Supplementary Material: Out-of-cluster prediction for model selection in regression with unsupervised clustering

Masao Ueki*

Supplementary Appendix

Regularity conditions

Denote by Ω the parameter space common to the m-dimensional parameter vector θ.

(C1) For a given $K \in \{K_{\text{true}}, \ldots, K_{\max}\}$ and for each $k = 1, \ldots, K$, the observations V_i are independent and identically distributed with probability density $f_k^{(K)}(V; \theta_k^{(K)})$ with respect to some measure μ. The density $f_k^{(K)}(V; \theta_k^{(K)})$ has a common support and the model is identifiable. Furthermore, the first and second logarithmic derivatives of f satisfy

$$E\left[\frac{\partial \log f_k^{(K)}(V; \theta_k^{(K)})}{\partial \theta_k^{(K)}}\right] = 0$$

and

$$I_k^{(K)} = E\left[\frac{\partial \log f_k^{(K)}(V; \theta_k^{(K)})}{\partial \theta_k^{(K)}} \frac{\partial \log f_k^{(K)}(V; \theta_k^{(K)})}{\partial \theta_k^{(K)T}}\right] = E\left[-\frac{\partial^2 \log f_k^{(K)}(V; \theta_k^{(K)})}{\partial \theta_k^{(K)} \partial \theta_k^{(K)T}}\right].$$

*uekimrsd@nifty.com. School of Information and Data Sciences, Nagasaki University, 1-14 Bunkyo-Machi, Nagasaki 852-8521, Japan / RIKEN Center for Advanced Intelligence Project, Nihonbashi 1-4-1 Nihonbashi, Chuo-ku, Tokyo 103-0027, Japan.

(C2) For a given $K \in \{K_{\text{true}}, \ldots, K_{\max}\}$ and for each $k = 1, \ldots, K$, the Fisher information matrix $I_k^{(K)}$ is finite and positive definite at the true parameter $\theta_k^{(K)}$.

(C3) For a given $K \in \{K_{\text{true}}, \ldots, K_{\max}\}$ and for each $k = 1, \ldots, K$, there exists an open subset $\omega \subset \Omega$ that contains the true parameter point $\theta_k^{(K)}$ such that, for almost all V, the density $f_k^{(K)}(V; \theta_k^{(K)})$ admits all third derivatives $\partial f_k^{(K)}(V; \theta_k^{(K)}) / \partial \theta_{k,j}^{(K)} \partial \theta_{k,j'}^{(K)} \partial \theta_{k,j''}^{(K)}$ for all $\theta_k^{(K)} \in \omega$. Furthermore, there exist functions $M_{jj'j''}$ such that

$$\left| \frac{\partial^3 \log f_k^{(K)}(V; \theta_k^{(K)})}{\partial \theta_{k,j}^{(K)} \partial \theta_{k,j'}^{(K)} \partial \theta_{k,j''}^{(K)}} \right| \leq M_{jj'j''}(V),$$

for all $\theta_k \in \omega$, where $m_{jj'j''} = E[M_{jj'j''}(V)] < \infty$ for any $j, j', j'' \in \{1, \ldots, m\}$.

(C4a) For each $K \in \{1, \ldots, K_{\max}\}$, there exist $\ell_\star(K)$ such that $\frac{1}{n} \sum_{k=1}^{K} \ell(\widehat{\theta}_k; y_{\mathcal{C}_k}, X_{\mathcal{C}_k}) \to \ell_\star(K)$ and $\tau_k \in (0, 1)$ such that $\frac{n_k}{n} \to \tau_k$ as $n \to \infty$. Furthermore, for any $K < K_{\text{true}}$, $n\{\ell_\star(K_{\text{true}}) - \ell_\star(K)\} \to \infty$.

(C4b) For each $K \in \{1, \ldots, K_{\max}\}$, there exist $\ell_\star(K)$ such that $\frac{1}{n} \sum_{k=1}^{K} \ell(\widehat{\theta}_k; y_{\mathcal{C}_k}, X_{\mathcal{C}_k}) \to \ell_\star(K)$ and $\tau_k \in (0, 1)$ such that $\frac{n_k}{n} \to \tau_k$ as $n \to \infty$. Furthermore, for any $K < K_{\text{true}}$, $\frac{n}{\log n}\{\ell_\star(K_{\text{true}}) - \ell_\star(K)\} \to \infty$ as $n \to \infty$.

(C5) For a given $K \in \{1, \ldots, K_{\max}\}$, as $n \to \infty$, for each $k \in \{1, \ldots, K\}$, there exist $\bar{\ell}_k(\theta)$ such that $\frac{1}{n_k} \ell(\theta; y_{\mathcal{C}_k}, X_{\mathcal{C}_k}) \to \bar{\ell}_k(\theta)$ and $\tau_k \in (0, 1)$ such that $\frac{n_k}{n} \to \tau_k$, where $n_k = |\mathcal{C}_k|$. Furthermore, for any $K \geq K_{\text{true}}$, there exists at least one pair (k, l) with $k \neq l$ such that $\theta_k^{(K)} \neq \theta_l^{(K)}$ and $\frac{n_k}{\log n_k}\{\bar{\ell}_k(\theta_k^{(K)}) - \bar{\ell}_k(\theta_l^{(K)})\} \to \infty$.

(C5') For a given $K \in \{1, \ldots, K_{\max}\}$, as $n \to \infty$, for each $k \in \{1, \ldots, K\}$, there exist $\bar{\ell}_k(\theta)$ such that $\frac{1}{n_k} \ell(\theta; y_{\mathcal{C}_k}, X_{\mathcal{C}_k}) \to \bar{\ell}_k(\theta)$ and $\tau_k \in (0, 1)$ such that $\frac{n_k}{n} \to \tau_k$, where $n_k = |\mathcal{C}_k|$. Furthermore, for any $K \geq K_{\text{true}}$, there exists at least one pair (k, l) with $k \neq l$ such that $\theta_k^{(K)} \neq \theta_l^{(K)}$, and $\min[\frac{n_k}{\log n_k}\{\bar{\ell}_k(\theta_k^{(K)}) - \bar{\ell}_k(\theta_l^{(K)})\}, \frac{n_l}{\log n_l}\{\bar{\ell}_l(\theta_l^{(K)}) - \bar{\ell}_l(\theta_k^{(K)})\}] \to \infty$.

The conditions (C1)–(C3) apply only to the true and overfitted models and are analogous to Conditions (A)–(C) of [1], applied separately to each K-cluster model. Conditions (C4a) and (C4b) distinguish the true model from

underfitted models for AIC and BIC, respectively, with the BIC condition being stronger. Condition (C5) requires that, in any model with $K \geq K_{\text{true}}$, at least one cluster parameter $\theta_k^{(K)}$ is adequately separated from the parameter of a different cluster $\theta_l^{(K)}$. Condition (C5') strengthens (C5) by requiring mutual separation: both the kth cluster and the lth cluster must be adequately separated from each other.

Lemmas

Here, we provide some lemmas needed to prove Theorem 2. The first claim in Theorem 1 shows that AIC avoids selecting underfitted models with probability approaching 1. In contrast, the second claim indicates that the AIC at $K = K_{\text{true}}$ is not necessarily minimized, even as $n \to \infty$. To address this issue, we introduce the following lemma concerning the exclusion procedures AICex1 and AICex2.

Overfitted models with a pair (k, l) such that $\theta_k^{(K)} = \theta_l^{(K)}$: The following lemma addresses the case $\theta_k^{(K)} = \theta_l^{(K)}$, which arises when $K > K_{\text{true}}$, i.e., in overfitted models.

Lemma 1. *Suppose the likelihood function satisfies conditions (C1)–(C3). For a given $K > K_{\text{true}}$ and a pair (k, l) with $k \neq l \in \{1, \ldots, K\}$ and $\theta_k^{(K)} = \theta_l^{(K)}$, we have*

$$P\left\{-2\ell(\widehat{\theta}_k^{(K)}; X_{\mathcal{C}_k}, y_{\mathcal{C}_k}) + m \log n_k > -2\ell(\widehat{\theta}_l^{(K)}; X_{\mathcal{C}_k}, y_{\mathcal{C}_k})\right\} \to 1,$$

where $n_k = |\mathcal{C}_k|$.

This result shows that condition (2.1) is satisfied, and thus the exclusion procedure removes overfitted models containing at least one pair (k, l) with $\theta_k^{(K)} = \theta_l^{(K)}$ and $k \neq l$ (corresponding to different clusters), with probability tending to 1 as n_k increases.

Models with a pair (k, l) such that $\theta_k^{(K)} \neq \theta_l^{(K)}$: The following lemma addresses the case where $\theta_k^{(K)} \neq \theta_l^{(K)}$.

Lemma 2. *Suppose that the likelihood function satisfies conditions (C1)–(C3) and (C5). For a given $K > K_{\text{true}}$ and any pair $k, l \in \{1, \ldots, K\}$ with $k \neq l$ and $\theta_k^{(K)} \neq \theta_l^{(K)}$, it holds that*

$$P\left\{-2\ell(\widehat{\theta}_k^{(K)}; X_{\mathcal{C}_k}, y_{\mathcal{C}_k}) + m \log n_k \leq -2\ell(\widehat{\theta}_l^{(K)}; X_{\mathcal{C}_k}, y_{\mathcal{C}_k})\right\} \to 1,$$

where $n_k = |\mathcal{C}_k|$.

From the above lemma, condition (2.1) is not satisfied for the pair (k, l) such that $\theta_k^{(K)} \neq \theta_l^{(K)}$, with probability tending to 1. We assume that the true model consists of K_{true} clusters, with $\theta_k^{(K_{\text{true}})} \neq \theta_l^{(K_{\text{true}})}$ for any distinct pair (k, l), $k, l \in \{1, \ldots, K_{\text{true}}\}$. Therefore, the exclusion procedure does not eliminate the true model.

Exclusion procedures: Let $\widehat{E}_1(K)$ and $\widehat{E}_2(K)$ denote the events that the exclusion procedures (2.1) and (2.2), respectively, remove the Kth model from the sequence $\{1, \ldots, K_{\max}\}$. We obtain the following result.

Lemma 3. *Suppose that conditions (C1)–(C3) and (C5) hold. Then,*

$$P\left[\cap_{K \in \Omega_+} \widehat{E}_1(K) \cap \{\widehat{E}_1(K_{\text{true}})\}^c\right] \to 1.$$

Furthermore, if condition (C5') in the Appendix holds, then

$$P\left[\cap_{K \in \Omega_+} \widehat{E}_2(K) \cap \{\widehat{E}_2(K_{\text{true}})\}^c\right] \to 1.$$

In words, the lemma shows that the exclusion procedures discard overfitted models while retaining the true model with high probability. Consequently, after applying the exclusion procedures, the candidate set for AIC asymptotically excludes all overfitted models Ω_+ while preserving Ω_0 and Ω_-. Now, by the first claim of Theorem 1, $P\left\{\min_{K \in \Omega_-} AIC(K) > AIC(K_{\text{true}})\right\} \to 1$, which implies that $AIC(K_{\text{true}})$ attains the minimum and the true model is selected; that is, model selection consistency holds. In summary, we obtain the following theorem.

Remark 2. Note that, from Lemma 3, the exclusion procedures remove the overfitted models Ω_+ from the candidate set and retain both the true model Ω_0 and the underfitted models Ω_-, with probability tending to 1. Therefore,

after the exclusion procedures, any method that selects Ω_0 from $\Omega_0 \cup \Omega_-$ with probability tending to 1 achieves model selection consistency. This situation arises, for example, when the evaluation criterion of Ω_0 (e.g., the maximized log-likelihood) is consistently better than that of Ω_-. While Lemma 3 holds in the large-sample setting, for finite samples the overfitted models may not be fully removed from the candidate set. Thus, we recommend combining the exclusion procedure with AIC to improve practical performance.

Proof of Theorem 1

For the first claim, for any $K \in \Omega_-$,

$$
\begin{aligned}
& P\left\{ \min_{K \in \Omega_-} AIC(K) > AIC(K_{\text{true}}) \right\} \\
&= P\left\{ \min_{K \in \Omega_-} \frac{-\ell(\widehat{\theta}^{(K)}) + mK}{n} - \frac{-\ell(\widehat{\theta}^{(K_{\text{true}})}) + mK_{\text{true}}}{n} > 0 \right\} \\
&\geq P\left\{ \min_{K \in \Omega_-} \frac{\ell(\widehat{\theta}^{(K_{\text{true}})}) - \ell(\widehat{\theta}^{(K)})}{n} - \frac{mK_{\text{true}}}{n} > 0 \right\} \\
&= P\left\{ \min_{K \in \Omega_-} n\{\ell_\star(K_{\text{true}}) - \ell_\star(K) + o_p(1)\} - mK_{\text{true}} > 0 \right\} \\
&\to 1,
\end{aligned}
$$

as $n \to \infty$, because condition (C4a) holds and mK_{true} is finite. Thus, the first claim follows.

For the second claim, for any $K \in \Omega_+$, we have $K > K_{\text{true}}$. Then,

$$
\begin{aligned}
& P\left\{ \min_{K \in \Omega_+} AIC(K) > AIC(K_{\text{true}}) \right\} \\
&= P\left\{ \min_{K \in \Omega_+} \{-2\ell(\widehat{\theta}^{(K)}) + 2mK\} > -2\ell(\widehat{\theta}^{(K_{\text{true}})}) + 2mK_{\text{true}} \right\} \\
&\leq P\left\{ -2\ell(\widehat{\theta}^{(K_{\text{true}}+1)}) + 2m(K_{\text{true}}+1) > -2\ell(\widehat{\theta}^{(K_{\text{true}})}) + 2mK_{\text{true}} \right\} \\
&= P\left\{ 2m > 2\ell(\widehat{\theta}^{(K_{\text{true}}+1)}) - 2\ell(\widehat{\theta}^{(K_{\text{true}})}) \right\}.
\end{aligned}
$$

The probability in the last display is strictly less than 1, because $2\{\ell(\widehat{\theta}^{(K_{\text{true}}+1)}) - \ell(\widehat{\theta}^{(K_{\text{true}})})\}$ is asymptotically distributed as χ^2_m as $n \to \infty$. This establishes the second claim.

Proof of Lemma 1

By a Taylor expansion around $\widehat{\theta}_k^{(K)}$, and using the first-order condition $\frac{\partial \ell(\widehat{\theta}_k^{(K)}; X_{\mathcal{C}_k}, y_{\mathcal{C}_k})}{\partial \theta_k^{(K)}} = 0$, we obtain

$$\begin{aligned}
&2\big\{\ell(\widehat{\theta}_k^{(K)}; X_{\mathcal{C}_k}, y_{\mathcal{C}_k}) - \ell(\widehat{\theta}_l^{(K)}; X_{\mathcal{C}_k}, y_{\mathcal{C}_k})\big\} \\
&= -\mathrm{tr}\left[(\widehat{\theta}_l^{(K)} - \widehat{\theta}_k^{(K)})(\widehat{\theta}_l^{(K)} - \widehat{\theta}_k^{(K)})^T \frac{\partial^2 \ell(\widehat{\theta}_k^{(K)}; X_{\mathcal{C}_k}, y_{\mathcal{C}_k})}{\partial \theta_k^{(K)}\, \partial \theta_k^{(K)T}}\right] + o_p(1).
\end{aligned}$$

The expression above is of order $O_p(1)$. This follows because, under the assumption $\theta_k^{(K)} = \theta_l^{(K)}$, we have

$$\widehat{\theta}_l^{(K)} - \widehat{\theta}_k^{(K)} = (\widehat{\theta}_l^{(K)} - \theta_l^{(K)}) - (\widehat{\theta}_k^{(K)} - \theta_k^{(K)}) = O_p\{\max(n_l^{-1/2}, n_k^{-1/2})\},$$

while the observed Hessian satisfies

$$\frac{\partial^2 \ell(\widehat{\theta}_k^{(K)}; X_{\mathcal{C}_k}, y_{\mathcal{C}_k})}{\partial \theta_k^{(K)}\, \partial \theta_k^{(K)T}} = O_p(n_k).$$

Therefore,

$$-2\big\{\ell(\widehat{\theta}_k^{(K)}; X_{\mathcal{C}_k}, y_{\mathcal{C}_k}) - \ell(\widehat{\theta}_l^{(K)}; X_{\mathcal{C}_k}, y_{\mathcal{C}_k})\big\} = O_p(1),$$

which implies that it is asymptotically bounded above by $\log n_k$ as $n_k \to \infty$. This establishes the claim.

Proof of Lemma 2

Since $\widehat{\theta}_k^{(K)} \to \theta_k^{(K)}$ and $\widehat{\theta}_l^{(K)} \to \theta_l^{(K)}$ in probability,

$$\begin{aligned}
&P\left\{-2\ell(\widehat{\theta}_k^{(K)}; X_{\mathcal{C}_k}, y_{\mathcal{C}_k}) + \log n_k < -2\ell(\widehat{\theta}_l^{(K)}; X_{\mathcal{C}_k}, y_{\mathcal{C}_k})\right\} \\
&= P\left\{\frac{\ell(\widehat{\theta}_k^{(K)}; X_{\mathcal{C}_k}, y_{\mathcal{C}_k}) - \ell(\widehat{\theta}_l^{(K)}; X_{\mathcal{C}_k}, y_{\mathcal{C}_k})}{n_k} > \frac{\log n_k}{2n_k}\right\} \\
&= P\left[\frac{n_k}{\log n_k}\left\{\bar{\ell}_k(\theta_k^{(K)}) - \bar{\ell}_k(\theta_l^{(K)}) + o_p(1)\right\} > \tfrac{1}{2}\right] \\
&\to 1,
\end{aligned}$$

where the last convergence follows from condition (C5) and the fact that $\theta_k^{(K)} \neq \theta_l^{(K)}$. Hence, the lemma is proved.

Detection probability under local asymptotic regime

In Lemma 2, we assume that $\theta_k^{(K)} \neq \theta_l^{(K)}$. To clarify how sensitive the proposed procedure is to the degree of separation between clusters, $\theta_k^{(K)} - \theta_l^{(K)}$, we consider the detection probability under local asymptotic regime to elaborate the detection probability in terms of the discrepancy between $\theta_k^{(K)}$ and $\theta_l^{(K)}$.. Throughout this subsection, we drop the superscript $^{(K)}$. For a pair of cluster indexes $k \neq l$, let the corresponding true parameters θ_k and θ_l satisfy

$$\theta_l = \theta_k + \frac{h}{\sqrt{n_k}}, \tag{S.1}$$

where h is the fixed difference. Let $\mathcal{C}_k$ and $\mathcal{C}_l$ be the corresponding index sets in n samples with sample sizes n_k and n_l. Denote the log-likelihood function for the conditional density in the kth cluster be

$$\ell_k(\theta) = \sum_{i \in \mathcal{C}_k} \log f(y_i | x_i; \theta),$$

where $f(y_i|x_i;\theta)$ is the conditional density. Let $\widehat{\theta}_k$ and $\widehat{\theta}_l$ be the maximum likelihood estimator based on the the samples in clusters $\mathcal{C}_k$ and $\mathcal{C}_l$, respectively. We consider

$$D_{kl} = \ell_k(\widehat{\theta}_k) - \ell_k(\widehat{\theta}_l).$$

In addition to Conditions (C1)–(C3) and suppose that $\frac{n_k}{n_l} \to \kappa \in (0, \infty)$ as $\min(n_k, n_l) \to \infty$. Let

$$S_k(\theta) = \frac{1}{\sqrt{n_k}} \sum_{i \in \mathcal{C}_k} \frac{\partial}{\partial \theta} \log f(y|x, \theta),$$

then $S_k = S_k(\theta_k)$ is asymptotically normal with mean zero and variance of $I_k(\theta_k)$ under Conditions (C1)–(C3) and (C5). By (S.1), we have

$$\ell_k(\theta_k) - \ell_k(\theta_l) = -h^T S_k + \frac{1}{2} h^T I_k(\theta_k) h + o_p(1). \tag{S.2}$$

Similarly, the difference between the maximum log-likelihood and the log-likelihood at θ_k is given by

$$\ell_k(\widehat{\theta}_k) - \ell_k(\theta_k) = \frac{1}{2} S_k^T I_k(\theta_k)^{-1} S_k + o_p(1). \tag{S.3}$$

On the other hand, $\widehat{\theta}_l$ is constructed from the samples in the cluster $\mathcal{C}_l$. Then, $W_l = \sqrt{n_l}(\widehat{\theta}_l - \theta_l)$ is asymptotically normal with zero mean and variance of $I_l(\theta_l)^{-1}$. Furthermore, W_l and S_k is mutually independent. A second-order Taylor expansion yields

$$\ell_k(\widehat{\theta}_l) - \ell_k(\theta_l) = \frac{1}{\sqrt{n_l}} W_l^T \sqrt{n_k} S_k(\theta_l) + \frac{1}{2n_l} W_l^T \frac{\partial^2}{\partial\theta\partial\theta^T} \ell_k(\theta_l) W_l + o_p(1).$$

By (S.1), $\frac{\partial^2}{\partial\theta\partial\theta^T}\ell_k(\theta_l) = n_k\{-I_k(\theta_k) + o_p(1)\}$, and

$$S_k(\theta_l) = S_k(\theta_k) + \frac{1}{\sqrt{n_k}} \sum_{i\in\mathcal{C}_k} \frac{\partial^2}{\partial\theta\partial\theta^T} \log f(y|x,\theta_k) \frac{h}{\sqrt{n_k}} = S_k - I_k(\theta_k)h + o_p(1).$$

Thus, the above expansion reduces to

$$\ell_k(\widehat{\theta}_l) - \ell_k(\theta_l) = \sqrt{\kappa} W_l^T\{S_k - I_k(\theta_k)h\} - \frac{\kappa}{2} W_l^T I_k(\theta_k) W_k + o_p(1). \quad \text{(S.4)}$$

Now define

$$D_{kl} = \ell_k(\widehat{\theta}_k) - \ell_k(\widehat{\theta}_l).$$

Then,

$$D_{kl} = \big\{\ell_k(\theta_k) - \ell_k(\theta_l)\big\} + \big\{\ell_k(\widehat{\theta}_k) - \ell_k(\theta_k)\big\} - \big\{\ell_k(\widehat{\theta}_l) - \ell_k(\theta_l)\big\} + o_p(1).$$

Substituting (S.2), (S.3) and (S.4) above yields

$$\begin{aligned} D_{kl} &= -h^T S_k + \frac{1}{2} h^T I_k(\theta_k) h + \frac{1}{2} S_k^T I_k(\theta_k)^{-1} S_k - \sqrt{\kappa} W_l^T \big\{S_k - I_k(\theta_k)h\big\} + \frac{\kappa}{2} W_l^T I_k(\theta_k) W_l + o_p(1) \\ &= \frac{1}{2}\big\{I_k(\theta_k)^{-1} S_k - h - \sqrt{\kappa} W_l\big\}^T I_k(\theta_k) \big\{I_k(\theta_k)^{-1} S_k - h - \sqrt{\kappa} W_l\big\} + o_p(1). \end{aligned}$$

Since $I_k(\theta_k)^{-1}S_k - h - \sqrt{\kappa}W_l$ is asymptotically normal with mean $-h$ and variance of $(1+\kappa)I_k(\theta_k)^{-1}$,

$$D_{kl} \xrightarrow{d} \frac{1}{2} Z^T I_k(\theta_k) Z, \qquad Z \sim N\big(-h, (1+\kappa) I_k(\theta_k)^{-1}\big).$$

or

$$D_{kl} \xrightarrow{d} \frac{1+\kappa}{2} \chi^2_m(\lambda), \qquad \lambda = \frac{h^T I_k(\theta_k) h}{1+\kappa}.$$

Therefore, the detection probability in Lemma 2 is expressed as

$$P\{-2\ell(\widehat{\theta}_k) + m\log n_k \le -2\ell(\widehat{\theta}_l)\} = P(2D_{kl} \ge m\log n_k) \xrightarrow{d} P\{\chi^2_m(\lambda) \ge m\log n_k/(1+\kappa)\}.$$

The last display goes to 1 if $||h||^2 \to \infty$ with a faster rate than $\log n_k$, and hence the difference between clusters k and l can be detected with probability tending to 1. This is an elaborated result of Lemma 2.

Proof of Lemma 3

To prove that

$$P\left[\cap_{K\in\Omega_+}\widehat{E}_1(K)\cap\{\widehat{E}_1(K_{\text{true}})\}^c\right]\to 1,$$

we show that

$$1-P\left[\cap_{K\in\Omega_+}\widehat{E}_1(K)\cap\{\widehat{E}_1(K_{\text{true}})\}^c\right]=P\left(\left[\cap_{K\in\Omega_+}\widehat{E}_1(K)\cap\{\widehat{E}_1(K_{\text{true}})\}^c\right]^c\right)\to 0.$$

By the assumption $K_{\max}<\infty$ and Bonferroni's inequality,

$$\begin{aligned}P\left(\left[\cap_{K\in\Omega_+}\widehat{E}_1(K)\cap\{\widehat{E}_1(K_{\text{true}})\}^c\right]^c\right)&=P\left[\cup_{K\in\Omega_+}\{\widehat{E}_1(K)\}^c\cup\widehat{E}_1(K_{\text{true}})\right]\\&\le\sum_{K\in\Omega_+}P[\{\widehat{E}_1(K)\}^c]+P\{\widehat{E}_1(K_{\text{true}})\}.\end{aligned}$$

For any $K\in\Omega_+$, there exists at least one pair (k,l) such that $\theta_k^{(K)}=\theta_l^{(K)}$ with $k\neq l\in\{1,\ldots,K\}$. Therefore, from Lemma 1, the probability that no such pair satisfies (2) tends to zero, which is equivalent to $P[\{\widehat{E}_1(K)\}^c]\to 0$. Hence, the first term converges to zero.

On the other hand, for the true model with K_{true} clusters, we have $\theta_k^{(K_{\text{true}})}=\theta_l^{(K_{\text{true}})}$ for any $k\neq l\in\{1,\ldots,K_{\text{true}}\}$. Therefore, from Lemma 2, the probability that all pairs do not satisfy (2) tends to one, and hence $P\{\widehat{E}_1(K_{\text{true}})\}\to 0$.

This proves

$$P\left(\left[\cap_{K\in\Omega_+}\widehat{E}_1(K)\cap\{\widehat{E}_1(K_{\text{true}})\}^c\right]^c\right)\to 0.$$

The second claim follows analogously.

Comparison with BIC

It is well known that BIC possesses model selection consistency. We provide the following theorem regarding BIC, which is essentially a known result from previous studies [10, 8].

Theorem 3. *Suppose that the likelihood function satisfies conditions (C1)–(C3) and (C4b). Then, as* $n\to\infty$*, we have* $P\left\{\min_{K\in\Omega_-}BIC(K)>BIC(K_{\text{true}})\right\}\to 1$ *and* $P\left\{\min_{K\in\Omega_+}BIC(K)>BIC(K_{\text{true}})\right\}\to 1$.

Note that condition (C4a) imposed in Theorem 1 is replaced by condition (C4b).

Remark 3. Both BIC and the exclusion procedures (AICex1 and AICex2) exhibit model selection consistency. However, they differ in the assumptions required, namely, condition (C4b) versus conditions (C5) or (C5'). We provide an example in the section "Comparison of conditions (C4b) and (C5) in Remark 3" where condition (C4b) does not hold but condition (C5) does. In this case, BIC fails to achieve model selection consistency, whereas the exclusion procedures succeed. In brief, the example considers a case in which one cluster has a stronger signal than the others but a smaller sample size. Intuitively, BIC relies on aggregated effects across clusters, while the exclusion procedures evaluate cluster-specific effects. Thus, the exclusion procedures are expected to have greater power to detect a single strong effect than BIC. This phenomenon is confirmed in our simulation studies with scenario "k1".

Remark 4. From Remark 2, once the exclusion procedure is applied, any method that selects Ω_0 from $\Omega_0 \cup \Omega_-$ with probability tending to 1 achieves model selection consistency. This observation suggests that BIC could also serve as an alternative to AIC. However, because BIC is already known to achieve model selection consistency, no further modification is required. On the other hand, BIC may fail to identify some models that can be recovered through the exclusion procedure described in Remark 3. Therefore, we propose the combined use of the exclusion procedure with AIC, rather than with BIC.

Proof of Theorem 3

For the first claim, consider any $K \in \Omega_-$. Then,

$$
\begin{aligned}
&P\left\{\min_{K\in\Omega_-} BIC(K) - BIC(K_{\text{true}}) > 0\right\} \\
&\quad = P\left\{\min_{K\in\Omega_-} \frac{-\ell(\widehat{\theta}^{(K)}) + mK\log n}{n} - \frac{-\ell(\widehat{\theta}^{(K_{\text{true}})}) + mK_{\text{true}}\log n}{n} > 0\right\} \\
&\quad \geq P\left\{\min_{K\in\Omega_-} \frac{\ell(\widehat{\theta}^{(K_{\text{true}})}) - \ell(\widehat{\theta}^{(K)})}{n} - \frac{mK_{\text{true}}\log n}{n} > 0\right\} \\
&\quad = P\left[\min_{K\in\Omega_-} \frac{n}{\log n}\{\ell_\star(K_{\text{true}}) - \ell_\star(K) + o_p(1)\} - mK_{\text{true}} > 0\right] \\
&\quad \to 1,
\end{aligned}
$$

as $n \to \infty$, by condition (C4b), noting also that mK_{true} is finite. Hence, the first claim holds.

For the second claim, take any $K \in \Omega_+$ so that $K > K_{\text{true}}$. Then,

$$
\begin{aligned}
&P\left\{\min_{K\in\Omega_+} BIC(K) > BIC(K_{\text{true}})\right\} \\
&\quad = P\left[\min_{K\in\Omega_+} \{-2\ell(\widehat{\theta}^{(K)}) + 2mK\log n\} > -2\ell(\widehat{\theta}^{(K_{\text{true}})}) + 2mK_{\text{true}}\log n\right] \\
&\quad \geq P\left[\min_{K\in\Omega_+} \{-2\ell(\widehat{\theta}^{(K)})\} + 2m(K_{\text{true}}+1)\log n > -2\ell(\widehat{\theta}^{(K_{\text{true}})}) + 2mK_{\text{true}}\log n\right] \\
&\quad \geq P\left\{2m\log n > 2\ell(\widehat{\theta}^{(K_{\max})}) - 2\ell(\widehat{\theta}^{(K_{\text{true}})})\right\} \\
&\quad \to 1,
\end{aligned}
$$

where we use the fact that, as $n \to \infty$,

$$
2\{\ell(\widehat{\theta}^{(K_{\max})}) - \ell(\widehat{\theta}^{(K_{\text{true}})})\} \sim \chi^2_{m(K_{\max}-K_{\text{true}})}.
$$

Thus, the second claim follows.

Comparison of conditions (C4b) and (C5) in Remark 3

We show that, under some situations, condition (C5) holds if condition (C4b) holds, while condition (C4b) does not always hold in case where condition (C5) holds. We omit the superscript (K) for brevity. Suppose that

$\sqrt{\frac{n_k}{\log n_k}}||\theta_k - \theta_l|| \to \infty$ and $\theta_k - \theta_l = o(1)$ as $n_k \to \infty$. Assume that there exists $\bar{\ell}_k(\theta_k)$ such that $n_k^{-1}\ell(\theta_k; y_{\mathcal{C}_k}, X_{\mathcal{C}_k}) \to \bar{\ell}_k(\theta_k)$ in probability, and that $-\nabla\nabla^T\bar{\ell}_k(\theta_k)$ is positive definite and of order $O(1)$, where $\nabla = \frac{\partial}{\partial\theta_k}$. Then, using $\nabla\bar{\ell}_k(\theta_k) = 0$,

$$\frac{n_k}{\log n_k}\{\bar{\ell}_k(\theta_k) - \bar{\ell}_k(\theta_l)\} = -\frac{n_k}{2\log n_k}(\theta_k - \theta_l)^T\nabla\nabla^T\bar{\ell}_k(\grave{\theta}_k)(\theta_k - \theta_l) \to \infty, \quad \text{(S.5)}$$

where $\grave{\theta}_k$ is between θ_k and θ_l. Therefore, condition (C5) holds.

Now, we assume that $\theta_k - \theta_l = o(1)$ for all $k \neq l$ with $K = K_{\text{true}} > 1$, and consider comparing single cluster with the true number of clusters, K in the following. Then, for a given θ,

$$\frac{1}{n}\sum_{k=1}^{K}\ell(\theta; y_{\mathcal{C}_k}, X_{\mathcal{C}_k}) = \sum_{k=1}^{K}\frac{n_k}{n}\frac{1}{n_k}\ell(\theta; y_{\mathcal{C}_k}, X_{\mathcal{C}_k}) \to \sum_{k=1}^{K}\frac{n_k}{n}\bar{\ell}_k(\theta).$$

The last display corresponds to $\ell_\star(1)$ in condition (C4b). For an arbitrarily θ such that $\theta_k - \theta = o(1)$ for all k, using $\nabla\bar{\ell}_k(\theta_k) = 0$,

$$\ell_\star(1) = \sum_{k=1}^{K}\frac{n_k}{n}\bar{\ell}_k(\theta) = \sum_{k=1}^{K}\frac{n_k}{n}\bar{\ell}_k(\theta_k) - \frac{1}{2}\sum_{k=1}^{K}(\theta - \theta_k)^T W_k(\theta - \theta_k) + o(||\theta - \theta_k||^2),$$

where $W_k = -\frac{n_k}{n}\nabla\nabla^T\bar{\ell}_k(\theta_k)$. Note that $\sum_{k=1}^{K}\frac{n_k}{n}\bar{\ell}_k(\theta_k)$ corresponds to $\ell_\star(K)$ in condition (C4b). The maximizer with respect to θ is the least false value and is given as

$$\bar{\theta} = \left(\sum_{k=1}^{K}W_k\right)^{-1}\sum_{k=1}^{K}W_k\theta_k + o(1),$$

and hence,

$$\begin{aligned}
\ell_\star(K) - \ell_\star(1) &= \frac{1}{2}\sum_{k=1}^{K}(\theta_k - \bar{\theta})^T W_k(\theta_k - \bar{\theta}) + o(||\theta_k - \bar{\theta}||^2) \\
&= \frac{1}{2}\sum_{k=1}^{K}(\theta_k^T W_k \theta_k - 2\bar{\theta}^T W_k \theta_k + \bar{\theta}^T W_k \bar{\theta}) + o(||\theta_k - \bar{\theta}||^2) \\
&= \frac{1}{2}\sum_{k=1}^{K}\theta_k^T W_k \theta_k - \frac{1}{2}\bar{\theta}^T\left(\sum_{k=1}^{K} W_k\right)\bar{\theta} + o(||\theta_k - \bar{\theta}||^2) \\
&\leq \frac{1}{2}\sum_{k=1}^{K}\theta_k^T W_k \theta_k + o(||\theta_k - \bar{\theta}||^2). \qquad \text{(S.6)}
\end{aligned}$$

First, suppose that condition (C4b) holds, then,

$$\frac{n}{\log n}\{\ell_\star(K) - \ell_\star(1)\} \to \infty.$$

Then, by (S.6), $\frac{n}{\log n}\sum_{k=1}^{K}(\theta_k - \bar{\theta})^T W_k(\theta_k - \bar{\theta}) \to \infty$. The left-hand side is

$$\frac{n}{\log n}\sum_{k=1}^{K}(\theta_k - \bar{\theta})^T W_k(\theta_k - \bar{\theta}) = \frac{1}{\log n}\sum_{k=1}^{K}||\sqrt{n}(\theta_k - \bar{\theta})||^2 O(\tfrac{n_k}{n}),$$

because $W_k = -\frac{n_k}{n}\nabla\nabla^T\bar{\ell}_k(\theta_k) = O(\frac{n_k}{n})$. Since K is finite, there exists at least an k such that

$$\frac{1}{\log n}||\sqrt{n_k}(\theta_k - \bar{\theta})||^2 \to \infty,$$

and

$$\sqrt{\frac{n_k}{\log n}}||\theta_k - \bar{\theta}|| = \sqrt{\frac{n_k}{\log n}}\left\|\theta_k - \left(\sum_{k=1}^{K} W_k\right)^{-1}\sum_{l=1}^{K} W_l\theta_l\right\| \to \infty,$$

Now, the middle term can be written as

$$\sqrt{\frac{n_k}{\log n}}\left\|\theta_k - \left(\sum_{k=1}^{K} W_k\right)^{-1}\sum_{k=1}^{K} W_k\theta_k\right\| = \sum_{l=1}^{K} O(\tfrac{n_l}{n})\sqrt{\frac{n_k}{\log n}}\left\|\theta_k - \theta_l\right\|.$$

Thus, there is at least an $l \neq k$ such that,

$$\frac{n_l}{n}\sqrt{\frac{n_k}{\log n}}\,\|\theta_k - \theta_l\| \to \infty,$$

and since $\frac{n_l}{n} < 1$,

$$\sqrt{\frac{n_k}{\log n}}\,\|\theta_k - \theta_l\| \to \infty.$$

This implies from the argument in (S.5) that condition (C5) holds.

