max}^t)^{-1}\mathbb{E}_{P_n}\sum_{k=1}^{K} z_k^*(\cdot;\boldsymbol{\Theta}^t)\|\cdot-\boldsymbol{\mu}_k^t\|^2 \\ &+ \alpha(-\frac{m}{2}\log 2\pi - \frac{m}{2}\log\delta_{\min}^t), \end{aligned}$$

it can be shown that $\mathbb{E}_{P_n}\sum_{k=1}^{K} z_k^*(\cdot;\boldsymbol{\Theta}^t)\|\cdot-\boldsymbol{\mu}_k^t\|^2 \geq c$ for some constant $c>0$, whenever any $(n-h)$ points of the data are not concentrated on less than or equal to $K$ points (García-Escudero *et al.*, 2008, Lemma A.2). It follows that

$$L(\boldsymbol{\Theta}^t,P_n) \leq \log K - \frac{m}{2}\log 2\pi - \frac{m}{2}\log\delta_{\min}^t - \frac{1}{2}(c\delta_{\min}^t)^{-1}c.$$

If $\delta_{\min}^t \to 0$, then $L(\boldsymbol{\Theta}^t,P_n) \to -\infty$, which again contradicts with (9). Therefore, $\boldsymbol{\Sigma}_k^t \to \boldsymbol{\Sigma}_k$, $1\leq k\leq K$.

If some $\pi_k = 0$, to complete the proof, we can trivially choose some $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ such that $\|\boldsymbol{\mu}_k\| < \infty$ and $\boldsymbol{\Sigma}_k$ satisfies the eigenvalue-ratio condition. It thus remains to show that when $\pi_k > 0$ for $k = 1,\ldots,K$, we must have $g = K$, so that the centers $\boldsymbol{\mu}_k^t$ all converge. That $g>0$ is obvious as otherwise $L(\boldsymbol{\Theta}^t,P_n) \to -\infty$. Following Lemma A.4 of García-Escudero *et al.* (2015), for $g>0$, we can show that for any $\mathbf{y}$,

$$\begin{aligned} 0 \leq& z(\mathbf{y};\boldsymbol{\Theta}^t)\log\{\sum_{k=1}^{K}\pi_k^t\phi(\mathbf{y};\boldsymbol{\mu}_k^t,\boldsymbol{\Sigma}_k^t)\} - z(\mathbf{y};\boldsymbol{\Theta}^t)\log\{\sum_{k=1}^{g}\pi_k^t\phi(\mathbf{y};\boldsymbol{\mu}_k^t,\boldsymbol{\Sigma}_k^t)\} \\ \leq& z(\mathbf{y};\boldsymbol{\Theta}^t)\log\left[1+\exp\{-\frac{1}{2}(\delta_{\min}^t)^{-1}\min_{g=k+1}^{K}\|\boldsymbol{\mu}_1^t-\boldsymbol{\mu}_g^t\|^2\}\{\sum_{g=k+1}^{K}\frac{\pi_g^t}{\pi_1^t}(\frac{\delta_{\max}^t}{\delta_{\min}^t})^{m/2}\}\exp\{\frac{1}{2}(\delta_{\min}^t)^{-1}\|\mathbf{y}-\boldsymbol{\mu}_1^t\|^2\}\right] \\ \to& 0 \end{aligned}$$

because of the established convergence results and the fact that $\min_{g=k+1}^{K} \|\boldsymbol{\mu}_1^t - \boldsymbol{\mu}_g^t\|^2 \to \infty$. Moreover, the above expression is uniformly dominated by a function of the form $c_1 + c_2\|\mathbf{y}\|^2$, by using the inequality $\log(1 + a\exp(y)) \leq y + \log(1 + a)$ for $y \geq 0$. It follows from the dominated convergence theorem that

$$\mathbb{E}_{P_n}\left[z(\cdot;\boldsymbol{\Theta}^t)\log\{\sum_{k=1}^{K}\pi_k^t\phi(\cdot;\boldsymbol{\mu}_k^t,\boldsymbol{\Sigma}_k^t)\}\right] - \mathbb{E}_{P_n}\left[z(\cdot;\boldsymbol{\Theta}^t)\log\{\sum_{k=1}^{g}\pi_k^t\phi(\cdot;\boldsymbol{\mu}_k^t,\boldsymbol{\Sigma}_k^t)\}\right] \to 0.$$

