Supplementary Material for "Clustering of Longitudinal Interval-Valued Data via Mixture Distribution under Covariance Separability"

1 Preliminary: parameter estimation in structured covariance matrices

Consider the following scalar function of a $q \times q$ positive symmetric matrix $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\theta)$ parametrized by θ

$$h(\theta; a, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_n) = a \log |\boldsymbol{\Sigma}| + \mathrm{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{S}), \tag{1}$$

where $a > 0$ and $\boldsymbol{S} = \sum_{i=1}^{n} \boldsymbol{z}_i \boldsymbol{z}_i^{\mathrm{T}}$ is a matrix derived from n vectors $\boldsymbol{z}_i$ of size q, which typically appears in the profiled likelihood of the covariance matrix $\boldsymbol{\Sigma}$ under Gaussianity. When temporal structure is posed on $\boldsymbol{\Sigma}$ such as in CS and AR models, the optimization function is much more simplified as described in what follows.

The modified Cholesky decomposition of the covariance model from an AR(1) model is the core of the computation, which is given by

$$\boldsymbol{\Sigma}_{AR}^{-1} = \sigma^{-2}\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L} = \frac{1}{\sigma^2(1-\rho^2)} \begin{bmatrix} \sqrt{1-\rho^2} & 0 & 0 & \cdots & 0 \\ -\rho & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & -\rho & 1 & 0 \\ 0 & \cdots & 0 & -\rho & 1 \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \sqrt{1-\rho^2} & 0 & 0 & \cdots & 0 \\ -\rho & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & -\rho & 1 & 0 \\ 0 & \cdots & 0 & -\rho & 1 \end{bmatrix}$$

where the (i,j)-th component of $\boldsymbol{\Sigma}_{AR} = \boldsymbol{\Sigma}_{AR}(\sigma^2, \rho)$ is $\sigma^2\rho^{|i-j|}$ with $|\rho| < 1$. However, we may let $\sigma^2 = 1$ since we do not lose any generality by fixing the $(1,1)$-th component of Σ_{AR} by 1 due to $\boldsymbol{U} \otimes \boldsymbol{V} = (a\boldsymbol{U}) \otimes (a^{-1}\boldsymbol{V})$ for $a \neq 0$. Therefore, if the likelihood is calculated over n copies of a Gaussian random variable $\boldsymbol{z} = (z_1, \ldots, z_q)^{\mathrm{T}}$, then the covariance estimation involves the optimization problem that minimizes (1) with $\theta = \rho$. Using the above decomposition, we have

$$h_{AR}(\rho\,; a, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_n) = a(q-1)\log(1-\rho^2) + (1-\rho^2)^{-1} \sum_{i=1}^{n} \left\{ z_{i1}^2(1-\rho^2) + \sum_{j=2}^{q} (z_{ij} - \rho z_{i,j-1})^2 \right\},$$

and thus, its first and second derivatives with respect to ρ are

$$\partial h_{AR}/\partial\rho = -\frac{2a(q-1)\rho}{1-\rho^2} - \frac{2(1+\rho^2)}{(1-\rho^2)^2}\sum_{i=1}^{n}\sum_{j=2}^{q} z_{ij}z_{i,j-1} + \frac{2\rho}{(1-\rho^2)^2}\sum_{i=1}^{n}\sum_{j=2}^{q}(z_{ij}^2 + z_{i,j-1}^2),$$

and

$$\partial^2 h_{AR}/\partial\rho^2 = -\frac{2a(q-1)(1+\rho^2)}{(1-\rho^2)^2} - \frac{4(3+\rho^2)}{(1-\rho^2)^3}\sum_{i=1}^{n}\sum_{j=2}^{q} z_{ij}z_{i,j-1} + \frac{2(1+3\rho^2)}{(1-\rho^2)^3}\sum_{i=1}^{n}\sum_{j=2}^{q}(z_{ij}^2 + z_{i,j-1}^2),$$

respectively.

