Statistica Sinica Preprint No: SS-2025-0466	
Title	Out-of-cluster Prediction for Model Selection in Regression with Unsupervised Clustering
Manuscript ID	SS-2025-0466
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202025.0466
Complete List of Authors	Masao Ueki
Corresponding Authors	Masao Ueki
E-mails	uekimrsd@nifty.com
Notice: Accepted author version.	

Statistica Sinica

OUT-OF-CLUSTER PREDICTION FOR MODEL SELECTION IN REGRESSION WITH UNSUPERVISED CLUSTERING

Masao Ueki

Nagasaki University / RIKEN Center for Advanced Intelligence Project

Abstract: In regression with unsupervised clustering, the explanatory variables are first clustered, and separate regression models are then built for each cluster. The resulting models are often evaluated using in-cluster prediction criteria, such as the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). This paper explores the usefulness of out-of-cluster prediction for evaluating regression models, particularly in selecting the number of clusters. In particular, we develop a model exclusion procedure that makes use of the reduced accuracy of out-of-cluster prediction compared with in-cluster prediction, under the assumption that regression models differ between clusters, to exclude redundant models before applying model selection. The model exclusion procedure is considered within standard regression frameworks, including generalized linear and Cox regression models. For Cox regression models, we propose a normalized partial log-likelihood to avoid divergence issues that arise when the standard partial log-likelihood is used for model selection. We show that selecting the number of clusters using AIC, combined with the proposed model exclusion procedure, achieves model selection consistency. We confirm the improved performance of

the proposed exclusion procedure through extensive simulation studies involving Gaussian linear, logistic, and Cox regression models combined with K-means clustering.

Key words and phrases: AIC, out-of-cluster prediction, normalized partial log-likelihood, regression with unsupervised clustering, model selection.

1. Introduction

In this paper, we study model selection for regression models combined with unsupervised clustering, where explanatory variables are first clustered without reference to the response variable, and regression models are then built separately for each cluster. Regression with unsupervised clustering has recently received increasing attention in several fields, including medicine, biomedicine, and other disciplines, reflecting the recognition of underlying heterogeneity and demonstrating that cluster-specific modeling can improve both interpretation and prediction (Choi et al., 2023; Teng et al., 2024). In many clustering methods, the number of clusters must be specified in advance. There has been considerable debate on how to determine the appropriate number of clusters in unsupervised learning. Existing approaches typically evaluate clustering quality using specific clustering metrics. For K-means clustering, numerous methods and metrics have been

proposed (Tibshirani et al., 2001; Batool and Hennig, 2021; Ueki, 2025).

When a response variable is available, a different approach can be taken. Recently, Katahira (2023) proposed a novel subtyping method that emphasizes prediction rather than clustering per se, and developed the cost of cluster mean-based prediction (CCMP) criterion for selecting the optimal number of clusters. This approach is closely related to standard model selection criteria such as the Akaike information criterion (AIC; Akaike, 1973) and the Bayesian information criterion (BIC; Schwarz, 1978). These criteria are widely used in model selection because their statistical properties are well established (Shao, 1997). See also Yang (2005) for further discussion of the properties of AIC and BIC. Prediction-based criteria such as CCMP are generally formulated using a loss function based on in-cluster predictions. Thus, CCMP is conceptually similar to AIC and BIC in that it relies on prediction-based evaluation.

In clustering settings, prediction can be evaluated not only within clusters (in-cluster prediction) but also across clusters (out-of-cluster prediction), which is rarely considered in standard model selection. Although out-of-cluster prediction has been largely overlooked in regression combined with unsupervised clustering, we investigate the utility of out-of-cluster prediction for evaluating regression models, particularly for selecting the num-

ber of clusters. The joint use of in-cluster and out-of-cluster samples has previously been applied to assess clustering quality (e.g., silhouette score; Rousseeuw, 1987). If regression models differ across clusters, out-of-cluster prediction should perform worse than in-cluster prediction. Motivated by this observation, we propose an exclusion procedure that removes models containing clusters in which out-of-cluster prediction is comparable to or better than in-cluster prediction, since such clusters provide little evidence of heterogeneity. After exclusion, we apply AIC to the remaining models to select the optimal one. We show that combining AIC with the proposed exclusion procedure achieves model selection consistency comparable to BIC. Unlike BIC, however, the exclusion procedure offers an additional advantage: it enables detailed identification of redundant clusters using out-of-cluster prediction and thereby improves model selection. The exclusion procedure can be incorporated into standard regression frameworks, including generalized linear and Cox regression models. In Cox regression models, Li et al. (2017) showed that the standard partial log-likelihood diverges to infinity as sample size increases, unlike the log-likelihood in generalized linear models. This divergence leads to overfitting when selecting models in Cox regression combined with unsupervised clustering. To address this issue, we propose a normalized partial log-likelihood that avoids divergence.