Second, we provide an example where condition (C4b) does not hold but condition (C5) or (C5') holds. For some integer $r \to \infty$, we set as $n_1 = r$, $n_l = \xi_r = \frac{r\lceil e^{(\log r)^\kappa}\rceil - r}{K-1}$ for $l > 1$, $\theta_1 = a_1\sqrt{\frac{\log \xi_r}{r}}$ and $\theta_l = a_l\sqrt{\frac{\log \xi_r}{\xi_r}}$ for $l > 1$, where $\kappa > 1$ is a constant, $\lceil\cdot\rceil$ is the ceiling function and $a_1, a_2, \ldots, a_K$ are some fixed constant vectors. Then, $n = n_1 + \sum_{l=2}^K n_l = r + (K-1)\xi_r = r\lceil e^{(\log r)^\kappa}\rceil$ and $\frac{n}{\log n} = \frac{n}{\log r + \log\lceil e^{(\log r)^\kappa}\rceil}$. By defining $q_l = -a_l^T \nabla\nabla^T \bar{\ell}_k(\theta_k) a_l$ for $l = 1, \ldots, K$,

$$\begin{aligned}
\frac{n}{\log n}\sum_{k=1}^K \theta_k^T W_k \theta_k &= -\frac{n}{\log n}\left\{\frac{n_1}{n}\theta_1^T \nabla\nabla^T \bar{\ell}_1(\theta_1)\theta_1 + \sum_{l=2}^K \frac{n_l}{n}\theta_l^T \nabla\nabla^T \bar{\ell}_l(\theta_l)\theta_l\right\} \\
&= \frac{n}{\log n}\left\{\frac{\log \xi_r}{r}\frac{r}{n}q_1 + \frac{\log \xi_r}{\xi_r}\frac{\xi_r}{n}\sum_{l=2}^K q_l\right\} \\
&= \frac{\log \xi_r}{\log n}q_1 + \frac{\log \xi_r}{\log n}\sum_{l=2}^K q_l.
\end{aligned}$$

Thus, by (S.6), $q_l = O(1)$ and $\kappa > 1$, as $r \to \infty$, it holds that $\frac{r}{\xi_r} = \frac{K-1}{\lceil e^{(\log r)^\kappa}\rceil - 1} \to 0$, $\frac{\log \xi_r}{\log n} = \frac{\log \xi_r}{\log\{r + (K-1)\xi_r\}} = \frac{\log \xi_r}{\log\{o(\xi_r) + (K-1)\xi_r\}} \to 1$ and

$$\frac{n}{\log n}\{\ell_\star(K) - \ell_\star(1)\} \leq \frac{n}{\log n}\left\{\sum_{k=1}^K \theta_k^T W_k \theta_k + o(1)\right\} = O(1),$$

which implies that condition (C4b) does not hold. On the other hand, as

$\frac{\log \xi_r}{\log r} = \frac{\log r + \log \frac{\lceil e^{(\log r)^\kappa} \rceil - 1}{K-1}}{\log r} = \frac{\log r + (\log r)^\kappa \{1+o(1)\}}{\log r} \to \infty$ due to $\kappa > 1$,

$$\begin{aligned}
\sqrt{\frac{n_1}{\log n_1}}||\theta_1 - \theta_l|| &= \sqrt{\frac{r}{\log r}} \left\| a_1 \sqrt{\frac{\log \xi_r}{r}} - a_l \sqrt{\frac{\log \xi_r}{\xi_r}} \right\| \\
&\geq \sqrt{\frac{r}{\log r}} \left\| a_1 \sqrt{\frac{\log \xi_r}{r}} \right\| - \sqrt{\frac{r}{\log r}} \left\| a_l \sqrt{\frac{\log \xi_r}{\xi_r}} \right\| \\
&\geq \sqrt{\frac{\log \xi_r}{\log r}} ||a_1|| - \sqrt{\frac{\log \xi_r}{\log r} \frac{r}{\xi_r}} ||a_l|| \\
&\to \infty,
\end{aligned}$$

and $\theta_1 - \theta_l = o(1)$ as $r = n_1 \to \infty$ hold and condition (C5) is satisfied. This shows that condition (C4b) does not always hold in case where condition (C5) holds. Moreover, for such an $l > 1$, as $\frac{\xi_r}{r} \to \infty$,

$$\begin{aligned}
\sqrt{\frac{n_l}{\log n_l}}||\theta_l - \theta_1|| &= \sqrt{\frac{\xi_r}{\log \xi_r}} \left\| a_1 \sqrt{\frac{\log \xi_r}{r}} - a_l \sqrt{\frac{\log \xi_r}{\xi_r}} \right\| \\
&\geq \sqrt{\frac{\xi_r}{\log \xi_r}} \left\| a_1 \sqrt{\frac{\log \xi_r}{r}} \right\| - \sqrt{\frac{\xi_r}{\log \xi_r}} \left\| a_l \sqrt{\frac{\log \xi_r}{\xi_r}} \right\| \\
&\geq \sqrt{\frac{\xi_r}{r}} ||a_1|| - ||a_l|| \\
&\to \infty,
\end{aligned}$$

which shows condition (C5') is also satisfied. This shows that condition (C4b) does not always hold in case where condition (C5') holds.

Normalized partial log-likelihood

We study the normalized partial log-likelihood function, which admits an intuitive interpretation. Let $\mu_i = x_i^T \beta$. Without loss of generality, assume that there are no ties among the observed y_i's, and that $y_1 < y_2 < \cdots < y_n$. Under this assumption, the partial log-likelihood function simplifies to

$$\ell_c(\beta; X, y, \delta) = \sum_{i=1}^n \delta_i \mu_i - \sum_{i=1}^n \delta_i \log \left(\sum_{j=i}^n e^{\mu_j} \right) = \sum_{i=1}^n \delta_i \log \left(\frac{e^{\mu_i}}{\sum_{j=i}^n e^{\mu_j}} \right). \tag{S.7}$$

We compare Cox regression models obtained from two different partitions: (i) the trivial partition $\widehat{\mathfrak{C}}^{(1)} = \{1, \ldots, n\}$, and (ii) a two-cluster partition $\widehat{\mathfrak{C}}^{(2)} = \widehat{\mathcal{C}}_1 \cup \widehat{\mathcal{C}}_2$. Assume that the observed y_i's within $\widehat{\mathcal{C}}_1$ and $\widehat{\mathcal{C}}_2$ remain ordered.

The sum of the two partial log-likelihood functions for $\widehat{\mathcal{C}}_1$ and $\widehat{\mathcal{C}}_2$ at a common parameter β is

$$\begin{aligned}
&\ell_c(\beta; X_{\widehat{\mathcal{C}}_1}, y_{\widehat{\mathcal{C}}_1}, \delta_{\widehat{\mathcal{C}}_1}) + \ell_c(\beta; X_{\widehat{\mathcal{C}}_2}, y_{\widehat{\mathcal{C}}_2}, \delta_{\widehat{\mathcal{C}}_2}) \\
&\quad = \sum_{i \in \widehat{\mathcal{C}}_1} \delta_i \log \left(\frac{e^{\mu_i}}{\sum_{j \geq i,\, j \in \widehat{\mathcal{C}}_1} e^{\mu_j}} \right) + \sum_{i \in \widehat{\mathcal{C}}_2} \delta_i \log \left(\frac{e^{\mu_i}}{\sum_{j \geq i,\, j \in \widehat{\mathcal{C}}_2} e^{\mu_j}} \right). \qquad \text{(S.8)}
\end{aligned}$$

Comparing (S.7) with (S.8), if $\sum_{j=i}^{n} e^{\mu_j}$ were equal to both $\sum_{j \geq i,\, j \in \widehat{\mathcal{C}}_1} e^{\mu_j}$ and $\sum_{j \geq i,\, j \in \widehat{\mathcal{C}}_2} e^{\mu_j}$, then the two expressions would coincide. However, the latter sums are always smaller than the former, since the full sum includes all terms as a superset. Consequently, although the parameter β is common, (S.8) is systematically larger than (S.7), which leads to divergence as n increases, as discussed in [6]. This situation differs markedly from that of the log-likelihood function for generalized linear models, where the sum of log-likelihood functions partitioned into disjoint sets is identical to the log-likelihood function for the full set $\{1, \ldots, n\}$ when evaluated at a common parameter value β.

By contrast, since $\widehat{\rho}_u n \log n = \log n \sum_{i=1}^{n} \delta_i$, the normalized version can be written as

$$\ell_{cc}(\beta; X, y, \delta) = \sum_{i=1}^{n} \delta_i \mu_i - \sum_{i=1}^{n} \delta_i \log \left(\frac{1}{n} \sum_{j=i}^{n} e^{\mu_j} \right) = \sum_{i=1}^{n} \delta_i \log \left(\frac{e^{\mu_i}}{\frac{1}{n} \sum_{j=i}^{n} e^{\mu_j}} \right), \qquad \text{(S.9)}$$

where the second term in the middle expression is of order $nO_p(1)$, by the argument in (A.7) of [6]. Similarly, the sum of the two normalized partial log-likelihoods for $\widehat{\mathcal{C}}_1$ and $\widehat{\mathcal{C}}_2$ is

$$\begin{aligned}
&\ell_{cc}(\beta; X_{\widehat{\mathcal{C}}_1}, y_{\widehat{\mathcal{C}}_1}, \delta_{\widehat{\mathcal{C}}_1}) + \ell_{cc}(\beta; X_{\widehat{\mathcal{C}}_2}, y_{\widehat{\mathcal{C}}_2}, \delta_{\widehat{\mathcal{C}}_2}) \\
&\quad = \sum_{i \in \widehat{\mathcal{C}}_1} \delta_i \mu_i - \sum_{i \in \widehat{\mathcal{C}}_1} \delta_i \log \left(\frac{1}{|\widehat{\mathcal{C}}_1|} \sum_{j \geq i,\, j \in \widehat{\mathcal{C}}_1} e^{\mu_j} \right) + \sum_{i \in \widehat{\mathcal{C}}_2} \delta_i \mu_i - \sum_{i \in \widehat{\mathcal{C}}_2} \delta_i \log \left(\frac{1}{|\widehat{\mathcal{C}}_2|} \sum_{j \geq i,\, j \in \widehat{\mathcal{C}}_2} e^{\mu_j} \right) \\
&\quad = \sum_{i \in \widehat{\mathcal{C}}_1} \delta_i \log \left(\frac{e^{\mu_i}}{\frac{1}{|\widehat{\mathcal{C}}_1|} \sum_{j \geq i,\, j \in \widehat{\mathcal{C}}_1} e^{\mu_j}} \right) + \sum_{i \in \widehat{\mathcal{C}}_2} \delta_i \log \left(\frac{e^{\mu_i}}{\frac{1}{|\widehat{\mathcal{C}}_2|} \sum_{j \geq i,\, j \in \widehat{\mathcal{C}}_2} e^{\mu_j}} \right), \qquad \text{(S.10)}
\end{aligned}$$

where the second and fourth terms on the second line are of order $|\widehat{\mathcal{C}}_1|O_p(1)$ and $|\widehat{\mathcal{C}}_2|O_p(1)$, respectively.

To summarize, the denominators in the logarithm are normalized by averages rather than sums, which prevents the systematic divergence caused by the increasing number of terms and allows the normalized partial log-likelihoods to be comparable across different sample sizes or partitions.

Theoretical results for normalized partial log-likelihood

We consider Lemmas 1 and 2 for the partial log-likelihood. Suppose that the assumptions (A)–(D) of [2] hold separately applied to K clusters as in the above regularity conditions (C1)–(C3). Then, the estimated regression coefficients of the Cox model for each cluster are asymptotically normal under both correct and overfitted models. Therefore, Lemmas 1 and 2 can be extended to the Cox model.

For Lemma 3, we simply apply the conditions (C4a), (C4b) and (C5) or (C5') to the normalized partial log-likelihood $\ell_{cc}(\widehat{\beta}_k; X_{\mathcal{C}_k}, Y_{\mathcal{C}_k}, \delta_{\mathcal{C}_k})$ instead of the log-likelihood as,

(D4a) For each $K \in \{1, \ldots, K_{\max}\}$, there exist $\ell_\star(K)$ such that $\frac{1}{n}\sum_{k=1}^{K} \ell_{cc}(\widehat{\beta}_k; X_{\mathcal{C}_k}, Y_{\mathcal{C}_k}, \delta_{\mathcal{C}_k}) \to \ell_\star(K)$ and $\tau_k \in (0,1)$ such that $\frac{n_k}{n} \to \tau_k$ as $n \to \infty$. Furthermore, for any $K < K_{\text{true}}$, $n\{\ell_\star(K_{\text{true}}) - \ell_\star(K)\} \to \infty$ as $n \to \infty$.

(D4b) For each $K \in \{1, \ldots, K_{\max}\}$, there exist $\ell_\star(K)$ such that $\frac{1}{n}\sum_{k=1}^{K} \ell_{cc}(\widehat{\beta}_k; X_{\mathcal{C}_k}, Y_{\mathcal{C}_k}, \delta_{\mathcal{C}_k}) \to \ell_\star(K)$ and $\tau_k \in (0,1)$ such that $\frac{n_k}{n} \to \tau_k$ as $n \to \infty$. Furthermore, for any $K < K_{\text{true}}$, $\frac{n}{\log n}\{\ell_\star(K_{\text{true}}) - \ell_\star(K)\} \to \infty$ as $n \to \infty$.

(D5) For a given $K \in \{1, \ldots, K_{\max}\}$, as $n \to \infty$, for each $k \in \{1, \ldots, K\}$, there exist $\bar{\ell}_k(\beta)$ such that $\frac{1}{n_k}\ell_{cc}(\beta; X_{\mathcal{C}_k}, Y_{\mathcal{C}_k}, \delta_{\mathcal{C}_k}) \to \bar{\ell}_k(\beta)$ and $\tau_k \in (0,1)$ such that $\frac{n_k}{n} \to \tau_k$, where $n_k = |\mathcal{C}_k|$. Furthermore, for any $K \geq K_{\text{true}}$, there exists at least a pair (k,l) such that $k \neq l \in \{1, \ldots, K\}$, and $\beta_k^{(K)} \neq \beta_l^{(K)}$ and $\frac{n_k}{\log n_k}\{\bar{\ell}_k(\beta_k^{(K)}) - \bar{\ell}_k(\beta_l^{(K)})\} \to \infty$.

(D5') For a given $K \in \{1, \ldots, K_{\max}\}$, as $n \to \infty$, for each $k \in \{1, \ldots, K\}$, there exist $\bar{\ell}_k(\beta)$ such that $\frac{1}{n_k}\ell_{cc}(\beta; X_{\mathcal{C}_k}, Y_{\mathcal{C}_k}, \delta_{\mathcal{C}_k}) \to \bar{\ell}_k(\beta)$ and $\tau_k \in (0,1)$ such that $\frac{n_k}{n} \to \tau_k$, where $n_k = |\mathcal{C}_k|$. Furthermore, for any $K \geq K_{\text{true}}$, there exists at least a pair (k,l) such that $k \neq l \in \{1, \ldots, K\}$, and $\beta_k^{(K)} \neq \beta_l^{(K)}$, $\min[\frac{n_k}{\log n_k}\{\bar{\ell}_k(\beta_k^{(K)}) - \bar{\ell}_k(\beta_l^{(K)})\}, \frac{n_l}{\log n_l}\{\bar{\ell}_l(\beta_l^{(K)}) - \bar{\ell}_l(\beta_k^{(K)})\}] \to \infty$.

Note that the above conditions are asymptotically analogous to the condition (E) of [6], which essentially implies that the normalized partial log-likelihood $\ell_c(\widehat{\beta}_k; X_{\mathcal{C}_k}, Y_{\mathcal{C}_k}, \delta_{\mathcal{C}_k}) + \rho_u \log n$ converges to some function, where ρ_u is the proportion of noncencosring in population.

Under the corresponding conditions among (D4a), (D4b), (D5) and (D5'), Lemmas 1–3 and Theorems 1–3 hold for the normalized partial log-likelihood.

We examine the consistency of the normalized partial log-likelihood in the K-cluster setting. Under the above conditions, equation (2.3) holds for each cluster $k = 1, \ldots, K$. Specifically, as $n \to \infty$,

$$\ell_c(\beta_k; X_{\mathcal{C}_k}, Y_{\mathcal{C}_k}, \delta_{\mathcal{C}_k}) = -\rho_u^{(k)} n_k \log n_k + n_k \bar{\ell}_k(\beta_k),$$

where $\bar{\ell}_k(\beta_k) = O(1)$ and $\rho_u^{(k)} = P(T_{\mathcal{C}_k} \leq C_{\mathcal{C}_k})$ for $k = 1, \ldots, K$. Therefore, the sum of the K normalized partial log-likelihoods,

$$\ell_{cc}(\mathfrak{C}^{(K)}) = \sum_{k=1}^{K} \left\{ \ell_c(\beta_k; X_{\mathcal{C}_k}, Y_{\mathcal{C}_k}, \delta_{\mathcal{C}_k}) + \widehat{\rho}_u^{(k)} n_k \log n_k \right\},$$

is asymptotically equivalent to $\sum_{k=1}^{K} n_k \bar{\ell}_k(\beta_k)$ and hence, the normalization remains valid as $n \to \infty$.

Equivalence between CCMP and Mallows' C_p

Let $z_{ik} = I_{\{i \in \widehat{\mathcal{C}}_k\}}$. Then,

$$\begin{aligned}
n\widehat{\sigma}^2 &= \sum_{i=1}^{n}(y_i - \bar{y})^2 \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K}(y_i - \bar{y})^2 z_{ik} \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K}(y_i - \bar{y}_k + \bar{y}_k - \bar{y})^2 z_{ik} \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K}\{(y_i - \bar{y}_k)^2 + 2(y_i - \bar{y}_k)(\bar{y}_k - \bar{y}) + (\bar{y}_k - \bar{y})^2\} z_{ik} \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K}(y_i - \bar{y}_k)^2 z_{ik} + \sum_{i=1}^{n}\sum_{k=1}^{K}(\bar{y}_k - \bar{y})^2 z_{ik} \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K}(y_i - \bar{y}_k)^2 z_{ik} + \sum_{k=1}^{K} n_k(\bar{y}_k - \bar{y})^2,
\end{aligned}$$

which gives

$$-\sum_{k=1}^{K} n_k(\bar{y}_k - \bar{y})^2 = \sum_{i=1}^{n}\sum_{k=1}^{K}(y_i - \bar{y}_k)^2 z_{ik} - n\widehat{\sigma}^2.$$

Therefore, CCMP is written as

$$CCMP(K) = -\sum_{k=1}^{K} n_k(\bar{y}_k - \bar{y})^2 + 2\sum_{k=1}^{K}\widehat{\sigma}_k^2 = \sum_{i=1}^{n}\sum_{k=1}^{K}(y_i - \bar{y}_k)^2 z_{ik} + 2\sum_{k=1}^{K}\widehat{\sigma}_k^2 - n\widehat{\sigma}^2,$$

where in the first two terms,

$$\sum_{i=1}^{n}\sum_{k=1}^{K}(y_i - \bar{y}_k)^2 z_{ik} + 2\sum_{k=1}^{K}\widehat{\sigma}_k^2 = \sum_{k=1}^{K}\left\{\sum_{i=1}^{n}(y_i - \bar{y}_k)^2 z_{ik} + 2\widehat{\sigma}_k^2\right\},$$

it is the sum of Mallows' C_p of K clusters,

$$\sum_{i=1}^{n}(y_i - \bar{y}_k)^2 z_{ik} + 2\widehat{\sigma}_k^2.$$

The third term $n\widehat{\sigma}^2$ is independent of the model, CCMP is equivalent to Mallows' C_p.

Mixture-of-Experts (MoE)

In our simulations in the main text, we consider first generating cluster label $\mathcal{C}_i \in \{1, \ldots, K\}$, and then generating cluster specific explanatory variable $x_{i,\mathcal{C}_i}$ from the multivariate normal distribution $N(\mu_{x,\mathcal{C}_i}, \Sigma_{x,\mathcal{C}_i})$, followed by generating y_i conditional on x_i and cluster label $\mathcal{C}_i$. This simulation model can be seen in the framework of the Mixture-of-Experts (MoE) model only if $\Sigma_{x,\mathcal{C}_i}$ does not differ across clusters.

To introduce the MoE model, consider a latent cluster variable $\mathcal{C} \in \{1, \ldots, K\}$ with prior probabilities $P(\mathcal{C} = k)$ for $k = 1, \ldots, K$. Suppose that, conditional on $\mathcal{C} = k$, the explanatory variable X and the response y are generated according to $X|(\mathcal{C} = k) \sim N_p(\mu_k, \Sigma_k)$, and $y|(X = x, \mathcal{C} = k)$ has the conditional distribution $f(y|x, \mathcal{C} = k)$. The resulting conditional distribution of y given $X = x$ can be written as:

$$f(y|x) = \sum_{k=1}^{K} f(y|x, \mathcal{C} = k) P(\mathcal{C} = k|X = x),$$

where $P(\mathcal{C} = k|X = x)$ is a multinomial logistic regression in x.

Since $p(x|\mathcal{C} = k) = (2\pi)^{-p/2}|\Sigma_k|^{-1/2} e^{-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)}$, by Bayes' theorem,

$$\begin{aligned} P(\mathcal{C} = k|X = x) &= \frac{P(\mathcal{C} = k) p(x|\mathcal{C} = k)}{\sum_{l=1}^{K} P(\mathcal{C} = l) p(x|\mathcal{C} = l)} \\ &= \frac{P(\mathcal{C} = k)|\Sigma_k|^{-1/2} e^{-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)}}{\sum_{l=1}^{K} P(\mathcal{C} = l)|\Sigma_l|^{-1/2} e^{-\frac{1}{2}(x-\mu_l)^T \Sigma_l^{-1}(x-\mu_l)}} \\ &= \frac{e^{\log\{P(\mathcal{C}=k)|\Sigma_k|^{-1/2}\} - \frac{1}{2}\mu_k^T \Sigma_k^{-1} \mu_k + \mu_k^T \Sigma_k^{-1} x - \frac{1}{2} x^T \Sigma_k^{-1} x}}{\sum_{l=1}^{K} e^{\log\{P(\mathcal{C}=l)|\Sigma_l|^{-1/2}\} - \frac{1}{2}\mu_l^T \Sigma_l^{-1} \mu_l + \mu_l^T \Sigma_l^{-1} x - \frac{1}{2} x^T \Sigma_l^{-1} x}}. \end{aligned}$$

This form does not generally coincide with the multinomial logistic regression. However, when $\Sigma_k = \Sigma$ for all k, the last display reduces to

$$\frac{e^{\log P(\mathcal{C}=k) - \frac{1}{2}\mu_k^T \Sigma^{-1} \mu_k + \mu_k^T \Sigma^{-1} x}}{\sum_{l=1}^{K} e^{\log P(C=l) - \frac{1}{2}\mu_l^T \Sigma^{-1} \mu_l + \mu_l^T \Sigma^{-1} x}}$$

Hence, by letting $a_{0,k} = \log P(\mathcal{C} = k) - \frac{1}{2}\mu_k^T \Sigma^{-1} \mu_k$ and $a_k = \Sigma^{-1}\mu_k$, the following multinomial logistic regression model emerges:

$$P(\mathcal{C} = k | X = x) = \frac{e^{a_{0,k} + x^T a_k}}{\sum_{l=1}^{K} e^{a_{0,l} + x^T a_l}}.$$

Therefore, the considered model can be the MoE with multinomial logistic gating function only when Σ_k is common for all clusters. Otherwise, it is not expressed as the MoE with multinomial logistic gating function.

Simulation study

Scenarios under MoE: We consider simulations under the MoE model with iid explanatory variables. The scenario is termed as "MoE". Specifically, for $i = 1, \ldots, n$, x_i is generated from the multivariate normal distribution with zero mean and variance matrix $\left(\frac{1}{4K_{\text{true}}}\right)^2 I_p$. Given x_i, cluster membership is generated from multinomial logistic regression model with probability

$$P(\mathcal{C}_i = k | x_i) = \frac{e^{A_k x_i}}{\sum_{l=1}^{K_{\text{true}}} e^{A_l x_i}},$$

for $k = 1, \ldots, K_{\text{true}}$, where A_k is the regression coefficients set as a submatrix of A_0, $A_k = A_{0,[1:K_{\text{true}},1:(p+1)]}$, where A_0 is the 9×10 matrix,

$$A_0 = 4 \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \\ 1 & 1 & -1 & 1 & 1 & 1 & 1 & -1 & 1 & 1 \\ 1 & 1 & -1 & -1 & -1 & 1 & 1 & -1 & -1 & -1 \\ 1 & 1 & 1 & -1 & 1 & 1 & 1 & 1 & -1 & 1 \\ 1 & 1 & 1 & -1 & -1 & 1 & 1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 & 1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & 1 & -1 & 1 & 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 & -1 & 1 & 1 & 1 & 1 & -1 \end{pmatrix}.$$

Then, the response variable is generated according to the same simulation model as the scenario "I" for y conditional on x and the cluster label with the enlarged regression coefficients $\beta_{p,K_{\text{true}},1}$ multiplied by 40.

Scenarios under intercept-only model: We consider simulations where the regression coefficients for the explanatory variables are zero and only the intercepts differ across cluster. The scenario is termed as "io". Specifically, the simulations are the same as those of "I" except for the true linear predictor is intercept-only, i.e. $\mu_{i,k}$ is set as the one of the K_{true} values equally-spaced between -3 and 3. For the case with $K_{\text{true}} = 1$, $\mu_{i,k} = -3$.

Logistic regression model with K-means clustering: The simulations for the logistic regression model follow the same design as those for the linear regression model, except that the mean function is generated from a logistic regression model. Specifically, explanatory variables x_i are generated in the same way as in the linear regression simulations, and the binary response variable $y_i \in \{0, 1\}$ for $i = 1, \ldots, n$ is generated according to $P(y_i = 1 \mid x_i) = \frac{1}{1+\exp(-\mu_{i,\mathcal{C}_i})}$, where $\mu_{i,k} = \beta^T_{p,K_{\text{true}},m_1,k} \operatorname{diag}(\Sigma_{x,k})^{-1/2}(x_i - \mu_{x,k})$.

We repeat the simulations 500 times. The evaluation uses three metrics, as in the linear regression case: (i) the proportion of correctly selecting the true number of clusters, (ii) the ARI, and (iii) the MSE of the linear predictor in the true regression function $\mu_{i,\mathcal{C}_i}$. The results are presented in Supplementary Tables S19–S36.

Overall, the findings resemble those from the Gaussian regression model simulations. AIC tends to overfit, whereas BIC tends to select the true model. AICex1 and AICex2 mitigate the overfitting observed with AIC, as reflected in the ARI values. When comparing $m_1 = 1$ with $m_1 = 2$, the ARI results suggest that AICex1 and AICex2 can detect the relevant cluster even in cases where BIC struggles, consistent with the Gaussian regression setting. The MSE results indicate that AIC often performs better than the other criteria, particularly compared with BIC, while AICex1 and AICex2 maintain low MSE values.

Behavior of MoE model: We examine the behavior of MoE under the setting of our simulation where explanatory variables are well-separated in the feature space. We consider the same setting of the simualtion "Dis" with $K_{\text{true}} = 3, m = 5$ and Gaussian model except that the effect size of the regression coeffcients varies from $b_0 = 0.1, 1, 2, 5$, and simulate the four datasets, where b_0 extends the $B^T_{(-1,0,1)}$ in Section 3.1 in the main text as:

$$B^T_{(-b_0,0,b_0)} = \begin{pmatrix} -b_0 & 0 & b_0 & -b_0 & 0 & b_0 & -b_0 & 0 & b_0 \\ -b_0 & -b_0 & -b_0 & 0 & 0 & 0 & b_0 & b_0 & b_0 \end{pmatrix},$$

that is, the experiment in "Dis" simulation is that under b_0. Figure S1 give the four simulated datasets shown in scatterplot matrix between response variable and explanatory varaibles together with the true labels of cluster membership (point style) and estimated cluster membership by the MoE (colored) where true number of clusters $K = 3$ was used. In the four simulated datasets, unsupervised clustering by K-means using explanatory variables alone can perfectly recover the true cluster membership because of the well-separated clusters, which is indicated by the adjusted Rand index (ARI) of 100% in the title of the figures. Thefore, the MoE with the resulted clusters as the initial cluster membership needs not be updated as it is the true cluster membership. Neverthelss, the resulted MoE failed to recover the true cluster membership except for the largest effect size dataset of $b_0 = 5$, and the ARI decreses as the effect size decreases. The failure of the MoE is not caused by the failure in optimization since the initial cluster membership is the true one. These imply that the MoE can degrade because of the inadequate effect size for the response variable, and the unsupervised clustering for explanatory variable alone is the better method in the considered simulations.

Timing experiments: Here, we have compared timing of the proposed approach with the MoE and BIC, and BIC under intercept-only model. We selected one of the simulated dataset with $p = 10$, $n = 1000$, $K_{\text{true}} = 3$ under the scenario "I" for Gaussian and logistic simulations. We consider K-means clustering by varing the number of clusters $K = 3, 9, 15$. The machine spec is Intel(R) Core(TM) i9-10940X CPU 3.30GHz. The timings (seconds) for the MoE, AICex, BIC, and BICi are given in Table S1.

Table S1: Timings (seconds) for MoE, AICex, BIC, and BICi by varying K.

Model	K	MoE	AICex	BIC	BICi
Gaussian	3	0.600	0.009	0.004	0.003
	9	0.117	0.013	0.010	0.008
	15	0.141	0.022	0.015	0.012
Logistic	3	2.449	0.012	0.010	0.012
	9	0.065	0.017	0.014	0.009
	15	0.226	0.029	0.023	0.014

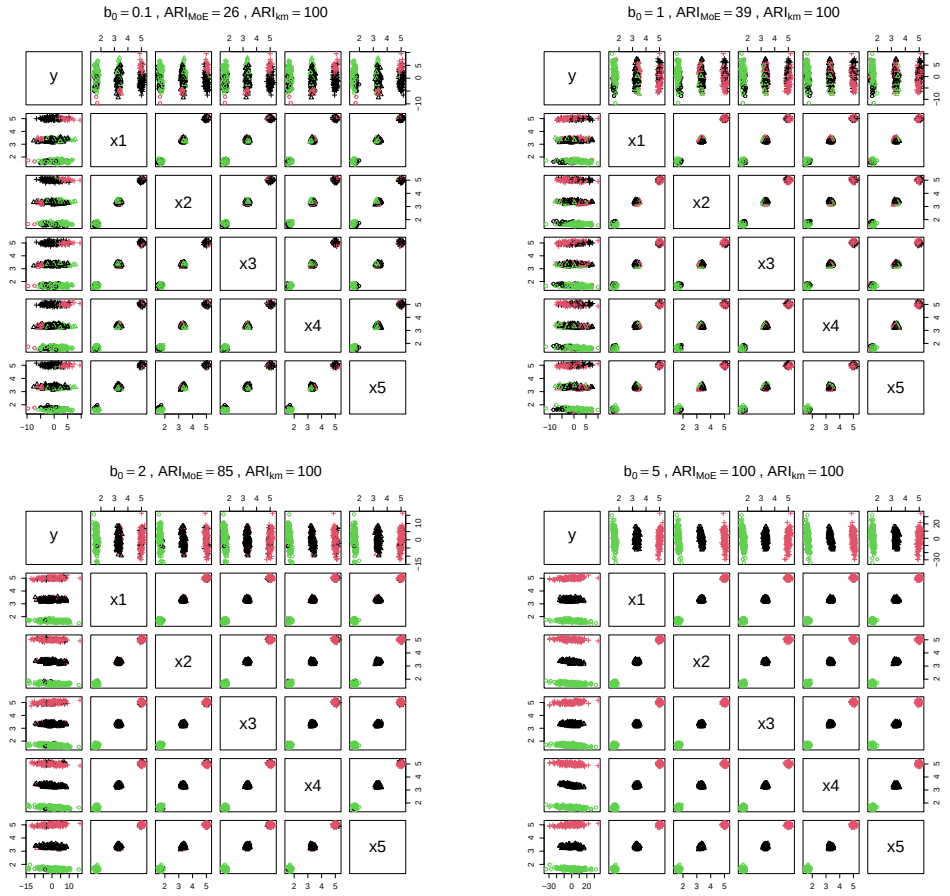

Figure S1: Simulated datasets under $b_0 = 0.1$ (top left), 1 (top right), 2 (bottom left), 5 (bottom right) with well-separated clusters in the explanatory variables along with $\mathrm{ARI}_{\mathrm{MoE}}$ (ARI for MoE) and $\mathrm{ARI}_{\mathrm{km}}$ (ARI for K-means using explanatory variables); Color shows the estimated cluster membership by MoE and point shape indicates the true cluster membership with the number of true clusters is 3.

Real data application

We illustrate the proposed approach using real data through three analyses corresponding to Gaussian, logistic, and Cox regression models. The real data applications also indicated that AIC and BIC under constant models tend to select a large number of clusters.

Gaussian regression model: We use the `plasma` dataset [7], available in the `bsamGP` [5] package for R. This dataset contains 314 observations on 14 variables. We consider a Gaussian linear model with plasma beta-carotene (`betaplasma`) as the response variable and 12 explanatory variables: Age (`age`), Sex (`sex`), Smoking status (`smoke`), BMI (`bmi`), Vitamin use (`vitas`), Calories consumed per day (`calories`), Fat consumed per day (`fat`), Fiber consumed per day (`fiber`), Alcoholic drinks per week (`alcohol`), Cholesterol intake (`cholesterol`), Dietary beta-carotene (`betadiet`), and Dietary retinol (`reedit`). Because the explanatory variables include both continuous and categorical types, we apply the K-prototypes algorithm [4], which accommodates mixed data, instead of the standard K-means algorithm [9].

The model selection results are as follows. AIC selects $K = 3$ as optimal, while BIC selects $K = 1$. For intercept-only models, both AIC and BIC select $K = 11$. AICex1 selects $K = 1$, whereas AICex2 selects $K = 2$.