From a compound symmetry model

$$\boldsymbol{\Sigma}_{CS}(\rho) = (1-\rho)\mathbf{I} + \rho\mathbf{1}\mathbf{1}^{\mathrm{T}},$$

which is parametrized only by ρ (i.e., $\sigma^2 = 1$) based on the same previous reasoning, we can obtain

$$\begin{aligned} h_{CS}(\rho\,; a, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_n) &= a\Big\{q\log(1-\rho) + \log\big(1 + \frac{q\rho}{1-\rho}\big)\Big\} \\ &\quad + (1-\rho)^{-1}\Big\{\sum_{i=1}^{n}\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{z}_i - \frac{\rho}{1+(q-1)\rho}\sum_{i=1}^{n}(\boldsymbol{z}_i^{\mathrm{T}}\mathbf{1})^2\Big\}, \end{aligned}$$

and its derivatives,

$$\partial h_{CS}/\partial\rho = \frac{a(q-1)}{1+(q-1)\rho} - \frac{a(q-1)}{1-\rho} + \frac{\sum_{i=1}^{n}\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{z}_i}{(1-\rho)^2} - \frac{\sum_{i=1}^{n}(\boldsymbol{z}_i^{\mathrm{T}}\mathbf{1})^2\big\{1+(q-1)\rho^2\big\}}{(1-\rho)^2\big\{1+(q-1)\rho\big\}^2}$$

and

$$\partial^2 h_{CS}/\partial\rho^2 = -\frac{a(q-1)^2}{\{1+(q-1)\rho\}^2} - \frac{a(q-1)}{(1-\rho)^2} + \frac{2\sum_{i=1}^{n}\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{z}_i}{(1-\rho)^3} - \frac{2\sum_{i=1}^{n}(\boldsymbol{z}_i^{\mathrm{T}}\mathbf{1})^2\; p_3(\rho)}{(1-\rho)^3\big\{1+(q-1)\rho\big\}^3},$$

where $p_3(\rho) = -(q-1)(1-\rho)^2 - \big\{1+(q-1)\rho\big\}\big\{1+(q-1)\rho^2\big\}$. Note that ρ should be in the interval $\big(-(q-1)^{-1}, 1\big)$ to have all positive eigenvalues of $\boldsymbol{\Sigma}_{CS}(\rho)$. The minimization problem associated with h for each covariance matrix can be solved iteratively using well-known constrained optimization techniques, one of which we use here is the log-barrier method.

2 The EM algorithm with temporally structured V

It can be seen that the only change in the EM algorithm under different covariance models given in Table 1 occurs in the covariance estimation at the M-step. First, assuming

heteroscedastic components along with K groups, we have the profiled log-likelihood of covariance matrices under separability for each group $k = 1, \ldots, K$ by

$$\sum_{i=1}^{n} w_{ik}^{(t)} \left\{ \log \hat{p}_k^{(t+1)} - \frac{q}{2} \log |\boldsymbol{U}_k| - \frac{p}{2} \log |\boldsymbol{V}_k| - \frac{1}{2} \mathrm{tr}\Big(\boldsymbol{U}_k^{-1} \big(\boldsymbol{Y}_i - \widehat{\boldsymbol{M}}_k^{(t+1)} \big) \boldsymbol{V}_k^{-1} \big(\boldsymbol{Y}_i - \widehat{\boldsymbol{M}}_k^{(t+1)} \big)^{\mathrm{T}} \Big) \right\}, \tag{2}$$

which has no explicit solutions in general. However, the alternating scheme (Dutilleul, 1999) can be applied here and has been empirically shown to work well with fast convergence based on our experiments. We alternate updating $\boldsymbol{U}_k$ and $\boldsymbol{V}_k$ as follows until convergence is met;

$$\begin{aligned} \mathrm{vec}(\boldsymbol{U}_k) &= \Big(\sum_{i=1}^{n} \boldsymbol{H}_{ik}^{(t)} \Big/ q \sum_{i=1}^{n} w_{ik}^{(t)} \Big) \mathrm{vec}(\boldsymbol{V}_k^{-1}), \\ \boldsymbol{V}_k &= \begin{cases} \mathrm{mat}\Big(\Big(\sum_{i=1}^{n} \boldsymbol{H}_{ik}^{(t)} \Big/ p \sum_{i=1}^{n} w_{ik}^{(t)} \Big)^{\mathrm{T}} \mathrm{vec}(\boldsymbol{U}_k^{-1}) \Big), & \text{if the model is UN,} \\ \boldsymbol{\Sigma}_{AR}(\hat{\rho}_k) & \text{if the model is AR,} \\ \quad \text{with } \hat{\rho}_k = \underset{\rho}{\mathrm{argmin}}\, h_{AR}\Big(\rho \,; p \textstyle\sum_{i=1}^{n} w_{ik}^{(t)}, \{\tilde{\boldsymbol{y}}_{ik\ell}\}_{1 \le i \le n, 1 \le \ell \le p} \Big), & \\ \boldsymbol{\Sigma}_{CS}(\hat{\rho}_k) & \text{if the model is CS,} \\ \quad \text{with } \hat{\rho}_k = \underset{\rho}{\mathrm{argmin}}\, h_{CS}\Big(\rho \,; p \textstyle\sum_{i=1}^{n} w_{ik}^{(t)}, \{\tilde{\boldsymbol{y}}_{ik\ell}\}_{1 \le i \le n, 1 \le \ell \le p} \Big), & \end{cases} \end{aligned} \tag{3}$$