So,

$$\begin{aligned}\lim_{t\to\infty}\sup L(\boldsymbol{\Theta}^t, P_n) &\leq \lim_{t\to\infty}\mathbb{E}_{P_n}\left[z(\cdot;\boldsymbol{\Theta}^t)\log\{\sum_{k=1}^{g}\pi_k^t\phi(\cdot;\boldsymbol{\mu}_k^t,\boldsymbol{\Sigma}_k^t)\} + \alpha\log\{\sum_{k=1}^{K}\pi_k^t\phi(\mathbf{0};\mathbf{0},\boldsymbol{\Sigma}_k^t)\}\right] \\ &= \mathbb{E}_{P_n}\left[z(\cdot;\widetilde{\boldsymbol{\Theta}})\log\{\sum_{k=1}^{g}\pi_k\phi(\cdot;\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\} + \alpha\log\{\sum_{k=1}^{K}\pi_k\phi(\mathbf{0};\mathbf{0},\boldsymbol{\Sigma}_k)\}\right],\end{aligned}$$

where $\widetilde{\boldsymbol{\Theta}}$ is a limit of the subsequence $\{\pi_1^t,\ldots,\pi_g^t;\boldsymbol{\mu}_1^t,\ldots,\boldsymbol{\mu}_g^t;\boldsymbol{\Sigma}_1^t,\ldots,\boldsymbol{\Sigma}_g^t\}$ and $z(\cdot;\widetilde{\boldsymbol{\Theta}})$ is the assignment function which would be derived when working with the first $g$ components. Note that the second term on the right hand side does not depend on $\boldsymbol{\mu}_k$. We can then properly choose some finite $\boldsymbol{\mu}_k$, $g+1 \leq k \leq K$, to result in a strictly larger objective value. This will lead to contradiction with the optimality condition in (9). It follows that it must be true that $g = K$. This completes the proof.

□

# 4 More Details on the Protein Data

The data consist of three sets of relative intensity levels of proteins along the entire secretomes of four disease-producing strains as well as one non-disease-producing strain of *C. perfringens*, all of which are *netB* positive. The experiment was replicated 3 times. Each set represents a biological replicate.

The proteins were digested into peptides and labeled with tandem mass tags (TMT- multiplex) (McAlister *et al.*, 2012; Weekes *et al.*, 2013). The labeled peptides were identified by analysis on an Orbitrap Fusion Mass Spectrometer followed by comparison of spectra using the SEQUEST algorithm against a Uniprot composite database derived from *C. perfringens*. The tandem mass spectrometry technology used in this study was untargeted (data-independent acquisition). The peptides were quantified based on mass-to-charge ratio (m/z) fragment reporter ions after the peptide ion was isolated and analyzed in a tandem mass spectrometry (MS2) experiment. MS2 fragment ions were co-isolated and co-fragmented for increasing the number of reporter ions in the MS3 spectrum 10-fold over the standard MS3 method (i.e., MultiNotch MS3). The Peptide-spectral matches were filtered to a 1% false discovery rate.

In our data, a protein may not be observed in all three replications. Missing values in mass spectrometry datasets occur widely and can originate from a number of sources, including from both technical and biological reasons (Karpievitch *et al.*, 2012; Webb-Robertson *et al.*, 2015). Similarly in this study, the source of missing data is a combination of technical and biological aspects. Technically speaking, sometimes the abundance of a peptide is below the instrument's detection limits and hence it appears as a missing value; or, a peptide cannot be balanced by the alignment of the precursor maps, leading to missing values (Lazar *et al.*, 2016). In our protein data, a small fraction of observations contained relative intensity values of zero. To avoid unbounded LIR values, these proteins were not used in the statistical analysis, and have to be directly examined by the biologists. Alternatively, one could replace the zero relative intensity values by a small positive number $\epsilon$ so that the LIR values would be finite, but the choice of the $\epsilon$ could be arbitrary. Since the main focus of the paper is on demonstrating the robust clustering methods, we did not pursue this approach in the paper.

# 5 Additional Simulation Results

We report additional simulation results from the null model without mean shift (Table 1), the simulation model in the paper with $\beta = 1$ (Table 2) and $\beta = 2$ (Table 3), and a simulation setup using multivariate $t$ mixture (Table 4).