We focus on the case $K = 2$ to illustrate the comparison between in-cluster and out-of-cluster prediction and to determine which clusters are merged or eliminated during the exclusion process. The in-cluster prediction errors, defined as the penalized negative log-likelihood on the left-hand side of equation (2.1), are 3077 for cluster 1 (cluster 1 $\rightarrow$ cluster 1) and 1133 for cluster 2 (cluster 2 $\rightarrow$ cluster 2). By contrast, the out-of-cluster prediction errors, defined as the negative cross-validated log-likelihood on the right-hand side of equation (2.1), are 3115 for the cluster 1 model applied to cluster 2 (cluster 1 $\rightarrow$ cluster 2) and 1084 for the cluster 2 model applied to cluster 1 (cluster 2 $\rightarrow$ cluster 1). Thus, the out-of-cluster prediction error for the cluster 1 model, 3115, is larger than its in-cluster prediction error, 3077, whereas the out-of-cluster prediction error for the cluster 2 model, 1084, is smaller than its in-cluster prediction error, 1133. This indicates a conflicting pattern of predictive performance between the cluster 1 and cluster 2 models. Since AICex1 excludes models aggressively, it excludes the $K = 2$ model. By contrast, AICex2 excludes a model only when both clusters perform worse in-cluster than out-of-cluster; hence, it retains $K = 2$ as the optimal model.

The discrepancy between AICex1 and AICex2 suggests the need for further investigation using regression summaries.

Regarding regression estimates, in the single-cluster model, `bmi` is negatively associated ($\beta = -5.9$, $P = 0.0003$), while `vitas` with "Yes fairly often" ($\beta = 79.3$, $P = 0.0007$), `fiber` ($\beta = 6.4$, $P = 0.02$), and `betadiet` ($\beta = 0.02$, $P = 0.04$) show positive associations. The adjusted R^2 is 0.15.

For the $K = 2$ model, cluster sizes are 235 and 79. Overall, regression summaries are broadly similar, but the effect of `fiber` differs between clusters. In cluster 1, `bmi` is negatively associated ($\beta = -5.0$, $P = 0.003$), while `vitas` with "Yes fairly often" ($\beta = 53.9$, $P = 0.02$), `fiber` ($\beta = 1.1$, $P = 0.72$), and `betadiet` ($\beta = 0.02$, $P = 0.24$) are positively associated. The adjusted R^2 is 0.09. In cluster 2, `bmi` is also negatively associated ($\beta = -5.9$, $P = 0.18$), while `vitas` with "Yes fairly often" ($\beta = 152.6$, $P = 0.03$), `fiber` ($\beta = 18.2$, $P = 0.009$), and `betadiet` ($\beta = 0.02$, $P = 0.33$) are positively associated. The adjusted R^2 is 0.19. Although the effect of `betadiet` is relatively weak, the scatterplot in Supplementary Figure S2 colored by clusters suggests that `betadiet` helps distinguish the two groups.

Logistic regression model: We apply the method to the Pima Indians Diabetes dataset `pima` from the `pdp` package [3] for R. We use the 392 complete cases with nine variables. The response variable is `diabetes`, which is binary, and the explanatory variables are the eight numeric variables: `pregnant`, `glucose`, `pressure`, `triceps`, `insulin`, `mass`, `pedigree`, and `age`. K-means clustering is applied to these eight variables. We evaluate model selection using AIC, BIC, AICex1, and AICex2, and all methods select $K = 1$ as optimal. By contrast, when applied under a constant model, both AIC and BIC select $K = 15$ as optimal.

Supplementary Tables

Linear regression with K-means clustering

K_{true} | n | p | s	**AIC**	**AICex1**	**AICex2**	**BIC**	**AICi**	**AICiex1**	**AICiex2**	**BICi**	**MoE**	**MoEk**
1 | 500 | 5 | I	98 | 98	100 | 100	100 | 100	100 | 100	0 | 0	23 | 23	21 | 21	0 | 0	100 | 100	100 | 100
1 | 500 | 5 | W	91 | 91	99 | 99	97 | 97	100 | 100	0 | 0	20 | 20	19 | 19	13 | 13	100 | 100	100 | 100
1 | 1000 | 5 | I	91 | 91	97 | 97	94 | 94	100 | 100	0 | 0	16 | 16	13 | 13	0 | 0	100 | 100	100 | 100
1 | 1000 | 5 | W	93 | 93	97 | 97	95 | 95	100 | 100	1 | 1	12 | 12	12 | 12	11 | 11	100 | 100	100 | 100
1 | 500 | 10 | I	96 | 96	100 | 100	96 | 96	100 | 100	0 | 0	24 | 24	22 | 22	4 | 4	100 | 100	100 | 100
1 | 500 | 10 | W	95 | 95	100 | 100	98 | 98	100 | 100	0 | 0	15 | 15	14 | 14	6 | 6	100 | 100	100 | 100
1 | 1000 | 10 | I	98 | 98	100 | 100	99 | 99	100 | 100	0 | 0	22 | 22	21 | 21	0 | 0	100 | 100	100 | 100
1 | 1000 | 10 | W	98 | 98	100 | 100	100 | 100	100 | 100	0 | 0	9 | 9	8 | 8	1 | 1	100 | 100	100 | 100
3 | 500 | 5 | I	98 | 97	98 | 99	99 | 99	95 | 99	34 | 26	1 | 0	2 | 2	18 | 14	11 | 18	26 | 37
3 | 500 | 5 | W	95 | 97	93 | 95	95 | 98	67 | 76	34 | 28	2 | 1	4 | 4	39 | 34	25 | 38	30 | 43
3 | 1000 | 5 | I	98 | 97	100 | 99	99 | 98	100 | 100	30 | 22	0 | 0	2 | 1	37 | 29	36 | 53	79 | 89
3 | 1000 | 5 | W	97 | 95	99 | 98	98 | 96	87 | 94	30 | 23	2 | 2	7 | 7	44 | 34	37 | 55	47 | 67
3 | 500 | 10 | I	99 | 98	100 | 100	99 | 98	89 | 100	31 | 23	1 | 0	3 | 2	15 | 13	8 | 14	14 | 21
3 | 500 | 10 | W	97 | 99	98 | 100	98 | 100	76 | 91	31 | 23	2 | 2	7 | 6	41 | 35	31 | 48	31 | 49
3 | 1000 | 10 | I	100 | 100	100 | 100	100 | 100	100 | 100	31 | 22	0 | 0	1 | 0	38 | 32	47 | 71	71 | 89
3 | 1000 | 10 | W	99 | 99	99 | 99	99 | 99	88 | 100	29 | 20	2 | 2	8 | 6	39 | 29	46 | 74	50 | 78
9 | 500 | 5 | I	98 | 98	47 | 43	97 | 98	0 | 0	49 | 46	0 | 0	0 | 0	0 | 0	0 | 0	0 | 0
9 | 500 | 5 | W	93 | 94	77 | 74	95 | 96	7 | 4	68 | 61	0 | 0	0 | 0	1 | 1	0 | 0	1 | 0
9 | 1000 | 5 | I	99 | 99	76 | 74	99 | 99	45 | 47	74 | 71	0 | 0	0 | 0	0 | 0	0 | 0	0 | 0
9 | 1000 | 5 | W	97 | 99	90 | 90	98 | 100	35 | 31	81 | 79	0 | 0	0 | 0	3 | 3	4 | 1	6 | 3
9 | 500 | 10 | I	92 | 90	45 | 43	95 | 94	7 | 6	66 | 60	0 | 0	1 | 0	3 | 1	0 | 0	0 | 0
9 | 500 | 10 | W	90 | 94	88 | 88	93 | 96	13 | 11	74 | 72	0 | 0	0 | 0	2 | 1	0 | 0	0 | 0
9 | 1000 | 10 | I	99 | 99	80 | 80	98 | 99	44 | 48	84 | 83	0 | 0	1 | 0	4 | 3	0 | 0	0 | 0
9 | 1000 | 10 | W	98 | 99	94 | 94	97 | 99	39 | 34	82 | 78	0 | 0	1 | 1	8 | 5	3 | 1	4 | 2

Table S2: Average of adjusted Rand index for all models | only for the first cluster (%) are given for simulations under Gaussian regression models for $K_{\text{true}} \in \{1, 3, 9\}$, $p \in \{5, 10\}$, $n \in \{500, 1000\}$, and the scenarios "I" and "W". Compared methods are AIC, BIC, AICex1, AICex2, AICi, BICi, AICiex1, AICiex2, MoE, and MoEk, where "i" stands for the intercept-only or the constant model.

$K_{\text{true}}\|n\|p\|s$	**AIC**	**AICex1**	**AICex2**	**BIC**	**AICi**	**AICiex1**	**AICiex2**	**BICi**	**MoE**	**MoEk**
1|500|5|I	0.10	0.10	0.10	0.10	2.33	4.02	3.98	2.87	0.10	0.10
1|500|5|W	0.09	0.09	0.09	0.09	1.62	3.00	3.00	2.13	0.09	0.09
1|1000|5|I	0.05	0.04	0.04	0.04	2.19	4.21	4.11	2.58	0.04	0.04
1|1000|5|W	0.05	0.05	0.05	0.04	1.47	2.71	2.58	1.74	0.04	0.04
1|500|10|I	0.19	0.19	0.19	0.19	6.19	9.06	9.02	7.04	0.19	0.19
1|500|10|W	0.20	0.19	0.20	0.19	4.66	6.96	6.81	5.60	0.19	0.19
1|1000|10|I	0.09	0.09	0.09	0.09	6.25	8.91	8.86	6.66	0.09	0.09
1|1000|10|W	0.08	0.08	0.08	0.08	4.59	7.36	7.34	5.13	0.08	0.08
3|500|5|I	0.33	0.32	0.32	0.32	2.64	3.98	3.98	3.78	2.74	2.46
3|500|5|W	0.34	0.33	0.33	0.46	2.07	3.84	3.85	2.86	1.38	1.30
3|1000|5|I	0.17	0.16	0.16	0.16	2.48	4.03	4.03	2.95	2.25	0.81
3|1000|5|W	0.17	0.16	0.17	0.19	1.95	4.09	4.00	2.34	1.03	0.77
3|500|10|I	0.59	0.58	0.59	0.62	6.24	8.42	8.41	8.16	5.66	5.61
3|500|10|W	0.61	0.61	0.61	0.79	5.18	7.99	7.99	6.49	2.19	2.15
3|1000|10|I	0.29	0.29	0.29	0.29	6.23	8.39	8.39	6.95	2.46	1.64
3|1000|10|W	0.30	0.29	0.30	0.31	5.21	8.12	8.10	5.79	1.41	1.33
9|500|5|I	1.04	2.88	1.04	3.43	3.19	3.41	3.41	3.40	3.44	3.44
9|500|5|W	1.05	1.10	1.03	2.82	2.86	3.23	3.23	3.23	3.01	2.99
9|1000|5|I	0.53	0.60	0.52	2.56	3.02	3.34	3.34	3.34	3.36	3.35
9|1000|5|W	0.52	0.55	0.52	2.04	2.73	3.25	3.25	3.24	2.92	2.76
9|500|10|I	2.10	5.80	1.99	6.73	6.30	6.72	6.72	6.71	6.74	6.74
9|500|10|W	2.13	1.84	1.87	5.47	5.90	6.57	6.57	6.57	5.98	5.92
9|1000|10|I	0.92	0.99	0.92	4.94	6.18	6.64	6.64	6.64	6.61	6.63
9|1000|10|W	0.95	0.97	0.96	3.26	5.53	6.50	6.50	6.49	5.41	5.33

Table S3: Mean squared errors for the linear predictor are given for simulations under Gaussian regression models for $K_{\text{true}} \in \{1, 3, 9\}$, $p \in \{5, 10\}$, $n \in \{500, 1000\}$, and the scenarios "I" and "W". Compared methods are AIC, BIC, AICex1, AICex2, AICi, BICi, AICiex1, AICiex2, MoE, and MoEk, where "i" stands for the intercept-only or the constant model.

$K_{\text{true}}\|n\|p\|s$	**AIC**	**AICex1**	**AICex2**	**BIC**	**AICi**	**AICiex1**	**AICiex2**	**BICi**	**MoE**	**MoEk**
1|500|5|I	98| 2| 0	100| 0| 0	100| 0| 0	100| 0| 0	0|100| 0	23| 77| 0	21| 79| 0	0|100| 0	100| 0| 0	100| 0| 0
1|500|5|k1	95| 5| 0	99| 1| 0	96| 4| 0	100| 0| 0	0|100| 0	27| 73| 0	27| 73| 0	1| 99| 0	100| 0| 0	100| 0| 0
1|1000|5|I	91| 9| 0	97| 3| 0	94| 6| 0	100| 0| 0	0|100| 0	16| 84| 0	13| 87| 0	0|100| 0	100| 0| 0	100| 0| 0
1|1000|5|k1	92| 8| 0	100| 0| 0	94| 6| 0	100| 0| 0	0|100| 0	22| 78| 0	20| 80| 0	0|100| 0	100| 0| 0	100| 0| 0
1|500|10|I	96| 4| 0	100| 0| 0	96| 4| 0	100| 0| 0	0|100| 0	24| 76| 0	22| 78| 0	4| 96| 0	100| 0| 0	100| 0| 0
1|500|10|k1	96| 4| 0	100| 0| 0	96| 4| 0	100| 0| 0	0|100| 0	32| 68| 0	32| 68| 0	9| 91| 0	100| 0| 0	100| 0| 0
1|1000|10|I	98| 2| 0	100| 0| 0	99| 1| 0	100| 0| 0	0|100| 0	22| 78| 0	21| 79| 0	0|100| 0	100| 0| 0	100| 0| 0
1|1000|10|k1	99| 1| 0	100| 0| 0	100| 0| 0	100| 0| 0	0|100| 0	27| 73| 0	24| 76| 0	0|100| 0	100| 0| 0	100| 0| 0
3|500|5|I	90| 10| 0	94| 1| 5	93| 7| 0	90| 0| 10	0|100| 0	0| 0|100	0| 0|100	0| 27| 73	10| 0| 90	22| 0| 78
3|500|5|k1	98| 2| 0	71| 1| 28	99| 1| 0	60| 0| 40	2| 95| 3	0| 0|100	0| 0|100	1| 9| 90	0| 0|100	2| 0| 98
3|1000|5|I	90| 10| 0	99| 1| 0	93| 7| 0	100| 0| 0	0|100| 0	0| 0|100	0| 0|100	0| 78| 22	23| 20| 57	80| 6| 14
3|1000|5|k1	95| 5| 0	87| 0| 13	96| 3| 1	100| 0| 0	0|100| 0	0| 0|100	0| 0|100	6| 23| 71	3| 0| 97	25| 1| 74
3|500|10|I	93| 7| 0	99| 1| 0	94| 6| 0	74| 0| 26	0|100| 0	0| 0|100	0| 1| 99	1| 23| 76	11| 0| 89	12| 0| 88
3|500|10|k1	94| 6| 0	75| 0| 25	96| 4| 0	83| 0| 17	1| 99| 0	0| 0|100	0| 0|100	25| 10| 65	0| 0|100	0| 0|100
3|1000|10|I	97| 3| 0	100| 0| 0	97| 3| 0	100| 0| 0	0|100| 0	0| 0|100	0| 0|100	0| 78| 22	41| 24| 35	60| 30| 10
3|1000|10|k1	98| 2| 0	81| 0| 19	99| 1| 0	100| 0| 0	0|100| 0	0| 0|100	0| 0|100	34| 48| 18	11| 1| 88	37| 2| 61
9|500|5|I	71| 23| 6	34| 3| 63	70| 22| 8	0| 0|100	1| 50| 49	0| 0|100	0| 0|100	0| 0|100	0| 0|100	0| 0|100
9|500|5|k1	36| 13| 51	8| 0| 92	30| 10| 60	0| 0|100	1| 9| 90	0| 0|100	0| 0|100	0| 0|100	0| 0|100	0| 0|100
9|1000|5|I	77| 23| 0	54| 4| 42	81| 15| 4	28| 0| 72	0| 87| 13	0| 0|100	0| 0|100	0| 0|100	0| 0|100	0| 0|100
9|1000|5|k1	75| 18| 7	20| 0| 80	65| 11| 24	0| 0|100	1| 29| 70	0| 0|100	0| 0|100	0| 0|100	0| 0|100	0| 0|100
9|500|10|I	34| 61| 5	29| 6| 65	45| 47| 8	0| 8| 92	1| 63| 36	0| 0|100	0| 0|100	0| 0|100	0| 0|100	0| 0|100
9|500|10|k1	23| 52| 25	16| 4| 80	35| 31| 34	0| 5| 95	2| 30| 68	0| 0|100	0| 0|100	0| 0|100	0| 0|100	0| 0|100
9|1000|10|I	76| 23| 1	63| 1| 36	80| 12| 8	20| 0| 80	2| 96| 2	0| 0|100	0| 0|100	0| 0|100	0| 0|100	0| 0|100
9|1000|10|k1	76| 19| 5	34| 0| 66	72| 12| 16	0| 0|100	6| 53| 41	0| 0|100	0| 0|100	0| 0|100	0| 0|100	0| 0|100

Table S4: Proportions of selected number of clusters (%), correct-|over-|under-estimations, are given for simulations under Gaussian regression models for $K_{\text{true}} \in \{1, 3, 9\}$, $p \in \{5, 10\}$, $n \in \{500, 1000\}$, and the scenarios "I" and "k1". Compared methods are AIC, BIC, AICex1, AICex2, AICi, BICi, AICiex1, AICiex2, MoE, and MoEk, where "i" stands for the intercept-only or the constant model.

$K_{\text{true}}\|n\|p\|s$	**AIC**	**AICex1**	**AICex2**	**BIC**	**AICi**	**AICiex1**	**AICiex2**	**BICi**	**MoE**	**MoEk**
1|500|5|I	98| 98	100|100	100|100	100|100	0| 0	23| 23	21| 21	0| 0	100|100	100|100
1|500|5|k1	95| 95	99| 99	96| 96	100|100	0| 0	27| 27	27| 27	1| 1	100|100	100|100
1|1000|5|I	91| 91	97| 97	94| 94	100|100	0| 0	16| 16	13| 13	0| 0	100|100	100|100
1|1000|5|k1	92| 92	100|100	94| 94	100|100	0| 0	22| 22	20| 20	0| 0	100|100	100|100
1|500|10|I	96| 96	100|100	96| 96	100|100	0| 0	24| 24	22| 22	4| 4	100|100	100|100
1|500|10|k1	96| 96	100|100	96| 96	100|100	0| 0	32| 32	32| 32	9| 9	100|100	100|100
1|1000|10|I	98| 98	100|100	99| 99	100|100	0| 0	22| 22	21| 21	0| 0	100|100	100|100
1|1000|10|k1	99| 99	100|100	100|100	100|100	0| 0	27| 27	24| 24	0| 0	100|100	100|100
3|500|5|I	98| 97	98| 99	99| 99	95| 99	34| 26	1| 0	2| 2	18| 14	11| 18	26| 37
3|500|5|k1	100|100	75| 72	100|100	60| 60	37| 42	0| 0	2| 0	6| 6	1| 3	6| 12
3|1000|5|I	98| 97	100| 99	99| 98	100|100	30| 22	0| 0	2| 1	37| 29	36| 53	79| 89
3|1000|5|k1	99|100	89| 87	99| 99	100|100	33| 36	0| 0	1| 0	18| 19	2| 4	22| 30
3|500|10|I	99| 98	100|100	99| 98	89|100	31| 23	1| 0	3| 2	15| 13	8| 14	14| 21
3|500|10|k1	99|100	77| 75	99|100	83| 83	35| 42	0| 0	1| 0	32| 32	2| 3	4| 9
3|1000|10|I	100|100	100|100	100|100	100|100	31| 22	0| 0	1| 0	38| 32	47| 71	71| 89
3|1000|10|k1	100|100	87| 82	100|100	100|100	32| 35	0| 0	0| 0	62| 69	9| 19	32| 54
9|500|5|I	98| 98	47| 43	97| 98	0| 0	49| 46	0| 0	0| 0	0| 0	0| 0	0| 0
9|500|5|k1	68| 67	11| 9	57| 54	0| 0	21| 17	0| 0	0| 0	0| 0	0| 0	0| 0
9|1000|5|I	99| 99	76| 74	99| 99	45| 47	74| 71	0| 0	0| 0	0| 0	0| 0	0| 0
9|1000|5|k1	95| 96	24| 21	89| 86	0| 0	35| 33	0| 0	0| 0	0| 0	0| 0	0| 0
9|500|10|I	92| 90	45| 43	95| 94	7| 6	66| 60	0| 0	1| 0	3| 1	0| 0	0| 0
9|500|10|k1	84| 90	23| 22	82| 83	4| 5	45| 41	0| 0	0| 0	0| 0	0| 0	0| 0
9|1000|10|I	99| 99	80| 80	98| 99	44| 48	84| 83	0| 0	1| 0	4| 3	0| 0	0| 0
9|1000|10|k1	98|100	36| 35	92| 91	0| 0	59| 61	0| 0	0| 0	0| 0	0| 0	0| 0

Table S5: Average of adjusted Rand index for all models | only for the first cluster (%) are given for simulations under Gaussian regression models for $K_{\text{true}} \in \{1, 3, 9\}$, $p \in \{5, 10\}$, $n \in \{500, 1000\}$, and the scenarios "I" and "k1". Compared methods are AIC, BIC, AICex1, AICex2, AICi, BICi, AICiex1, AICiex2, MoE, and MoEk, where "i" stands for the intercept-only or the constant model.

$K_{\text{true}}\|n\|p\|s$	**AIC**	**AICex1**	**AICex2**	**BIC**	**AICi**	**AICiex1**	**AICiex2**	**BICi**	**MoE**	**MoEk**
1|500|5|I	0.10	0.10	0.10	0.10	2.33	4.02	3.98	2.87	0.10	0.10
1|500|5|k1	0.11	0.10	0.10	0.10	2.32	4.09	4.06	2.76	0.10	0.10
1|1000|5|I	0.05	0.04	0.04	0.04	2.19	4.21	4.11	2.58	0.04	0.04
1|1000|5|k1	0.05	0.05	0.05	0.05	2.20	4.08	4.04	2.64	0.05	0.05
1|500|10|I	0.19	0.19	0.19	0.19	6.19	9.06	9.02	7.04	0.19	0.19
1|500|10|k1	0.19	0.19	0.19	0.19	6.12	9.24	9.20	6.93	0.19	0.19
1|1000|10|I	0.09	0.09	0.09	0.09	6.25	8.91	8.86	6.66	0.09	0.09
1|1000|10|k1	0.10	0.10	0.10	0.10	6.33	9.19	8.97	6.73	0.10	0.10
3|500|5|I	0.33	0.32	0.32	0.32	2.64	3.98	3.98	3.78	2.74	2.46
3|500|5|k1	0.32	0.36	0.32	0.44	1.45	1.95	1.96	1.92	1.64	1.59
3|1000|5|I	0.17	0.16	0.16	0.16	2.48	4.03	4.03	2.95	2.25	0.81
3|1000|5|k1	0.16	0.16	0.16	0.16	1.32	1.91	1.91	1.86	1.58	1.51
3|500|10|I	0.59	0.58	0.59	0.62	6.24	8.42	8.41	8.16	5.66	5.61
3|500|10|k1	0.59	0.65	0.58	0.61	3.11	4.01	4.01	3.95	3.29	3.27
3|1000|10|I	0.29	0.29	0.29	0.29	6.23	8.39	8.39	6.95	2.46	1.64
3|1000|10|k1	0.28	0.32	0.28	0.28	3.18	3.99	3.99	3.82	3.13	1.51
9|500|5|I	1.04	2.88	1.04	3.43	3.19	3.41	3.41	3.40	3.44	3.44
9|500|5|k1	1.11	1.21	1.14	1.21	1.15	1.14	1.14	1.14	1.21	1.21
9|1000|5|I	0.53	0.60	0.52	2.56	3.02	3.34	3.34	3.34	3.36	3.35
9|1000|5|k1	0.49	1.12	0.52	1.16	1.12	1.13	1.13	1.13	1.16	1.16
9|500|10|I	2.10	5.80	1.99	6.73	6.30	6.72	6.72	6.71	6.74	6.74
9|500|10|k1	2.24	2.31	2.10	2.37	2.26	2.24	2.24	2.24	2.35	2.35
9|1000|10|I	0.92	0.99	0.92	4.94	6.18	6.64	6.64	6.64	6.61	6.63
9|1000|10|k1	0.91	2.16	0.92	2.31	2.21	2.23	2.23	2.23	2.31	2.31

Table S6: Mean squared errors for the linear predictor are given for simulations under Gaussian regression models for $K_{\text{true}} \in \{1, 3, 9\}$, $p \in \{5, 10\}$, $n \in \{500, 1000\}$, and the scenarios "I" and "k1". Compared methods are AIC, BIC, AICex1, AICex2, AICi, BICi, AICiex1, AICiex2, MoE, and MoEk, where "i" stands for the intercept-only or the constant model.

$K_{\text{true}}\|n\|p\|s$	**AIC**	**AICex1**	**AICex2**	**BIC**	**AICi**	**AICiex1**	**AICiex2**	**BICi**	**MoE**	**MoEk**
1|500|5|I	98| 2| 0	100| 0| 0	100| 0| 0	100| 0| 0	0|100| 0	23| 77| 0	21| 79| 0	0|100| 0	100| 0| 0	100| 0| 0
1|500|5|Imb	91| 9| 0	98| 2| 0	95| 5| 0	100| 0| 0	0|100| 0	26| 74| 0	23| 77| 0	1| 99| 0	100| 0| 0	100| 0| 0
1|1000|5|I	91| 9| 0	97| 3| 0	94| 6| 0	100| 0| 0	0|100| 0	16| 84| 0	13| 87| 0	0|100| 0	100| 0| 0	100| 0| 0
1|1000|5|Imb	94| 6| 0	100| 0| 0	98| 2| 0	100| 0| 0	0|100| 0	23| 77| 0	20| 80| 0	0|100| 0	100| 0| 0	100| 0| 0
1|500|10|I	96| 4| 0	100| 0| 0	96| 4| 0	100| 0| 0	0|100| 0	24| 76| 0	22| 78| 0	4| 96| 0	100| 0| 0	100| 0| 0
1|500|10|Imb	96| 4| 0	100| 0| 0	97| 3| 0	100| 0| 0	0|100| 0	28| 72| 0	26| 74| 0	5| 95| 0	100| 0| 0	100| 0| 0
1|1000|10|I	98| 2| 0	100| 0| 0	99| 1| 0	100| 0| 0	0|100| 0	22| 78| 0	21| 79| 0	0|100| 0	100| 0| 0	100| 0| 0
1|1000|10|Imb	100| 0| 0	100| 0| 0	100| 0| 0	100| 0| 0	0|100| 0	31| 69| 0	29| 71| 0	0|100| 0	100| 0| 0	100| 0| 0
3|500|5|I	90| 10| 0	94| 1| 5	93| 7| 0	90| 0| 10	0|100| 0	0| 0|100	0| 0|100	0| 27| 73	10| 0| 90	22| 0| 78
3|500|5|Imb	92| 8| 0	96| 0| 4	96| 4| 0	85| 0| 15	0|100| 0	0| 0|100	1| 0| 99	0| 59| 41	0| 0|100	1| 0| 99
3|1000|5|I	90| 10| 0	99| 1| 0	93| 7| 0	100| 0| 0	0|100| 0	0| 0|100	0| 0|100	0| 78| 22	23| 20| 57	80| 6| 14
3|1000|5|Imb	92| 8| 0	98| 2| 0	94| 6| 0	100| 0| 0	0|100| 0	0| 0|100	0| 2| 98	0| 99| 1	0| 0|100	6| 1| 93
3|500|10|I	93| 7| 0	99| 1| 0	94| 6| 0	74| 0| 26	0|100| 0	0| 0|100	0| 1| 99	1| 23| 76	11| 0| 89	12| 0| 88
3|500|10|Imb	96| 4| 0	77| 0| 23	96| 4| 0	98| 0| 2	0|100| 0	0| 0|100	0| 0|100	0| 44| 56	0| 0|100	0| 0|100
3|1000|10|I	97| 3| 0	100| 0| 0	97| 3| 0	100| 0| 0	0|100| 0	0| 0|100	0| 0|100	0| 78| 22	41| 24| 35	60| 30| 10
3|1000|10|Imb	99| 1| 0	97| 0| 3	100| 0| 0	100| 0| 0	0|100| 0	0| 0|100	0| 2| 98	0| 94| 6	0| 0|100	6| 0| 94
9|500|5|I	71| 23| 6	34| 3| 63	70| 22| 8	0| 0|100	1| 50| 49	0| 0|100	0| 0|100	0| 0|100	0| 0|100	0| 0|100
9|500|5|Imb	28| 44| 28	11| 1| 88	37| 26| 37	0| 5| 95	1| 54| 45	0| 0|100	0| 0|100	0| 0|100	0| 0|100	0| 0|100
9|1000|5|I	77| 23| 0	54| 4| 42	81| 15| 4	28| 0| 72	0| 87| 13	0| 0|100	0| 0|100	0| 0|100	0| 0|100	0| 0|100
9|1000|5|Imb	34| 51| 15	18| 1| 81	33| 17| 50	0| 1| 99	2| 87| 11	0| 0|100	0| 0|100	0| 0|100	0| 0|100	0| 0|100
9|500|10|I	34| 61| 5	29| 6| 65	45| 47| 8	0| 8| 92	1| 63| 36	0| 0|100	0| 0|100	0| 0|100	0| 0|100	0| 0|100
9|500|10|Imb	29| 55| 16	13| 3| 84	31| 31| 38	2| 12| 86	1| 66| 33	0| 0|100	0| 0|100	0| 0|100	0| 0|100	0| 0|100
9|1000|10|I	76| 23| 1	63| 1| 36	80| 12| 8	20| 0| 80	2| 96| 2	0| 0|100	0| 0|100	0| 0|100	0| 0|100	0| 0|100
9|1000|10|Imb	26| 58| 16	7| 2| 91	28| 14| 58	0| 2| 98	2| 92| 6	0| 0|100	0| 0|100	0| 2| 98	0| 0|100	0| 0|100

Table S7: Proportions of selected number of clusters (%), correct-|over-|under-estimations, are given for simulations under Gaussian regression models for $K_{\text{true}} \in \{1, 3, 9\}$, $p \in \{5, 10\}$, $n \in \{500, 1000\}$, and the scenarios "I" and "Imb". Compared methods are AIC, BIC, AICex1, AICex2, AICi, BICi, AICiex1, AICiex2, MoE, and MoEk, where "i" stands for the intercept-only or the constant model.

$K_{\text{true}}\|n\|p\|s$	**AIC**	**AICex1**	**AICex2**	**BIC**	**AICi**	**AICiex1**	**AICiex2**	**BICi**	**MoE**	**MoEk**
1|500|5|I	98| 98	100|100	100|100	100|100	0| 0	23| 23	21| 21	0| 0	100|100	100|100
1|500|5|Imb	91| 91	98| 98	95| 95	100|100	0| 0	26| 26	23| 23	1| 1	100|100	100|100
1|1000|5|I	91| 91	97| 97	94| 94	100|100	0| 0	16| 16	13| 13	0| 0	100|100	100|100
1|1000|5|Imb	94| 94	100|100	98| 98	100|100	0| 0	23| 23	20| 20	0| 0	100|100	100|100
1|500|10|I	96| 96	100|100	96| 96	100|100	0| 0	24| 24	22| 22	4| 4	100|100	100|100
1|500|10|Imb	96| 96	100|100	97| 97	100|100	0| 0	28| 28	26| 26	5| 5	100|100	100|100
1|1000|10|I	98| 98	100|100	99| 99	100|100	0| 0	22| 22	21| 21	0| 0	100|100	100|100
1|1000|10|Imb	100|100	100|100	100|100	100|100	0| 0	31| 31	29| 29	0| 0	100|100	100|100
3|500|5|I	98| 97	98| 99	99| 99	95| 99	34| 26	1| 0	2| 2	18| 14	11| 18	26| 37
3|500|5|Imb	96|100	99| 98	98|100	89| 87	22| 58	1| 0	2| 1	24| 55	2| 2	7| 5
3|1000|5|I	98| 97	100| 99	99| 98	100|100	30| 22	0| 0	2| 1	37| 29	36| 53	79| 89
3|1000|5|Imb	96|100	99|100	97|100	100|100	17| 56	2| 1	6| 4	30| 87	15| 11	34| 23
3|500|10|I	99| 98	100|100	99| 98	89|100	31| 23	1| 0	3| 2	15| 13	8| 14	14| 21
3|500|10|Imb	98|100	90| 83	98|100	98| 98	21| 61	1| 0	3| 1	17| 42	4| 3	5| 5
3|1000|10|I	100|100	100|100	100|100	100|100	31| 22	0| 0	1| 0	38| 32	47| 71	71| 89
3|1000|10|Imb	99|100	100| 98	100|100	100|100	18| 63	0| 0	3| 3	29| 85	33| 29	54| 46
9|500|5|I	98| 98	47| 43	97| 98	0| 0	49| 46	0| 0	0| 0	0| 0	0| 0	0| 0
9|500|5|Imb	94| 91	31| 25	93| 86	5| 5	49| 58	0| 0	1| 0	0| 0	0| 0	0| 0
9|1000|5|I	99| 99	76| 74	99| 99	45| 47	74| 71	0| 0	0| 0	0| 0	0| 0	0| 0
9|1000|5|Imb	95| 95	54| 41	96| 86	66| 35	68| 89	0| 0	1| 0	2| 1	0| 0	3| 0
9|500|10|I	92| 90	45| 43	95| 94	7| 6	66| 60	0| 0	1| 0	3| 1	0| 0	0| 0
9|500|10|Imb	91| 95	32| 27	93| 89	12| 13	68| 79	1| 0	1| 0	3| 2	0| 0	0| 0
9|1000|10|I	99| 99	80| 80	98| 99	44| 48	84| 83	0| 0	1| 0	4| 3	0| 0	0| 0
9|1000|10|Imb	93| 96	41| 27	92| 79	74| 42	77| 95	0| 0	0| 0	12| 9	0| 0	1| 0

Table S8: Average of adjusted Rand index for all models | only for the first cluster (%) are given for simulations under Gaussian regression models for $K_{\text{true}} \in \{1, 3, 9\}$, $p \in \{5, 10\}$, $n \in \{500, 1000\}$, and the scenarios "I" and "Imb". Compared methods are AIC, BIC, AICex1, AICex2, AICi, BICi, AICiex1, AICiex2, MoE, and MoEk, where "i" stands for the intercept-only or the constant model.