where the operator "mat" is an inversion of vectorization associated with "vec", and let $\boldsymbol{H}_{ik}^{(t)} = \sum_{i=1}^{n} w_{ik}^{(t)} (\boldsymbol{Y}_i - \widehat{\boldsymbol{M}}_k) \otimes (\boldsymbol{Y}_i - \widehat{\boldsymbol{M}}_k)$ and a $q \times 1$ vector $\tilde{\boldsymbol{y}}_{ik\ell}$, $\ell = 1, \ldots, p$ be the ℓ-th row vector from $\sqrt{w_{ik}^{(t)}} (\boldsymbol{Y}_i - \widehat{\boldsymbol{M}}_k)$ post-multiplied by $\boldsymbol{U}_k^{-1/2}$ after transposition. It should be remarked that another loop to optimize temporal parameters ρ_k is embedded within the outer loop of the EM algorithm, which may cause the overall computation to be slow. However, an additional experiment not shown here reports that running all covariance models with $K \in \{1, \ldots, 5\}$ only takes a moderate amount of execution time, approximately an average of 94.12 with a standard deviation of 17.9 in seconds over 100 repetitions (data are generated from one of the cases in the simulation study).

When group components are homoscedastic, the estimates for common covariances are

given by

$$
\begin{aligned}
\mathrm{vec}(\boldsymbol{U}) &= \Big(\sum_{i=1}^{n}\sum_{k=1}^{K}\boldsymbol{H}_{ik}^{(t)}/nq\Big)\mathrm{vec}(\boldsymbol{V}^{-1}),\\
\boldsymbol{V} &= \begin{cases}
\mathrm{mat}\Big(\Big(\sum_{i=1}^{n}\sum_{k=1}^{K}\boldsymbol{H}_{ik}^{(t)}/np\Big)^{\mathrm{T}}\mathrm{vec}(\boldsymbol{U}^{-1})\Big), & \text{if the model is UN,}\\
\boldsymbol{\Sigma}_{AR}(\hat{\rho}) & \text{if the model is AR,}\\
\quad \text{with } \hat{\rho} = \underset{\rho}{\mathrm{argmin}}\, h_{AR}\Big(\rho\,; np, \{\tilde{\boldsymbol{y}}_{ik\ell}\}_{1\le i\le n, 1\le k\le K, 1\le \ell\le p}\Big), & \\
\boldsymbol{\Sigma}_{CS}(\hat{\rho}) & \text{if the model is CS,}\\
\quad \text{with } \hat{\rho} = \underset{\rho}{\mathrm{argmin}}\, h_{CS}\Big(\rho\,; np, \{\tilde{\boldsymbol{y}}_{ik\ell}\}_{1\le i\le n, 1\le k\le K, 1\le \ell\le p}\Big), &
\end{cases}
\end{aligned}
\tag{4}
$$

where relevant notations are defined as before.

3 Results of Numerical Study

We provide the results from the numerical study, which are omitted in the main body of the paper due to limited space. First, we present contingency tables for model selection under various settings. We mention some details commonly applied to all of the following figures. Covariance candidates are listed in rows, and the number of clusters in columns. Covariance models not selected at all through 50 repetitions are omitted in the table. The greater the frequency in each cell, the darker the blue it shows. The column labels indicate types of a true covariance model. The row label denotes the modulus of a mean vector, or c. In each figure, a sub-figure on the top is for balanced-sized clusters and the other on the bottom is for unbalanced-size clusters.