Table 1: Simulation: the null model without mean-shifted points.

| | | Oracle | PenN-Mix | PenN-Mix(0) | Uni-Mix | K-Means | T-Mix | N-Mix | TrimN-Mix |
|---|---|---|---|---|---|---|---|---|---|
| | | $\eta\% = 0\%$ | | | | | | | |
| MSE($\boldsymbol{\mu}$) | mean | 0.004 | 0.004 | 0.004 | 0.147 | 0.051 | 0.005 | 0.006 | 0.005 |
| | sd | 0.002 | 0.002 | 0.002 | 0.145 | 0.020 | 0.003 | 0.003 | 0.003 |
| MSE($\boldsymbol{\Sigma}$) | mean | 0.003 | 0.003 | 0.003 | 0.260 | 0.317 | 0.280 | 0.281 | 0.289 |
| | sd | 0.001 | 0.001 | 0.001 | 0.017 | 0.011 | 0.013 | 0.013 | 0.013 |
| FNR | mean | 2.50% | 2.05% | 2.50% | 69.16% | 0.81% | 4.30% | 3.59% | 4.27% |
| | sd | 0.90% | 0.80% | 0.90% | 4.82% | 0.48% | 1.25% | 1.08% | 1.26% |
| FPR | mean | 0.44% | 0.56% | 0.44% | 0.07% | 8.13% | 0.91% | 1.17% | 0.93% |
| | sd | 0.19% | 0.21% | 0.19% | 0.08% | 1.27% | 0.30% | 0.34% | 0.32% |

Table 2: Simulation: the magnitude of the mean-shift is set to $\beta = 1$ and the probability that the data experiences a mean shift is varied, $\eta\% \in \{5\%, 10\%, 15\%\}$.

| | | Oracle | PenN-Mix | PenN-Mix(0) | Uni-Mix | K-Means | T-Mix | N-Mix | TrimN-Mix |
|---|---|---|---|---|---|---|---|---|---|
| | | $\eta\% = 5\%$, $\beta = 1$ | | | | | | | |
| MSE($\boldsymbol{\mu}$) | mean | 0.005 | 0.010 | 0.011 | 0.229 | 0.061 | 0.393 | 0.424 | 0.008 |
| | sd | 0.002 | 0.006 | 0.006 | 0.200 | 0.031 | 0.150 | 0.109 | 0.004 |
| MSE($\boldsymbol{\Sigma}$) | mean | 0.003 | 0.005 | 0.006 | 0.260 | 0.314 | 0.387 | 0.398 | 0.357 |
| | sd | 0.001 | 0.002 | 0.002 | 0.019 | 0.018 | 0.077 | 0.080 | 0.015 |
| FNR | mean | 2.40% | 1.80% | 2.40% | 71.13% | 0.85% | 2.89% | 2.60% | 14.70% |
| | sd | 1.19% | 1.00% | 1.17% | 7.23% | 0.50% | 1.15% | 1.15% | 2.00% |
| FPR | mean | 0.41% | 1.86% | 1.73% | 0.10% | 9.02% | 8.43% | 9.08% | 0.90% |
| | sd | 0.19% | 0.61% | 0.53% | 0.09% | 2.94% | 2.47% | 1.39% | 0.39% |
| | | $\eta\% = 10\%$, $\beta = 1$ | | | | | | | |
| MSE($\boldsymbol{\mu}$) | mean | 0.004 | 0.014 | 0.022 | 0.325 | 0.053 | 0.779 | 0.766 | 0.009 |
| | sd | 0.002 | 0.008 | 0.017 | 0.234 | 0.027 | 0.167 | 0.149 | 0.005 |
| MSE($\boldsymbol{\Sigma}$) | mean | 0.003 | 0.006 | 0.009 | 0.250 | 0.302 | 0.496 | 0.481 | 0.389 |
| | sd | 0.001 | 0.002 | 0.004 | 0.021 | 0.017 | 0.099 | 0.110 | 0.015 |
| FNR | mean | 2.44% | 1.79% | 2.60% | 70.89% | 0.87% | 4.68% | 4.09% | 23.42% |
| | sd | 0.92% | 0.78% | 0.87% | 8.72% | 0.49% | 2.05% | 2.42% | 3.36% |
| FPR | mean | 0.43% | 2.55% | 2.54% | 0.10% | 8.38% | 15.00% | 15.36% | 0.80% |
| | sd | 0.17% | 0.84% | 0.97% | 0.11% | 2.79% | 2.13% | 1.27% | 0.35% |
| | | $\eta\% = 15\%$, $\beta = 1$ | | | | | | | |
| MSE($\boldsymbol{\mu}$) | mean | 0.004 | 0.014 | 0.025 | 0.387 | 0.051 | 1.069 | 1.146 | 0.014 |
| | sd | 0.002 | 0.009 | 0.024 | 0.271 | 0.027 | 0.929 | 0.714 | 0.014 |
| MSE($\boldsymbol{\Sigma}$) | mean | 0.003 | 0.006 | 0.011 | 0.247 | 0.296 | 0.569 | 0.580 | 0.420 |
| | sd | 0.001 | 0.002 | 0.007 | 0.017 | 0.016 | 0.403 | 0.339 | 0.015 |
| FNR | mean | 2.38% | 1.84% | 2.83% | 69.79% | 0.84% | 11.94% | 8.99% | 31.33% |
| | sd | 1.02% | 0.83% | 1.13% | 9.48% | 0.46% | 15.80% | 12.99% | 4.12% |
| FPR | mean | 0.41% | 2.65% | 2.74% | 0.10% | 8.12% | 17.64% | 20.91% | 0.64% |
| | sd | 0.16% | 1.03% | 1.49% | 0.12% | 3.20% | 7.95% | 3.79% | 0.34% |