Through extensive simulations involving Gaussian linear, logistic, and Cox regression models combined with K-means clustering on explanatory variables, and under the assumption that the latent subgroup structure is primarily reflected in the explanatory variables and can be reasonably recovered via unsupervised clustering, we find that the proposed exclusion procedure, when used in conjunction with AIC, improves the identification of the correct number of clusters compared with AIC alone, while better preserving the accuracy of regression function estimates than BIC. Applications to real datasets further illustrate the practical value of the proposed procedure.

2. Methods

2.1 Regression with unsupervised clustering

We consider a regression problem for a dataset of size n, $(X_1, y_1), \ldots, (X_n, y_n)$, where $X_i = (X_{i,1}, \ldots, X_{i,p})$ is a p-dimensional vector of explanatory variables for subject i, and y_i is the associated response variable, for $i = 1, \ldots, n$.

Unsupervised clustering: For a given number of clusters $K > 1$, we apply an unsupervised clustering method, such as K-means, to the ob-

servations $X_1, \ldots, X_n$ in order to partition the n samples into K disjoint clusters, that is, $\widehat{\mathcal{C}}_1 \cup \cdots \cup \widehat{\mathcal{C}}_K = \{1, \ldots, n\}$, with $\widehat{\mathcal{C}}_k \cap \widehat{\mathcal{C}}_{k'} = \emptyset$ for any $k \neq k' \in \{1, \ldots, K\}$.

Regression model: Given a partition of the samples into K clusters, $\widehat{\mathcal{C}}_1 \cup \cdots \cup \widehat{\mathcal{C}}_K$, we construct K regression models, $\widehat{\mathcal{M}}(K) = \{\widehat{M}_1, \ldots, \widehat{M}_K\}$, independently within each cluster. That is, the kth regression model $\widehat{M}_k$ is obtained by regressing $y_{\widehat{\mathcal{C}}_k}$ onto $X_{\widehat{\mathcal{C}}_k}$, where $X_{\widehat{\mathcal{C}}_k} = (X_i : i \in \widehat{\mathcal{C}}_k)$ and $y_{\widehat{\mathcal{C}}_k} = (y_i : i \in \widehat{\mathcal{C}}_k)$. Assume that the regression model admits a log-likelihood function $\ell(\theta; X, y) = \sum_{i=1}^{n} \ell(y_i \mid X_i; \theta)$, where $\ell(y_i \mid X_i; \theta)$ denotes the log of the conditional density (for continuous y_i) or probability mass function (for discrete y_i), and θ is an m-dimensional parameter vector including p regression coefficients (with $m \geq p$). Then, the cluster-specific maximum likelihood estimates are $\widehat{\theta}_k = \mathrm{argmax}_\theta \, \ell(\theta; X_{\widehat{\mathcal{C}}_k}, y_{\widehat{\mathcal{C}}_k})$, $k = 1, \ldots, K$.

2.2 Selection of the number of clusters

To perform regression with unsupervised clustering, it is necessary to determine the number of clusters K in advance. Several approaches are available for the selection. Among them, AIC and BIC are standard model selection criteria. In addition, a recently developed criterion has been proposed by Katahira (2023).

AIC: The AIC for the regression model with unsupervised clustering, $\widehat{\mathcal{M}}(K)$, is defined as $AIC(K) = \sum_{k=1}^{K}\{-2\ell(\widehat{\theta}_k; X_{\widehat{\mathcal{C}}_k}, y_{\widehat{\mathcal{C}}_k}) + 2m\}$. For the set of candidate models $\widehat{\mathcal{M}}(K)$ with $K = 1, 2, 3, \ldots, K_{\max}$, where $K_{\max}$ is the prespecified maximum number of clusters to search, the selected K is the value that minimizes $AIC(K)$.

BIC: The BIC for the regression model with unsupervised clustering, $\widehat{\mathcal{M}}(K)$, is defined as $BIC(K) = \sum_{k=1}^{K}\{-2\ell(\widehat{\theta}_k; X_{\widehat{\mathcal{C}}_k}, y_{\widehat{\mathcal{C}}_k}) + \log(n)m\}$. Similarly, for the set of candidate models $\widehat{\mathcal{M}}(K)$ with $K = 1, 2, 3, \ldots, K_{\max}$, the selected K is the value that minimizes $BIC(K)$.