We can similarly interpret contingency tables of model selection given in Figure 1, 2, and 3, regarding covariance models and the modulus of mean vectors (i.e., c), as in Section 3. It seems slightly harder to choose the correct mixture models (covariance models and the number of groups) if the cluster size varies, which, however, is not significant to be generalized. We note that signals are more condensed in μ^{step} than in μ^{one}, so a smaller magnitude ($c = 2, 4$) of the length of the mean vector μ^{step} is sufficient to distinguish samples. When more than two components comprise a mixture model, the level of separation should be larger than before since more overlapping area would be expected among them. This is why c is set to be larger in the $K = 5$ case than in $K = 2$. However, selection of the covariance matrix is less focused in the true model, which we conjecture occurs because there are more local minima. The general solution of this phenomenon is to initialize the EM algorithm from multiple starting points and to use one of the results based on the information criterion.

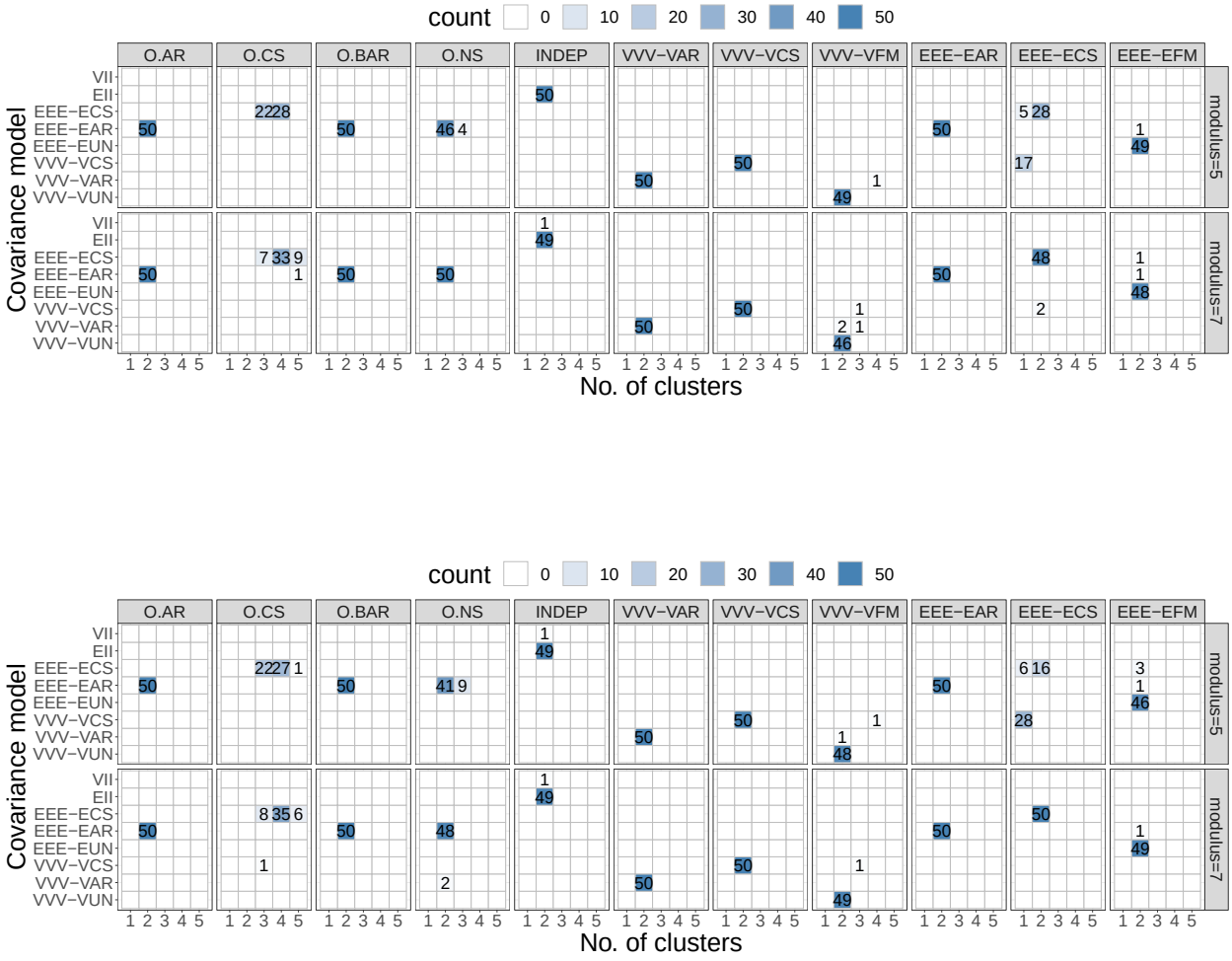

Figure 1: Contingency table of selected models when $K = 2, \mu^{one}(c)$, and balanced (top) or unbalanced (bottom) groups are used.