Table 3: Simulation: the magnitude of the mean-shift is set to $\beta = 2$ and the probability that the data experiences a mean shift is varied, $\eta\% \in \{5\%, 10\%, 15\%\}$.

| | | Oracle | PenN-Mix | PenN-Mix(0) | Uni-Mix | K-Means | T-Mix | N-Mix | TrimN-Mix |
|---|---|---|---|---|---|---|---|---|---|
| | | $\eta\% = 5\%$, $\beta = 2$ | | | | | | | |
| MSE($\boldsymbol{\mu}$) | mean | 0.004 | 0.033 | 0.133 | 0.106 | 0.075 | 0.050 | 0.293 | 0.007 |
| | sd | 0.002 | 0.023 | 0.043 | 0.083 | 0.187 | 0.021 | 0.406 | 0.003 |
| MSE($\boldsymbol{\Sigma}$) | mean | 0.003 | 0.051 | 0.179 | 0.268 | 0.291 | 0.392 | 0.324 | 0.327 |
| | sd | 0.001 | 0.035 | 0.056 | 0.017 | 0.016 | 0.224 | 0.272 | 0.017 |
| FNR | mean | 2.44% | 2.17% | 2.86% | 64.65% | 0.77% | 2.50% | 5.02% | 8.93% |
| | sd | 0.93% | 0.88% | 1.08% | 7.93% | 0.47% | 1.18% | 8.30% | 2.50% |
| FPR | mean | 0.41% | 3.19% | 5.44% | 0.15% | 9.71% | 6.82% | 6.87% | 0.83% |
| | sd | 0.17% | 1.02% | 0.56% | 0.12% | 10.76% | 0.83% | 0.75% | 0.32% |
| | | $\eta\% = 10\%$, $\beta = 2$ | | | | | | | |
| MSE($\boldsymbol{\mu}$) | mean | 0.004 | 0.017 | 0.389 | 0.096 | 0.204 | 1.796 | 1.086 | 0.008 |
| | sd | 0.002 | 0.023 | 0.069 | 0.065 | 0.479 | 1.632 | 1.318 | 0.005 |
| MSE($\boldsymbol{\Sigma}$) | mean | 0.003 | 0.025 | 0.424 | 0.275 | 0.270 | 1.502 | 1.024 | 0.348 |
| | sd | 0.001 | 0.034 | 0.071 | 0.017 | 0.020 | 1.483 | 1.146 | 0.019 |
| FNR | mean | 2.43% | 2.53% | 3.76% | 60.78% | 0.81% | 27.35% | 16.96% | 13.39% |
| | sd | 1.07% | 1.05% | 1.50% | 8.06% | 0.52% | 26.15% | 21.68% | 3.43% |
| FPR | mean | 0.44% | 2.33% | 10.59% | 0.21% | 16.70% | 28.74% | 9.56% | 0.76% |
| | sd | 0.20% | 1.16% | 0.81% | 0.16% | 25.37% | 35.22% | 7.06% | 0.34% |
| | | $\eta\% = 15\%$, $\beta = 2$ | | | | | | | |
| MSE($\boldsymbol{\mu}$) | mean | 0.004 | 0.010 | 0.629 | 0.087 | 0.378 | 2.008 | 0.981 | 0.009 |
| | sd | 0.002 | 0.008 | 0.108 | 0.049 | 0.674 | 1.378 | 1.124 | 0.005 |
| MSE($\boldsymbol{\Sigma}$) | mean | 0.003 | 0.013 | 0.615 | 0.282 | 0.248 | 2.277 | 0.857 | 0.371 |
| | sd | 0.001 | 0.013 | 0.134 | 0.015 | 0.034 | 3.819 | 0.895 | 0.016 |
| FNR | mean | 2.44% | 3.06% | 4.23% | 57.07% | 0.68% | 33.27% | 15.32% | 18.23% |
| | sd | 0.98% | 1.04% | 1.55% | 7.31% | 0.54% | 18.65% | 17.86% | 3.71% |
| FPR | mean | 0.44% | 1.69% | 15.50% | 0.28% | 25.73% | 69.91% | 13.74% | 0.71% |
| | sd | 0.19% | 0.85% | 0.83% | 0.20% | 35.21% | 39.28% | 15.96% | 0.33% |