$K_{\text{true}}\|n\|p\|s$	**AIC**	**AICex1**	**AICex2**	**BIC**	**AICi**	**AICiex1**	**AICiex2**	**BICi**	**MoE**	**MoEk**
1|500|5|I	0.10	0.10	0.10	0.10	2.33	4.02	3.98	2.87	0.10	0.10
1|500|5|Imb	0.11	0.11	0.11	0.11	2.28	4.21	4.11	2.86	0.11	0.11
1|1000|5|I	0.05	0.04	0.04	0.04	2.19	4.21	4.11	2.58	0.04	0.04
1|1000|5|Imb	0.04	0.04	0.04	0.04	2.21	4.15	4.05	2.67	0.04	0.04
1|500|10|I	0.19	0.19	0.19	0.19	6.19	9.06	9.02	7.04	0.19	0.19
1|500|10|Imb	0.19	0.19	0.19	0.19	6.16	9.06	9.04	6.97	0.19	0.19
1|1000|10|I	0.09	0.09	0.09	0.09	6.25	8.91	8.86	6.66	0.09	0.09
1|1000|10|Imb	0.09	0.09	0.09	0.09	6.35	9.12	9.10	6.67	0.09	0.09
3|500|5|I	0.33	0.32	0.32	0.32	2.64	3.98	3.98	3.78	2.74	2.46
3|500|5|Imb	0.32	0.31	0.31	0.33	2.75	4.42	4.42	3.47	2.03	2.01
3|1000|5|I	0.17	0.16	0.16	0.16	2.48	4.03	4.03	2.95	2.25	0.81
3|1000|5|Imb	0.16	0.16	0.16	0.16	2.62	4.49	4.48	3.00	1.70	1.66
3|500|10|I	0.59	0.58	0.59	0.62	6.24	8.42	8.41	8.16	5.66	5.61
3|500|10|Imb	0.62	0.67	0.62	0.62	6.48	9.01	9.01	8.36	3.78	3.75
3|1000|10|I	0.29	0.29	0.29	0.29	6.23	8.39	8.39	6.95	2.46	1.64
3|1000|10|Imb	0.29	0.29	0.29	0.29	6.64	9.17	9.14	7.05	2.68	2.53
9|500|5|I	1.04	2.88	1.04	3.43	3.19	3.41	3.41	3.40	3.44	3.44
9|500|5|Imb	1.09	2.95	1.08	3.12	2.96	3.14	3.14	3.14	3.15	3.15
9|1000|5|I	0.53	0.60	0.52	2.56	3.02	3.34	3.34	3.34	3.36	3.35
9|1000|5|Imb	0.58	1.97	0.61	0.86	2.71	3.13	3.13	3.13	3.06	3.05
9|500|10|I	2.10	5.80	1.99	6.73	6.30	6.72	6.72	6.71	6.74	6.74
9|500|10|Imb	2.04	5.71	2.01	6.06	5.82	6.26	6.26	6.26	6.12	6.12
9|1000|10|I	0.92	0.99	0.92	4.94	6.18	6.64	6.64	6.64	6.61	6.63
9|1000|10|Imb	1.01	5.59	1.17	1.53	5.72	6.27	6.27	6.25	6.14	6.14

Table S9: Mean squared errors for the linear predictor are given for simulations under Gaussian regression models for $K_{\text{true}} \in \{1, 3, 9\}$, $p \in \{5, 10\}$, $n \in \{500, 1000\}$, and the scenarios "I" and "Imb". Compared methods are AIC, BIC, AICex1, AICex2, AICi, BICi, AICiex1, AICiex2, MoE, and MoEk, where "i" stands for the intercept-only or the constant model.

$K_{\text{true}}\|n\|p\|s$	**AIC**	**AICex1**	**AICex2**	**BIC**	**AICi**	**AICiex1**	**AICiex2**	**BICi**	**MoE**	**MoEk**
1|500|5|I	98| 2| 0	100| 0| 0	100| 0| 0	100| 0| 0	0|100| 0	23| 77| 0	21| 79| 0	0|100| 0	100| 0| 0	100| 0| 0
1|500|5|Dis	95| 5| 0	99| 1| 0	96| 4| 0	100| 0| 0	0|100| 0	30| 70| 0	25| 75| 0	0|100| 0	100| 0| 0	100| 0| 0
1|1000|5|I	91| 9| 0	97| 3| 0	94| 6| 0	100| 0| 0	0|100| 0	16| 84| 0	13| 87| 0	0|100| 0	100| 0| 0	100| 0| 0
1|1000|5|Dis	90| 10| 0	100| 0| 0	93| 7| 0	100| 0| 0	0|100| 0	15| 85| 0	14| 86| 0	0|100| 0	100| 0| 0	100| 0| 0
1|500|10|I	96| 4| 0	100| 0| 0	96| 4| 0	100| 0| 0	0|100| 0	24| 76| 0	22| 78| 0	4| 96| 0	100| 0| 0	100| 0| 0
1|500|10|Dis	97| 3| 0	100| 0| 0	100| 0| 0	100| 0| 0	0|100| 0	26| 74| 0	24| 76| 0	5| 95| 0	100| 0| 0	100| 0| 0
1|1000|10|I	98| 2| 0	100| 0| 0	99| 1| 0	100| 0| 0	0|100| 0	22| 78| 0	21| 79| 0	0|100| 0	100| 0| 0	100| 0| 0
1|1000|10|Dis	97| 3| 0	100| 0| 0	100| 0| 0	100| 0| 0	0|100| 0	22| 78| 0	20| 80| 0	1| 99| 0	100| 0| 0	100| 0| 0
3|500|5|I	90| 10| 0	94| 1| 5	93| 7| 0	90| 0| 10	0|100| 0	0| 0|100	0| 0|100	0| 27| 73	10| 0| 90	22| 0| 78
3|500|5|Dis	88| 11| 1	98| 1| 1	91| 8| 1	54| 0| 46	0|100| 0	0| 0|100	0| 0|100	0| 36| 64	1| 0| 99	2| 1| 97
3|1000|5|I	90| 10| 0	99| 1| 0	93| 7| 0	100| 0| 0	0|100| 0	0| 0|100	0| 0|100	0| 78| 22	23| 20| 57	80| 6| 14
3|1000|5|Dis	96| 4| 0	98| 2| 0	96| 4| 0	97| 0| 3	0|100| 0	0| 0|100	0| 0|100	0| 65| 35	0| 0|100	1| 10| 89
3|500|10|I	93| 7| 0	99| 1| 0	94| 6| 0	74| 0| 26	0|100| 0	0| 0|100	0| 1| 99	1| 23| 76	11| 0| 89	12| 0| 88
3|500|10|Dis	94| 6| 0	99| 1| 0	96| 4| 0	73| 0| 27	0|100| 0	0| 0|100	0| 0|100	0| 23| 77	2| 0| 98	2| 0| 98
3|1000|10|I	97| 3| 0	100| 0| 0	97| 3| 0	100| 0| 0	0|100| 0	0| 0|100	0| 0|100	0| 78| 22	41| 24| 35	60| 30| 10
3|1000|10|Dis	97| 3| 0	100| 0| 0	99| 1| 0	100| 0| 0	0|100| 0	0| 0|100	0| 0|100	0| 72| 28	1| 10| 89	4| 18| 78
9|500|5|I	71| 23| 6	34| 3| 63	70| 22| 8	0| 0|100	1| 50| 49	0| 0|100	0| 0|100	0| 0|100	0| 0|100	0| 0|100
9|500|5|Dis	51| 37| 12	54| 7| 39	51| 26| 23	0| 0|100	2| 51| 47	0| 0|100	0| 0|100	0| 0|100	0| 0|100	0| 0|100
9|1000|5|I	77| 23| 0	54| 4| 42	81| 15| 4	28| 0| 72	0| 87| 13	0| 0|100	0| 0|100	0| 0|100	0| 0|100	0| 0|100
9|1000|5|Dis	74| 24| 2	76| 2| 22	76| 15| 9	11| 0| 89	0| 84| 16	0| 0|100	0| 0|100	0| 0|100	0| 0|100	0| 0|100
9|500|10|I	34| 61| 5	29| 6| 65	45| 47| 8	0| 8| 92	1| 63| 36	0| 0|100	0| 0|100	0| 0|100	0| 0|100	0| 0|100
9|500|10|Dis	36| 59| 5	60| 11| 29	50| 43| 7	0| 7| 93	0| 67| 33	0| 0|100	0| 0|100	0| 0|100	0| 0|100	0| 0|100
9|1000|10|I	76| 23| 1	63| 1| 36	80| 12| 8	20| 0| 80	2| 96| 2	0| 0|100	0| 0|100	0| 0|100	0| 0|100	0| 0|100
9|1000|10|Dis	75| 24| 1	73| 1| 26	76| 11| 13	17| 0| 83	3| 92| 5	0| 0|100	0| 0|100	0| 0|100	0| 0|100	0| 0|100

Table S10: Proportions of selected number of clusters (%), correct-|over-|under-estimations, are given for simulations under Gaussian regression models for $K_{\text{true}} \in \{1, 3, 9\}$, $p \in \{5, 10\}$, $n \in \{500, 1000\}$, and the scenarios "I" and "Dis". Compared methods are AIC, BIC, AICex1, AICex2, AICi, BICi, AICiex1, AICiex2, MoE, and MoEk, where "i" stands for the intercept-only or the constant model.

$K_{\text{true}}\|n\|p\|s$	**AIC**	**AICex1**	**AICex2**	**BIC**	**AICi**	**AICiex1**	**AICiex2**	**BICi**	**MoE**	**MoEk**
1|500|5|I	98| 98	100|100	100|100	100|100	0| 0	23| 23	21| 21	0| 0	100|100	100|100
1|500|5|Dis	95| 95	99| 99	96| 96	100|100	0| 0	30| 30	25| 25	0| 0	100|100	100|100
1|1000|5|I	91| 91	97| 97	94| 94	100|100	0| 0	16| 16	13| 13	0| 0	100|100	100|100
1|1000|5|Dis	90| 90	100|100	93| 93	100|100	0| 0	15| 15	14| 14	0| 0	100|100	100|100
1|500|10|I	96| 96	100|100	96| 96	100|100	0| 0	24| 24	22| 22	4| 4	100|100	100|100
1|500|10|Dis	97| 97	100|100	100|100	100|100	0| 0	26| 26	24| 24	5| 5	100|100	100|100
1|1000|10|I	98| 98	100|100	99| 99	100|100	0| 0	22| 22	21| 21	0| 0	100|100	100|100
1|1000|10|Dis	97| 97	100|100	100|100	100|100	0| 0	22| 22	20| 20	1| 1	100|100	100|100
3|500|5|I	98| 97	98| 99	99| 99	95| 99	34| 26	1| 0	2| 2	18| 14	11| 18	26| 37
3|500|5|Dis	97| 96	99| 99	98| 97	81|100	32| 23	1| 2	4| 4	21| 15	1| 1	3| 4
3|1000|5|I	98| 97	100| 99	99| 98	100|100	30| 22	0| 0	2| 1	37| 29	36| 53	79| 89
3|1000|5|Dis	99| 99	100|100	99| 99	99|100	31| 22	0| 0	1| 0	30| 23	10| 12	23| 26
3|500|10|I	99| 98	100|100	99| 98	89|100	31| 23	1| 0	3| 2	15| 13	8| 14	14| 21
3|500|10|Dis	99| 98	100|100	99| 99	89|100	32| 24	0| 0	4| 3	16| 12	3| 4	1| 2
3|1000|10|I	100|100	100|100	100|100	100|100	31| 22	0| 0	1| 0	38| 32	47| 71	71| 89
3|1000|10|Dis	99| 99	100|100	100| 99	100|100	31| 21	0| 0	2| 3	33| 24	28| 39	37| 49
9|500|5|I	98| 98	47| 43	97| 98	0| 0	49| 46	0| 0	0| 0	0| 0	0| 0	0| 0
9|500|5|Dis	95| 96	72| 71	95| 95	1| 1	51| 49	1| 0	2| 0	0| 0	0| 0	0| 0
9|1000|5|I	99| 99	76| 74	99| 99	45| 47	74| 71	0| 0	0| 0	0| 0	0| 0	0| 0
9|1000|5|Dis	98| 98	85| 84	98| 99	44| 48	74| 69	0| 0	0| 0	0| 0	0| 0	0| 0
9|500|10|I	92| 90	45| 43	95| 94	7| 6	66| 60	0| 0	1| 0	3| 1	0| 0	0| 0
9|500|10|Dis	93| 92	76| 76	96| 97	6| 6	66| 59	0| 0	0| 0	1| 1	0| 0	0| 0
9|1000|10|I	99| 99	80| 80	98| 99	44| 48	84| 83	0| 0	1| 0	4| 3	0| 0	0| 0
9|1000|10|Dis	99| 99	85| 83	98| 97	53| 59	81| 79	0| 0	1| 0	2| 1	0| 0	0| 0

Table S11: Average of adjusted Rand index for all models | only for the first cluster (%) are given for simulations under Gaussian regression models for $K_{\text{true}} \in \{1, 3, 9\}$, $p \in \{5, 10\}$, $n \in \{500, 1000\}$, and the scenarios "I" and "Dis". Compared methods are AIC, BIC, AICex1, AICex2, AICi, BICi, AICiex1, AICiex2, MoE, and MoEk, where "i" stands for the intercept-only or the constant model.

$K_{\text{true}}\|n\|p\|s$	**AIC**	**AICex1**	**AICex2**	**BIC**	**AICi**	**AICiex1**	**AICiex2**	**BICi**	**MoE**	**MoEk**
1|500|5|I	0.10	0.10	0.10	0.10	2.33	4.02	3.98	2.87	0.10	0.10
1|500|5|Dis	0.10	0.10	0.10	0.09	2.31	3.96	3.92	2.85	0.09	0.09
1|1000|5|I	0.05	0.04	0.04	0.04	2.19	4.21	4.11	2.58	0.04	0.04
1|1000|5|Dis	0.05	0.04	0.05	0.04	2.19	3.92	3.92	2.61	0.04	0.04
1|500|10|I	0.19	0.19	0.19	0.19	6.19	9.06	9.02	7.04	0.19	0.19
1|500|10|Dis	0.21	0.20	0.20	0.20	6.21	8.97	8.93	6.93	0.20	0.20
1|1000|10|I	0.09	0.09	0.09	0.09	6.25	8.91	8.86	6.66	0.09	0.09
1|1000|10|Dis	0.10	0.10	0.10	0.10	6.35	9.14	9.11	6.69	0.10	0.10
3|500|5|I	0.33	0.32	0.32	0.32	2.64	3.98	3.98	3.78	2.74	2.46
3|500|5|Dis	0.28	0.27	0.28	0.45	2.59	4.02	4.02	3.73	2.88	2.88
3|1000|5|I	0.17	0.16	0.16	0.16	2.48	4.03	4.03	2.95	2.25	0.81
3|1000|5|Dis	0.15	0.15	0.15	0.15	2.49	3.97	3.97	2.98	2.72	2.62
3|500|10|I	0.59	0.58	0.59	0.62	6.24	8.42	8.41	8.16	5.66	5.61
3|500|10|Dis	0.55	0.54	0.55	0.64	6.34	8.40	8.43	8.03	5.80	5.84
3|1000|10|I	0.29	0.29	0.29	0.29	6.23	8.39	8.39	6.95	2.46	1.64
3|1000|10|Dis	0.30	0.30	0.30	0.30	6.37	8.30	8.30	6.99	4.85	4.68
9|500|5|I	1.04	2.88	1.04	3.43	3.19	3.41	3.41	3.40	3.44	3.44
9|500|5|Dis	1.11	1.17	1.10	3.38	3.19	3.37	3.38	3.36	3.39	3.39
9|1000|5|I	0.53	0.60	0.52	2.56	3.02	3.34	3.34	3.34	3.36	3.35
9|1000|5|Dis	0.50	0.51	0.50	2.61	3.00	3.31	3.31	3.31	3.33	3.33
9|500|10|I	2.10	5.80	1.99	6.73	6.30	6.72	6.72	6.71	6.74	6.74
9|500|10|Dis	2.11	1.90	1.95	6.51	6.17	6.62	6.62	6.62	6.64	6.64
9|1000|10|I	0.92	0.99	0.92	4.94	6.18	6.64	6.64	6.64	6.61	6.63
9|1000|10|Dis	0.90	0.92	0.91	1.55	6.12	6.63	6.63	6.63	6.66	6.66

Table S12: Mean squared errors for the linear predictor are given for simulations under Gaussian regression models for $K_{\text{true}} \in \{1, 3, 9\}$, $p \in \{5, 10\}$, $n \in \{500, 1000\}$, and the scenarios "I" and "Dis". Compared methods are AIC, BIC, AICex1, AICex2, AICi, BICi, AICiex1, AICiex2, MoE, and MoEk, where "i" stands for the intercept-only or the constant model.

$K_{\text{true}}\|n\|p\|s$	**AIC**	**AICex1**	**AICex2**	**BIC**	**AICi**	**AICiex1**	**AICiex2**	**BICi**	**MoE**	**MoEk**
1|500|5|I	98| 2| 0	100| 0| 0	100| 0| 0	100| 0| 0	0|100| 0	23| 77| 0	21| 79| 0	0|100| 0	100| 0| 0	100| 0| 0
1|500|5|MoE	90| 10| 0	100| 0| 0	94| 6| 0	100| 0| 0	0|100| 0	12| 88| 0	12| 88| 0	0|100| 0	100| 0| 0	100| 0| 0
1|1000|5|I	91| 9| 0	97| 3| 0	94| 6| 0	100| 0| 0	0|100| 0	16| 84| 0	13| 87| 0	0|100| 0	100| 0| 0	100| 0| 0
1|1000|5|MoE	96| 4| 0	98| 2| 0	98| 2| 0	100| 0| 0	0|100| 0	3| 97| 0	3| 97| 0	0|100| 0	100| 0| 0	100| 0| 0
1|500|10|I	96| 4| 0	100| 0| 0	96| 4| 0	100| 0| 0	0|100| 0	24| 76| 0	22| 78| 0	4| 96| 0	100| 0| 0	100| 0| 0
1|500|10|MoE	98| 2| 0	99| 1| 0	99| 1| 0	100| 0| 0	0|100| 0	18| 82| 0	11| 89| 0	0|100| 0	100| 0| 0	100| 0| 0
1|1000|10|I	98| 2| 0	100| 0| 0	99| 1| 0	100| 0| 0	0|100| 0	22| 78| 0	21| 79| 0	0|100| 0	100| 0| 0	100| 0| 0
1|1000|10|MoE	99| 1| 0	100| 0| 0	100| 0| 0	100| 0| 0	0|100| 0	12| 88| 0	10| 90| 0	0|100| 0	100| 0| 0	100| 0| 0
3|500|5|I	90| 10| 0	94| 1| 5	93| 7| 0	90| 0| 10	0|100| 0	0| 0|100	0| 0|100	0| 27| 73	10| 0| 90	22| 0| 78
3|500|5|MoE	4| 94| 2	1| 1| 98	16| 10| 74	2| 1| 97	1| 99| 0	24| 8| 68	27| 12| 61	11| 87| 2	75| 0| 25	54| 15| 31
3|1000|5|I	90| 10| 0	99| 1| 0	93| 7| 0	100| 0| 0	0|100| 0	0| 0|100	0| 0|100	0| 78| 22	23| 20| 57	80| 6| 14
3|1000|5|MoE	4| 96| 0	5| 1| 94	18| 10| 72	11| 7| 82	0|100| 0	34| 11| 55	33| 17| 50	1| 97| 2	100| 0| 0	81| 15| 4
3|500|10|I	93| 7| 0	99| 1| 0	94| 6| 0	74| 0| 26	0|100| 0	0| 0|100	0| 1| 99	1| 23| 76	11| 0| 89	12| 0| 88
3|500|10|MoE	15| 63| 22	0| 0|100	11| 2| 87	0| 0|100	2| 97| 1	16| 3| 81	20| 8| 72	12| 81| 7	77| 0| 23	63| 14| 23
3|1000|10|I	97| 3| 0	100| 0| 0	97| 3| 0	100| 0| 0	0|100| 0	0| 0|100	0| 0|100	0| 78| 22	41| 24| 35	60| 30| 10
3|1000|10|MoE	6| 90| 4	0| 0|100	7| 2| 91	1| 2| 97	1| 98| 1	21| 13| 66	20| 18| 62	7| 90| 3	100| 0| 0	79| 20| 1
9|500|5|I	71| 23| 6	34| 3| 63	70| 22| 8	0| 0|100	1| 50| 49	0| 0|100	0| 0|100	0| 0|100	0| 0|100	0| 0|100
9|500|5|MoE	0| 0|100	0| 0|100	0| 0|100	0| 0|100	3| 5| 92	0| 0|100	0| 0|100	0| 0|100	0| 0|100	0| 0|100
9|1000|5|I	77| 23| 0	54| 4| 42	81| 15| 4	28| 0| 72	0| 87| 13	0| 0|100	0| 0|100	0| 0|100	0| 0|100	0| 0|100
9|1000|5|MoE	0| 0|100	0| 0|100	0| 0|100	0| 0|100	0| 0|100	6| 13| 81	0| 0|100	0| 0|100	0| 0|100	0| 0|100	0| 0|100
9|500|10|I	34| 61| 5	29| 6| 65	45| 47| 8	0| 8| 92	1| 63| 36	0| 0|100	0| 0|100	0| 0|100	0| 0|100	0| 0|100
9|500|10|MoE	0| 0|100	0| 0|100	0| 0|100	0| 0|100	0| 10| 90	0| 0|100	0| 0|100	0| 0|100	0| 0|100	0| 0|100
9|1000|10|I	76| 23| 1	63| 1| 36	80| 12| 8	20| 0| 80	2| 96| 2	0| 0|100	0| 0|100	0| 0|100	0| 0|100	0| 0|100
9|1000|10|MoE	0| 0|100	0| 0|100	0| 0|100	0| 0|100	2| 13| 85	0| 0|100	0| 0|100	0| 0|100	0| 0|100	0| 0|100

Table S13: Proportions of selected number of clusters (%), correct-|over-|under-estimations, are given for simulations under Gaussian regression models for $K_{\text{true}} \in \{1, 3, 9\}$, $p \in \{5, 10\}$, $n \in \{500, 1000\}$, and the scenarios "I" and "MoE". Compared methods are AIC, BIC, AICex1, AICex2, AICi, BICi, AICiex1, AICiex2, MoE, and MoEk, where "i" stands for the intercept-only or the constant model.

$K_{\text{true}}\|n\|p\|s$	**AIC**	**AICex1**	**AICex2**	**BIC**	**AICi**	**AICiex1**	**AICiex2**	**BICi**	**MoE**	**MoEk**
1|500|5|I	98| 98	100|100	100|100	100|100	0| 0	23| 23	21| 21	0| 0	100|100	100|100
1|500|5|MoE	90| 90	100|100	94| 94	100|100	0| 0	12| 12	12| 12	0| 0	100|100	100|100
1|1000|5|I	91| 91	97| 97	94| 94	100|100	0| 0	16| 16	13| 13	0| 0	100|100	100|100
1|1000|5|MoE	96| 96	98| 98	98| 98	100|100	0| 0	3| 3	3| 3	0| 0	100|100	100|100
1|500|10|I	96| 96	100|100	96| 96	100|100	0| 0	24| 24	22| 22	4| 4	100|100	100|100
1|500|10|MoE	98| 98	99| 99	99| 99	100|100	0| 0	18| 18	11| 11	0| 0	100|100	100|100
1|1000|10|I	98| 98	100|100	99| 99	100|100	0| 0	22| 22	21| 21	0| 0	100|100	100|100
1|1000|10|MoE	99| 99	100|100	100|100	100|100	0| 0	12| 12	10| 10	0| 0	100|100	100|100
3|500|5|I	98| 97	98| 99	99| 99	95| 99	34| 26	1| 0	2| 2	18| 14	11| 18	26| 37
3|500|5|MoE	4| 7	1| 1	2| 3	1| 1	3| 6	3| 3	3| 3	4| 8	28| 37	26| 35
3|1000|5|I	98| 97	100| 99	99| 98	100|100	30| 22	0| 0	2| 1	37| 29	36| 53	79| 89
3|1000|5|MoE	3| 7	1| 1	2| 3	2| 2	3| 6	3| 3	3| 4	3| 6	33| 42	32| 41
3|500|10|I	99| 98	100|100	99| 98	89|100	31| 23	1| 0	3| 2	15| 13	8| 14	14| 21
3|500|10|MoE	4| 6	0| 0	2| 1	1| 0	3| 7	2| 2	3| 3	4| 7	38| 49	37| 48
3|1000|10|I	100|100	100|100	100|100	100|100	31| 22	0| 0	1| 0	38| 32	47| 71	71| 89
3|1000|10|MoE	3| 6	1| 1	2| 1	1| 1	3| 6	3| 3	3| 3	3| 7	46| 54	46| 54
9|500|5|I	98| 98	47| 43	97| 98	0| 0	49| 46	0| 0	0| 0	0| 0	0| 0	0| 0
9|500|5|MoE	0| 0	0| 0	0| 0	0| 0	0| 1	0| 0	0| 0	0| 0	0| 0	0| 0
9|1000|5|I	99| 99	76| 74	99| 99	45| 47	74| 71	0| 0	0| 0	0| 0	0| 0	0| 0
9|1000|5|MoE	0| 0	0| 0	0| 0	0| 0	0| 2	0| 0	0| 0	0| 0	0| 0	0| 0
9|500|10|I	92| 90	45| 43	95| 94	7| 6	66| 60	0| 0	1| 0	3| 1	0| 0	0| 0
9|500|10|MoE	0| 0	0| 0	0| 0	0| 0	0| 1	0| 0	0| 0	0| 0	0| 0	0| 0
9|1000|10|I	99| 99	80| 80	98| 99	44| 48	84| 83	0| 0	1| 0	4| 3	0| 0	0| 0
9|1000|10|MoE	0| 0	0| 0	0| 0	0| 0	0| 2	0| 0	0| 0	0| 0	0| 0	0| 0

Table S14: Average of adjusted Rand index for all models | only for the first cluster (%) are given for simulations under Gaussian regression models for $K_{\text{true}} \in \{1, 3, 9\}$, $p \in \{5, 10\}$, $n \in \{500, 1000\}$, and the scenarios "I" and "MoE". Compared methods are AIC, BIC, AICex1, AICex2, AICi, BICi, AICiex1, AICiex2, MoE, and MoEk, where "i" stands for the intercept-only or the constant model.

K_{true}|n|p|s	**AIC**	**AICex1**	**AICex2**	**BIC**	**AICi**	**AICiex1**	**AICiex2**	**BICi**	**MoE**	**MoEk**
1|500|5|I	0.10	0.10	0.10	0.10	2.33	4.02	3.98	2.87	0.10	0.10
1|500|5|MoE	0.11	0.10	0.11	0.10	188.57	354.49	341.10	191.77	0.10	0.10
1|1000|5|I	0.05	0.04	0.04	0.04	2.19	4.21	4.11	2.58	0.04	0.04
1|1000|5|MoE	0.05	0.05	0.05	0.05	201.01	321.84	317.76	201.65	0.05	0.05
1|500|10|I	0.19	0.19	0.19	0.19	6.19	9.06	9.02	7.04	0.19	0.19
1|500|10|MoE	0.16	0.16	0.16	0.16	572.76	889.02	875.45	614.84	0.16	0.16
1|1000|10|I	0.09	0.09	0.09	0.09	6.25	8.91	8.86	6.66	0.09	0.09
1|1000|10|MoE	0.10	0.10	0.10	0.10	610.49	890.75	885.11	614.89	0.10	0.10
3|500|5|I	0.33	0.32	0.32	0.32	2.64	3.98	3.98	3.78	2.74	2.46
3|500|5|MoE	16.61	19.05	18.57	19.37	28.11	38.99	38.97	30.29	4.80	5.10
3|1000|5|I	0.17	0.16	0.16	0.16	2.48	4.03	4.03	2.95	2.25	0.81
3|1000|5|MoE	17.65	19.33	19.02	19.29	29.06	38.21	38.00	29.28	4.30	4.40
3|500|10|I	0.59	0.58	0.59	0.62	6.24	8.42	8.41	8.16	5.66	5.61
3|500|10|MoE	26.81	31.38	30.53	31.38	65.61	85.63	84.45	69.73	5.11	5.24
3|1000|10|I	0.29	0.29	0.29	0.29	6.23	8.39	8.39	6.95	2.46	1.64
3|1000|10|MoE	27.09	30.74	30.61	30.83	67.38	85.65	85.33	69.12	3.93	3.96
9|500|5|I	1.04	2.88	1.04	3.43	3.19	3.41	3.41	3.40	3.44	3.44
9|500|5|MoE	4.09	4.09	4.09	4.09	4.04	4.03	4.03	4.03	4.09	4.09
9|1000|5|I	0.53	0.60	0.52	2.56	3.02	3.34	3.34	3.34	3.36	3.35
9|1000|5|MoE	4.15	4.15	4.15	4.15	4.15	4.14	4.14	4.14	4.15	4.15
9|500|10|I	2.10	5.80	1.99	6.73	6.30	6.72	6.72	6.71	6.74	6.74
9|500|10|MoE	8.30	8.30	8.30	8.30	8.34	8.35	8.35	8.35	8.30	8.30
9|1000|10|I	0.92	0.99	0.92	4.94	6.18	6.64	6.64	6.64	6.61	6.63
9|1000|10|MoE	8.17	8.17	8.17	8.17	8.20	8.20	8.20	8.20	8.17	8.17

Table S15: Mean squared errors for the linear predictor are given for simulations under Gaussian regression models for $K_{\text{true}} \in \{1, 3, 9\}$, $p \in \{5, 10\}$, $n \in \{500, 1000\}$, and the scenarios "I" and "MoE". Compared methods are AIC, BIC, AICex1, AICex2, AICi, BICi, AICiex1, AICiex2, MoE, and MoEk, where "i" stands for the intercept-only or the constant model.

$K_{\text{true}}\|n\|p\|s$	**AIC**	**AICex1**	**AICex2**	**BIC**	**AICi**	**AICiex1**	**AICiex2**	**BICi**	**MoE**	**MoEk**
1|500|5|I	98| 2| 0	100| 0| 0	100| 0| 0	100| 0| 0	0|100| 0	23| 77| 0	21| 79| 0	0|100| 0	100| 0| 0	100| 0| 0
1|500|5|io	93| 7| 0	98| 2| 0	94| 6| 0	100| 0| 0	68| 32| 0	98| 2| 0	95| 5| 0	100| 0| 0	100| 0| 0	100| 0| 0
1|1000|5|I	91| 9| 0	97| 3| 0	94| 6| 0	100| 0| 0	0|100| 0	16| 84| 0	13| 87| 0	0|100| 0	100| 0| 0	100| 0| 0
1|1000|5|io	93| 7| 0	99| 1| 0	97| 3| 0	100| 0| 0	74| 26| 0	99| 1| 0	98| 2| 0	100| 0| 0	100| 0| 0	100| 0| 0
1|500|10|I	96| 4| 0	100| 0| 0	96| 4| 0	100| 0| 0	0|100| 0	24| 76| 0	22| 78| 0	4| 96| 0	100| 0| 0	100| 0| 0
1|500|10|io	96| 4| 0	99| 1| 0	97| 3| 0	100| 0| 0	70| 30| 0	97| 3| 0	97| 3| 0	100| 0| 0	100| 0| 0	100| 0| 0
1|1000|10|I	98| 2| 0	100| 0| 0	99| 1| 0	100| 0| 0	0|100| 0	22| 78| 0	21| 79| 0	0|100| 0	100| 0| 0	100| 0| 0
1|1000|10|io	95| 5| 0	100| 0| 0	97| 3| 0	100| 0| 0	76| 24| 0	97| 3| 0	97| 3| 0	100| 0| 0	100| 0| 0	100| 0| 0
3|500|5|I	90| 10| 0	94| 1| 5	93| 7| 0	90| 0| 10	0|100| 0	0| 0|100	0| 0|100	0| 27| 73	10| 0| 90	22| 0| 78
3|500|5|io	11| 1| 88	10| 0| 90	11| 1| 88	0| 0|100	81| 19| 0	98| 2| 0	97| 3| 0	100| 0| 0	0| 0|100	0| 0|100
3|1000|5|I	90| 10| 0	99| 1| 0	93| 7| 0	100| 0| 0	0|100| 0	0| 0|100	0| 0|100	0| 78| 22	23| 20| 57	80| 6| 14
3|1000|5|io	32| 3| 65	34| 0| 66	34| 0| 66	0| 0|100	78| 22| 0	99| 1| 0	99| 1| 0	100| 0| 0	0| 0|100	0| 0|100
3|500|10|I	93| 7| 0	99| 1| 0	94| 6| 0	74| 0| 26	0|100| 0	0| 0|100	0| 1| 99	1| 23| 76	11| 0| 89	12| 0| 88
3|500|10|io	2| 1| 97	3| 0| 97	2| 1| 97	0| 0|100	77| 23| 0	99| 1| 0	98| 2| 0	100| 0| 0	0| 0|100	0| 0|100
3|1000|10|I	97| 3| 0	100| 0| 0	97| 3| 0	100| 0| 0	0|100| 0	0| 0|100	0| 0|100	0| 78| 22	41| 24| 35	60| 30| 10
3|1000|10|io	2| 0| 98	4| 0| 96	2| 0| 98	0| 0|100	71| 29| 0	100| 0| 0	100| 0| 0	99| 1| 0	0| 0|100	0| 0|100
9|500|5|I	71| 23| 6	34| 3| 63	70| 22| 8	0| 0|100	1| 50| 49	0| 0|100	0| 0|100	0| 0|100	0| 0|100	0| 0|100
9|500|5|io	0| 1| 99	0| 0|100	0| 0|100	0| 0|100	4| 5| 91	0| 0|100	0| 0|100	0| 0|100	0| 0|100	0| 0|100
9|1000|5|I	77| 23| 0	54| 4| 42	81| 15| 4	28| 0| 72	0| 87| 13	0| 0|100	0| 0|100	0| 0|100	0| 0|100	0| 0|100
9|1000|5|io	0| 0|100	0| 0|100	0| 0|100	0| 0|100	16| 11| 73	0| 0|100	0| 0|100	0| 0|100	0| 0|100	0| 0|100
9|500|10|I	34| 61| 5	29| 6| 65	45| 47| 8	0| 8| 92	1| 63| 36	0| 0|100	0| 0|100	0| 0|100	0| 0|100	0| 0|100
9|500|10|io	0| 7| 93	0| 0|100	0| 1| 99	0| 0|100	5| 8| 87	0| 0|100	0| 0|100	0| 0|100	0| 0|100	0| 0|100
9|1000|10|I	76| 23| 1	63| 1| 36	80| 12| 8	20| 0| 80	2| 96| 2	0| 0|100	0| 0|100	0| 0|100	0| 0|100	0| 0|100
9|1000|10|io	0| 0|100	0| 0|100	0| 0|100	0| 0|100	11| 23| 66	0| 0|100	0| 0|100	0| 0|100	0| 0|100	0| 0|100

Table S16: Proportions of selected number of clusters (%), correct-|over-|under-estimations, are given for simulations under Gaussian regression models for $K_{\text{true}} \in \{1, 3, 9\}$, $p \in \{5, 10\}$, $n \in \{500, 1000\}$, and the scenarios "I" and "io". Compared methods are AIC, BIC, AICex1, AICex2, AICi, BICi, AICiex1, AICiex2, MoE, and MoEk, where "i" stands for the intercept-only or the constant model.