CCMP: The CCMP is a recently developed criterion proposed by Katahira (2023). For a given clustering $\widehat{\mathcal{C}}_1 \cup \cdots \cup \widehat{\mathcal{C}}_K = \{1, \ldots, n\}$, the CCMP is defined as $CCMP(K) = -\sum_{k=1}^{K}\{n_k(\bar{y}_k - \bar{y})^2 - 2\widehat{\sigma}_k^2\}$, where $n_k = |\widehat{\mathcal{C}}_k|$, $\bar{y}_k = n_k^{-1}\sum_{i\in\widehat{\mathcal{C}}_k} y_i$, $\widehat{\sigma}_k^2 = (n_k - 1)^{-1}\sum_{i\in\widehat{\mathcal{C}}_k}(y_i - \bar{y}_k)^2$, and $\bar{y} = n^{-1}\sum_{i=1}^{n} y_i$. The optimal number of clusters K is the minimizer of $CCMP(K)$.

This criterion is derived as an unbiased estimator of the mean squared error between the true mean function and the fitted intercept-only model (or constant model) within each cluster. Because the AIC is also an unbiased criterion for the Kullback–Leibler loss, which coincides with the mean squared error in Gaussian regression, the CCMP is expected to behave sim-

ilarly to the AIC. In Supplementary Appendix, we show that the CCMP is essentially equivalent to Mallows' C_p criterion (Mallows, 1973), which is known to approximate the AIC (Shao, 1997). Thus, the CCMP can be regarded as an analogue of the AIC for intercept-only models within clusters. Throughout this paper, we evaluate the performance of clustering procedures in comparison with such intercept-only models.

2.3 Utility of out-of-cluster prediction

It can be seen that AIC and BIC involve the term $\sum_{k=1}^{K} \ell(\widehat{\theta}_k; X_{\widehat{\mathcal{C}}_k}, y_{\widehat{\mathcal{C}}_k})$, which is the sum of K log-likelihood functions evaluated at the parameter $\widehat{\theta}_k$ estimated from samples within its own cluster $\widehat{\mathcal{C}}_k$, for $k = 1, \ldots, K$. In other words, this represents the sum of in-cluster predictions across the K clusters.

In addition to these in-cluster predictions, one can also consider out-of-cluster samples. Specifically, the loss function in the kth cluster can be evaluated using parameters estimated from a different cluster, $l \neq k$, $\ell(\widehat{\theta}_l; X_{\widehat{\mathcal{C}}_k}, y_{\widehat{\mathcal{C}}_k})$. The availability of such out-of-cluster evaluations is unique to clustering problems and is generally absent in other model selection settings. This perspective provides additional information for assessing model adequacy. Our proposal is to incorporate out-of-cluster prediction as a

complementary criterion for model evaluation.

If the regression models differ substantially across clusters, out-of-cluster predictions are expected to perform poorly compared to in-cluster predictions, assuming correct model specification. Conversely, if the regression models yield similar accuracy even for out-of-cluster samples, this suggests that the model may be essentially homogeneous and contain redundant clusters. By quantifying out-of-cluster prediction, we can detect and exclude such non-heterogeneous regression models. In practice, this involves inspecting the $K(K-1)$ out-of-cluster prediction values under K clusters.

Although it is informative to examine these values directly, we further propose using them systematically to eliminate redundant regression models, thereby facilitating the automatic selection of the optimal number of clusters. To this end, we present two procedures, termed exclusion procedures 1 and 2. The first procedure tends to remove redundant models aggressively when at least one such redundant cluster is present, whereas the second moderates this by requiring the presence of at least one redundant cluster pair. Subsequently, AIC is applied to the reduced set of candidate models.

Exclusion procedure 1: The first procedure excludes a regression model if there exists a pair (k, l) with $k \neq l$ such that the out-of-cluster prediction is better than the in-cluster prediction. This condition is expressed as

$$-2\ell(\widehat{\theta}_k; X_{\widehat{\mathcal{C}}_k}, y_{\widehat{\mathcal{C}}_k}) + m \log |\widehat{\mathcal{C}}_k| > -2\ell(\widehat{\theta}_l; X_{\widehat{\mathcal{C}}_k}, y_{\widehat{\mathcal{C}}_k}), \tag{2.1}$$