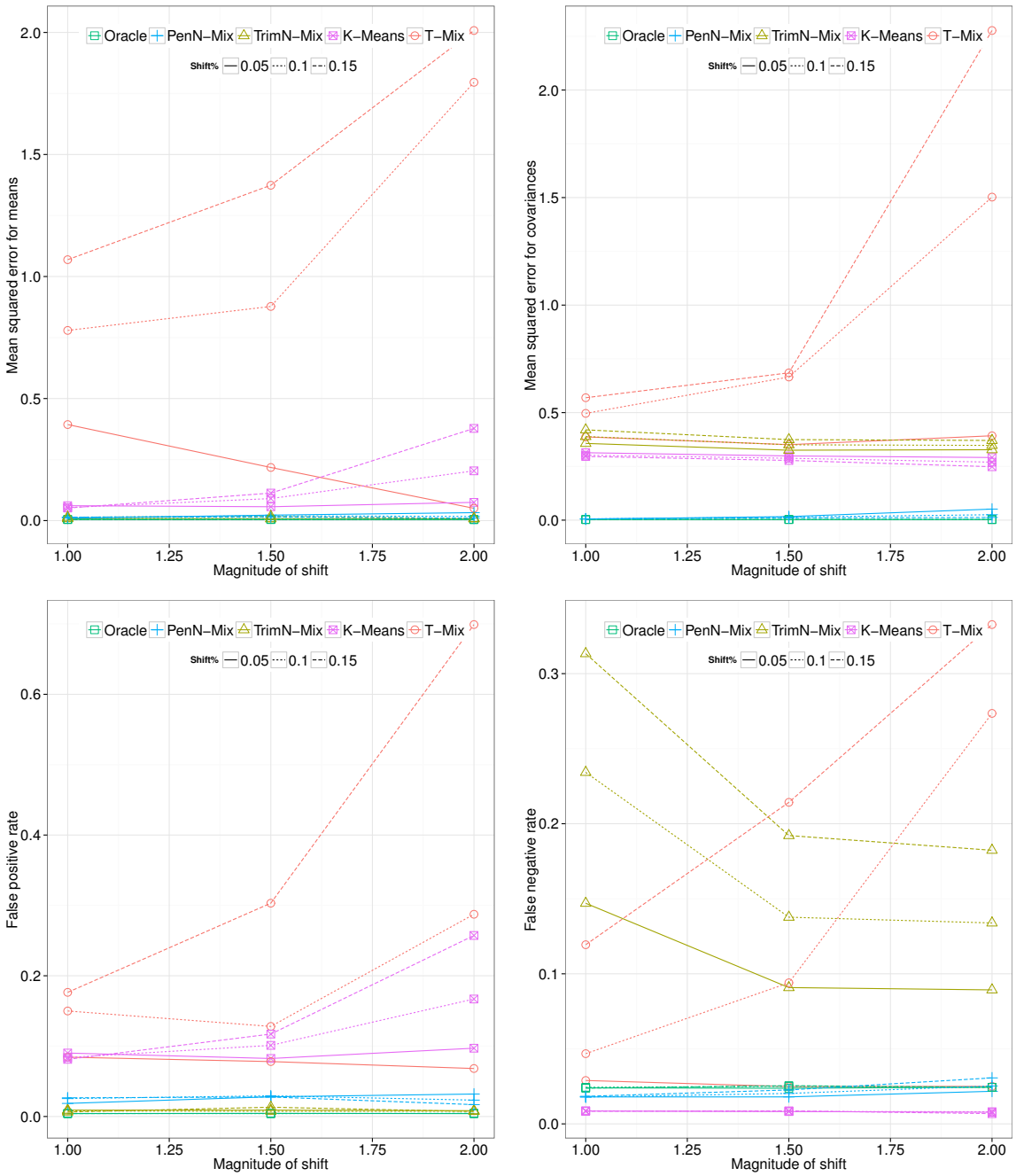


Figure 1: Simulation: performance of different methods with varying shift proportions and shift magnitudes. Oracle: the three-component normal mixture model fitted without proteins of discordant change; PenN-Mix: the proposed penalized and constrained normal mixture model; TrimN-Mix: the trimmed normal mixture model; K-Means: the K-means clustering method; T-Mix: the multivariate $t$-mixture model.

Table 4: Simulation: data are generated from multivariate $t$ distributions. Specifically, the data from the reference distribution are from multivariate $t$ with 5 degrees of freedom, while the data from the other two components are from multivariate $t$ with 10 degrees of freedom. The rest of setup is exactly the same as the multivariate normal simulation model in the paper. The magnitude of the mean-shift is set to $\beta = 2$ and the probability that the data experiences a mean shift is varied, $\eta\% \in \{5\%, 10\%, 15\%\}$.

| | | Oracle | PenN-Mix | PenN-Mix(0) | Uni-Mix | K-Means | T-Mix | N-Mix | TrimN-Mix |
|---|---|---|---|---|---|---|---|---|---|
| | | $\eta\% = 5\%$, $\beta = 2$ | | | | | | | |
| MSE($\boldsymbol{\mu}$) | mean | 0.043 | 0.137 | 0.229 | 1.251 | 0.126 | 0.212 | 0.800 | 0.013 |
| | sd | 0.015 | 0.034 | 0.048 | 0.919 | 0.217 | 0.086 | 1.013 | 0.010 |
| MSE($\boldsymbol{\Sigma}$) | mean | 0.255 | 0.434 | 0.537 | 0.228 | 0.227 | 0.249 | 0.577 | 0.228 |
| | sd | 0.063 | 0.081 | 0.084 | 0.078 | 0.013 | 0.098 | 0.672 | 0.018 |
| FNR | mean | 6.49% | 5.55% | 7.42% | 74.94% | 1.46% | 2.66% | 11.23% | 8.16% |