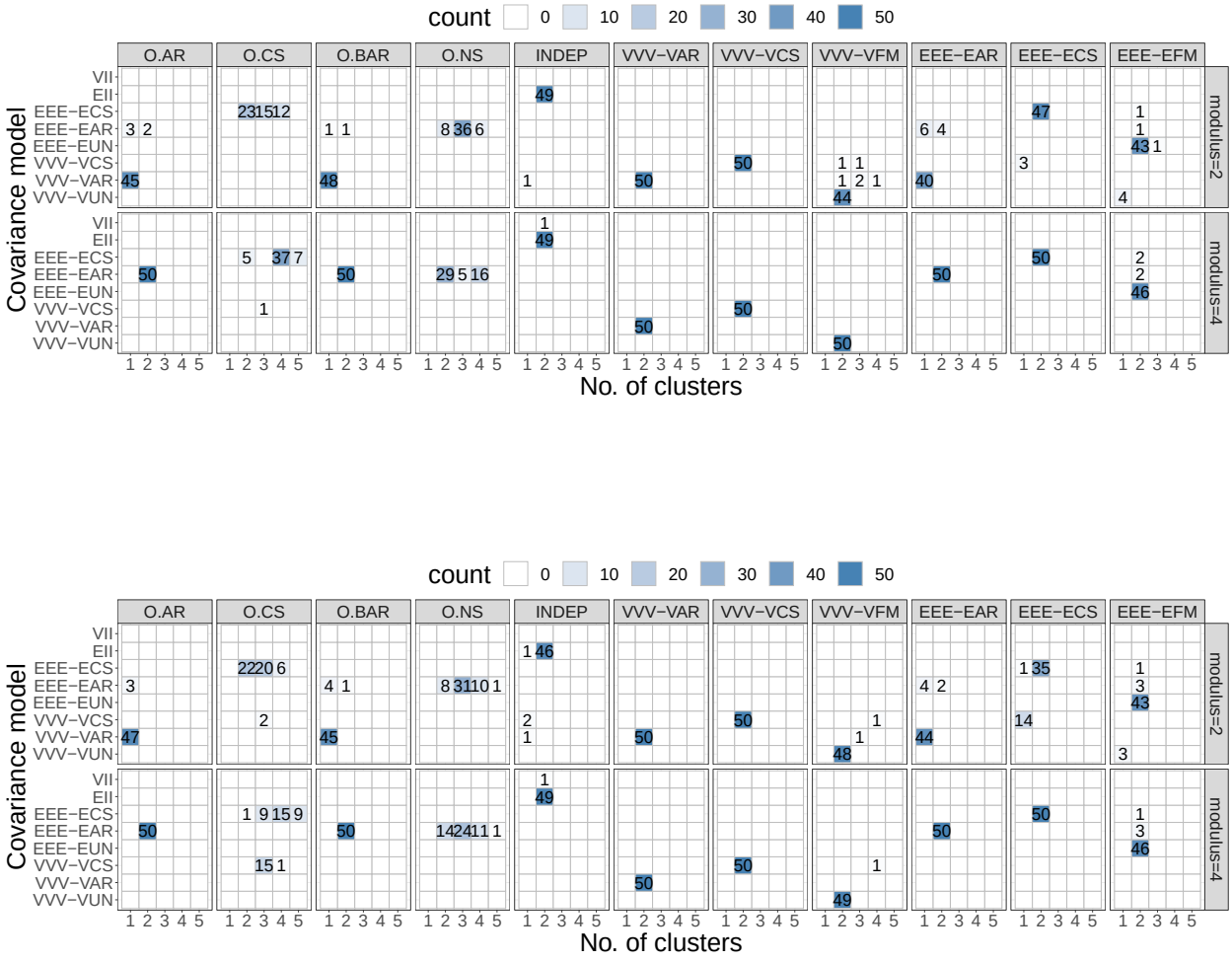

Figure 2: Contingency table of selected models when $K = 2, \mu^{step}(c)$, and balanced (top) or unbalanced (bottom) groups are used.