where $|\widehat{\mathcal{C}}_k|$ denotes the cardinality of $\widehat{\mathcal{C}}_k$. The term $m \log |\widehat{\mathcal{C}}_k|$ is the BIC penalty, which is introduced to make the left-hand side comparable to the right-hand side. On the right-hand side, $\ell(\widehat{\theta}_l; X_{\widehat{\mathcal{C}}_k}, y_{\widehat{\mathcal{C}}_k})$, the dataset $\widehat{\mathcal{C}}_k$ used for evaluation is independent of the dataset used to obtain the model or parameter estimate $\widehat{\theta}_l$. In contrast, on the left-hand side, the same dataset $\widehat{\mathcal{C}}_k$ is used both to evaluate the model and to estimate the parameter. Ideally, the left-hand side serves as an alternative to $\ell(\widehat{\theta}_k; X'_{\widehat{\mathcal{C}}_k}, y'_{\widehat{\mathcal{C}}_k})$, where $(X'_{\widehat{\mathcal{C}}_k}, y'_{\widehat{\mathcal{C}}_k})$ is a dataset independent of $(X_{\widehat{\mathcal{C}}_k}, y_{\widehat{\mathcal{C}}_k})$ but is unavailable. The BIC penalty plays the role of correcting the optimism in $\ell(\widehat{\theta}_k; X_{\widehat{\mathcal{C}}_k}, y_{\widehat{\mathcal{C}}_k})$ and also ensuring model selection consistency.

We initialize the exclusion set as $\mathcal{E}_1 = \emptyset$. For a given set of candidate models $\widehat{\mathcal{M}}(K)$ with $K = 2, 3, \ldots, K_{\max}$, where $K_{\max}$ is the prespecified maximum number of clusters, if at least one pair satisfies (2.1), then K is added to the exclusion set $\mathcal{E}_1$. The final number of clusters is chosen as the value of K that minimizes the AIC over $K \in \{1, 2, \ldots, K_{\max}\} \setminus \mathcal{E}_1$. We refer

to this criterion as AICex1.

Exclusion procedure 2: There is concern that AICex1 may be too aggressive in excluding candidate models, because it relies on the presence of any cluster that produces indistinguishable predictions from out-of-cluster samples. This increases the chance of false exclusions when many comparisons are made. To address this issue, we present an alternative approach.

Instead of (2.1), the second procedure excludes regression models if there exists at least one pair (k, l) with $k \neq l$ such that

$$-2\ell(\widehat{\theta}_k; X_{\widehat{\mathcal{C}}_k}, y_{\widehat{\mathcal{C}}_k}) + m \log |\widehat{\mathcal{C}}_k| > -2\ell(\widehat{\theta}_l; X_{\widehat{\mathcal{C}}_k}, y_{\widehat{\mathcal{C}}_k}) \quad \text{and} \quad -2\ell(\widehat{\theta}_l; X_{\widehat{\mathcal{C}}_l}, y_{\widehat{\mathcal{C}}_l}) + m \log |\widehat{\mathcal{C}}_l| > -2\ell(\widehat{\theta}_k; X_{\widehat{\mathcal{C}}_l}, y_{\widehat{\mathcal{C}}_l}). \tag{2.2}$$

Unlike (2.1), this criterion requires agreement from the perspectives of both clusters l and k before concluding that the model is non-heterogeneous. This symmetry ensures that if model l produces predictions similar to those of k, then the converse also holds.

The exclusion set is initialized as $\mathcal{E}_2 = \emptyset$. For a given set $\widehat{\mathcal{M}}(K)$ with $K = 2, 3, \ldots, K_{\max}$, where $K_{\max}$ is the prespecified maximum number of clusters, add K to the exclusion set $\mathcal{E}_2$ if at least one pair satisfies (2.2). The selected K is then defined as the one that minimizes AIC over $K \in \{1, 2, \ldots, K_{\max}\} \setminus \mathcal{E}_2$, which we term AICex2.

2.4 Theoretical study

We investigate theoretical properties concerning the selection of the number of clusters in a large-sample setting. Suppose we have a sequence of partitions of n samples into $K = 1, 2, \ldots, K_{\max}$ clusters: $\mathfrak{C}^{(1)}, \ldots, \mathfrak{C}^{(K_{\max})}$, where $\mathfrak{C}^{(K)} = \mathcal{C}_1^{(K)} \cup \cdots \cup \mathcal{C}_K^{(K)} = \{1, \ldots, n\}$ is the Kth partition, with $\mathcal{C}_k^{(K)} \cap \mathcal{C}_l^{(K)} = \emptyset$ for $k \neq l$ and $K = 1, \ldots, K_{\max}$. We assume that $K_{\max}$ is fixed.