| | sd | 1.54% | 1.44% | 1.74% | 16.16% | 0.63% | 1.92% | 16.46% | 2.27% |
| FPR | mean | 6.79% | 12.06% | 13.44% | 1.26% | 18.59% | 15.68% | 15.40% | 5.75% |
| | sd | 0.83% | 1.13% | 1.02% | 1.11% | 11.58% | 2.27% | 1.74% | 1.26% |
| | | $\eta\% = 10\%$, $\beta = 2$ | | | | | | | |
| MSE($\boldsymbol{\mu}$) | mean | 0.040 | 0.115 | 0.456 | 0.888 | 0.218 | 3.039 | 2.243 | 0.012 |
| | sd | 0.015 | 0.044 | 0.089 | 0.730 | 0.426 | 1.480 | 1.576 | 0.015 |
| MSE($\boldsymbol{\Sigma}$) | mean | 0.242 | 0.433 | 0.747 | 0.212 | 0.205 | 1.224 | 1.651 | 0.259 |
| | sd | 0.054 | 0.098 | 0.094 | 0.046 | 0.015 | 1.034 | 1.113 | 0.021 |
| FNR | mean | 6.25% | 5.53% | 8.50% | 63.93% | 1.36% | 42.54% | 34.87% | 11.67% |
| | sd | 1.48% | 1.31% | 2.26% | 14.30% | 0.62% | 22.14% | 24.94% | 2.75% |
| FPR | mean | 6.45% | 11.73% | 18.58% | 1.87% | 23.11% | 59.56% | 16.33% | 4.85% |
| | sd | 0.87% | 1.61% | 1.10% | 1.38% | 21.60% | 35.16% | 3.24% | 1.12% |
| | | $\eta\% = 15\%$, $\beta = 2$ | | | | | | | |
| MSE($\boldsymbol{\mu}$) | mean | 0.038 | 0.072 | 0.700 | 0.680 | 0.302 | 3.444 | 2.029 | 0.123 |
| | sd | 0.015 | 0.030 | 0.107 | 0.383 | 0.536 | 0.918 | 1.456 | 0.660 |
| MSE($\boldsymbol{\Sigma}$) | mean | 0.234 | 0.378 | 0.899 | 0.218 | 0.191 | 1.864 | 1.436 | 0.286 |
| | sd | 0.054 | 0.090 | 0.104 | 0.076 | 0.021 | 1.095 | 1.013 | 0.035 |
| FNR | mean | 5.96% | 5.92% | 9.48% | 56.44% | 1.49% | 47.82% | 32.06% | 16.82% |
| | sd | 1.44% | 1.42% | 1.90% | 16.52% | 0.66% | 15.15% | 24.53% | 8.77% |
| FPR | mean | 6.29% | 10.02% | 23.74% | 2.27% | 27.38% | 87.12% | 21.05% | 6.27% |
| | sd | 0.81% | 1.61% | 1.23% | 1.23% | 27.33% | 14.37% | 14.05% | 12.92% |