$K_{\text{true}}\|n\|p\|s$	**AIC**	**AICex1**	**AICex2**	**BIC**	**AICi**	**AICiex1**	**AICiex2**	**BICi**	**MoE**	**MoEk**
1|500|5|I	98| 98	100|100	100|100	100|100	0| 0	23| 23	21| 21	0| 0	100|100	100|100
1|500|5|io	93| 93	98| 98	94| 94	100|100	68| 68	98| 98	95| 95	100|100	100|100	100|100
1|1000|5|I	91| 91	97| 97	94| 94	100|100	0| 0	16| 16	13| 13	0| 0	100|100	100|100
1|1000|5|io	93| 93	99| 99	97| 97	100|100	74| 74	99| 99	98| 98	100|100	100|100	100|100
1|500|10|I	96| 96	100|100	96| 96	100|100	0| 0	24| 24	22| 22	4| 4	100|100	100|100
1|500|10|io	96| 96	99| 99	97| 97	100|100	70| 70	97| 97	97| 97	100|100	100|100	100|100
1|1000|10|I	98| 98	100|100	99| 99	100|100	0| 0	22| 22	21| 21	0| 0	100|100	100|100
1|1000|10|io	95| 95	100|100	97| 97	100|100	76| 76	97| 97	97| 97	100|100	100|100	100|100
3|500|5|I	98| 97	98| 99	99| 99	95| 99	34| 26	1| 0	2| 2	18| 14	11| 18	26| 37
3|500|5|io	14| 12	11| 10	14| 12	0| 0	92| 91	100| 99	100| 99	100|100	0| 0	0| 0
3|1000|5|I	98| 97	100| 99	99| 98	100|100	30| 22	0| 0	2| 1	37| 29	36| 53	79| 89
3|1000|5|io	37| 38	34| 34	37| 37	0| 0	92| 91	100| 99	100| 99	100|100	0| 0	0| 0
3|500|10|I	99| 98	100|100	99| 98	89|100	31| 23	1| 0	3| 2	15| 13	8| 14	14| 21
3|500|10|io	7| 7	3| 3	7| 7	0| 0	92| 89	100| 99	100| 99	100|100	0| 0	0| 0
3|1000|10|I	100|100	100|100	100|100	100|100	31| 22	0| 0	1| 0	38| 32	47| 71	71| 89
3|1000|10|io	5| 4	4| 4	5| 4	0| 0	90| 87	100|100	100|100	100|100	0| 0	0| 0
9|500|5|I	98| 98	47| 43	97| 98	0| 0	49| 46	0| 0	0| 0	0| 0	0| 0	0| 0
9|500|5|io	3| 1	1| 0	1| 0	0| 0	68| 65	51| 41	54| 45	47| 37	0| 0	0| 0
9|1000|5|I	99| 99	76| 74	99| 99	45| 47	74| 71	0| 0	0| 0	0| 0	0| 0	0| 0
9|1000|5|io	3| 1	1| 0	2| 1	0| 0	82| 80	58| 56	61| 58	52| 45	0| 0	0| 0
9|500|10|I	92| 90	45| 43	95| 94	7| 6	66| 60	0| 0	1| 0	3| 1	0| 0	0| 0
9|500|10|io	6| 6	0| 0	1| 1	0| 0	71| 69	50| 40	53| 43	45| 34	0| 0	0| 0
9|1000|10|I	99| 99	80| 80	98| 99	44| 48	84| 83	0| 0	1| 0	4| 3	0| 0	0| 0
9|1000|10|io	1| 0	0| 0	1| 0	0| 0	83| 83	58| 55	62| 61	52| 44	0| 0	0| 0

Table S17: Average of adjusted Rand index for all models | only for the first cluster (%) are given for simulations under Gaussian regression models for $K_{\text{true}} \in \{1, 3, 9\}$, $p \in \{5, 10\}$, $n \in \{500, 1000\}$, and the scenarios "I" and "io". Compared methods are AIC, BIC, AICex1, AICex2, AICi, BICi, AICiex1, AICiex2, MoE, and MoEk, where "i" stands for the intercept-only or the constant model.

$K_{\text{true}}\|n\|p\|s$	**AIC**	**AICex1**	**AICex2**	**BIC**	**AICi**	**AICiex1**	**AICiex2**	**BICi**	**MoE**	**MoEk**
1|500|5|I	0.10	0.10	0.10	0.10	2.33	4.02	3.98	2.87	0.10	0.10
1|500|5|io	0.10	0.09	0.10	0.09	0.01	0.01	0.01	0.01	0.09	0.09
1|1000|5|I	0.05	0.04	0.04	0.04	2.19	4.21	4.11	2.58	0.04	0.04
1|1000|5|io	0.05	0.05	0.05	0.05	0.01	0.00	0.00	0.00	0.05	0.05
1|500|10|I	0.19	0.19	0.19	0.19	6.19	9.06	9.02	7.04	0.19	0.19
1|500|10|io	0.19	0.18	0.18	0.18	0.02	0.01	0.01	0.01	0.18	0.18
1|1000|10|I	0.09	0.09	0.09	0.09	6.25	8.91	8.86	6.66	0.09	0.09
1|1000|10|io	0.10	0.09	0.10	0.09	0.01	0.01	0.01	0.01	0.09	0.09
3|500|5|I	0.33	0.32	0.32	0.32	2.64	3.98	3.98	3.78	2.74	2.46
3|500|5|io	0.22	0.21	0.22	0.20	0.06	0.05	0.05	0.05	0.20	0.20
3|1000|5|I	0.17	0.16	0.16	0.16	2.48	4.03	4.03	2.95	2.25	0.81
3|1000|5|io	0.16	0.16	0.16	0.16	0.03	0.02	0.02	0.02	0.16	0.16
3|500|10|I	0.59	0.58	0.59	0.62	6.24	8.42	8.41	8.16	5.66	5.61
3|500|10|io	0.26	0.26	0.26	0.26	0.05	0.04	0.04	0.04	0.26	0.26
3|1000|10|I	0.29	0.29	0.29	0.29	6.23	8.39	8.39	6.95	2.46	1.64
3|1000|10|io	0.16	0.16	0.16	0.15	0.04	0.03	0.03	0.03	0.15	0.15
9|500|5|I	1.04	2.88	1.04	3.43	3.19	3.41	3.41	3.40	3.44	3.44
9|500|5|io	0.11	0.11	0.11	0.10	0.22	0.28	0.26	0.38	0.10	0.10
9|1000|5|I	0.53	0.60	0.52	2.56	3.02	3.34	3.34	3.34	3.36	3.35
9|1000|5|io	0.06	0.05	0.06	0.05	0.13	0.19	0.17	0.25	0.05	0.05
9|500|10|I	2.10	5.80	1.99	6.73	6.30	6.72	6.72	6.71	6.74	6.74
9|500|10|io	0.20	0.19	0.20	0.19	0.21	0.29	0.27	0.39	0.19	0.19
9|1000|10|I	0.92	0.99	0.92	4.94	6.18	6.64	6.64	6.64	6.61	6.63
9|1000|10|io	0.10	0.10	0.10	0.10	0.13	0.21	0.17	0.25	0.10	0.10

Table S18: Mean squared errors for the linear predictor are given for simulations under Gaussian regression models for $K_{\text{true}} \in \{1, 3, 9\}$, $p \in \{5, 10\}$, $n \in \{500, 1000\}$, and the scenarios "I" and "io". Compared methods are AIC, BIC, AICex1, AICex2, AICi, BICi, AICiex1, AICiex2, MoE, and MoEk, where "i" stands for the intercept-only or the constant model.

Logistic regression with K-means clustering

K_{true}	n	p	s	**AIC**			**AICex1**			**AICex2**			**BIC**			**AICi**			**AICiex1**			**AICiex2**			**BICi**			**MoE**			**MoEk**		
1	500	5	I	90	10	0	97	3	0	92	8	0	100	0	0	0	100	0	15	85	0	14	86	0	0	100	0	100	0	0	100	0	0
1	500	5	W	93	7	0	99	1	0	94	6	0	100	0	0	1	99	0	13	87	0	12	88	0	3	97	0	100	0	0	100	0	0
1	1000	5	I	94	6	0	98	2	0	95	5	0	100	0	0	0	100	0	7	93	0	7	93	0	0	100	0	100	0	0	100	0	0
1	1000	5	W	95	5	0	97	3	0	96	4	0	100	0	0	0	100	0	7	93	0	5	95	0	4	96	0	100	0	0	100	0	0
1	500	10	I	95	5	0	100	0	0	98	2	0	100	0	0	0	100	0	15	85	0	13	87	0	0	100	0	100	0	0	100	0	0
1	500	10	W	90	10	0	97	3	0	94	6	0	100	0	0	0	100	0	12	88	0	11	89	0	5	95	0	100	0	0	100	0	0
1	1000	10	I	96	4	0	100	0	0	98	2	0	100	0	0	0	100	0	10	90	0	10	90	0	0	100	0	100	0	0	100	0	0
1	1000	10	W	96	4	0	99	1	0	96	4	0	100	0	0	0	100	0	4	96	0	4	96	0	0	100	0	100	0	0	100	0	0
3	500	5	I	92	8	0	95	5	0	94	6	0	95	0	5	0	100	0	0	0	100	0	2	98	0	97	3	26	3	71	39	1	60
3	500	5	W	76	23	1	84	10	6	80	17	3	63	0	37	0	100	0	0	6	94	0	9	91	0	92	8	5	0	95	2	0	98
3	1000	5	I	88	12	0	94	6	0	91	9	0	100	0	0	0	100	0	0	4	96	0	7	93	0	100	0	39	13	48	66	3	31
3	1000	5	W	94	6	0	95	3	2	93	6	1	92	0	8	0	100	0	0	7	93	0	14	86	0	99	1	17	0	83	18	0	82
3	500	10	I	81	19	0	94	6	0	88	12	0	85	0	15	0	100	0	0	0	100	0	0	100	0	88	12	2	0	98	7	0	93
3	500	10	W	81	16	3	91	2	7	85	8	7	57	0	43	0	100	0	0	2	98	0	6	94	0	87	13	1	0	99	2	0	98
3	1000	10	I	95	5	0	97	3	0	95	5	0	100	0	0	0	100	0	0	5	95	0	8	92	0	99	1	30	4	66	61	6	33
3	1000	10	W	96	4	0	94	1	5	93	2	5	73	0	27	0	100	0	0	8	92	0	12	88	0	99	1	7	0	93	3	0	97
9	500	5	I	65	29	6	68	13	19	76	17	7	1	0	99	0	83	17	0	0	100	0	0	100	0	1	99	0	0	100	0	0	100
9	500	5	W	43	46	11	39	12	49	43	12	45	0	0	100	0	94	6	0	0	100	0	0	100	0	9	91	0	0	100	0	0	100
9	1000	5	I	84	16	0	87	10	3	86	14	0	80	1	19	0	98	2	0	0	100	0	0	100	0	23	77	0	0	100	0	0	100
9	1000	5	W	68	26	6	61	10	29	59	16	25	32	0	68	0	98	2	0	0	100	0	0	100	0	36	64	0	0	100	0	0	100
9	500	10	I	23	46	31	22	4	74	25	4	71	0	0	100	0	81	19	0	0	100	0	0	100	0	0	100	0	0	100	0	0	100
9	500	10	W	20	43	37	7	0	93	7	0	93	0	0	100	0	89	11	0	0	100	0	0	100	0	1	99	0	0	100	0	0	100
9	1000	10	I	52	48	0	70	17	13	65	32	3	36	0	64	1	95	4	0	0	100	0	0	100	0	1	99	0	0	100	0	0	100
9	1000	10	W	65	31	4	57	6	37	56	10	34	2	0	98	0	97	3	0	0	100	0	0	100	0	23	77	0	0	100	0	0	100

Table S19: Proportions of selected number of clusters (%), correct-|over-|under-estimations, are given for simulations under logistic regression models for $K_{\text{true}} \in \{1, 3, 9\}$, $p \in \{5, 10\}$, $n \in \{500, 1000\}$, and the scenarios "I" and "W". Compared methods are AIC, BIC, AICex1, AICex2, AICi, BICi, AICiex1, AICiex2, MoE, and MoEk, where "i" stands for the intercept-only or the constant model.

$K_{\text{true}}\|n\|p\|s$	**AIC**	**AICex1**	**AICex2**	**BIC**	**AICi**	**AICiex1**	**AICiex2**	**BICi**	**MoE**	**MoEk**
1|500|5|I	90| 90	97| 97	92| 92	100|100	0| 0	15| 15	14| 14	0| 0	100|100	100|100
1|500|5|W	93| 93	99| 99	94| 94	100|100	1| 1	13| 13	12| 12	3| 3	100|100	100|100
1|1000|5|I	94| 94	98| 98	95| 95	100|100	0| 0	7| 7	7| 7	0| 0	100|100	100|100
1|1000|5|W	95| 95	97| 97	96| 96	100|100	0| 0	7| 7	5| 5	4| 4	100|100	100|100
1|500|10|I	95| 95	100|100	98| 98	100|100	0| 0	15| 15	13| 13	0| 0	100|100	100|100
1|500|10|W	90| 90	97| 97	94| 94	100|100	0| 0	12| 12	11| 11	5| 5	100|100	100|100
1|1000|10|I	96| 96	100|100	98| 98	100|100	0| 0	10| 10	10| 10	0| 0	100|100	100|100
1|1000|10|W	96| 96	99| 99	96| 96	100|100	0| 0	4| 4	4| 4	0| 0	100|100	100|100
3|500|5|I	97| 97	99| 99	99| 98	98|100	30| 24	1| 0	7| 5	39| 31	32| 54	54| 81
3|500|5|W	93| 93	96| 98	95| 97	82| 93	30| 23	13| 13	17| 17	38| 30	27| 38	36| 49
3|1000|5|I	98| 95	99| 98	99| 97	100|100	30| 22	6| 7	13| 10	34| 25	59| 90	76| 95
3|1000|5|W	99| 99	99| 98	99| 98	97| 97	29| 23	10| 7	19| 14	33| 26	42| 58	51| 66
3|500|10|I	91| 88	99| 98	97| 95	94|100	31| 22	7| 4	11| 6	37| 28	15| 23	20| 30
3|500|10|W	93| 93	97| 96	95| 95	80| 93	30| 22	5| 6	13| 16	33| 26	25| 35	34| 48
3|1000|10|I	99|100	100|100	99|100	100|100	30| 21	8| 7	15| 12	35| 26	45| 65	67| 85
3|1000|10|W	99| 98	96| 96	96| 96	89| 99	29| 21	9| 8	14| 12	33| 26	43| 59	45| 60
9|500|5|I	97| 95	89| 89	98| 97	9| 9	69| 66	1| 0	2| 0	1| 1	0| 0	0| 0
9|500|5|W	94| 94	80| 82	88| 91	18| 14	79| 76	2| 1	2| 1	8| 7	1| 0	2| 0
9|1000|5|I	99| 98	98| 97	99| 98	97|100	81| 80	1| 0	2| 0	19| 19	0| 0	0| 0
9|1000|5|W	98| 98	91| 93	94| 95	72| 73	81| 82	1| 0	2| 0	30| 30	6| 3	11| 9
9|500|10|I	86| 85	50| 46	73| 71	0| 0	70| 69	3| 1	3| 1	1| 1	0| 0	0| 0
9|500|10|W	87| 88	60| 55	69| 68	4| 2	75| 71	2| 1	2| 1	1| 1	0| 0	1| 0
9|1000|10|I	97| 96	95| 93	98| 96	76| 83	80| 79	1| 0	1| 0	1| 1	0| 0	0| 0
9|1000|10|W	97| 98	91| 91	93| 92	47| 44	81| 75	3| 0	3| 0	20| 17	2| 2	4| 3

Table S20: Average of adjusted Rand index for all models | only for the first cluster (%) are given for simulations under logistic regression models for $K_{\text{true}} \in \{1, 3, 9\}$, $p \in \{5, 10\}$, $n \in \{500, 1000\}$, and the scenarios "I" and "io". Compared methods are AIC, BIC, AICex1, AICex2, AICi, BICi, AICiex1, AICiex2, MoE, and MoEk, where "i" stands for the intercept-only or the constant model.

K_{true}|n|p|s	**AIC**	**AICex1**	**AICex2**	**BIC**	**AICi**	**AICiex1**	**AICiex2**	**BICi**	**MoE**	**MoEk**
1|500|5|I	0.10	0.09	0.10	0.09	2.38	3.88	3.88	2.62	0.09	3.91
1|500|5|W	0.09	0.09	0.09	0.09	2.12	2.52	2.50	2.13	0.09	3.30
1|1000|5|I	0.05	0.05	0.05	0.05	2.30	3.91	3.91	2.41	0.05	3.94
1|1000|5|W	0.04	0.04	0.04	0.04	1.71	2.67	2.54	1.67	0.04	3.28
1|500|10|I	0.29	0.27	0.27	0.27	6.66	8.91	8.89	6.97	0.27	8.14
1|500|10|W	0.24	0.22	0.23	0.22	5.67	6.73	6.56	5.72	0.22	7.41
1|1000|10|I	0.13	0.12	0.12	0.12	6.86	9.07	9.01	6.95	0.12	8.17
1|1000|10|W	0.13	0.12	0.13	0.12	5.68	7.65	7.57	5.89	0.12	8.48
3|500|5|I	0.32	0.32	0.32	0.31	2.48	3.99	3.98	2.67	13.68	6.80
3|500|5|W	0.40	0.39	0.40	0.42	2.32	3.66	3.66	2.40	7.91	6.71
3|1000|5|I	0.16	0.15	0.16	0.15	2.59	3.99	3.99	2.62	5.69	2.71
3|1000|5|W	0.16	0.16	0.16	0.17	2.05	3.83	3.74	2.10	10.30	12.64
3|500|10|I	1.36	1.17	1.28	1.31	6.47	8.33	8.33	6.75	6.56	8.33
3|500|10|W	1.21	1.18	1.20	1.24	5.95	8.29	8.26	6.34	11.93	29.52
3|1000|10|I	0.43	0.43	0.43	0.41	6.65	8.32	8.26	6.71	185.67	30.86
3|1000|10|W	0.42	0.42	0.43	0.50	6.06	8.53	8.49	6.11	45.20	40.93
9|500|5|I	2.00	1.81	1.77	3.27	2.93	3.28	3.28	3.28	3.29	3.53
9|500|5|W	3.05	1.94	1.88	2.56	2.73	3.33	3.33	3.25	3.15	3.45
9|1000|5|I	0.61	0.60	0.60	0.62	2.96	3.33	3.33	3.27	3.33	3.59
9|1000|5|W	0.68	0.69	0.70	1.01	2.76	3.35	3.35	2.97	3.15	2.95
9|500|10|I	2611.26	6.86	8.58	6.70	6.13	6.74	6.74	6.74	6.70	6.90
9|500|10|W	2082.72	4.38	4.69	5.68	5.67	6.44	6.44	6.42	5.91	6.48
9|1000|10|I	3.52	2.54	2.78	2.72	6.23	6.66	6.66	6.66	6.65	6.90
9|1000|10|W	2.53	2.20	2.21	3.45	5.79	6.53	6.53	6.22	6.17	6.28

Table S21: Mean squared errors for the linear predictor are given for simulations under logistic regression models for $K_{\text{true}} \in \{1, 3, 9\}$, $p \in \{5, 10\}$, $n \in \{500, 1000\}$, and the scenarios "I" and "io". Compared methods are AIC, BIC, AICex1, AICex2, AICi, BICi, AICiex1, AICiex2, MoE, and MoEk, where "i" stands for the intercept-only or the constant model.

$K_{\text{true}}\|n\|p\|s$	**AIC**	**AICex1**	**AICex2**	**BIC**	**AICi**	**AICiex1**	**AICiex2**	**BICi**	**MoE**	**MoEk**
1|500|5|I	90| 10| 0	97| 3| 0	92| 8| 0	100| 0| 0	0|100| 0	15| 85| 0	14| 86| 0	0|100| 0	100| 0| 0	100| 0| 0
1|500|5|k1	91| 9| 0	98| 2| 0	91| 9| 0	100| 0| 0	0|100| 0	11| 89| 0	7| 93| 0	0|100| 0	100| 0| 0	100| 0| 0
1|1000|5|I	94| 6| 0	98| 2| 0	95| 5| 0	100| 0| 0	0|100| 0	7| 93| 0	7| 93| 0	0|100| 0	100| 0| 0	100| 0| 0
1|1000|5|k1	94| 6| 0	98| 2| 0	95| 5| 0	100| 0| 0	0|100| 0	7| 93| 0	7| 93| 0	0|100| 0	100| 0| 0	100| 0| 0
1|500|10|I	95| 5| 0	100| 0| 0	98| 2| 0	100| 0| 0	0|100| 0	15| 85| 0	13| 87| 0	0|100| 0	100| 0| 0	100| 0| 0
1|500|10|k1	95| 5| 0	98| 2| 0	97| 3| 0	100| 0| 0	0|100| 0	17| 83| 0	15| 85| 0	0|100| 0	100| 0| 0	100| 0| 0
1|1000|10|I	96| 4| 0	100| 0| 0	98| 2| 0	100| 0| 0	0|100| 0	10| 90| 0	10| 90| 0	0|100| 0	100| 0| 0	100| 0| 0
1|1000|10|k1	96| 4| 0	100| 0| 0	98| 2| 0	100| 0| 0	0|100| 0	10| 90| 0	10| 90| 0	0|100| 0	100| 0| 0	100| 0| 0
3|500|5|I	92| 8| 0	95| 5| 0	94| 6| 0	95| 0| 5	0|100| 0	0| 0|100	0| 2| 98	0| 97| 3	26| 3| 71	39| 1| 60
3|500|5|k1	87| 13| 0	93| 5| 2	87| 13| 0	99| 0| 1	0|100| 0	0| 0|100	0| 0|100	0| 57| 43	1| 0| 99	8| 0| 92
3|1000|5|I	88| 12| 0	94| 6| 0	91| 9| 0	100| 0| 0	0|100| 0	0| 4| 96	0| 7| 93	0|100| 0	39| 13| 48	66| 3| 31
3|1000|5|k1	92| 8| 0	96| 3| 1	94| 6| 0	100| 0| 0	0|100| 0	0| 3| 97	0| 5| 95	0| 95| 5	17| 1| 82	52| 2| 46
3|500|10|I	81| 19| 0	94| 6| 0	88| 12| 0	85| 0| 15	0|100| 0	0| 0|100	0| 0|100	0| 88| 12	2| 0| 98	7| 0| 93
3|500|10|k1	91| 9| 0	92| 1| 7	95| 5| 0	90| 0| 10	0|100| 0	0| 1| 99	0| 1| 99	0| 46| 54	0| 0|100	0| 0|100
3|1000|10|I	95| 5| 0	97| 3| 0	95| 5| 0	100| 0| 0	0|100| 0	0| 5| 95	0| 8| 92	0| 99| 1	30| 4| 66	61| 6| 33
3|1000|10|k1	95| 5| 0	97| 2| 1	95| 5| 0	100| 0| 0	0|100| 0	0| 0|100	0| 0|100	0| 80| 20	2| 0| 98	23| 0| 77
9|500|5|I	65| 29| 6	68| 13| 19	76| 17| 7	1| 0| 99	0| 83| 17	0| 0|100	0| 0|100	0| 1| 99	0| 0|100	0| 0|100
9|500|5|k1	55| 33| 12	27| 4| 69	63| 24| 13	0| 0|100	0| 54| 46	0| 0|100	0| 0|100	0| 0|100	0| 0|100	0| 0|100
9|1000|5|I	84| 16| 0	87| 10| 3	86| 14| 0	80| 1| 19	0| 98| 2	0| 0|100	0| 0|100	0| 23| 77	0| 0|100	0| 0|100
9|1000|5|k1	83| 16| 1	52| 6| 42	85| 12| 3	0| 0|100	0| 82| 18	0| 0|100	0| 0|100	0| 0|100	0| 0|100	0| 0|100
9|500|10|I	23| 46| 31	22| 4| 74	25| 4| 71	0| 0|100	0| 81| 19	0| 0|100	0| 0|100	0| 0|100	0| 0|100	0| 0|100
9|500|10|k1	23| 57| 20	37| 4| 59	52| 16| 32	0| 0|100	1| 45| 54	0| 0|100	0| 0|100	0| 0|100	0| 0|100	0| 0|100
9|1000|10|I	52| 48| 0	70| 17| 13	65| 32| 3	36| 0| 64	1| 95| 4	0| 0|100	0| 0|100	0| 1| 99	0| 0|100	0| 0|100
9|1000|10|k1	59| 38| 3	44| 10| 46	60| 34| 6	0| 0|100	1| 78| 21	0| 0|100	0| 0|100	0| 0|100	0| 0|100	0| 0|100

Table S22: Proportions of selected number of clusters (%), correct-|over-|under-estimations, are given for simulations under logistic regression models for $K_{\text{true}} \in \{1, 3, 9\}$, $p \in \{5, 10\}$, $n \in \{500, 1000\}$, and the scenarios "I" and "k1". Compared methods are AIC, BIC, AICex1, AICex2, AICi, BICi, AICiex1, AICiex2, MoE, and MoEk, where "i" stands for the intercept-only or the constant model.

$K_{\text{true}}\|n\|p\|s$	**AIC**	**AICex1**	**AICex2**	**BIC**	**AICi**	**AICiex1**	**AICiex2**	**BICi**	**MoE**	**MoEk**
1|500|5|I	90| 90	97| 97	92| 92	100|100	0| 0	15| 15	14| 14	0| 0	100|100	100|100
1|500|5|k1	91| 91	98| 98	91| 91	100|100	0| 0	11| 11	7| 7	0| 0	100|100	100|100
1|1000|5|I	94| 94	98| 98	95| 95	100|100	0| 0	7| 7	7| 7	0| 0	100|100	100|100
1|1000|5|k1	94| 94	98| 98	95| 95	100|100	0| 0	7| 7	7| 7	0| 0	100|100	100|100
1|500|10|I	95| 95	100|100	98| 98	100|100	0| 0	15| 15	13| 13	0| 0	100|100	100|100
1|500|10|k1	95| 95	98| 98	97| 97	100|100	0| 0	17| 17	15| 15	0| 0	100|100	100|100
1|1000|10|I	96| 96	100|100	98| 98	100|100	0| 0	10| 10	10| 10	0| 0	100|100	100|100
1|1000|10|k1	96| 96	100|100	98| 98	100|100	0| 0	10| 10	10| 10	0| 0	100|100	100|100
3|500|5|I	97| 97	99| 99	99| 98	98|100	30| 24	1| 0	7| 5	39| 31	32| 54	54| 81
3|500|5|k1	97| 99	98| 98	97| 99	99| 99	31| 36	2| 0	4| 0	25| 29	3| 5	11| 20
3|1000|5|I	98| 95	99| 98	99| 97	100|100	30| 22	6| 7	13| 10	34| 25	59| 90	76| 95
3|1000|5|k1	98|100	98| 99	99|100	100|100	29| 33	4| 3	10| 5	35| 39	12| 24	54| 81
3|500|10|I	91| 88	99| 98	97| 95	94|100	31| 22	7| 4	11| 6	37| 28	15| 23	20| 30
3|500|10|k1	96| 97	94| 93	98| 99	90| 90	32| 37	6| 1	10| 2	23| 27	1| 2	4| 7
3|1000|10|I	99|100	100|100	99|100	100|100	30| 21	8| 7	15| 12	35| 26	45| 65	67| 85
3|1000|10|k1	99|100	99| 99	99|100	100|100	31| 37	5| 0	7| 1	33| 41	8| 12	27| 44
9|500|5|I	97| 95	89| 89	98| 97	9| 9	69| 66	1| 0	2| 0	1| 1	0| 0	0| 0
9|500|5|k1	94| 96	43| 38	96| 94	0| 0	49| 52	2| 0	2| 0	1| 0	0| 0	0| 0
9|1000|5|I	99| 98	98| 97	99| 98	97|100	81| 80	1| 0	2| 0	19| 19	0| 0	0| 0
9|1000|5|k1	99|100	70| 64	98| 99	1| 0	68| 75	1| 0	1| 0	0| 0	0| 0	0| 0
9|500|10|I	86| 85	50| 46	73| 71	0| 0	70| 69	3| 1	3| 1	1| 1	0| 0	0| 0
9|500|10|k1	88| 88	53| 49	91| 84	0| 0	43| 43	2| 0	3| 0	0| 0	0| 0	0| 0
9|1000|10|I	97| 96	95| 93	98| 96	76| 83	80| 79	1| 0	1| 0	1| 1	0| 0	0| 0
9|1000|10|k1	97| 98	66| 60	97| 97	0| 0	68| 73	2| 0	2| 0	0| 0	0| 0	0| 0

Table S23: Average of adjusted Rand index for all models | only for the first cluster (%) are given for simulations under logistic regression models for $K_{\text{true}} \in \{1, 3, 9\}$, $p \in \{5, 10\}$, $n \in \{500, 1000\}$, and the scenarios "I" and "k1". Compared methods are AIC, BIC, AICex1, AICex2, AICi, BICi, AICiex1, AICiex2, MoE, and MoEk, where "i" stands for the intercept-only or the constant model.

$K_{\text{true}}\|n\|p\|s$	**AIC**	**AICex1**	**AICex2**	**BIC**	**AICi**	**AICiex1**	**AICiex2**	**BICi**	**MoE**	**MoEk**
1|500|5|I	0.10	0.09	0.10	0.09	2.38	3.88	3.88	2.62	0.09	3.91
1|500|5|k1	0.10	0.09	0.10	0.09	2.31	4.09	4.00	2.60	0.09	3.94
1|1000|5|I	0.05	0.05	0.05	0.05	2.30	3.91	3.91	2.41	0.05	3.94
1|1000|5|k1	0.05	0.05	0.05	0.05	2.30	3.91	3.91	2.41	0.05	3.94
1|500|10|I	0.29	0.27	0.27	0.27	6.66	8.91	8.89	6.97	0.27	8.14
1|500|10|k1	0.24	0.24	0.24	0.24	6.73	8.90	8.88	6.93	0.24	8.09
1|1000|10|I	0.13	0.12	0.12	0.12	6.86	9.07	9.01	6.95	0.12	8.17
1|1000|10|k1	0.13	0.12	0.12	0.12	6.86	9.07	9.01	6.95	0.12	8.17
3|500|5|I	0.32	0.32	0.32	0.31	2.48	3.99	3.98	2.67	13.68	6.80
3|500|5|k1	0.26	0.25	0.26	0.25	1.29	1.95	1.95	1.60	1.61	2.09
3|1000|5|I	0.16	0.15	0.16	0.15	2.59	3.99	3.99	2.62	5.69	2.71
3|1000|5|k1	0.12	0.12	0.12	0.12	1.23	1.97	1.96	1.29	1.67	3.46
3|500|10|I	1.36	1.17	1.28	1.31	6.47	8.33	8.33	6.75	6.56	8.33
3|500|10|k1	0.85	0.82	0.81	0.87	3.17	3.94	3.94	3.68	3.23	3.92
3|1000|10|I	0.43	0.43	0.43	0.41	6.65	8.32	8.26	6.71	185.67	30.86
3|1000|10|k1	0.29	0.29	0.29	0.28	3.20	3.97	3.97	3.37	3.30	14.73
9|500|5|I	2.00	1.81	1.77	3.27	2.93	3.28	3.28	3.28	3.29	3.53
9|500|5|k1	1.25	1.09	0.98	1.14	1.10	1.12	1.12	1.12	1.14	1.35
9|1000|5|I	0.61	0.60	0.60	0.62	2.96	3.33	3.33	3.27	3.33	3.59
9|1000|5|k1	0.36	0.58	0.37	1.14	1.07	1.13	1.13	1.13	1.14	1.38
9|500|10|I	2611.26	6.86	8.58	6.70	6.13	6.74	6.74	6.74	6.70	6.90
9|500|10|k1	1778.58	2.70	35.45	2.24	2.20	2.22	2.22	2.22	2.24	2.45
9|1000|10|I	3.52	2.54	2.78	2.72	6.23	6.66	6.66	6.66	6.65	6.90
9|1000|10|k1	1.28	1.95	1.28	2.23	2.12	2.22	2.22	2.22	2.23	2.46

Table S24: Mean squared errors for the linear predictor are given for simulations under logistic regression models for $K_{\text{true}} \in \{1, 3, 9\}$, $p \in \{5, 10\}$, $n \in \{500, 1000\}$, and the scenarios "I" and "k1". Compared methods are AIC, BIC, AICex1, AICex2, AICi, BICi, AICiex1, AICiex2, MoE, and MoEk, where "i" stands for the intercept-only or the constant model.

$K_{\text{true}}\|n\|p\|s$	**AIC**	**AICex1**	**AICex2**	**BIC**	**AICi**	**AICiex1**	**AICiex2**	**BICi**	**MoE**	**MoEk**
1|500|5|I	90| 10| 0	97| 3| 0	92| 8| 0	100| 0| 0	0|100| 0	15| 85| 0	14| 86| 0	0|100| 0	100| 0| 0	100| 0| 0
1|500|5|Imb	86| 14| 0	95| 5| 0	90| 10| 0	100| 0| 0	0|100| 0	13| 87| 0	13| 87| 0	0|100| 0	100| 0| 0	100| 0| 0
1|1000|5|I	94| 6| 0	98| 2| 0	95| 5| 0	100| 0| 0	0|100| 0	7| 93| 0	7| 93| 0	0|100| 0	100| 0| 0	100| 0| 0
1|1000|5|Imb	94| 6| 0	99| 1| 0	98| 2| 0	100| 0| 0	0|100| 0	3| 97| 0	3| 97| 0	0|100| 0	100| 0| 0	100| 0| 0
1|500|10|I	95| 5| 0	100| 0| 0	98| 2| 0	100| 0| 0	0|100| 0	15| 85| 0	13| 87| 0	0|100| 0	100| 0| 0	100| 0| 0
1|500|10|Imb	94| 6| 0	99| 1| 0	96| 4| 0	100| 0| 0	0|100| 0	20| 80| 0	17| 83| 0	1| 99| 0	100| 0| 0	100| 0| 0
1|1000|10|I	96| 4| 0	100| 0| 0	98| 2| 0	100| 0| 0	0|100| 0	10| 90| 0	10| 90| 0	0|100| 0	100| 0| 0	100| 0| 0
1|1000|10|Imb	95| 5| 0	100| 0| 0	95| 5| 0	100| 0| 0	0|100| 0	7| 93| 0	6| 94| 0	0|100| 0	100| 0| 0	100| 0| 0
3|500|5|I	92| 8| 0	95| 5| 0	94| 6| 0	95| 0| 5	0|100| 0	0| 0|100	0| 2| 98	0| 97| 3	26| 3| 71	39| 1| 60
3|500|5|Imb	88| 12| 0	95| 5| 0	91| 9| 0	100| 0| 0	0|100| 0	1| 0| 99	3| 2| 95	0|100| 0	2| 0| 98	7| 0| 93
3|1000|5|I	88| 12| 0	94| 6| 0	91| 9| 0	100| 0| 0	0|100| 0	0| 4| 96	0| 7| 93	0|100| 0	39| 13| 48	66| 3| 31
3|1000|5|Imb	97| 3| 0	99| 1| 0	97| 3| 0	100| 0| 0	0|100| 0	0| 2| 98	1| 3| 96	0|100| 0	4| 0| 96	46| 1| 53
3|500|10|I	81| 19| 0	94| 6| 0	88| 12| 0	85| 0| 15	0|100| 0	0| 0|100	0| 0|100	0| 88| 12	2| 0| 98	7| 0| 93
3|500|10|Imb	87| 13| 0	97| 1| 2	96| 4| 0	100| 0| 0	0|100| 0	0| 1| 99	1| 1| 98	0| 97| 3	0| 0|100	0| 0|100
3|1000|10|I	95| 5| 0	97| 3| 0	95| 5| 0	100| 0| 0	0|100| 0	0| 5| 95	0| 8| 92	0| 99| 1	30| 4| 66	61| 6| 33
3|1000|10|Imb	95| 5| 0	99| 1| 0	96| 4| 0	100| 0| 0	0|100| 0	0| 1| 99	1| 3| 96	0|100| 0	5| 0| 95	7| 0| 93
9|500|5|I	65| 29| 6	68| 13| 19	76| 17| 7	1| 0| 99	0| 83| 17	0| 0|100	0| 0|100	0| 1| 99	0| 0|100	0| 0|100
9|500|5|Imb	27| 44| 29	19| 10| 71	21| 24| 55	0| 0|100	2| 93| 5	0| 0|100	0| 0|100	0| 1| 99	0| 0|100	0| 0|100
9|1000|5|I	84| 16| 0	87| 10| 3	86| 14| 0	80| 1| 19	0| 98| 2	0| 0|100	0| 0|100	0| 23| 77	0| 0|100	0| 0|100
9|1000|5|Imb	34| 58| 8	23| 9| 68	32| 39| 29	4| 0| 96	0| 99| 1	0| 0|100	0| 0|100	1| 26| 73	0| 0|100	0| 0|100
9|500|10|I	23| 46| 31	22| 4| 74	25| 4| 71	0| 0|100	0| 81| 19	0| 0|100	0| 0|100	0| 0|100	0| 0|100	0| 0|100
9|500|10|Imb	20| 47| 33	3| 0| 97	4| 0| 96	0| 0|100	2| 83| 15	0| 0|100	0| 0|100	1| 2| 97	0| 0|100	0| 0|100
9|1000|10|I	52| 48| 0	70| 17| 13	65| 32| 3	36| 0| 64	1| 95| 4	0| 0|100	0| 0|100	0| 1| 99	0| 0|100	0| 0|100
9|1000|10|Imb	21| 47| 32	8| 0| 92	13| 14| 73	2| 0| 98	3| 92| 5	0| 0|100	0| 0|100	1| 6| 93	0| 0|100	0| 0|100

Table S25: Proportions of selected number of clusters (%), correct-|over-|under-estimations, are given for simulations under logistic regression models for $K_{\text{true}} \in \{1, 3, 9\}$, $p \in \{5, 10\}$, $n \in \{500, 1000\}$, and the scenarios "I" and "Imb". Compared methods are AIC, BIC, AICex1, AICex2, AICi, BICi, AICiex1, AICiex2, MoE, and MoEk, where "i" stands for the intercept-only or the constant model.