Throughout this subsection, we assume that the sequence of partitions contains the unique true partition $\mathfrak{C}^{(K_{\text{true}})} = \mathcal{C}_1^{(K_{\text{true}})} \cup \cdots \cup \mathcal{C}_{K_{\text{true}}}^{(K_{\text{true}})} = \{1, \ldots, n\}$, with the true number of clusters K_{true} ($1 \leq K_{\text{true}} \leq K_{\max}$). For each $k \in \{1, \ldots, K_{\text{true}}\}$, let $V_{\mathcal{C}_k^{(K_{\text{true}})}} = (V_i : i \in \mathcal{C}_k^{(K_{\text{true}})})$ denote independently and identically distributed random vectors, where $V_i = (X_i, y_i)$, X_i is a p-dimensional explanatory vector, and y_i is a response variable. Here, p is assumed to be fixed.

For $k = 1, \ldots, K_{\text{true}}$, we assume that the true model for samples in the kth cluster, $V_{\mathcal{C}_k^{(K_{\text{true}})}}$, has density $f(y|x; \theta_k^{(K_{\text{true}})})\, c_k^{(K_{\text{true}})}(x)$, where $\theta_k^{(K_{\text{true}})}$ is an m-dimensional vector of unknown parameters (including p regression coefficients and possibly $m - p$ additional parameters), $c_k^{(K_{\text{true}})}(x)$ is the marginal density of X, and $f(y|x; \theta_k^{(K_{\text{true}})})$ is the conditional density of y given $X = x$. The functional form of $f(y|x; \theta)$ is assumed to be common

to all K_{true} clusters, and the differences among clusters arise only from the parameter values. That is, for any $k \neq l \in \{1, \ldots, K_{\text{true}}\}$, it holds that $\theta_k^{(K_{\text{true}})} \neq \theta_l^{(K_{\text{true}})}$. The regularity conditions imposed in the theoretical analyses are given in Supplementary Appendix and a brief summary is provided in Table 1.

Table 1: Brief summary of Conditions (C1)–(C3), (C4a), (C4b), (C5) and (C5') given in Supplementary Appendix.

	Brief description
(C1)	Within each cluster, the observations are i.i.d. from an identifiable parametric model with common support. The usual score identity and information equality hold.
(C2)	The Fisher information matrix is finite and positive definite at the true parameter for each cluster.
(C3)	The log-density is locally three-times differentiable around the true parameter, and its third derivatives are dominated by integrable functions.
(C4a)	For each candidate K, the average log-likelihood converges: $\frac{1}{n}\sum_{k=1}^{K} \ell(\widehat{\theta}_k; y_{\mathcal{C}_k}, X_{\mathcal{C}_k}) \to \ell_\star(K)$. Moreover, underfitted models ($K < K_{\text{true}}$) are separated from the true model: $n\{\ell_\star(K_{\text{true}}) - \ell_\star(K)\} \to \infty$.
(C4b)	For each candidate K, the average log-likelihood converges: $\frac{1}{n}\sum_{k=1}^{K} \ell(\widehat{\theta}_k; y_{\mathcal{C}_k}, X_{\mathcal{C}_k}) \to \ell_\star(K)$. Moreover, underfitted models ($K < K_{\text{true}}$) are separated from the true model: $\frac{n}{\log n}\{\ell_\star(K_{\text{true}}) - \ell_\star(K)\} \to \infty$.
(C5)	For each candidate K and each cluster k, the log-likelihood converges: $\frac{1}{n_k}\ell(\theta; y_{\mathcal{C}_k}, X_{\mathcal{C}_k}) \to \bar{\ell}_k(\theta)$. Moreover, for any $K \geq K_{\text{true}}$, at least one pair (k, l) is separated: $\frac{n_k}{\log n_k}\{\bar{\ell}_k(\theta_k^{(K)}) - \bar{\ell}_k(\theta_l^{(K)})\} \to \infty$.
(C5')	For each candidate K and each cluster k, the log-likelihood converges: $\frac{1}{n_k}\ell(\theta; y_{\mathcal{C}_k}, X_{\mathcal{C}_k}) \to \bar{\ell}_k(\theta)$. Moreover, for any $K \geq K_{\text{true}}$, at least one pair (k, l) is mutually separated: $\min[\frac{n_k}{\log n_k}\{\bar{\ell}_k(\theta_k^{(K)}) - \bar{\ell}_k(\theta_l^{(K)})\}, \frac{n_l}{\log n_l}\{\bar{\ell}_l(\theta_l^{(K)}) - \bar{\ell}_l(\theta_k^{(K)})\}] \to \infty$.