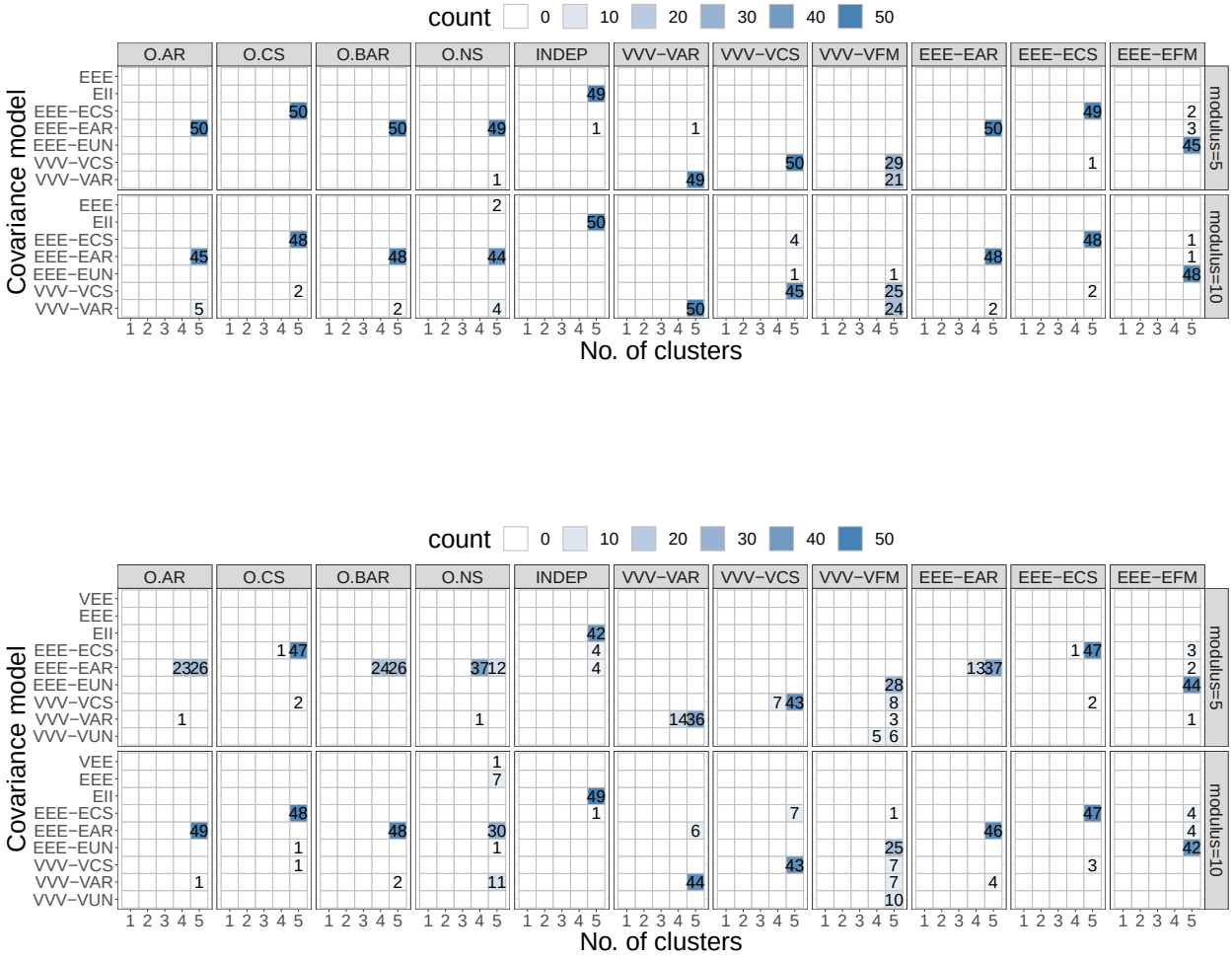

Figure 3: Contingency table of selected models when $K = 5$ and balanced (top) or unbalanced (bottom) groups are used.

Next, results for the identification of clusters are given in the following Figure 4, 5, and 6. As noted in Section 3, we compute the best accuracy up to relabeling of the estimated membership. We mention some details commonly applied to all of the following figures. Comparative methods are "Mclust" from Fraley et al. (2012) (red, left) and "SEP" proposed by this paper (blue, right). The column label indicates a type of true covariance model. The row label denotes the modulus of a mean vector, or c. In each figure, a sub-figure on the top is for balanced-sized clusters and the other on the bottom is for unbalanced-size clusters.

Under covariance separability, our model performs uniformly better than the other model. When nonseparable covariance models are assumed, the proposed model shows comparative or even higher accuracy, except (C2) O.CS with a mean vector μ^{one} and (C4) O.NS with a

mean vector μ^{step}. The exceptional cases occur when our model does not correctly specify the number of clusters. In other words, once the number of groups is consistently estimated, the separable structure of the covariance matrix can be an alternative robust option in mixture model-based clustering.