K_{true}	n	p	s	**AIC**	**AICex1**	**AICex2**	**BIC**	**AICi**	**AICiex1**	**AICiex2**	**BICi**	**MoE**	**MoEk**
1	500	5	I	90 | 90	97 | 97	92 | 92	100 | 100	0 | 0	15 | 15	14 | 14	0 | 0	100 | 100	100 | 100
1	500	5	Imb	86 | 86	95 | 95	90 | 90	100 | 100	0 | 0	13 | 13	13 | 13	0 | 0	100 | 100	100 | 100
1	1000	5	I	94 | 94	98 | 98	95 | 95	100 | 100	0 | 0	7 | 7	7 | 7	0 | 0	100 | 100	100 | 100
1	1000	5	Imb	94 | 94	99 | 99	98 | 98	100 | 100	0 | 0	3 | 3	3 | 3	0 | 0	100 | 100	100 | 100
1	500	10	I	95 | 95	100 | 100	98 | 98	100 | 100	0 | 0	15 | 15	13 | 13	0 | 0	100 | 100	100 | 100
1	500	10	Imb	94 | 94	99 | 99	96 | 96	100 | 100	0 | 0	20 | 20	17 | 17	1 | 1	100 | 100	100 | 100
1	1000	10	I	96 | 96	100 | 100	98 | 98	100 | 100	0 | 0	10 | 10	10 | 10	0 | 0	100 | 100	100 | 100
1	1000	10	Imb	95 | 95	100 | 100	95 | 95	100 | 100	0 | 0	7 | 7	6 | 6	0 | 0	100 | 100	100 | 100
3	500	5	I	97 | 97	99 | 99	99 | 98	98 | 100	30 | 24	1 | 0	7 | 5	39 | 31	32 | 54	54 | 81
3	500	5	Imb	93 | 99	98 | 100	96 | 100	100 | 100	19 | 50	4 | 2	12 | 9	26 | 67	14 | 13	39 | 39
3	1000	5	I	98 | 95	99 | 98	99 | 97	100 | 100	30 | 22	6 | 7	13 | 10	34 | 25	59 | 90	76 | 95
3	1000	5	Imb	99 | 100	100 | 100	99 | 100	100 | 100	16 | 51	7 | 5	17 | 11	19 | 59	29 | 26	69 | 72
3	500	10	I	91 | 88	99 | 98	97 | 95	94 | 100	31 | 22	7 | 4	11 | 6	37 | 28	15 | 23	20 | 30
3	500	10	Imb	92 | 98	99 | 99	98 | 100	100 | 100	19 | 57	5 | 3	13 | 7	26 | 73	6 | 7	18 | 19
3	1000	10	I	99 | 100	100 | 100	99 | 100	100 | 100	30 | 21	8 | 7	15 | 12	35 | 26	45 | 65	67 | 85
3	1000	10	Imb	97 | 100	100 | 100	98 | 100	100 | 100	17 | 59	3 | 2	14 | 9	23 | 75	25 | 29	55 | 58
9	500	5	I	97 | 95	89 | 89	98 | 97	9 | 9	69 | 66	1 | 0	2 | 0	1 | 1	0 | 0	0 | 0
9	500	5	Imb	93 | 90	64 | 55	92 | 82	16 | 9	72 | 94	1 | 0	4 | 1	1 | 1	0 | 0	0 | 0
9	1000	5	I	99 | 98	98 | 97	99 | 98	97 | 100	81 | 80	1 | 0	2 | 0	19 | 19	0 | 0	0 | 0
9	1000	5	Imb	94 | 98	81 | 61	95 | 90	93 | 52	73 | 97	1 | 0	4 | 2	20 | 27	0 | 1	14 | 2
9	500	10	I	86 | 85	50 | 46	73 | 71	0 | 0	70 | 69	3 | 1	3 | 1	1 | 1	0 | 0	0 | 0
9	500	10	Imb	90 | 87	45 | 29	81 | 56	1 | 0	67 | 84	1 | 0	4 | 1	3 | 3	0 | 0	0 | 0
9	1000	10	I	97 | 96	95 | 93	98 | 96	76 | 83	80 | 79	1 | 0	1 | 0	1 | 1	0 | 0	0 | 0
9	1000	10	Imb	92 | 87	58 | 38	89 | 69	81 | 45	74 | 94	1 | 1	4 | 1	6 | 7	0 | 0	1 | 0

Table S26: Average of adjusted Rand index for all models | only for the first cluster (%) are given for simulations under logistic regression models for $K_{\text{true}} \in \{1, 3, 9\}$, $p \in \{5, 10\}$, $n \in \{500, 1000\}$, and the scenarios "I" and "Imb". Compared methods are AIC, BIC, AICex1, AICex2, AICi, BICi, AICiex1, AICiex2, MoE, and MoEk, where "i" stands for the intercept-only or the constant model.

K_{true}|n|p|s	**AIC**	**AICex1**	**AICex2**	**BIC**	**AICi**	**AICiex1**	**AICiex2**	**BICi**	**MoE**	**MoEk**
1|500|5|I	0.10	0.09	0.10	0.09	2.38	3.88	3.88	2.62	0.09	3.91
1|500|5|Imb	0.10	0.09	0.10	0.09	2.47	3.84	3.80	2.61	0.09	3.97
1|1000|5|I	0.05	0.05	0.05	0.05	2.30	3.91	3.91	2.41	0.05	3.94
1|1000|5|Imb	0.05	0.05	0.05	0.05	2.28	3.78	3.78	2.37	0.05	3.92
1|500|10|I	0.29	0.27	0.27	0.27	6.66	8.91	8.89	6.97	0.27	8.14
1|500|10|Imb	0.26	0.26	0.26	0.25	6.67	9.16	9.08	6.94	0.25	8.17
1|1000|10|I	0.13	0.12	0.12	0.12	6.86	9.07	9.01	6.95	0.12	8.17
1|1000|10|Imb	0.13	0.12	0.13	0.12	6.93	8.92	8.91	7.02	0.12	8.22
3|500|5|I	0.32	0.32	0.32	0.31	2.48	3.99	3.98	2.67	13.68	6.80
3|500|5|Imb	0.36	0.35	0.36	0.35	2.65	4.43	4.43	2.83	2.41	3.91
3|1000|5|I	0.16	0.15	0.16	0.15	2.59	3.99	3.99	2.62	5.69	2.71
3|1000|5|Imb	0.15	0.15	0.15	0.15	2.62	4.45	4.45	2.66	7.27	2.75
3|500|10|I	1.36	1.17	1.28	1.31	6.47	8.33	8.33	6.75	6.56	8.33
3|500|10|Imb	1.45	1.40	1.40	1.37	6.82	8.98	8.98	7.17	4.61	8.13
3|1000|10|I	0.43	0.43	0.43	0.41	6.65	8.32	8.26	6.71	185.67	30.86
3|1000|10|Imb	0.48	0.47	0.48	0.47	7.10	9.16	9.14	7.19	32.99	3.24
9|500|5|I	2.00	1.81	1.77	3.27	2.93	3.28	3.28	3.28	3.29	3.53
9|500|5|Imb	202.68	3.03	2.41	3.04	2.74	3.18	3.18	3.18	3.17	3.38
9|1000|5|I	0.61	0.60	0.60	0.62	2.96	3.33	3.33	3.27	3.33	3.59
9|1000|5|Imb	1.00	0.89	0.90	0.74	2.69	3.15	3.15	3.04	3.24	3.19
9|500|10|I	2611.26	6.86	8.58	6.70	6.13	6.74	6.74	6.74	6.70	6.90
9|500|10|Imb	2416.09	5.74	4.81	6.11	5.69	6.26	6.25	6.22	6.11	6.47
9|1000|10|I	3.52	2.54	2.78	2.72	6.23	6.66	6.66	6.66	6.65	6.90
9|1000|10|Imb	564.27	5.00	2.61	1.86	5.70	6.18	6.18	6.11	6.32	6.38

Table S27: Mean squared errors for the linear predictor are given for simulations under logistic regression models for $K_{\text{true}} \in \{1, 3, 9\}$, $p \in \{5, 10\}$, $n \in \{500, 1000\}$, and the scenarios "I" and "Imb". Compared methods are AIC, BIC, AICex1, AICex2, AICi, BICi, AICiex1, AICiex2, MoE, and MoEk, where "i" stands for the intercept-only or the constant model.

$K_{\text{true}}\|n\|p\|s$	**AIC**	**AICex1**	**AICex2**	**BIC**	**AICi**	**AICiex1**	**AICiex2**	**BICi**	**MoE**	**MoEk**
1|500|5|I	90| 10| 0	97| 3| 0	92| 8| 0	100| 0| 0	0|100| 0	15| 85| 0	14| 86| 0	0|100| 0	100| 0| 0	100| 0| 0
1|500|5|Dis	85| 15| 0	94| 6| 0	88| 12| 0	100| 0| 0	0|100| 0	16| 84| 0	14| 86| 0	0|100| 0	100| 0| 0	100| 0| 0
1|1000|5|I	94| 6| 0	98| 2| 0	95| 5| 0	100| 0| 0	0|100| 0	7| 93| 0	7| 93| 0	0|100| 0	100| 0| 0	100| 0| 0
1|1000|5|Dis	89| 11| 0	97| 3| 0	92| 8| 0	100| 0| 0	0|100| 0	7| 93| 0	7| 93| 0	0|100| 0	100| 0| 0	100| 0| 0
1|500|10|I	95| 5| 0	100| 0| 0	98| 2| 0	100| 0| 0	0|100| 0	15| 85| 0	13| 87| 0	0|100| 0	100| 0| 0	100| 0| 0
1|500|10|Dis	94| 6| 0	100| 0| 0	96| 4| 0	100| 0| 0	0|100| 0	15| 85| 0	13| 87| 0	0|100| 0	100| 0| 0	100| 0| 0
1|1000|10|I	96| 4| 0	100| 0| 0	98| 2| 0	100| 0| 0	0|100| 0	10| 90| 0	10| 90| 0	0|100| 0	100| 0| 0	100| 0| 0
1|1000|10|Dis	96| 4| 0	100| 0| 0	98| 2| 0	100| 0| 0	0|100| 0	13| 87| 0	12| 88| 0	0|100| 0	100| 0| 0	100| 0| 0
3|500|5|I	92| 8| 0	95| 5| 0	94| 6| 0	95| 0| 5	0|100| 0	0| 0|100	0| 2| 98	0| 97| 3	26| 3| 71	39| 1| 60
3|500|5|Dis	82| 18| 0	92| 8| 0	84| 16| 0	76| 0| 24	0|100| 0	0| 0|100	0| 3| 97	0| 96| 4	10| 2| 88	28| 3| 69
3|1000|5|I	88| 12| 0	94| 6| 0	91| 9| 0	100| 0| 0	0|100| 0	0| 4| 96	0| 7| 93	0|100| 0	39| 13| 48	66| 3| 31
3|1000|5|Dis	84| 16| 0	89| 11| 0	84| 16| 0	100| 0| 0	0|100| 0	0| 3| 97	0| 4| 96	0|100| 0	15| 19| 66	44| 19| 37
3|500|10|I	81| 19| 0	94| 6| 0	88| 12| 0	85| 0| 15	0|100| 0	0| 0|100	0| 0|100	0| 88| 12	2| 0| 98	7| 0| 93
3|500|10|Dis	88| 12| 0	97| 3| 0	93| 7| 0	86| 0| 14	0|100| 0	0| 0|100	0| 2| 98	0| 89| 11	1| 0| 99	4| 0| 96
3|1000|10|I	95| 5| 0	97| 3| 0	95| 5| 0	100| 0| 0	0|100| 0	0| 5| 95	0| 8| 92	0| 99| 1	30| 4| 66	61| 6| 33
3|1000|10|Dis	90| 10| 0	92| 8| 0	90| 10| 0	100| 0| 0	0|100| 0	0| 2| 98	0| 5| 95	0| 99| 1	2| 6| 92	21| 3| 76
9|500|5|I	65| 29| 6	68| 13| 19	76| 17| 7	1| 0| 99	0| 83| 17	0| 0|100	0| 0|100	0| 1| 99	0| 0|100	0| 0|100
9|500|5|Dis	46| 51| 3	61| 18| 21	61| 27| 12	0| 0|100	1| 85| 14	0| 0|100	0| 0|100	0| 0|100	0| 0|100	0| 0|100
9|1000|5|I	84| 16| 0	87| 10| 3	86| 14| 0	80| 1| 19	0| 98| 2	0| 0|100	0| 0|100	0| 23| 77	0| 0|100	0| 0|100
9|1000|5|Dis	62| 38| 0	68| 11| 21	64| 30| 6	44| 3| 53	0| 98| 2	0| 0|100	0| 0|100	0| 8| 92	0| 0|100	0| 0|100
9|500|10|I	23| 46| 31	22| 4| 74	25| 4| 71	0| 0|100	0| 81| 19	0| 0|100	0| 0|100	0| 0|100	0| 0|100	0| 0|100
9|500|10|Dis	17| 54| 29	11| 3| 86	13| 3| 84	0| 0|100	1| 75| 24	0| 0|100	0| 0|100	0| 2| 98	0| 0|100	0| 0|100
9|1000|10|I	52| 48| 0	70| 17| 13	65| 32| 3	36| 0| 64	1| 95| 4	0| 0|100	0| 0|100	0| 1| 99	0| 0|100	0| 0|100
9|1000|10|Dis	50| 50| 0	69| 13| 18	68| 26| 6	26| 1| 73	0| 93| 7	0| 0|100	0| 0|100	0| 5| 95	0| 0|100	0| 0|100

Table S28: Proportions of selected number of clusters (%), correct-|over-|under-estimations, are given for simulations under logistic regression models for $K_{\text{true}} \in \{1, 3, 9\}$, $p \in \{5, 10\}$, $n \in \{500, 1000\}$, and the scenarios "I" and "Dis". Compared methods are AIC, BIC, AICex1, AICex2, AICi, BICi, AICiex1, AICiex2, MoE, and MoEk, where "i" stands for the intercept-only or the constant model.

$K_{\text{true}}\|n\|p\|s$	**AIC**	**AICex1**	**AICex2**	**BIC**	**AICi**	**AICiex1**	**AICiex2**	**BICi**	**MoE**	**MoEk**
1|500|5|I	90| 90	97| 97	92| 92	100|100	0| 0	15| 15	14| 14	0| 0	100|100	100|100
1|500|5|Dis	85| 85	94| 94	88| 88	100|100	0| 0	16| 16	14| 14	0| 0	100|100	100|100
1|1000|5|I	94| 94	98| 98	95| 95	100|100	0| 0	7| 7	7| 7	0| 0	100|100	100|100
1|1000|5|Dis	89| 89	97| 97	92| 92	100|100	0| 0	7| 7	7| 7	0| 0	100|100	100|100
1|500|10|I	95| 95	100|100	98| 98	100|100	0| 0	15| 15	13| 13	0| 0	100|100	100|100
1|500|10|Dis	94| 94	100|100	96| 96	100|100	0| 0	15| 15	13| 13	0| 0	100|100	100|100
1|1000|10|I	96| 96	100|100	98| 98	100|100	0| 0	10| 10	10| 10	0| 0	100|100	100|100
1|1000|10|Dis	96| 96	100|100	98| 98	100|100	0| 0	13| 13	12| 12	0| 0	100|100	100|100
3|500|5|I	97| 97	99| 99	99| 98	98|100	30| 24	1| 0	7| 5	39| 31	32| 54	54| 81
3|500|5|Dis	95| 96	99| 99	96| 98	90|100	31| 23	6| 6	11| 11	39| 30	16| 23	28| 41
3|1000|5|I	98| 95	99| 98	99| 97	100|100	30| 22	6| 7	13| 10	34| 25	59| 90	76| 95
3|1000|5|Dis	97| 95	98| 98	97| 95	100|100	30| 21	5| 6	13| 15	35| 26	34| 53	58| 80
3|500|10|I	91| 88	99| 98	97| 95	94|100	31| 22	7| 4	11| 6	37| 28	15| 23	20| 30
3|500|10|Dis	96| 94	99| 99	98| 97	94|100	30| 22	4| 2	6| 5	37| 28	4| 4	9| 10
3|1000|10|I	99|100	100|100	99|100	100|100	30| 21	8| 7	15| 12	35| 26	45| 65	67| 85
3|1000|10|Dis	98| 98	99| 98	98| 98	100|100	30| 21	4| 5	13| 17	35| 26	30| 36	40| 49
9|500|5|I	97| 95	89| 89	98| 97	9| 9	69| 66	1| 0	2| 0	1| 1	0| 0	0| 0
9|500|5|Dis	95| 96	86| 86	97| 98	10| 10	72| 70	2| 1	3| 1	0| 0	0| 0	0| 0
9|1000|5|I	99| 98	98| 97	99| 98	97|100	81| 80	1| 0	2| 0	19| 19	0| 0	0| 0
9|1000|5|Dis	98| 96	93| 92	97| 97	92| 99	81| 79	2| 0	3| 1	7| 6	0| 0	0| 0
9|500|10|I	86| 85	50| 46	73| 71	0| 0	70| 69	3| 1	3| 1	1| 1	0| 0	0| 0
9|500|10|Dis	87| 87	33| 30	63| 60	0| 0	64| 62	1| 0	2| 1	2| 1	0| 0	0| 0
9|1000|10|I	97| 96	95| 93	98| 96	76| 83	80| 79	1| 0	1| 0	1| 1	0| 0	0| 0
9|1000|10|Dis	96| 96	92| 92	98| 99	77| 82	78| 76	2| 1	3| 1	4| 4	0| 0	0| 0

Table S29: Average of adjusted Rand index for all models | only for the first cluster (%) are given for simulations under logistic regression models for $K_{\text{true}} \in \{1, 3, 9\}$, $p \in \{5, 10\}$, $n \in \{500, 1000\}$, and the scenarios "I" and "Dis". Compared methods are AIC, BIC, AICex1, AICex2, AICi, BICi, AICiex1, AICiex2, MoE, and MoEk, where "i" stands for the intercept-only or the constant model.

$K_{\text{true}}\|n\|p\|s$	**AIC**	**AICex1**	**AICex2**	**BIC**	**AICi**	**AICiex1**	**AICiex2**	**BICi**	**MoE**	**MoEk**
1|500|5|I	0.10	0.09	0.10	0.09	2.38	3.88	3.88	2.62	0.09	3.91
1|500|5|Dis	0.14	0.13	0.14	0.11	2.33	4.19	4.18	2.61	0.11	3.91
1|1000|5|I	0.05	0.05	0.05	0.05	2.30	3.91	3.91	2.41	0.05	3.94
1|1000|5|Dis	0.06	0.05	0.06	0.05	2.26	3.62	3.54	2.40	0.05	3.93
1|500|10|I	0.29	0.27	0.27	0.27	6.66	8.91	8.89	6.97	0.27	8.14
1|500|10|Dis	0.28	0.26	0.27	0.26	6.68	9.06	9.06	6.88	0.26	8.12
1|1000|10|I	0.13	0.12	0.12	0.12	6.86	9.07	9.01	6.95	0.12	8.17
1|1000|10|Dis	0.13	0.13	0.13	0.13	6.89	9.00	8.95	6.99	0.13	8.15
3|500|5|I	0.32	0.32	0.32	0.31	2.48	3.99	3.98	2.67	13.68	6.80
3|500|5|Dis	0.38	0.34	0.37	0.41	2.58	3.97	3.95	2.69	3.23	4.21
3|1000|5|I	0.16	0.15	0.16	0.15	2.59	3.99	3.99	2.62	5.69	2.71
3|1000|5|Dis	0.15	0.15	0.15	0.14	2.54	3.98	3.97	2.61	18.20	160.86
3|500|10|I	1.36	1.17	1.28	1.31	6.47	8.33	8.33	6.75	6.56	8.33
3|500|10|Dis	1.10	1.08	1.08	1.11	6.43	8.38	8.38	6.71	6.34	8.14
3|1000|10|I	0.43	0.43	0.43	0.41	6.65	8.32	8.26	6.71	185.67	30.86
3|1000|10|Dis	0.42	0.42	0.42	0.39	6.60	8.20	8.18	6.68	6.89	8.08
9|500|5|I	2.00	1.81	1.77	3.27	2.93	3.28	3.28	3.28	3.29	3.53
9|500|5|Dis	2.87	2.13	1.93	3.24	2.97	3.31	3.31	3.31	3.31	3.55
9|1000|5|I	0.61	0.60	0.60	0.62	2.96	3.33	3.33	3.27	3.33	3.59
9|1000|5|Dis	0.59	0.62	0.61	0.62	2.94	3.34	3.34	3.31	3.34	3.59
9|500|10|I	2611.26	6.86	8.58	6.70	6.13	6.74	6.74	6.74	6.70	6.90
9|500|10|Dis	4244.78	6.43	6.01	6.63	6.24	6.72	6.72	6.71	6.63	6.95
9|1000|10|I	3.52	2.54	2.78	2.72	6.23	6.66	6.66	6.66	6.65	6.90
9|1000|10|Dis	3.47	2.47	2.38	2.88	6.19	6.62	6.62	6.53	6.57	6.85

Table S30: Mean squared errors for the linear predictor are given for simulations under logistic regression models for $K_{\text{true}} \in \{1, 3, 9\}$, $p \in \{5, 10\}$, $n \in \{500, 1000\}$, and the scenarios "I" and "Dis". Compared methods are AIC, BIC, AICex1, AICex2, AICi, BICi, AICiex1, AICiex2, MoE, and MoEk, where "i" stands for the intercept-only or the constant model.

$K_{\text{true}}\|n\|p\|s$	**AIC**	**AICex1**	**AICex2**	**BIC**	**AICi**	**AICiex1**	**AICiex2**	**BICi**	**MoE**	**MoEk**
1|500|5|I	90| 10| 0	97| 3| 0	92| 8| 0	100| 0| 0	0|100| 0	15| 85| 0	14| 86| 0	0|100| 0	100| 0| 0	100| 0| 0
1|500|5|MoE	83| 17| 0	88| 12| 0	85| 15| 0	100| 0| 0	0|100| 0	9| 91| 0	8| 92| 0	0|100| 0	100| 0| 0	100| 0| 0
1|1000|5|I	94| 6| 0	98| 2| 0	95| 5| 0	100| 0| 0	0|100| 0	7| 93| 0	7| 93| 0	0|100| 0	100| 0| 0	100| 0| 0
1|1000|5|MoE	85| 15| 0	97| 3| 0	91| 9| 0	100| 0| 0	0|100| 0	2| 98| 0	2| 98| 0	0|100| 0	100| 0| 0	100| 0| 0
1|500|10|I	95| 5| 0	100| 0| 0	98| 2| 0	100| 0| 0	0|100| 0	15| 85| 0	13| 87| 0	0|100| 0	100| 0| 0	100| 0| 0
1|500|10|MoE	84| 16| 0	97| 3| 0	84| 16| 0	84| 16| 0	0|100| 0	13| 87| 0	13| 87| 0	0|100| 0	99| 1| 0	99| 1| 0
1|1000|10|I	96| 4| 0	100| 0| 0	98| 2| 0	100| 0| 0	0|100| 0	10| 90| 0	10| 90| 0	0|100| 0	100| 0| 0	100| 0| 0
1|1000|10|MoE	91| 9| 0	93| 7| 0	91| 9| 0	100| 0| 0	0|100| 0	13| 87| 0	12| 88| 0	0|100| 0	100| 0| 0	100| 0| 0
3|500|5|I	92| 8| 0	95| 5| 0	94| 6| 0	95| 0| 5	0|100| 0	0| 0|100	0| 2| 98	0| 97| 3	26| 3| 71	39| 1| 60
3|500|5|MoE	10| 46| 44	5| 0| 95	8| 15| 77	0| 0|100	1| 96| 3	40| 12| 48	39| 15| 46	8| 87| 5	0| 0|100	0| 0|100
3|1000|5|I	88| 12| 0	94| 6| 0	91| 9| 0	100| 0| 0	0|100| 0	0| 4| 96	0| 7| 93	0|100| 0	39| 13| 48	66| 3| 31
3|1000|5|MoE	18| 76| 6	10| 0| 90	24| 16| 60	0| 0|100	1| 99| 0	39| 24| 37	41| 25| 34	4| 95| 1	0| 0|100	1| 0| 99
3|500|10|I	81| 19| 0	94| 6| 0	88| 12| 0	85| 0| 15	0|100| 0	0| 0|100	0| 0|100	0| 88| 12	2| 0| 98	7| 0| 93
3|500|10|MoE	7| 26| 67	1| 0| 99	3| 8| 89	0| 0|100	0| 99| 1	22| 8| 70	25| 8| 67	2| 96| 2	0| 0|100	0| 0|100
3|1000|10|I	95| 5| 0	97| 3| 0	95| 5| 0	100| 0| 0	0|100| 0	0| 5| 95	0| 8| 92	0| 99| 1	30| 4| 66	61| 6| 33
3|1000|10|MoE	7| 21| 72	0| 0|100	5| 4| 91	0| 0|100	0| 99| 1	29| 20| 51	32| 21| 47	1| 98| 1	0| 0|100	0| 0|100
9|500|5|I	65| 29| 6	68| 13| 19	76| 17| 7	1| 0| 99	0| 83| 17	0| 0|100	0| 0|100	0| 1| 99	0| 0|100	0| 0|100
9|500|5|MoE	0| 0|100	0| 0|100	0| 0|100	0| 0|100	2| 8| 90	0| 0|100	0| 0|100	0| 0|100	0| 0|100	0| 0|100
9|1000|5|I	84| 16| 0	87| 10| 3	86| 14| 0	80| 1| 19	0| 98| 2	0| 0|100	0| 0|100	0| 23| 77	0| 0|100	0| 0|100
9|1000|5|MoE	0| 0|100	0| 0|100	0| 0|100	0| 0|100	3| 7| 90	0| 0|100	0| 0|100	0| 0|100	0| 0|100	0| 0|100
9|500|10|I	23| 46| 31	22| 4| 74	25| 4| 71	0| 0|100	0| 81| 19	0| 0|100	0| 0|100	0| 0|100	0| 0|100	0| 0|100
9|500|10|MoE	0| 1| 99	0| 0|100	0| 0|100	0| 0|100	1| 9| 90	0| 0|100	0| 0|100	0| 0|100	0| 0|100	0| 0|100
9|1000|10|I	52| 48| 0	70| 17| 13	65| 32| 3	36| 0| 64	1| 95| 4	0| 0|100	0| 0|100	0| 1| 99	0| 0|100	0| 0|100
9|1000|10|MoE	0| 0|100	0| 0|100	0| 0|100	0| 0|100	4| 15| 81	0| 0|100	0| 0|100	0| 0|100	0| 0|100	0| 0|100

Table S31: Proportions of selected number of clusters (%), correct-|over-|under-estimations, are given for simulations under logistic regression models for $K_{\text{true}} \in \{1, 3, 9\}$, $p \in \{5, 10\}$, $n \in \{500, 1000\}$, and the scenarios "I" and "MoE". Compared methods are AIC, BIC, AICex1, AICex2, AICi, BICi, AICiex1, AICiex2, MoE, and MoEk, where "i" stands for the intercept-only or the constant model.

$K_{\text{true}}\|n\|p\|s$	**AIC**	**AICex1**	**AICex2**	**BIC**	**AICi**	**AICiex1**	**AICiex2**	**BICi**	**MoE**	**MoEk**
1|500|5|I	90| 90	97| 97	92| 92	100|100	0| 0	15| 15	14| 14	0| 0	100|100	100|100
1|500|5|MoE	83| 83	88| 88	85| 85	100|100	0| 0	9| 9	8| 8	0| 0	100|100	100|100
1|1000|5|I	94| 94	98| 98	95| 95	100|100	0| 0	7| 7	7| 7	0| 0	100|100	100|100
1|1000|5|MoE	85| 85	97| 97	91| 91	100|100	0| 0	2| 2	2| 2	0| 0	100|100	100|100
1|500|10|I	95| 95	100|100	98| 98	100|100	0| 0	15| 15	13| 13	0| 0	100|100	100|100
1|500|10|MoE	84| 84	97| 97	84| 84	84| 84	0| 0	13| 13	13| 13	0| 0	99| 99	99| 99
1|1000|10|I	96| 96	100|100	98| 98	100|100	0| 0	10| 10	10| 10	0| 0	100|100	100|100
1|1000|10|MoE	91| 91	93| 93	91| 91	100|100	0| 0	13| 13	12| 12	0| 0	100|100	100|100
3|500|5|I	97| 97	99| 99	99| 98	98|100	30| 24	1| 0	7| 5	39| 31	32| 54	54| 81
3|500|5|MoE	3| 4	2| 1	2| 2	0| 0	3| 6	4| 4	4| 4	4| 6	2| 0	1| 0
3|1000|5|I	98| 95	99| 98	99| 97	100|100	30| 22	6| 7	13| 10	34| 25	59| 90	76| 95
3|1000|5|MoE	4| 6	2| 2	3| 3	0| 0	3| 6	4| 4	4| 4	3| 6	9| 4	5| 2
3|500|10|I	91| 88	99| 98	97| 95	94|100	31| 22	7| 4	11| 6	37| 28	15| 23	20| 30
3|500|10|MoE	1| 2	0| 0	1| 1	0| 0	3| 7	2| 2	2| 2	4| 8	0| 0	0| 0
3|1000|10|I	99|100	100|100	99|100	100|100	30| 21	8| 7	15| 12	35| 26	45| 65	67| 85
3|1000|10|MoE	2| 2	0| 0	1| 1	0| 0	3| 6	2| 3	3| 3	3| 6	2| 0	1| 0
9|500|5|I	97| 95	89| 89	98| 97	9| 9	69| 66	1| 0	2| 0	1| 1	0| 0	0| 0
9|500|5|MoE	0| 0	0| 0	0| 0	0| 0	0| 2	0| 0	0| 0	0| 0	0| 0	0| 0
9|1000|5|I	99| 98	98| 97	99| 98	97|100	81| 80	1| 0	2| 0	19| 19	0| 0	0| 0
9|1000|5|MoE	0| 0	0| 0	0| 0	0| 0	0| 1	0| 0	0| 0	0| 0	0| 0	0| 0
9|500|10|I	86| 85	50| 46	73| 71	0| 0	70| 69	3| 1	3| 1	1| 1	0| 0	0| 0
9|500|10|MoE	0| 0	0| 0	0| 0	0| 0	0| 2	0| 0	0| 0	0| 0	0| 0	0| 0
9|1000|10|I	97| 96	95| 93	98| 96	76| 83	80| 79	1| 0	1| 0	1| 1	0| 0	0| 0
9|1000|10|MoE	0| 0	0| 0	0| 0	0| 0	0| 2	0| 0	0| 0	0| 0	0| 0	0| 0

Table S32: Average of adjusted Rand index for all models | only for the first cluster (%) are given for simulations under logistic regression models for $K_{\text{true}} \in \{1, 3, 9\}$, $p \in \{5, 10\}$, $n \in \{500, 1000\}$, and the scenarios "I" and "MoE". Compared methods are AIC, BIC, AICex1, AICex2, AICi, BICi, AICiex1, AICiex2, MoE, and MoEk, where "i" stands for the intercept-only or the constant model.

K_{true}|n|p|s	**AIC**	**AICex1**	**AICex2**	**BIC**	**AICi**	**AICiex1**	**AICiex2**	**BICi**	**MoE**	**MoEk**
1|500|5|I	0.10	0.09	0.10	0.09	2.38	3.88	3.88	2.62	0.09	3.91
1|500|5|MoE	31.87	27.63	29.44	14.22	317.79	471.01	471.01	337.30	14.22	478.69
1|1000|5|I	0.05	0.05	0.05	0.05	2.30	3.91	3.91	2.41	0.05	3.94
1|1000|5|MoE	7.18	5.96	7.06	5.49	352.11	469.18	466.86	362.35	5.49	483.21
1|500|10|I	0.29	0.27	0.27	0.27	6.66	8.91	8.89	6.97	0.27	8.14
1|500|10|MoE	402.48	402.48	402.48	402.48	911.92	975.84	975.84	922.73	402.48	964.87
1|1000|10|I	0.13	0.12	0.12	0.12	6.86	9.07	9.01	6.95	0.12	8.17
1|1000|10|MoE	51.47	48.12	51.47	40.62	924.69	979.41	979.41	926.94	40.62	969.88
3|500|5|I	0.32	0.32	0.32	0.31	2.48	3.99	3.98	2.67	13.68	6.80
3|500|5|MoE	27.85	27.86	27.70	28.01	36.11	40.39	40.34	36.34	28.15	41.33
3|1000|5|I	0.16	0.15	0.16	0.15	2.59	3.99	3.99	2.62	5.69	2.71
3|1000|5|MoE	26.51	27.88	27.62	28.06	35.18	40.31	40.22	35.64	22.78	39.38
3|500|10|I	1.36	1.17	1.28	1.31	6.47	8.33	8.33	6.75	6.56	8.33
3|500|10|MoE	57.95	57.35	57.09	57.60	80.75	89.45	89.28	82.44	57.60	87.54
3|1000|10|I	0.43	0.43	0.43	0.41	6.65	8.32	8.26	6.71	185.67	30.86
3|1000|10|MoE	56.33	58.10	57.16	58.31	82.43	88.58	88.50	82.74	58.10	87.40
9|500|5|I	2.00	1.81	1.77	3.27	2.93	3.28	3.28	3.28	3.29	3.53
9|500|5|MoE	4.04	4.04	4.04	4.04	4.06	4.07	4.07	4.07	4.04	4.30
9|1000|5|I	0.61	0.60	0.60	0.62	2.96	3.33	3.33	3.27	3.33	3.59
9|1000|5|MoE	4.08	4.08	4.08	4.08	4.09	4.10	4.10	4.10	4.08	4.32
9|500|10|I	2611.26	6.86	8.58	6.70	6.13	6.74	6.74	6.74	6.70	6.90
9|500|10|MoE	8.12	8.12	8.12	8.12	8.11	8.21	8.21	8.21	8.12	8.40
9|1000|10|I	3.52	2.54	2.78	2.72	6.23	6.66	6.66	6.66	6.65	6.90
9|1000|10|MoE	8.15	8.16	8.16	8.16	8.14	8.20	8.20	8.20	8.16	8.40

Table S33: Mean squared errors for the linear predictor are given for simulations under logistic regression models for $K_{\text{true}} \in \{1, 3, 9\}$, $p \in \{5, 10\}$, $n \in \{500, 1000\}$, and the scenarios "I" and "MoE". Compared methods are AIC, BIC, AICex1, AICex2, AICi, BICi, AICiex1, AICiex2, MoE, and MoEk, where "i" stands for the intercept-only or the constant model.