The log-likelihood function based on the conditional density in the kth cluster at parameter θ is $\ell_k^{(K_{\text{true}})}(V_{\mathcal{C}_k^{(K_{\text{true}})}}; \theta) = \sum_{i \in \mathcal{C}_k^{(K_{\text{true}})}} \log f(y_i | x_i; \theta)$.

AIC: First, we present a theoretical result on AIC, which is consistent with findings from previous studies (Yang, 2005; Shao, 1997).

Theorem 1. *Suppose that the likelihood function satisfies conditions (C1)–(C3) and (C4a) in Supplementary Appendix. Then we have* $P\left\{\min_{K\in\Omega_-} AIC(K) > AIC(K_{\text{true}})\right\} \to 1$ *and* $P\left\{\min_{K\in\Omega_+} AIC(K) > AIC(K_{\text{true}})\right\} < 1$ *asymptotically.*

The proof is given in Supplementary Appendix. The first claim 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 $n \to \infty$.

Exclusion procedures:

Here, we show that the proposed exclusion procedures, AICex1 and AICex2, address the aforementioned limitation of AIC. Lemmas 1–3 in the Supplementary Appendix show 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.

Theorem 2. *Suppose that conditions (C1)–(C3) and (C5) in Supplementary Appendix hold. Then, AICex1 selects the true model with probability tending to 1. Additionally, if condition (C5') in Supplementary Appendix holds, AICex2 also selects the true model with probability tending to 1.*

Remark 1. Throughout this subsection, we assume that the partitions $\mathfrak{C}^{(1)}, \ldots, \mathfrak{C}^{(K_{\max})}$ are given, and that one of them coincides with the true partition $\mathfrak{C}^{(K_{\text{true}})}$. The preceding results also hold when the partition is estimated as $\widehat{\mathcal{C}}_1 \cup \cdots \cup \widehat{\mathcal{C}}_{K_{\text{true}}} = \{1, \ldots, n\}$ by unsupervised clustering based solely on explanatory variables (without using the response), provided that $P(\widehat{\mathfrak{C}}^{(K_{\text{true}})} = \mathfrak{C}^{(K_{\text{true}})}) \to 1$ as $n \to \infty$. This situation may occur when the clusters are well separated and can be reliably estimated from explanatory variables alone.

It is well known that BIC possesses model selection consistency. A comparison with BIC, provided in the Supplementary Appendix, suggests that the exclusion procedures are expected to have greater power than BIC.

2.5 Cox regression model

The AIC and BIC for the regression model with unsupervised clustering, $\widehat{\mathcal{M}}(K)$, can be extended to Cox regression by replacing the likelihood with the partial likelihood. However, a technical challenge arises. A recent study by Li et al. (2017) shows that the partial log-likelihood diverges to ∞ at the rate of $n \log n$ as $n \to \infty$. This dependence on n does not pose problems when comparing nested models based on the same sample. In contrast, it may affect comparisons between combined models estimated from different samples.

Let T denote the survival time and $X = (X_1, \ldots, X_p)^T$ the associated p-dimensional vector of explanatory variables. Consider the Cox proportional hazards regression model, $h(t|x) = h_0(t)e^{x^T\beta}$, where β is the regression coefficient vector, $h(t|x)$ is the conditional hazard function of T given $X = x$, and $h_0(t)$ is an unspecified baseline hazard function. Suppose that $(T_1, x_1), \ldots, (T_n, x_n)$ is a random sample of (T, X), and that the observed right-censored survival data are $(Y_1, \delta_1, x_1), \ldots, (Y_n, \delta_n, x_n)$, where $Y_i = \min(T_i, C_i)$, $\delta_i = I_{\{T_i \le C_i\}}$, and C_i is the censoring time, assumed independent of T_i given $X = x_i$. Without loss of generality, assume that there are no ties among the observed continuous variables Y_i. The partial log-likelihood function of the observed data is $\ell_c(\beta; X, Y, \delta) =$

$\sum_{i=1}^n \delta_i x_i^T\beta - \sum_{i=1}^n \delta_i \log\left(\sum_{j=1}^n I_{\{Y_j\geq Y_i\}} e^{x_j^T\beta}\right)$. Unlike the log-likelihood in generalized linear models, Li et al. (2017) (Theorem 1) shows that

$$\ell_c(\beta; X, Y, \delta) = -\rho_u n \log n + n\bar{\ell}(\beta), \tag{2.3}$$

as $n \to \infty$, where $\bar{\ell}(\beta)$ is the quantity of order $O_p(1)$ provided that $E(X)$ and β are of order $O(1)$, and $\rho_u = P(T \leq C)$ is the proportion of uncensored observations. This divergence does not affect the use of partial likelihood in AIC or BIC as long as model comparison is restricted to the common sample $\{1, \ldots, n\}$, as in the variable selection framework considered in Li et al. (2017).