Figure 4: Boxplot of 50 accuracy values for cluster membership when $K = 2, \mu^{one}(c)$, and balanced (top) or unbalanced (bottom) groups are used.

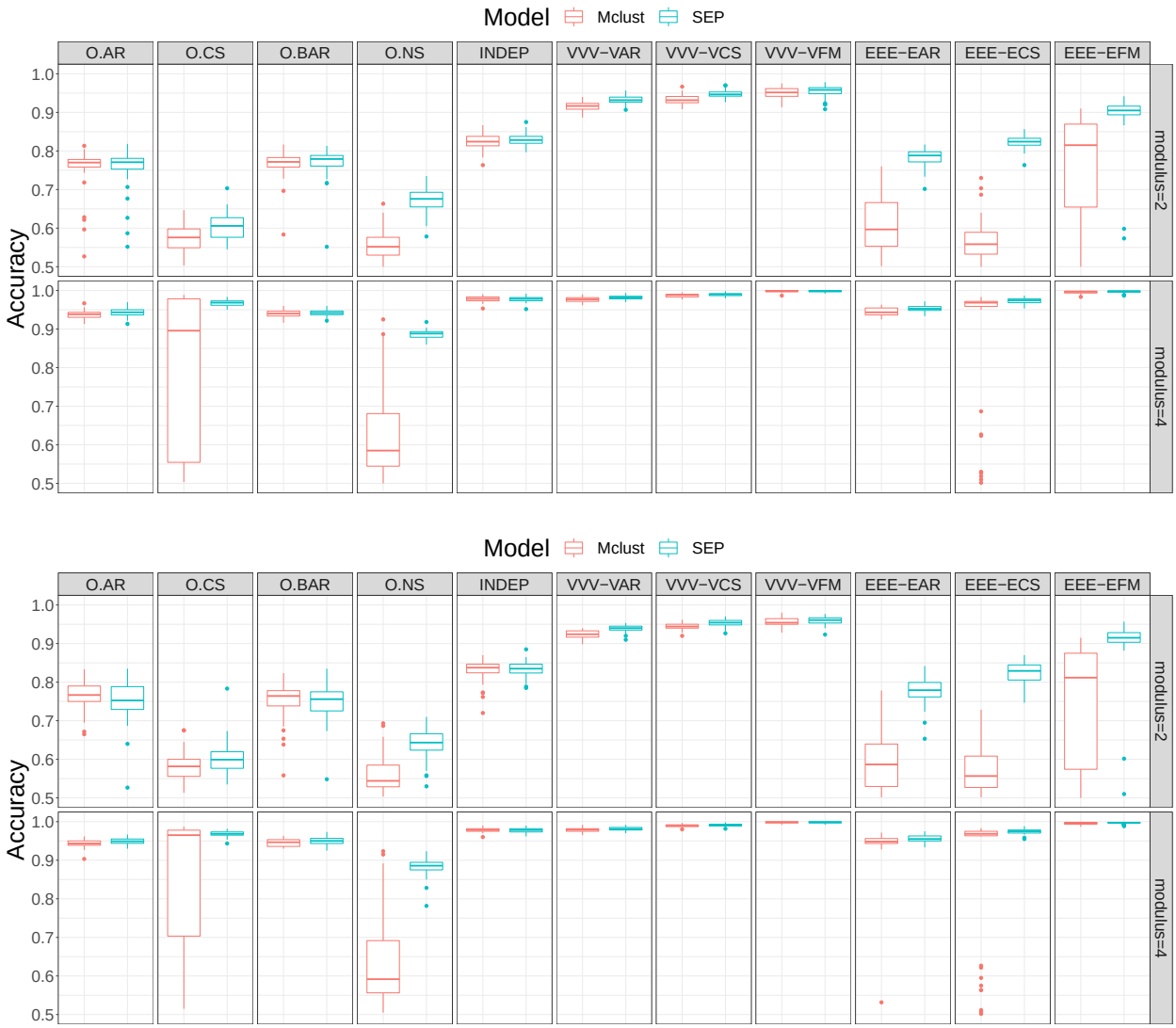

Figure 5: Boxplot of 50 accuracy values for cluster membership when $K = 2, \mu^{step}(c)$, and balanced (top) or unbalanced (bottom) groups are used.