$K_{\text{true}}\|n\|p\|s$	**AIC**	**AICex1**	**AICex2**	**BIC**	**AICi**	**AICiex1**	**AICiex2**	**BICi**	**MoE**	**MoEk**
1|500|5|I	90| 10| 0	97| 3| 0	92| 8| 0	100| 0| 0	0|100| 0	15| 85| 0	14| 86| 0	0|100| 0	100| 0| 0	100| 0| 0
1|500|5|io	76| 24| 0	97| 3| 0	93| 7| 0	100| 0| 0	58| 42| 0	90| 10| 0	89| 11| 0	99| 1| 0	100| 0| 0	100| 0| 0
1|1000|5|I	94| 6| 0	98| 2| 0	95| 5| 0	100| 0| 0	0|100| 0	7| 93| 0	7| 93| 0	0|100| 0	100| 0| 0	100| 0| 0
1|1000|5|io	87| 13| 0	97| 3| 0	92| 8| 0	100| 0| 0	51| 49| 0	93| 7| 0	91| 9| 0	99| 1| 0	100| 0| 0	100| 0| 0
1|500|10|I	95| 5| 0	100| 0| 0	98| 2| 0	100| 0| 0	0|100| 0	15| 85| 0	13| 87| 0	0|100| 0	100| 0| 0	100| 0| 0
1|500|10|io	70| 30| 0	99| 1| 0	94| 6| 0	100| 0| 0	47| 53| 0	95| 5| 0	93| 7| 0	100| 0| 0	100| 0| 0	100| 0| 0
1|1000|10|I	96| 4| 0	100| 0| 0	98| 2| 0	100| 0| 0	0|100| 0	10| 90| 0	10| 90| 0	0|100| 0	100| 0| 0	100| 0| 0
1|1000|10|io	87| 13| 0	100| 0| 0	99| 1| 0	100| 0| 0	53| 47| 0	96| 4| 0	94| 6| 0	100| 0| 0	100| 0| 0	100| 0| 0
3|500|5|I	92| 8| 0	95| 5| 0	94| 6| 0	95| 0| 5	0|100| 0	0| 0|100	0| 2| 98	0| 97| 3	26| 3| 71	39| 1| 60
3|500|5|io	13| 10| 77	15| 0| 85	16| 4| 80	0| 0|100	58| 42| 0	87| 13| 0	82| 18| 0	99| 1| 0	0| 0|100	0| 0|100
3|1000|5|I	88| 12| 0	94| 6| 0	91| 9| 0	100| 0| 0	0|100| 0	0| 4| 96	0| 7| 93	0|100| 0	39| 13| 48	66| 3| 31
3|1000|5|io	56| 6| 38	51| 4| 45	56| 5| 39	0| 0|100	51| 49| 0	94| 6| 0	91| 9| 0	100| 0| 0	0| 0|100	0| 0|100
3|500|10|I	81| 19| 0	94| 6| 0	88| 12| 0	85| 0| 15	0|100| 0	0| 0|100	0| 0|100	0| 88| 12	2| 0| 98	7| 0| 93
3|500|10|io	2| 2| 96	2| 0| 98	1| 0| 99	0| 0|100	60| 40| 0	91| 9| 0	89| 11| 0	98| 2| 0	0| 0|100	0| 0|100
3|1000|10|I	95| 5| 0	97| 3| 0	95| 5| 0	100| 0| 0	0|100| 0	0| 5| 95	0| 8| 92	0| 99| 1	30| 4| 66	61| 6| 33
3|1000|10|io	6| 0| 94	7| 0| 93	6| 0| 94	0| 0|100	53| 47| 0	91| 9| 0	82| 18| 0	100| 0| 0	0| 0|100	0| 0|100
9|500|5|I	65| 29| 6	68| 13| 19	76| 17| 7	1| 0| 99	0| 83| 17	0| 0|100	0| 0|100	0| 1| 99	0| 0|100	0| 0|100
9|500|5|io	0| 1| 99	0| 0|100	0| 0|100	0| 0|100	14| 31| 55	0| 0|100	0| 0|100	0| 0|100	0| 0|100	0| 0|100
9|1000|5|I	84| 16| 0	87| 10| 3	86| 14| 0	80| 1| 19	0| 98| 2	0| 0|100	0| 0|100	0| 23| 77	0| 0|100	0| 0|100
9|1000|5|io	0| 0|100	0| 0|100	0| 0|100	0| 0|100	37| 29| 34	1| 0| 99	1| 0| 99	3| 0| 97	0| 0|100	0| 0|100
9|500|10|I	23| 46| 31	22| 4| 74	25| 4| 71	0| 0|100	0| 81| 19	0| 0|100	0| 0|100	0| 0|100	0| 0|100	0| 0|100
9|500|10|io	0| 0|100	0| 0|100	0| 0|100	0| 0|100	9| 30| 61	0| 0|100	0| 0|100	0| 0|100	0| 0|100	0| 0|100
9|1000|10|I	52| 48| 0	70| 17| 13	65| 32| 3	36| 0| 64	1| 95| 4	0| 0|100	0| 0|100	0| 1| 99	0| 0|100	0| 0|100
9|1000|10|io	0| 0|100	0| 0|100	0| 0|100	0| 0|100	31| 44| 25	1| 0| 99	3| 0| 97	2| 0| 98	0| 0|100	0| 0|100

Table S34: Proportions of selected number of clusters (%), correct-|over-|under-estimations, are given for simulations under logistic regression models for $K_{\text{true}} \in \{1, 3, 9\}$, $p \in \{5, 10\}$, $n \in \{500, 1000\}$, and the scenarios "I" and "io". Compared methods are AIC, BIC, AICex1, AICex2, AICi, BICi, AICiex1, AICiex2, MoE, and MoEk, where "i" stands for the intercept-only or the constant model.

$K_{\text{true}}\|n\|p\|s$	**AIC**	**AICex1**	**AICex2**	**BIC**	**AICi**	**AICiex1**	**AICiex2**	**BICi**	**MoE**	**MoEk**
1|500|5|I	90| 90	97| 97	92| 92	100|100	0| 0	15| 15	14| 14	0| 0	100|100	100|100
1|500|5|io	76| 76	97| 97	93| 93	100|100	58| 58	90| 90	89| 89	99| 99	100|100	100|100
1|1000|5|I	94| 94	98| 98	95| 95	100|100	0| 0	7| 7	7| 7	0| 0	100|100	100|100
1|1000|5|io	87| 87	97| 97	92| 92	100|100	51| 51	93| 93	91| 91	99| 99	100|100	100|100
1|500|10|I	95| 95	100|100	98| 98	100|100	0| 0	15| 15	13| 13	0| 0	100|100	100|100
1|500|10|io	70| 70	99| 99	94| 94	100|100	47| 47	95| 95	93| 93	100|100	100|100	100|100
1|1000|10|I	96| 96	100|100	98| 98	100|100	0| 0	10| 10	10| 10	0| 0	100|100	100|100
1|1000|10|io	87| 87	100|100	99| 99	100|100	53| 53	96| 96	94| 94	100|100	100|100	100|100
3|500|5|I	97| 97	99| 99	99| 98	98|100	30| 24	1| 0	7| 5	39| 31	32| 54	54| 81
3|500|5|io	20| 19	15| 15	20| 19	0| 0	83| 79	98| 98	97| 96	100|100	0| 0	0| 0
3|1000|5|I	98| 95	99| 98	99| 97	100|100	30| 22	6| 7	13| 10	34| 25	59| 90	76| 95
3|1000|5|io	61| 61	54| 54	61| 61	0| 0	76| 73	99| 99	98| 98	100|100	0| 0	0| 0
3|500|10|I	91| 88	99| 98	97| 95	94|100	31| 22	7| 4	11| 6	37| 28	15| 23	20| 30
3|500|10|io	6| 6	2| 2	3| 3	0| 0	79| 76	99| 97	98| 97	100| 99	0| 0	0| 0
3|1000|10|I	99|100	100|100	99|100	100|100	30| 21	8| 7	15| 12	35| 26	45| 65	67| 85
3|1000|10|io	10| 8	7| 7	10| 8	0| 0	78| 74	99| 98	97| 96	100|100	0| 0	0| 0
9|500|5|I	97| 95	89| 89	98| 97	9| 9	69| 66	1| 0	2| 0	1| 1	0| 0	0| 0
9|500|5|io	3| 2	0| 0	2| 1	0| 0	84| 82	59| 54	61| 57	55| 47	0| 0	0| 0
9|1000|5|I	99| 98	98| 97	99| 98	97|100	81| 80	1| 0	2| 0	19| 19	0| 0	0| 0
9|1000|5|io	3| 1	1| 0	2| 0	0| 0	91| 90	67| 62	68| 65	66| 61	0| 0	0| 0
9|500|10|I	86| 85	50| 46	73| 71	0| 0	70| 69	3| 1	3| 1	1| 1	0| 0	0| 0
9|500|10|io	1| 0	0| 0	1| 0	0| 0	81| 80	59| 52	61| 56	57| 49	0| 0	0| 0
9|1000|10|I	97| 96	95| 93	98| 96	76| 83	80| 79	1| 0	1| 0	1| 1	0| 0	0| 0
9|1000|10|io	1| 1	0| 0	0| 0	0| 0	92| 90	67| 62	70| 68	68| 62	0| 0	0| 0

Table S35: Average of adjusted Rand index for all models | only for the first cluster (%) are given for simulations under logistic regression models for $K_{\text{true}} \in \{1, 3, 9\}$, $p \in \{5, 10\}$, $n \in \{500, 1000\}$, and the scenarios "I" and "io". Compared methods are AIC, BIC, AICex1, AICex2, AICi, BICi, AICiex1, AICiex2, MoE, and MoEk, where "i" stands for the intercept-only or the constant model.

$K_{\text{true}}\|n\|p\|s$	**AIC**	**AICex1**	**AICex2**	**BIC**	**AICi**	**AICiex1**	**AICiex2**	**BICi**	**MoE**	**MoEk**
1|500|5|I	0.10	0.09	0.10	0.09	2.38	3.88	3.88	2.62	0.09	3.91
1|500|5|io	0.30	0.22	0.23	0.21	0.10	0.03	0.03	0.02	0.21	9.28
1|1000|5|I	0.05	0.05	0.05	0.05	2.30	3.91	3.91	2.41	0.05	3.94
1|1000|5|io	0.13	0.11	0.12	0.11	0.08	0.01	0.01	0.01	0.11	9.28
1|500|10|I	0.29	0.27	0.27	0.27	6.66	8.91	8.89	6.97	0.27	8.14
1|500|10|io	0.69	0.50	0.54	0.50	0.17	0.02	0.02	0.02	0.50	9.29
1|1000|10|I	0.13	0.12	0.12	0.12	6.86	9.07	9.01	6.95	0.12	8.17
1|1000|10|io	0.24	0.22	0.22	0.22	0.06	0.01	0.01	0.01	0.22	9.29
3|500|5|I	0.32	0.32	0.32	0.31	2.48	3.99	3.98	2.67	13.68	6.80
3|500|5|io	0.28	0.25	0.26	0.23	0.14	0.08	0.09	0.08	0.23	4.59
3|1000|5|I	0.16	0.15	0.16	0.15	2.59	3.99	3.99	2.62	5.69	2.71
3|1000|5|io	0.28	0.24	0.27	0.17	0.11	0.05	0.05	0.04	0.17	4.59
3|500|10|I	1.36	1.17	1.28	1.31	6.47	8.33	8.33	6.75	6.56	8.33
3|500|10|io	0.26	0.25	0.26	0.25	0.14	0.07	0.07	0.06	0.25	4.56
3|1000|10|I	0.43	0.43	0.43	0.41	6.65	8.32	8.26	6.71	185.67	30.86
3|1000|10|io	0.16	0.16	0.16	0.15	0.09	0.04	0.04	0.04	0.15	4.58
9|500|5|I	2.00	1.81	1.77	3.27	2.93	3.28	3.28	3.28	3.29	3.53
9|500|5|io	0.12	0.11	0.12	0.11	0.30	0.25	0.23	0.29	0.11	2.84
9|1000|5|I	0.61	0.60	0.60	0.62	2.96	3.33	3.33	3.27	3.33	3.59
9|1000|5|io	0.05	0.05	0.05	0.05	0.13	0.16	0.16	0.17	0.05	2.83
9|500|10|I	2611.26	6.86	8.58	6.70	6.13	6.74	6.74	6.74	6.70	6.90
9|500|10|io	0.19	0.19	0.19	0.19	0.25	0.26	0.24	0.26	0.19	2.81
9|1000|10|I	3.52	2.54	2.78	2.72	6.23	6.66	6.66	6.66	6.65	6.90
9|1000|10|io	0.09	0.09	0.09	0.09	0.15	0.18	0.17	0.20	0.09	2.83

Table S36: Mean squared errors for the linear predictor are given for simulations under logistic regression models for $K_{\text{true}} \in \{1, 3, 9\}$, $p \in \{5, 10\}$, $n \in \{500, 1000\}$, and the scenarios "I" and "io". Compared methods are AIC, BIC, AICex1, AICex2, AICi, BICi, AICiex1, AICiex2, MoE, and MoEk, where "i" stands for the intercept-only or the constant model.

Cox regression with K-means clustering

K_{true}	n	p	s	**nAIC**	**nAICex1**	**nAICex2**	**nBIC**	**AIC**	**AICex1**	**AICex2**	**BIC**
1	500	5	I	100|100	100|100	100|100	100|100	0| 0	100|100	92| 92	0| 0
1	500	5	W	99| 99	99| 99	99| 99	100|100	0| 0	99| 99	89| 89	0| 0
1	1000	5	I	100|100	100|100	100|100	100|100	0| 0	99| 99	91| 91	0| 0
1	1000	5	W	100|100	100|100	100|100	100|100	0| 0	100|100	100|100	0| 0
1	500	10	I	100|100	100|100	100|100	100|100	0| 0	100|100	95| 95	0| 0
1	500	10	W	100|100	100|100	100|100	100|100	0| 0	100|100	88| 88	0| 0
1	1000	10	I	98| 98	100|100	98| 98	100|100	0| 0	100|100	97| 97	0| 0
1	1000	10	W	100|100	100|100	100|100	100|100	0| 0	100|100	98| 98	0| 0
3	500	5	I	100|100	100|100	100|100	100|100	26| 19	100| 99	96| 95	26| 20
3	500	5	W	99| 99	98| 98	99| 99	98|100	27| 22	99| 98	95| 93	28| 22
3	1000	5	I	100|100	100|100	100|100	100|100	25| 18	100|100	99| 98	26| 18
3	1000	5	W	99|100	99|100	99|100	99|100	27| 21	99| 99	97| 96	27| 21
3	500	10	I	100|100	100|100	100|100	100|100	26| 19	100| 99	91| 88	31| 23
3	500	10	W	100|100	100| 99	100| 99	99|100	27| 21	99| 99	87| 83	33| 26
3	1000	10	I	100|100	100|100	100|100	100|100	26| 18	100| 99	98| 98	26| 18
3	1000	10	W	100|100	100|100	100|100	99|100	26| 19	100|100	98| 98	26| 20
9	500	5	I	99|100	85| 84	97| 97	99|100	79| 83	85| 84	92| 92	79| 82
9	500	5	W	99| 99	83| 83	97| 97	96| 99	79| 81	83| 83	92| 91	80| 84
9	1000	5	I	100|100	99| 99	99|100	100|100	79| 75	99| 99	98| 98	79| 76
9	1000	5	W	99|100	95| 95	99|100	99|100	79| 77	95| 95	98| 97	79| 77
9	500	10	I	97| 98	91| 91	98| 99	96| 99	80| 83	90| 89	85| 83	87| 87
9	500	10	W	97| 97	96| 95	97| 96	89| 95	81| 78	95| 94	88| 86	90| 88
9	1000	10	I	99|100	95| 95	99| 99	99|100	79| 77	95| 94	94| 93	79| 80
9	1000	10	W	99| 99	97| 97	98| 98	97|100	80| 77	97| 97	94| 93	80| 79

Table S37: Average of adjusted Rand index for all models | only for the first cluster (%) are given for simulations under logistic regression models for $K_{\text{true}} \in \{1, 3, 9\}$, $p \in \{5, 10\}$, $n \in \{500, 1000\}$, and the scenarios "I" and "W". Compared methods are nAIC, nBIC, nAICex1, nAICex2, AIC, BIC, AICex1, and AICex2, where "n" stands for the normalized partial log-likelihood.

K_{true}	n	p	s	**nAIC**	**nAICex1**	**nAICex2**	**nBIC**	**AIC**	**AICex1**	**AICex2**	**BIC**
1	500	5	I	0.03	0.03	0.03	0.03	4.66	0.03	0.03	4.38
1	500	5	W	0.02	0.02	0.02	0.02	8.18	0.02	0.03	6.51
1	1000	5	I	0.02	0.02	0.02	0.02	3.28	0.02	0.02	3.25
1	1000	5	W	0.01	0.01	0.01	0.01	2.82	0.01	0.01	2.82
1	500	10	I	0.07	0.07	0.07	0.07	1549.69	0.07	0.07	1095.89
1	500	10	W	0.05	0.05	0.05	0.05	1574.73	0.05	0.06	409.49
1	1000	10	I	0.04	0.03	0.04	0.03	4.99	0.03	0.04	4.68
1	1000	10	W	0.03	0.03	0.03	0.03	6.16	0.03	0.03	6.14
3	500	5	I	0.09	0.09	0.09	0.09	2.48	0.09	0.12	2.35
3	500	5	W	0.09	0.09	0.09	0.09	4.98	0.09	0.11	3.66
3	1000	5	I	0.04	0.04	0.04	0.04	1.80	0.04	0.05	1.77
3	1000	5	W	0.05	0.05	0.05	0.05	2.28	0.05	0.05	2.24
3	500	10	I	0.22	0.21	0.22	0.21	662.65	0.22	0.32	11.65
3	500	10	W	0.20	0.20	0.20	0.21	1205.57	0.20	0.28	49.83
3	1000	10	I	0.09	0.09	0.09	0.09	2.73	0.09	0.10	2.52
3	1000	10	W	0.09	0.09	0.09	0.09	3.95	0.09	0.10	3.62
9	500	5	I	0.34	0.38	0.35	0.34	1.40	0.40	0.58	1.31
9	500	5	W	0.37	0.56	0.38	0.39	1.54	0.57	0.70	1.49
9	1000	5	I	0.14	0.14	0.14	0.14	0.65	0.14	0.16	0.58
9	1000	5	W	0.14	0.15	0.14	0.14	0.84	0.17	0.17	0.80
9	500	10	I	1.90	1.52	1.52	1.41	1133.77	1.90	82.06	220.39
9	500	10	W	1.30	1.24	1.22	1.30	784.39	1.48	3.86	5.41
9	1000	10	I	0.40	0.40	0.40	0.40	1.49	0.42	0.59	1.30
9	1000	10	W	0.35	0.36	0.36	0.40	1.69	0.39	0.62	1.54

Table S38: Mean squared errors for the linear predictor are given for simulations under logistic regression models for $K_{\text{true}} \in \{1, 3, 9\}$, $p \in \{5, 10\}$, $n \in \{500, 1000\}$, and the scenarios "I" and "W". Compared methods are nAIC, nBIC, nAICex1, nAICex2, AIC, BIC, AICex1, and AICex2, where "n" stands for the normalized partial log-likelihood.

K_{true}|n|p|s	**nAIC**	**nAICex1**	**nAICex2**	**nBIC**	**AIC**	**AICex1**	**AICex2**	**BIC**
1|500|5|I	100| 0| 0	100| 0| 0	100| 0| 0	100| 0| 0	0|100| 0	100| 0| 0	92| 8| 0	0|100| 0
1|500|5|k1	100| 0| 0	100| 0| 0	100| 0| 0	100| 0| 0	0|100| 0	100| 0| 0	92| 8| 0	0|100| 0
1|1000|5|I	100| 0| 0	100| 0| 0	100| 0| 0	100| 0| 0	0|100| 0	99| 1| 0	91| 9| 0	0|100| 0
1|1000|5|k1	100| 0| 0	100| 0| 0	100| 0| 0	100| 0| 0	0|100| 0	99| 1| 0	91| 9| 0	0|100| 0
1|500|10|I	100| 0| 0	100| 0| 0	100| 0| 0	100| 0| 0	0|100| 0	100| 0| 0	95| 5| 0	0|100| 0
1|500|10|k1	100| 0| 0	100| 0| 0	100| 0| 0	100| 0| 0	0|100| 0	100| 0| 0	95| 5| 0	0|100| 0
1|1000|10|I	98| 2| 0	100| 0| 0	98| 2| 0	100| 0| 0	0|100| 0	100| 0| 0	97| 3| 0	0|100| 0
1|1000|10|k1	98| 2| 0	100| 0| 0	98| 2| 0	100| 0| 0	0|100| 0	100| 0| 0	97| 3| 0	0|100| 0
3|500|5|I	99| 1| 0	100| 0| 0	99| 1| 0	100| 0| 0	0|100| 0	97| 3| 0	80| 20| 0	0|100| 0
3|500|5|k1	99| 1| 0	99| 1| 0	99| 1| 0	100| 0| 0	0|100| 0	94| 6| 0	83| 17| 0	0|100| 0
3|1000|5|I	99| 1| 0	99| 1| 0	99| 1| 0	100| 0| 0	0|100| 0	99| 1| 0	91| 9| 0	0|100| 0
3|1000|5|k1	99| 1| 0	99| 1| 0	99| 1| 0	100| 0| 0	0|100| 0	99| 1| 0	94| 6| 0	0|100| 0
3|500|10|I	99| 1| 0	100| 0| 0	99| 1| 0	100| 0| 0	0|100| 0	97| 3| 0	64| 36| 0	0|100| 0
3|500|10|k1	98| 2| 0	100| 0| 0	98| 2| 0	100| 0| 0	0|100| 0	97| 3| 0	78| 22| 0	0|100| 0
3|1000|10|I	100| 0| 0	100| 0| 0	100| 0| 0	100| 0| 0	0|100| 0	99| 1| 0	91| 9| 0	0|100| 0
3|1000|10|k1	99| 1| 0	99| 1| 0	99| 1| 0	100| 0| 0	0|100| 0	99| 1| 0	92| 8| 0	0|100| 0
9|500|5|I	84| 16| 0	67| 3| 30	84| 9| 7	87| 4| 9	0|100| 0	60| 10| 30	32| 63| 5	0|100| 0
9|500|5|k1	82| 16| 2	2| 0| 98	16| 2| 82	37| 1| 62	0|100| 0	2| 0| 98	11| 8| 81	0|100| 0
9|1000|5|I	92| 8| 0	94| 1| 5	93| 4| 3	95| 5| 0	0|100| 0	89| 6| 5	70| 28| 2	0|100| 0
9|1000|5|k1	86| 14| 0	36| 1| 63	56| 6| 38	90| 5| 5	0|100| 0	36| 1| 63	46| 16| 38	0|100| 0
9|500|10|I	55| 45| 0	69| 15| 16	68| 32| 0	61| 4| 35	0|100| 0	51| 34| 15	3| 97| 0	8| 92| 0
9|500|10|k1	44| 52| 4	9| 0| 91	42| 43| 15	0| 0|100	0|100| 0	9| 0| 91	4| 83| 13	0|100| 0
9|1000|10|I	86| 14| 0	86| 1| 13	86| 11| 3	86| 11| 3	0|100| 0	81| 6| 13	40| 57| 3	0|100| 0
9|1000|10|k1	75| 25| 0	55| 1| 44	76| 15| 9	74| 13| 13	0|100| 0	54| 2| 44	40| 52| 8	0|100| 0

Table S39: Proportions of selected number of clusters (%), correct-|over-|under-estimations, are given for simulations under Cox regression models for $K_{\text{true}} \in \{1, 3, 9\}$, $p \in \{5, 10\}$, $n \in \{500, 1000\}$, and the scenarios "I" and "k1". Compared methods are nAIC, nBIC, nAICex1, nAICex2, AIC, BIC, AICex1, and AICex2, where "n" stands for the normalized partial log-likelihood.

$K_{\text{true}}\|n\|p\|s$	**nAIC**	**nAICex1**	**nAICex2**	**nBIC**	**AIC**	**AICex1**	**AICex2**	**BIC**
1|500|5|I	100|100	100|100	100|100	100|100	0| 0	100|100	92| 92	0| 0
1|500|5|k1	100|100	100|100	100|100	100|100	0| 0	100|100	92| 92	0| 0
1|1000|5|I	100|100	100|100	100|100	100|100	0| 0	99| 99	91| 91	0| 0
1|1000|5|k1	100|100	100|100	100|100	100|100	0| 0	99| 99	91| 91	0| 0
1|500|10|I	100|100	100|100	100|100	100|100	0| 0	100|100	95| 95	0| 0
1|500|10|k1	100|100	100|100	100|100	100|100	0| 0	100|100	95| 95	0| 0
1|1000|10|I	98| 98	100|100	98| 98	100|100	0| 0	100|100	97| 97	0| 0
1|1000|10|k1	98| 98	100|100	98| 98	100|100	0| 0	100|100	97| 97	0| 0
3|500|5|I	100|100	100|100	100|100	100|100	26| 19	100| 99	96| 95	26| 20
3|500|5|k1	100|100	100|100	100|100	100|100	25| 29	99|100	96|100	25| 29
3|1000|5|I	100|100	100|100	100|100	100|100	25| 18	100|100	99| 98	26| 18
3|1000|5|k1	100|100	100|100	100|100	100|100	25| 29	100|100	99|100	25| 29
3|500|10|I	100|100	100|100	100|100	100|100	26| 19	100| 99	91| 88	31| 23
3|500|10|k1	100|100	100|100	100|100	100|100	25| 32	100|100	93| 95	28| 36
3|1000|10|I	100|100	100|100	100|100	100|100	26| 18	100| 99	98| 98	26| 18
3|1000|10|k1	100|100	100|100	100|100	100|100	25| 29	100|100	99|100	25| 30
9|500|5|I	99|100	85| 84	97| 97	99|100	79| 83	85| 84	92| 92	79| 82
9|500|5|k1	99|100	23| 11	55| 45	52| 52	78| 92	24| 11	56| 44	78| 93
9|1000|5|I	100|100	99| 99	99|100	100|100	79| 75	99| 99	98| 98	79| 76
9|1000|5|k1	99|100	60| 50	83| 75	99|100	78| 96	60| 50	82| 75	78| 96
9|500|10|I	97| 98	91| 91	98| 99	96| 99	80| 83	90| 89	85| 83	87| 87
9|500|10|k1	95| 99	18| 13	86| 88	3| 3	79| 96	18| 13	79| 87	84| 97
9|1000|10|I	99|100	95| 95	99| 99	99|100	79| 77	95| 94	94| 93	79| 80
9|1000|10|k1	99|100	71| 63	97| 96	98| 99	78| 95	70| 63	94| 95	79| 98

Table S40: Average of adjusted Rand index for all models | only for the first cluster (%) are given for simulations under logistic regression models for $K_{\text{true}} \in \{1, 3, 9\}$, $p \in \{5, 10\}$, $n \in \{500, 1000\}$, and the scenarios "I" and "k1". Compared methods are nAIC, nBIC, nAICex1, nAICex2, AIC, BIC, AICex1, and AICex2, where "n" stands for the normalized partial log-likelihood.

$K_{\text{true}}\|n\|p\|s$	**nAIC**	**nAICex1**	**nAICex2**	**nBIC**	**AIC**	**AICex1**	**AICex2**	**BIC**
1|500|5|I	0.03	0.03	0.03	0.03	4.66	0.03	0.03	4.38
1|500|5|k1	0.03	0.03	0.03	0.03	4.66	0.03	0.03	4.38
1|1000|5|I	0.02	0.02	0.02	0.02	3.28	0.02	0.02	3.25
1|1000|5|k1	0.02	0.02	0.02	0.02	3.28	0.02	0.02	3.25
1|500|10|I	0.07	0.07	0.07	0.07	1549.69	0.07	0.07	1095.89
1|500|10|k1	0.07	0.07	0.07	0.07	1549.69	0.07	0.07	1095.89
1|1000|10|I	0.04	0.03	0.04	0.03	4.99	0.03	0.04	4.68
1|1000|10|k1	0.04	0.03	0.04	0.03	4.99	0.03	0.04	4.68
3|500|5|I	0.09	0.09	0.09	0.09	2.48	0.09	0.12	2.35
3|500|5|k1	0.07	0.07	0.07	0.07	1.33	0.07	0.08	1.33
3|1000|5|I	0.04	0.04	0.04	0.04	1.80	0.04	0.05	1.77
3|1000|5|k1	0.04	0.04	0.04	0.04	0.93	0.04	0.04	0.92
3|500|10|I	0.22	0.21	0.22	0.21	662.65	0.22	0.32	11.65
3|500|10|k1	0.18	0.18	0.18	0.18	9.05	0.18	0.21	3.72
3|1000|10|I	0.09	0.09	0.09	0.09	2.73	0.09	0.10	2.52
3|1000|10|k1	0.07	0.07	0.07	0.07	1.35	0.07	0.07	1.31
9|500|5|I	0.34	0.38	0.35	0.34	1.40	0.40	0.58	1.31
9|500|5|k1	0.24	1.03	0.83	0.59	0.69	1.03	0.83	0.67
9|1000|5|I	0.14	0.14	0.14	0.14	0.65	0.14	0.16	0.58
9|1000|5|k1	0.09	0.82	0.12	0.09	0.27	0.82	0.15	0.26
9|500|10|I	1.90	1.52	1.52	1.41	1133.77	1.90	82.06	220.39
9|500|10|k1	1.06	2.06	1.08	2.14	8.01	2.06	2.23	3.48
9|1000|10|I	0.40	0.40	0.40	0.40	1.49	0.42	0.59	1.30
9|1000|10|k1	0.24	0.30	0.26	0.24	0.64	0.30	0.30	0.59

Table S41: Mean squared errors for the linear predictor are given for simulations under logistic regression models for $K_{\text{true}} \in \{1, 3, 9\}$, $p \in \{5, 10\}$, $n \in \{500, 1000\}$, and the scenarios "I" and "k1". Compared methods are nAIC, nBIC, nAICex1, nAICex2, AIC, BIC, AICex1, and AICex2, where "n" stands for the normalized partial log-likelihood.

$K_{\text{true}}\|n\|p\|s$	**nAIC**	**nAICex1**	**nAICex2**	**nBIC**	**AIC**	**AICex1**	**AICex2**	**BIC**
1|500|5|I	100| 0| 0	100| 0| 0	100| 0| 0	100| 0| 0	0|100| 0	100| 0| 0	92| 8| 0	0|100| 0
1|500|5|Imb	100| 0| 0	100| 0| 0	100| 0| 0	100| 0| 0	0|100| 0	100| 0| 0	92| 8| 0	0|100| 0
1|1000|5|I	100| 0| 0	100| 0| 0	100| 0| 0	100| 0| 0	0|100| 0	99| 1| 0	91| 9| 0	0|100| 0
1|1000|5|Imb	100| 0| 0	100| 0| 0	100| 0| 0	100| 0| 0	0|100| 0	99| 1| 0	91| 9| 0	0|100| 0
1|500|10|I	100| 0| 0	100| 0| 0	100| 0| 0	100| 0| 0	0|100| 0	100| 0| 0	95| 5| 0	0|100| 0
1|500|10|Imb	100| 0| 0	100| 0| 0	100| 0| 0	100| 0| 0	0|100| 0	100| 0| 0	95| 5| 0	0|100| 0
1|1000|10|I	98| 2| 0	100| 0| 0	98| 2| 0	100| 0| 0	0|100| 0	100| 0| 0	97| 3| 0	0|100| 0
1|1000|10|Imb	98| 2| 0	100| 0| 0	98| 2| 0	100| 0| 0	0|100| 0	100| 0| 0	97| 3| 0	0|100| 0
3|500|5|I	99| 1| 0	100| 0| 0	99| 1| 0	100| 0| 0	0|100| 0	97| 3| 0	80| 20| 0	0|100| 0
3|500|5|Imb	100| 0| 0	100| 0| 0	100| 0| 0	100| 0| 0	0|100| 0	99| 1| 0	88| 12| 0	0|100| 0
3|1000|5|I	99| 1| 0	99| 1| 0	99| 1| 0	100| 0| 0	0|100| 0	99| 1| 0	91| 9| 0	0|100| 0
3|1000|5|Imb	99| 1| 0	99| 1| 0	99| 1| 0	100| 0| 0	0|100| 0	98| 2| 0	95| 5| 0	0|100| 0
3|500|10|I	99| 1| 0	100| 0| 0	99| 1| 0	100| 0| 0	0|100| 0	97| 3| 0	64| 36| 0	0|100| 0
3|500|10|Imb	99| 1| 0	100| 0| 0	99| 1| 0	100| 0| 0	0|100| 0	100| 0| 0	85| 15| 0	0|100| 0
3|1000|10|I	100| 0| 0	100| 0| 0	100| 0| 0	100| 0| 0	0|100| 0	99| 1| 0	91| 9| 0	0|100| 0
3|1000|10|Imb	100| 0| 0	100| 0| 0	100| 0| 0	100| 0| 0	0|100| 0	100| 0| 0	97| 3| 0	0|100| 0
9|500|5|I	84| 16| 0	67| 3| 30	84| 9| 7	87| 4| 9	0|100| 0	60| 10| 30	32| 63| 5	0|100| 0
9|500|5|Imb	48| 37| 15	17| 0| 83	44| 18| 38	29| 3| 68	0|100| 0	16| 1| 83	22| 50| 28	0|100| 0
9|1000|5|I	92| 8| 0	94| 1| 5	93| 4| 3	95| 5| 0	0|100| 0	89| 6| 5	70| 28| 2	0|100| 0
9|1000|5|Imb	44| 51| 5	35| 1| 64	44| 12| 44	45| 19| 36	0|100| 0	32| 4| 64	38| 20| 42	0|100| 0
9|500|10|I	55| 45| 0	69| 15| 16	68| 32| 0	61| 4| 35	0|100| 0	51| 34| 15	3| 97| 0	8| 92| 0
9|500|10|Imb	27| 51| 22	11| 9| 80	26| 30| 44	14| 2| 84	3| 87| 10	11| 12| 77	15| 57| 28	3| 86| 11
9|1000|10|I	86| 14| 0	86| 1| 13	86| 11| 3	86| 11| 3	0|100| 0	81| 6| 13	40| 57| 3	0|100| 0
9|1000|10|Imb	27| 64| 9	25| 3| 72	33| 20| 47	20| 7| 73	0|100| 0	25| 3| 72	32| 32| 36	0|100| 0

Table S42: Proportions of selected number of clusters (%), correct-|over-|under-estimations, are given for simulations under Cox regression models for $K_{\text{true}} \in \{1, 3, 9\}$, $p \in \{5, 10\}$, $n \in \{500, 1000\}$, and the scenarios "I" and "Imb". Compared methods are nAIC, nBIC, nAICex1, nAICex2, AIC, BIC, AICex1, and AICex2, where "n" stands for the normalized partial log-likelihood.