In contrast, we show that the partial likelihood cannot be used to compare Cox regression models when unsupervised clustering is applied. To see this, consider Cox regression with an unsupervised clustering of X. Given a K-partition of the n samples, $\widehat{\mathfrak{C}}^{(K)} = \widehat{\mathcal{C}}_1 \cup \cdots \cup \widehat{\mathcal{C}}_K$, we construct K separate Cox regression models, $\widehat{\mathcal{M}}(K) = \{\widehat{M}_1, \ldots, \widehat{M}_K\}$, each estimated from the samples in one cluster as $\widehat{\beta}_k = \text{argmax}_\beta\, \ell_c(\beta; X_{\widehat{\mathcal{C}}_k}, Y_{\widehat{\mathcal{C}}_k}, \delta_{\widehat{\mathcal{C}}_k})$, for $k = 1, \ldots, K$. For simplicity, let us compare Cox regression models obtained from two different partitions: (i) the trivial partition $\widehat{\mathfrak{C}}^{(1)} = \{1, \ldots, n\}$, and (ii) the two-cluster partition $\widehat{\mathfrak{C}}^{(2)} = \widehat{\mathcal{C}}_1 \cup \widehat{\mathcal{C}}_2$. Applying (2.3), we obtain the partial log-likelihood for the first model as $\ell_c(\widehat{\mathfrak{C}}^{(1)}) = \ell_c(\widehat{\beta}; X, Y, \delta) = -\rho_u n \log n + O_p(n)$, while that for the second model is, under

the assumption that the uncensoring proportions are common across clusters, $\ell_c(\widehat{\mathfrak{C}}^{(2)}) = \ell_c(\widehat{\beta}_1; X_{\widehat{\mathcal{C}}_1}, Y_{\widehat{\mathcal{C}}_1}, \delta_{\widehat{\mathcal{C}}_1}) + \ell_c(\widehat{\beta}_2; X_{\widehat{\mathcal{C}}_2}, Y_{\widehat{\mathcal{C}}_2}, \delta_{\widehat{\mathcal{C}}_2}) = -\rho_u\big(n_1 \log n_1 + n_2 \log n_2\big) + O_p(n)$, where $n_1 = |\widehat{\mathcal{C}}_1|$ and $n_2 = |\widehat{\mathcal{C}}_2|$, with $\min(n_1, n_2) \to \infty$. Since $n_1 \log n_1 + n_2 \log n_2 < n \log n$ with $n = n_1 + n_2$ by superadditivity, we obtain $\ell_c(\widehat{\mathfrak{C}}^{(1)}) < \ell_c(\widehat{\mathfrak{C}}^{(2)})$ asymptotically, provided that the regression coefficients and $E(X)$ remain of order $O(1)$.

This phenomenon also extends to AIC and BIC based on the partial log-likelihood. The AIC for the first model is $-2\ell_c(\widehat{\mathfrak{C}}^{(1)}) + 2p = 2\rho_u n \log n + O_p(n)$, and that for the second model is $-2\ell_c(\widehat{\mathfrak{C}}^{(2)}) + 4p = 2\rho_u(n_1 \log n_1 + n_2 \log n_2) + O_p(n)$, since the penalty term of AIC is of lower order than the leading term (2.3) of the partial log-likelihood. An analogous conclusion holds for BIC.

More generally, for partitions with more than two clusters, $\ell_c(\widehat{\mathfrak{C}}^{(K)}) < \ell_c(\widehat{\mathfrak{C}}^{(K')})$ for any $K < K'$ in probability, and the same conclusion applies to AIC and BIC based on the partial log-likelihood. Thus, the partial log-likelihood always favors finer partitions, irrespective of model fit. We therefore conclude that it is unsuitable for comparing Cox regression models with unsupervised clustering.