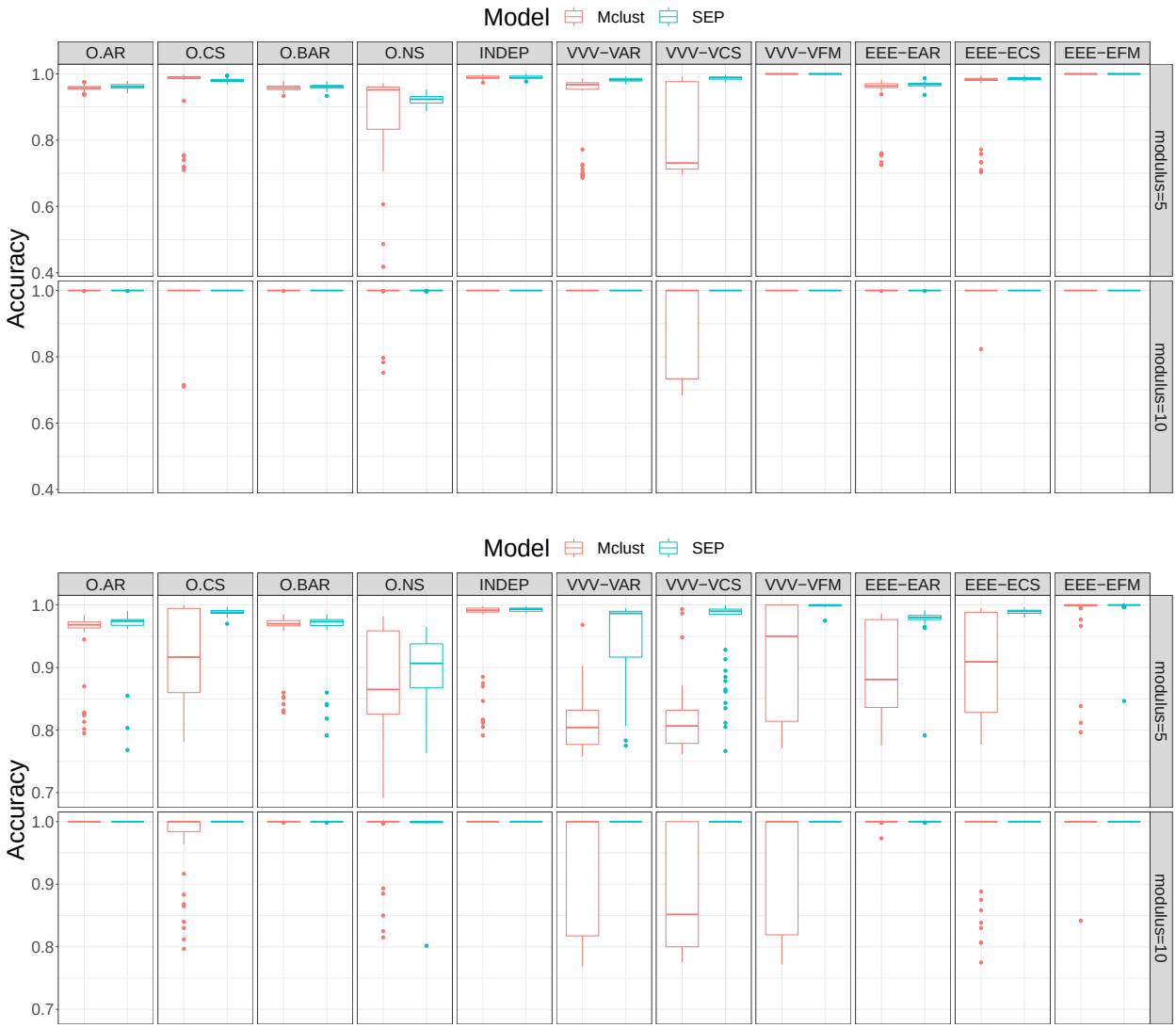

Figure 6: Boxplot of 50 accuracy values for cluster membership when $K = 5$ and balanced (top) or unbalanced (bottom) groups are used.

References

Dutilleul P. (1999). The mle algorithm for the matrix normal distribution. *Journal of Statistical Computation and Simulation*, **64**, 105-123.

Fraley, C., Raftery, A., Murphy, T., Scrucca, L. (2012). mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation. Technical Report No.597, Department of Statistics, University of Washington.