K_{true}	n	p	s	**nAIC**	**nAICex1**	**nAICex2**	**nBIC**	**AIC**	**AICex1**	**AICex2**	**BIC**
1	500	5	I	100|100	100|100	100|100	100|100	0| 0	100|100	92| 92	0| 0
1	500	5	Imb	100|100	100|100	100|100	100|100	0| 0	100|100	92| 92	0| 0
1	1000	5	I	100|100	100|100	100|100	100|100	0| 0	99| 99	91| 91	0| 0
1	1000	5	Imb	100|100	100|100	100|100	100|100	0| 0	99| 99	91| 91	0| 0
1	500	10	I	100|100	100|100	100|100	100|100	0| 0	100|100	95| 95	0| 0
1	500	10	Imb	100|100	100|100	100|100	100|100	0| 0	100|100	95| 95	0| 0
1	1000	10	I	98| 98	100|100	98| 98	100|100	0| 0	100|100	97| 97	0| 0
1	1000	10	Imb	98| 98	100|100	98| 98	100|100	0| 0	100|100	97| 97	0| 0
3	500	5	I	100|100	100|100	100|100	100|100	26| 19	100| 99	96| 95	26| 20
3	500	5	Imb	100|100	100|100	100|100	100|100	16| 43	100|100	95|100	16| 45
3	1000	5	I	100|100	100|100	100|100	100|100	25| 18	100|100	99| 98	26| 18
3	1000	5	Imb	100|100	100|100	100|100	100|100	14| 48	99|100	98|100	14| 48
3	500	10	I	100|100	100|100	100|100	100|100	26| 19	100| 99	91| 88	31| 23
3	500	10	Imb	100|100	100|100	100|100	100|100	16| 48	100|100	91| 99	19| 58
3	1000	10	I	100|100	100|100	100|100	100|100	26| 18	100| 99	98| 98	26| 18
3	1000	10	Imb	100|100	100|100	100|100	100|100	14| 51	100|100	99|100	14| 53
9	500	5	I	99|100	85| 84	97| 97	99|100	79| 83	85| 84	92| 92	79| 82
9	500	5	Imb	97| 98	61| 47	94| 89	97| 80	68| 98	61| 47	90| 90	68| 98
9	1000	5	I	100|100	99| 99	99|100	100|100	79| 75	99| 99	98| 98	79| 76
9	1000	5	Imb	96| 99	85| 64	94| 84	98| 88	68|100	84| 64	93| 84	68|100
9	500	10	I	97| 98	91| 91	98| 99	96| 99	80| 83	90| 89	85| 83	87| 87
9	500	10	Imb	94| 95	54| 43	90| 85	95| 72	74| 96	53| 43	82| 87	77| 96
9	1000	10	I	99|100	95| 95	99| 99	99|100	79| 77	95| 94	94| 93	79| 80
9	1000	10	Imb	95| 99	81| 62	96| 87	95| 72	68|100	81| 62	94| 87	68|100

Table S43: Average of adjusted Rand index for all models | only for the first cluster (%) are given for simulations under logistic regression models for $K_{\text{true}} \in \{1, 3, 9\}$, $p \in \{5, 10\}$, $n \in \{500, 1000\}$, and the scenarios "I" and "Imb". Compared methods are nAIC, nBIC, nAICex1, nAICex2, AIC, BIC, AICex1, and AICex2, where "n" stands for the normalized partial log-likelihood.

K_{true}	n	p	s	**nAIC**	**nAICex1**	**nAICex2**	**nBIC**	**AIC**	**AICex1**	**AICex2**	**BIC**
1	500	5	I	0.03	0.03	0.03	0.03	4.66	0.03	0.03	4.38
1	500	5	Imb	0.03	0.03	0.03	0.03	4.66	0.03	0.03	4.38
1	1000	5	I	0.02	0.02	0.02	0.02	3.28	0.02	0.02	3.25
1	1000	5	Imb	0.02	0.02	0.02	0.02	3.28	0.02	0.02	3.25
1	500	10	I	0.07	0.07	0.07	0.07	1549.69	0.07	0.07	1095.89
1	500	10	Imb	0.07	0.07	0.07	0.07	1549.69	0.07	0.07	1095.89
1	1000	10	I	0.04	0.03	0.04	0.03	4.99	0.03	0.04	4.68
1	1000	10	Imb	0.04	0.03	0.04	0.03	4.99	0.03	0.04	4.68
3	500	5	I	0.09	0.09	0.09	0.09	2.48	0.09	0.12	2.35
3	500	5	Imb	0.10	0.10	0.10	0.10	3.04	0.10	0.10	2.84
3	1000	5	I	0.04	0.04	0.04	0.04	1.80	0.04	0.05	1.77
3	1000	5	Imb	0.05	0.05	0.05	0.04	2.24	0.05	0.05	2.24
3	500	10	I	0.22	0.21	0.22	0.21	662.65	0.22	0.32	11.65
3	500	10	Imb	0.27	0.27	0.27	0.27	1566.23	0.27	0.28	12.87
3	1000	10	I	0.09	0.09	0.09	0.09	2.73	0.09	0.10	2.52
3	1000	10	Imb	0.10	0.10	0.10	0.10	3.53	0.10	0.10	3.41
9	500	5	I	0.34	0.38	0.35	0.34	1.40	0.40	0.58	1.31
9	500	5	Imb	0.58	1.75	0.62	0.62	1.56	1.75	0.74	1.45
9	1000	5	I	0.14	0.14	0.14	0.14	0.65	0.14	0.16	0.58
9	1000	5	Imb	0.19	0.43	0.28	0.23	0.63	0.43	0.29	0.62
9	500	10	I	1.90	1.52	1.52	1.41	1133.77	1.90	82.06	220.39
9	500	10	Imb	494.37	5.79	6.21	1.75	988.40	5.79	12.84	578.85
9	1000	10	I	0.40	0.40	0.40	0.40	1.49	0.42	0.59	1.30
9	1000	10	Imb	0.70	0.97	0.79	0.91	1.51	0.97	0.93	1.41

Table S44: Mean squared errors for the linear predictor are given for simulations under logistic regression models for $K_{\text{true}} \in \{1, 3, 9\}$, $p \in \{5, 10\}$, $n \in \{500, 1000\}$, and the scenarios "I" and "Imb". Compared methods are nAIC, nBIC, nAICex1, nAICex2, AIC, BIC, AICex1, and AICex2, where "n" stands for the normalized partial log-likelihood.

K_{true}	n	p	s	**nAIC**			**nAICex1**			**nAICex2**			**nBIC**			**AIC**			**AICex1**			**AICex2**			**BIC**		
1	500	5	I	100	0	0	100	0	0	100	0	0	100	0	0	0	100	0	100	0	0	92	8	0	0	100	0
1	500	5	Dis	100	0	0	100	0	0	100	0	0	100	0	0	0	100	0	100	0	0	92	8	0	0	100	0
1	1000	5	I	100	0	0	100	0	0	100	0	0	100	0	0	0	100	0	99	1	0	91	9	0	0	100	0
1	1000	5	Dis	100	0	0	100	0	0	100	0	0	100	0	0	0	100	0	99	1	0	91	9	0	0	100	0
1	500	10	I	100	0	0	100	0	0	100	0	0	100	0	0	0	100	0	100	0	0	95	5	0	0	100	0
1	500	10	Dis	100	0	0	100	0	0	100	0	0	100	0	0	0	100	0	100	0	0	95	5	0	0	100	0
1	1000	10	I	98	2	0	100	0	0	98	2	0	100	0	0	0	100	0	100	0	0	97	3	0	0	100	0
1	1000	10	Dis	98	2	0	100	0	0	98	2	0	100	0	0	0	100	0	100	0	0	97	3	0	0	100	0
3	500	5	I	99	1	0	100	0	0	99	1	0	100	0	0	0	100	0	97	3	0	80	20	0	0	100	0
3	500	5	Dis	99	1	0	100	0	0	100	0	0	100	0	0	0	100	0	98	2	0	85	15	0	0	100	0
3	1000	5	I	99	1	0	99	1	0	99	1	0	100	0	0	0	100	0	99	1	0	91	9	0	0	100	0
3	1000	5	Dis	99	1	0	99	1	0	99	1	0	100	0	0	0	100	0	98	2	0	90	10	0	0	100	0
3	500	10	I	99	1	0	100	0	0	99	1	0	100	0	0	0	100	0	97	3	0	64	36	0	0	100	0
3	500	10	Dis	97	3	0	100	0	0	98	2	0	100	0	0	0	100	0	99	1	0	62	38	0	0	100	0
3	1000	10	I	100	0	0	100	0	0	100	0	0	100	0	0	0	100	0	99	1	0	91	9	0	0	100	0
3	1000	10	Dis	100	0	0	100	0	0	100	0	0	100	0	0	0	100	0	100	0	0	92	8	0	0	100	0
9	500	5	I	84	16	0	67	3	30	84	9	7	87	4	9	0	100	0	60	10	30	32	63	5	0	100	0
9	500	5	Dis	72	28	0	61	3	36	71	15	14	71	13	16	0	100	0	54	10	36	17	73	10	0	100	0
9	1000	5	I	92	8	0	94	1	5	93	4	3	95	5	0	0	100	0	89	6	5	70	28	2	0	100	0
9	1000	5	Dis	82	18	0	82	2	16	82	4	14	83	12	5	0	100	0	79	5	16	58	31	11	0	100	0
9	500	10	I	55	45	0	69	15	16	68	32	0	61	4	35	0	100	0	51	34	15	3	97	0	8	92	0
9	500	10	Dis	44	56	0	61	19	20	56	43	1	64	5	31	0	100	0	43	39	18	4	96	0	6	94	0
9	1000	10	I	86	14	0	86	1	13	86	11	3	86	11	3	0	100	0	81	6	13	40	57	3	0	100	0
9	1000	10	Dis	79	21	0	80	3	17	79	15	6	81	14	5	0	100	0	75	8	17	42	56	2	0	100	0

Table S45: Proportions of selected number of clusters (%), correct-|over-|under-estimations, are given for simulations under Cox regression models for $K_{\text{true}} \in \{1, 3, 9\}$, $p \in \{5, 10\}$, $n \in \{500, 1000\}$, and the scenarios "I" and "Dis". Compared methods are nAIC, nBIC, nAICex1, nAICex2, AIC, BIC, AICex1, and AICex2, where "n" stands for the normalized partial log-likelihood.

K_{true}	n	p	s	**nAIC**	**nAICex1**	**nAICex2**	**nBIC**	**AIC**	**AICex1**	**AICex2**	**BIC**
1	500	5	I	100|100	100|100	100|100	100|100	0| 0	100|100	92| 92	0| 0
1	500	5	Dis	100|100	100|100	100|100	100|100	0| 0	100|100	92| 92	0| 0
1	1000	5	I	100|100	100|100	100|100	100|100	0| 0	99| 99	91| 91	0| 0
1	1000	5	Dis	100|100	100|100	100|100	100|100	0| 0	99| 99	91| 91	0| 0
1	500	10	I	100|100	100|100	100|100	100|100	0| 0	100|100	95| 95	0| 0
1	500	10	Dis	100|100	100|100	100|100	100|100	0| 0	100|100	95| 95	0| 0
1	1000	10	I	98| 98	100|100	98| 98	100|100	0| 0	100|100	97| 97	0| 0
1	1000	10	Dis	98| 98	100|100	98| 98	100|100	0| 0	100|100	97| 97	0| 0
3	500	5	I	100|100	100|100	100|100	100|100	26| 19	100| 99	96| 95	26| 20
3	500	5	Dis	100|100	100|100	100|100	100|100	26| 19	100| 99	97| 96	26| 20
3	1000	5	I	100|100	100|100	100|100	100|100	25| 18	100|100	99| 98	26| 18
3	1000	5	Dis	100|100	100|100	100|100	100|100	25| 18	100|100	99| 98	26| 18
3	500	10	I	100|100	100|100	100|100	100|100	26| 19	100| 99	91| 88	31| 23
3	500	10	Dis	100| 99	100|100	100| 99	100|100	26| 19	100| 99	90| 87	32| 24
3	1000	10	I	100|100	100|100	100|100	100|100	26| 18	100| 99	98| 98	26| 18
3	1000	10	Dis	100|100	100|100	100|100	100|100	26| 18	100|100	99| 98	26| 18
9	500	5	I	99|100	85| 84	97| 97	99|100	79| 83	85| 84	92| 92	79| 82
9	500	5	Dis	99| 99	84| 82	97| 98	98| 99	79| 80	84| 82	93| 94	80| 81
9	1000	5	I	100|100	99| 99	99|100	100|100	79| 75	99| 99	98| 98	79| 76
9	1000	5	Dis	99|100	96| 96	98| 98	99|100	79| 75	95| 96	97| 97	79| 76
9	500	10	I	97| 98	91| 91	98| 99	96| 99	80| 83	90| 89	85| 83	87| 87
9	500	10	Dis	95| 96	92| 91	97| 96	96| 99	80| 82	91| 90	85| 85	87| 87
9	1000	10	I	99|100	95| 95	99| 99	99|100	79| 77	95| 94	94| 93	79| 80
9	1000	10	Dis	99|100	96| 96	99|100	99|100	79| 78	96| 95	95| 94	79| 80

Table S46: Average of adjusted Rand index for all models | only for the first cluster (%) are given for simulations under logistic regression models for $K_{\text{true}} \in \{1, 3, 9\}$, $p \in \{5, 10\}$, $n \in \{500, 1000\}$, and the scenarios "I" and "Dis". Compared methods are nAIC, nBIC, nAICex1, nAICex2, AIC, BIC, AICex1, and AICex2, where "n" stands for the normalized partial log-likelihood.

$K_{\text{true}}\|n\|p\|s$	**nAIC**	**nAICex1**	**nAICex2**	**nBIC**	**AIC**	**AICex1**	**AICex2**	**BIC**
1|500|5|I	0.03	0.03	0.03	0.03	4.66	0.03	0.03	4.38
1|500|5|Dis	0.03	0.03	0.03	0.03	4.66	0.03	0.03	4.38
1|1000|5|I	0.02	0.02	0.02	0.02	3.28	0.02	0.02	3.25
1|1000|5|Dis	0.02	0.02	0.02	0.02	3.28	0.02	0.02	3.25
1|500|10|I	0.07	0.07	0.07	0.07	1549.69	0.07	0.07	1095.89
1|500|10|Dis	0.07	0.07	0.07	0.07	1549.69	0.07	0.07	1095.89
1|1000|10|I	0.04	0.03	0.04	0.03	4.99	0.03	0.04	4.68
1|1000|10|Dis	0.04	0.03	0.04	0.03	4.99	0.03	0.04	4.68
3|500|5|I	0.09	0.09	0.09	0.09	2.48	0.09	0.12	2.35
3|500|5|Dis	0.09	0.09	0.09	0.09	2.51	0.09	0.11	2.43
3|1000|5|I	0.04	0.04	0.04	0.04	1.80	0.04	0.05	1.77
3|1000|5|Dis	0.04	0.04	0.04	0.04	1.83	0.04	0.05	1.81
3|500|10|I	0.22	0.21	0.22	0.21	662.65	0.22	0.32	11.65
3|500|10|Dis	0.22	0.21	0.22	0.21	690.56	0.22	0.34	7.15
3|1000|10|I	0.09	0.09	0.09	0.09	2.73	0.09	0.10	2.52
3|1000|10|Dis	0.09	0.09	0.09	0.09	2.74	0.09	0.10	2.43
9|500|5|I	0.34	0.38	0.35	0.34	1.40	0.40	0.58	1.31
9|500|5|Dis	0.35	0.39	0.37	0.37	1.29	0.43	0.58	1.25
9|1000|5|I	0.14	0.14	0.14	0.14	0.65	0.14	0.16	0.58
9|1000|5|Dis	0.14	0.14	0.14	0.14	0.68	0.15	0.17	0.66
9|500|10|I	1.90	1.52	1.52	1.41	1133.77	1.90	82.06	220.39
9|500|10|Dis	2.62	1.52	1.66	1.42	1403.00	1.82	98.02	916.34
9|1000|10|I	0.40	0.40	0.40	0.40	1.49	0.42	0.59	1.30
9|1000|10|Dis	0.40	0.42	0.42	0.40	1.41	0.44	0.61	1.33

Table S47: Mean squared errors for the linear predictor are given for simulations under logistic regression models for $K_{\text{true}} \in \{1, 3, 9\}$, $p \in \{5, 10\}$, $n \in \{500, 1000\}$, and the scenarios "I" and "Dis". Compared methods are nAIC, nBIC, nAICex1, nAICex2, AIC, BIC, AICex1, and AICex2, where "n" stands for the normalized partial log-likelihood.

$K_{\text{true}}\|n\|p\|s$	**nAIC**	**nAICex1**	**nAICex2**	**nBIC**	**AIC**	**AICex1**	**AICex2**	**BIC**
1|500|5|I	100| 0| 0	100| 0| 0	100| 0| 0	100| 0| 0	0|100| 0	100| 0| 0	92| 8| 0	0|100| 0
1|500|5|MoE	100| 0| 0	100| 0| 0	100| 0| 0	100| 0| 0	0|100| 0	99| 1| 0	71| 29| 0	0|100| 0
1|1000|5|I	100| 0| 0	100| 0| 0	100| 0| 0	100| 0| 0	0|100| 0	99| 1| 0	91| 9| 0	0|100| 0
1|1000|5|MoE	100| 0| 0	100| 0| 0	100| 0| 0	100| 0| 0	0|100| 0	100| 0| 0	80| 20| 0	0|100| 0
1|500|10|I	100| 0| 0	100| 0| 0	100| 0| 0	100| 0| 0	0|100| 0	100| 0| 0	95| 5| 0	0|100| 0
1|500|10|MoE	100| 0| 0	100| 0| 0	100| 0| 0	100| 0| 0	0|100| 0	99| 1| 0	78| 22| 0	0|100| 0
1|1000|10|I	98| 2| 0	100| 0| 0	98| 2| 0	100| 0| 0	0|100| 0	100| 0| 0	97| 3| 0	0|100| 0
1|1000|10|MoE	100| 0| 0	100| 0| 0	100| 0| 0	100| 0| 0	0|100| 0	100| 0| 0	92| 8| 0	0|100| 0
3|500|5|I	99| 1| 0	100| 0| 0	99| 1| 0	100| 0| 0	0|100| 0	97| 3| 0	80| 20| 0	0|100| 0
3|500|5|MoE	13| 15| 72	5| 1| 94	11| 3| 86	2| 0| 98	0|100| 0	7| 1| 92	30| 11| 59	0|100| 0
3|1000|5|I	99| 1| 0	99| 1| 0	99| 1| 0	100| 0| 0	0|100| 0	99| 1| 0	91| 9| 0	0|100| 0
3|1000|5|MoE	15| 23| 62	10| 0| 90	10| 7| 83	5| 2| 93	0|100| 0	21| 1| 78	31| 26| 43	0|100| 0
3|500|10|I	99| 1| 0	100| 0| 0	99| 1| 0	100| 0| 0	0|100| 0	97| 3| 0	64| 36| 0	0|100| 0
3|500|10|MoE	16| 26| 58	1| 1| 98	18| 6| 76	0| 0|100	0|100| 0	1| 1| 98	23| 27| 50	0|100| 0
3|1000|10|I	100| 0| 0	100| 0| 0	100| 0| 0	100| 0| 0	0|100| 0	99| 1| 0	91| 9| 0	0|100| 0
3|1000|10|MoE	14| 36| 50	3| 0| 97	8| 4| 88	2| 0| 98	0|100| 0	5| 0| 95	22| 13| 65	0|100| 0
9|500|5|I	84| 16| 0	67| 3| 30	84| 9| 7	87| 4| 9	0|100| 0	60| 10| 30	32| 63| 5	0|100| 0
9|500|5|MoE	0| 1| 99	0| 0|100	0| 0|100	0| 0|100	0|100| 0	0| 0|100	0| 0|100	0|100| 0
9|1000|5|I	92| 8| 0	94| 1| 5	93| 4| 3	95| 5| 0	0|100| 0	89| 6| 5	70| 28| 2	0|100| 0
9|1000|5|MoE	0| 2| 98	0| 0|100	0| 0|100	0| 0|100	0|100| 0	0| 0|100	0| 0|100	0|100| 0
9|500|10|I	55| 45| 0	69| 15| 16	68| 32| 0	61| 4| 35	0|100| 0	51| 34| 15	3| 97| 0	8| 92| 0
9|500|10|MoE	0| 0|100	0| 0|100	0| 0|100	0| 0|100	0|100| 0	0| 0|100	0| 0|100	0|100| 0
9|1000|10|I	86| 14| 0	86| 1| 13	86| 11| 3	86| 11| 3	0|100| 0	81| 6| 13	40| 57| 3	0|100| 0
9|1000|10|MoE	0| 0|100	0| 0|100	0| 0|100	0| 0|100	0|100| 0	0| 0|100	0| 0|100	0|100| 0

Table S48: Proportions of selected number of clusters (%), correct-|over-|under-estimations, are given for simulations under Cox regression models for $K_{\text{true}} \in \{1, 3, 9\}$, $p \in \{5, 10\}$, $n \in \{500, 1000\}$, and the scenarios "I" and "MoE". Compared methods are nAIC, nBIC, nAICex1, nAICex2, AIC, BIC, AICex1, and AICex2, where "n" stands for the normalized partial log-likelihood.

$K_{\text{true}}\|n\|p\|s$	**nAIC**	**nAICex1**	**nAICex2**	**nBIC**	**AIC**	**AICex1**	**AICex2**	**BIC**
1|500|5|I	100|100	100|100	100|100	100|100	0| 0	100|100	92| 92	0| 0
1|500|5|MoE	100|100	100|100	100|100	100|100	0| 0	99| 99	71| 71	0| 0
1|1000|5|I	100|100	100|100	100|100	100|100	0| 0	99| 99	91| 91	0| 0
1|1000|5|MoE	100|100	100|100	100|100	100|100	0| 0	100|100	80| 80	0| 0
1|500|10|I	100|100	100|100	100|100	100|100	0| 0	100|100	95| 95	0| 0
1|500|10|MoE	100|100	100|100	100|100	100|100	0| 0	99| 99	78| 78	0| 0
1|1000|10|I	98| 98	100|100	98| 98	100|100	0| 0	100|100	97| 97	0| 0
1|1000|10|MoE	100|100	100|100	100|100	100|100	0| 0	100|100	92| 92	0| 0
3|500|5|I	100|100	100|100	100|100	100|100	26| 19	100| 99	96| 95	26| 20
3|500|5|MoE	2| 3	2| 1	2| 2	1| 1	2| 6	2| 1	4| 4	3| 6
3|1000|5|I	100|100	100|100	100|100	100|100	25| 18	100|100	99| 98	26| 18
3|1000|5|MoE	2| 3	1| 1	2| 2	1| 1	2| 5	2| 2	3| 4	2| 5
3|500|10|I	100|100	100|100	100|100	100|100	26| 19	100| 99	91| 88	31| 23
3|500|10|MoE	2| 3	1| 1	2| 2	0| 0	2| 6	1| 1	2| 3	3| 6
3|1000|10|I	100|100	100|100	100|100	100|100	26| 18	100| 99	98| 98	26| 18
3|1000|10|MoE	2| 3	1| 0	1| 1	0| 0	2| 5	1| 1	2| 2	2| 5
9|500|5|I	99|100	85| 84	97| 97	99|100	79| 83	85| 84	92| 92	79| 82
9|500|5|MoE	0| 0	0| 0	0| 0	0| 0	0| 7	0| 0	0| 0	0| 7
9|1000|5|I	100|100	99| 99	99|100	100|100	79| 75	99| 99	98| 98	79| 76
9|1000|5|MoE	0| 0	0| 0	0| 0	0| 0	0| 5	0| 0	0| 0	0| 5
9|500|10|I	97| 98	91| 91	98| 99	96| 99	80| 83	90| 89	85| 83	87| 87
9|500|10|MoE	0| 0	0| 0	0| 0	0| 0	0| 7	0| 0	0| 0	0| 7
9|1000|10|I	99|100	95| 95	99| 99	99|100	79| 77	95| 94	94| 93	79| 80
9|1000|10|MoE	0| 0	0| 0	0| 0	0| 0	0| 5	0| 0	0| 0	0| 5

Table S49: Average of adjusted Rand index for all models | only for the first cluster (%) are given for simulations under logistic regression models for $K_{\text{true}} \in \{1, 3, 9\}$, $p \in \{5, 10\}$, $n \in \{500, 1000\}$, and the scenarios "I" and "MoE". Compared methods are nAIC, nBIC, nAICex1, nAICex2, AIC, BIC, AICex1, and AICex2, where "n" stands for the normalized partial log-likelihood.

K_{true}	n	p	s	**nAIC**	**nAICex1**	**nAICex2**	**nBIC**	**AIC**	**AICex1**	**AICex2**	**BIC**
1	500	5	I	0.03	0.03	0.03	0.03	4.66	0.03	0.03	4.38
1	500	5	MoE	1.39	1.39	1.39	1.39	11996.62	1.39	2.46	11136.39
1	1000	5	I	0.02	0.02	0.02	0.02	3.28	0.02	0.02	3.25
1	1000	5	MoE	1.05	1.05	1.05	1.05	3183.84	1.05	1.25	2409.07
1	500	10	I	0.07	0.07	0.07	0.07	1549.69	0.07	0.07	1095.89
1	500	10	MoE	3.94	3.94	3.94	3.94	27858.36	4.26	6.90	11533.87
1	1000	10	I	0.04	0.03	0.04	0.03	4.99	0.03	0.04	4.68
1	1000	10	MoE	1.70	1.70	1.70	1.70	15395.22	1.70	2.04	10192.31
3	500	5	I	0.09	0.09	0.09	0.09	2.48	0.09	0.12	2.35
3	500	5	MoE	31.54	31.51	31.51	31.55	41.59	31.51	32.37	39.32
3	1000	5	I	0.04	0.04	0.04	0.04	1.80	0.04	0.05	1.77
3	1000	5	MoE	31.72	31.78	31.78	31.56	37.14	32.05	32.95	37.11
3	500	10	I	0.22	0.21	0.22	0.21	662.65	0.22	0.32	11.65
3	500	10	MoE	64.96	66.14	64.96	66.38	4515.83	66.02	66.00	1534.61
3	1000	10	I	0.09	0.09	0.09	0.09	2.73	0.09	0.10	2.52
3	1000	10	MoE	65.18	66.13	66.02	66.27	69.59	66.47	66.84	69.15
9	500	5	I	0.34	0.38	0.35	0.34	1.40	0.40	0.58	1.31
9	500	5	MoE	4.00	4.04	4.04	4.04	3.63	4.04	4.04	3.63
9	1000	5	I	0.14	0.14	0.14	0.14	0.65	0.14	0.16	0.58
9	1000	5	MoE	4.06	4.07	4.07	4.07	3.86	4.07	4.07	3.86
9	500	10	I	1.90	1.52	1.52	1.41	1133.77	1.90	82.06	220.39
9	500	10	MoE	8.11	8.16	8.16	8.16	6.71	8.16	8.16	6.70
9	1000	10	I	0.40	0.40	0.40	0.40	1.49	0.42	0.59	1.30
9	1000	10	MoE	8.15	8.20	8.20	8.20	7.44	8.20	8.20	7.43

Table S50: Mean squared errors for the linear predictor are given for simulations under logistic regression models for $K_{\text{true}} \in \{1, 3, 9\}$, $p \in \{5, 10\}$, $n \in \{500, 1000\}$, and the scenarios "I" and "MoE". Compared methods are nAIC, nBIC, nAICex1, nAICex2, AIC, BIC, AICex1, and AICex2, where "n" stands for the normalized partial log-likelihood.

Supplementary Figures

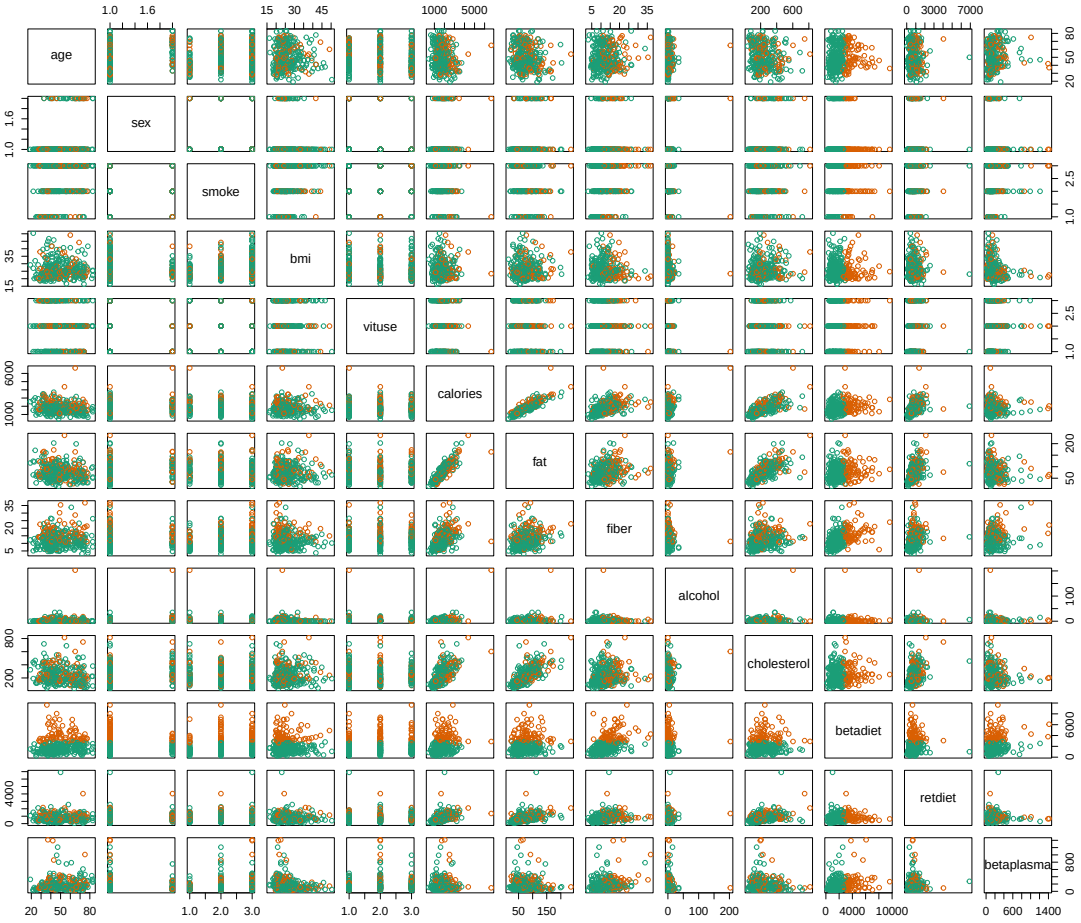

Figure S2: Scatterplot of plasma data. Optimal clusters are colored differently.

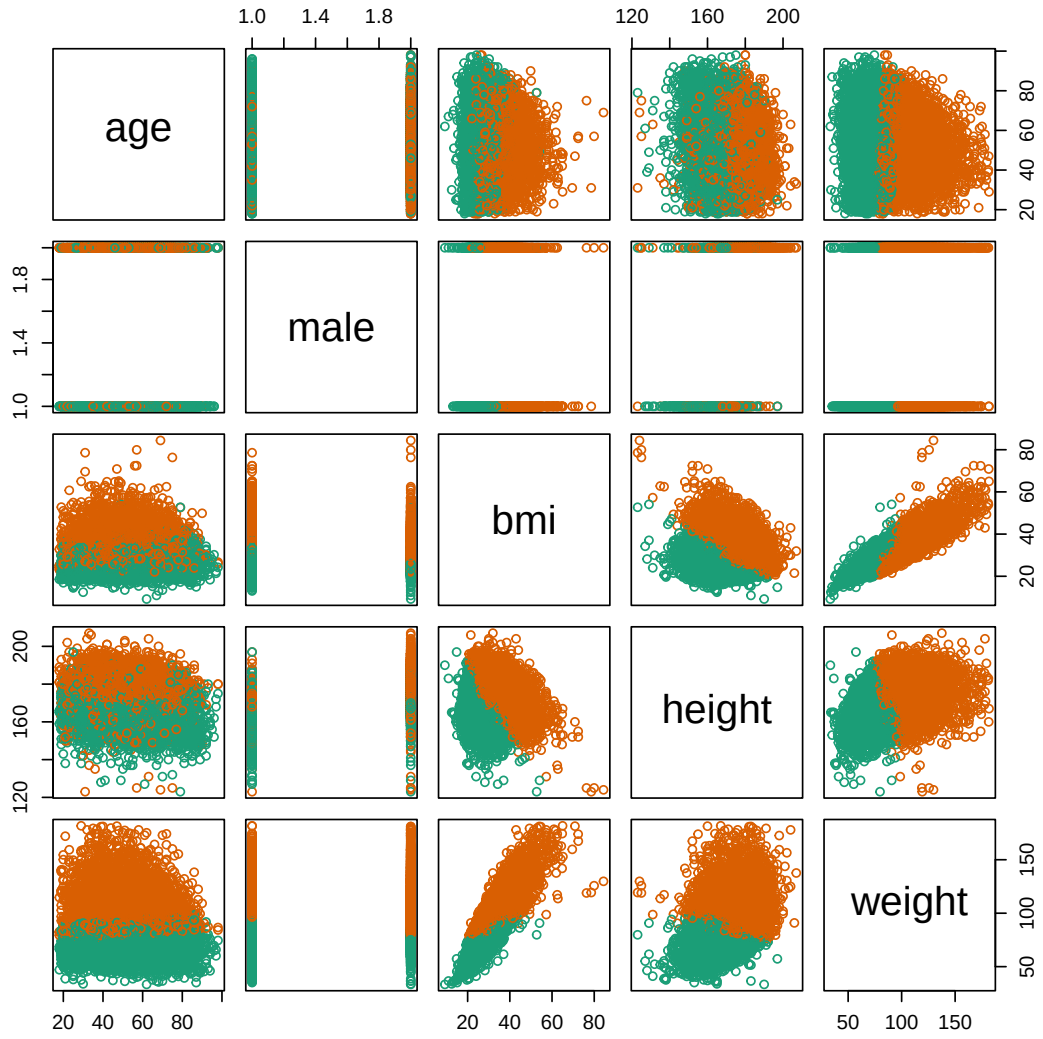

Figure S3: Scatterplot of NAFLD1 data. Optimal clusters are colored differently.

References

[1] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

[2] Jianqing Fan and Runze Li. Variable selection for Cox's proportional hazards model and frailty model. *The Annals of Statistics*, 30(1), 2002.

[3] Brandon M. Greenwell. *pdp: Partial Dependence Plots*, 2024. R package version 0.8.2.

[4] Zhexue Huang. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery*, 2(3):283–304, 1998.

[5] Seongil Jo, Taeryon Choi, and Beomjo Park. *bsamGP: Bayesian Spectral Analysis Models using Gaussian Process Priors*, 2025. R package version 1.2.7.

[6] Runze Li, Jian-Jian Ren, Guangren Yang, and Ye Yu. Asymptotic behavior of Cox's partial likelihood and its application to variable selection. *Statistica Sinica*, 28(4):2713–2731, 2017.

[7] David W Nierenberg, Therese A Stukel, John A Baron, Bradley J Dain, E Robert Greenberg, and Skin Cancer Prevention Study Group. Determinants of plasma levels of beta-carotene and retinol. *American Journal of Epidemiology*, 130(3):511–521, 1989.

[8] Jun Shao. An asymptotic theory for linear model selection. *Statistica sinica*, 7(2):221–242, 1997.

[9] Hon Yiu So, Jinhui Ma, Lauren E. Griffith, and Narayanaswamy Balakrishnan. Application of machine learning methods in the imputation of heterogeneous co-missing data. *Japanese Journal of Statistics and Data Science*, 8(1):691–720, 2025.

[10] Yuhong Yang. Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation. *Biometrika*, 92(4):937–950, 2005.