# References


Fraley, C. and Raftery, A. E. (1998) How many clusters? Which clustering method? Answers via model-based cluster analysis. *The Computer Journal*, **41**, 578–588.

García-Escudero, L. A., Gordaliza, A., Matrán, C. and Mayo-Iscar, A. (2008) A general trimming approach to robust cluster analysis. *The Annals of Statistics*, **36**, 1324–1345.

García-Escudero, L. A., Gordaliza, A., Matrán, C. and Mayo-Iscar, A. (2015) Avoiding spurious local maximizers in mixture modeling. *Statistics and Computing*, **25**, 619–633.

Huang, J., Breheny, P. and Ma, S. (2012) A selective review of group selection in high dimensional models. *Statistical Science*, **27**, 481–499.

Karpievitch, Y. V., Dabney, A. R. and Smith, R. D. (2012) Normalization and missing value imputation for label-free lc-ms analysis. *BMC Bioinformatics*, **13**, S5–S5.

Lazar, C., Gatto, L., Ferro, M., Bruley, C. and Burger, T. (2016) Accounting for the multiple natures of missing values in label-free quantitative proteomics data sets to compare imputation strategies. *Journal of Proteome Research*, **15**, 1116–1125.

McAlister, G. C., Huttlin, E. L., Haas, W., Ting, L., Jedrychowski, M. P., Rogers, J. C., Kuhn, K., Pike, I., Grothe, R. A., Blethrow, J. D. and Gygi, S. P. (2012) Increasing the multiplexing capacity of tmts using reporter ion isotopologues with isobaric masses. *Analytical Chemistry*, **84**, 7469–7478.

Webb-Robertson, B.-J. M., Wiberg, H. K., Matzke, M. M., Brown, J. N., Wang, J., McDermott, J. E., Smith, R. D., Rodland, K. D., Metz, T. O., Pounds, J. G. and Waters, K. M. (2015) Review, evaluation, and discussion of the challenges of missing value imputation for mass spectrometry-based label-free global proteomics. *Journal of proteome research*, **14**, 1993–2001.

Weekes, M. P., Tan, S. Y. L., Poole, E., Talbot, S., Antrobus, R., Smith, D. L., Montag, C., Gygi, S. P., Sinclair, J. H. and Lehner, P. J. (2013) Latency-associated degradation of the mrp1 drug transporter during latent human cytomegalovirus infection. *Science*, **340**, 199–202.

Yuan, M. and Lin, Y. (2006) Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B*, **68**, 49–67.