To address this issue, we propose the following simple normalization of the partial log-likelihood: $\ell_{cc}(\beta; X, Y, \delta) = \ell_c(\beta; X, Y, \delta) + \widehat{\rho}_u n \log n$, where

$\widehat{\rho}_u = \frac{1}{n}\sum_{i=1}^{n}\delta_i$, i.e., the observed uncensoring proportion in the n samples, and $\ell_{cc}(\beta; X, Y, \delta)$ is asymptotically equivalent to $n\bar{\ell}(\beta)$ in (2.3). Similarly, for the K-partition, $\widehat{\mathfrak{C}}^{(K)}$, we define: $\ell_{cc}(\widehat{\mathfrak{C}}^{(K)}) = \sum_{k=1}^{K}\left\{\ell_c(\widehat{\beta}_k; X_{\widehat{\mathcal{C}}_k}, Y_{\widehat{\mathcal{C}}_k}, \delta_{\widehat{\mathcal{C}}_k}) + \widehat{\rho}_u^{(k)} n_k \log n_k\right\}$, where $\widehat{\rho}_u^{(k)} = \frac{1}{n_k}\sum_{i\in\widehat{\mathcal{C}}_k}\delta_i$. We evaluate AIC and BIC by substituting the above normalized partial log-likelihood in place of the original one. Details and further interpretation are provided in Supplementary Appendix.

Finally, we remark that the statements of Lemmas 1–3 and Theorem 2 remain valid if the exclusion procedures are applied to the partial log-likelihood under the regularity conditions of Li et al. (2017). These results also hold for the normalized version, since the out-of-cluster prediction is evaluated on the same samples. Additional details are given in Supplementary Appendix.

3. Simulation studies

In this section, we present simulation studies to evaluate the performance of the proposed approaches (AICex1 and AICex2) in comparison with AIC, BIC, AICi, and BICi under Gaussian, logistic, and Cox regression models. Here, AICi and BICi refer to the AIC and BIC of intercept-only (constant) models fitted separately within each cluster. Cluster membership is estimated by K-means based on the explanatory variables for a specified

number of clusters. We also compare the mixture-of-experts (MoE) model as a competitor in the Gaussian and logistic regression simulations. The MoE is a finite mixture of regression models with a specified model for cluster assignment. We employ a multinomial logistic regression model for cluster membership conditional on the explanatory variables. The optimal number of clusters is selected using BIC. We fit the MoE using the implementation in the `flexmix` package in R (Leisch, 2004; Grün and Leisch, 2007). We consider two MoE approaches: one using random initial clusters (MoE) and the other using K-means clustering of the explanatory variables to define the initial clusters (MoEk). In addition, the timing comparisons with MoE in Supplementary Table S1 show that the proposed approaches have substantially lower computational cost than MoE.

Generation of cluster membership and explanatory variables: We first generate p-dimensional explanatory variables for K_{true} clusters from a multivariate normal distribution. The mean vector of the kth cluster is $\mu_{x,k} = \frac{k}{K_{\text{true}}} 1_p$, where 1_p denotes the p-dimensional vector of ones. The covariance matrix of the kth cluster is denoted by $\Sigma_{x,k}$. Let $\Sigma_x = \left(\frac{1}{4K_{\text{true}}}\right)^2 I_p$, where I_p is the $p \times p$ identity matrix. The scenario termed "I" assumes that $\Sigma_{x,k} = \Sigma_x$ for each k. We additionally consider the scenario termed "W",

in which the covariance matrices vary randomly according to $\Sigma_{x,k} = \frac{1}{p}W_k$ $(k = 1, \ldots, K_{\text{true}})$, where W_ks are realizations from the Wishart distribution $W_p(\Sigma_x, p)$.

Given the sample size n, cluster membership $\mathcal{C}_i \in \{1, \ldots, K_{\text{true}}\}$ is generated from a multinomial distribution with probabilities $\frac{1}{1+(K_{\text{true}}-1)\sqrt{m_1}}(1, \sqrt{m_1}, \ldots, \sqrt{m_1})$, where m_1 is a constant that inflates the frequency of clusters other than the first one. This constant also scales down the regression coefficients in the subsequent step. The parameter m_1 corresponds to the setting described in Remark 3, under which BIC often fails to detect the true structure, whereas AICex1 and AICex2 succeed. We consider the scenario with constant $m_1 = 2$, termed "k1". We additionally consider the scenario termed "Imb", which introduces cluster imbalance under a multinomial distribution with probabilities proportional to the K_{true}-vector whose first $[K_{\text{true}}/2]$ elements are $n^{4/5}$ and whose remaining elements are n. To examine the behavior under well-separated clusters in the explanatory variable space, we consider the scenario "Dis", in which the mean vector $\mu_{x,k}$ is replaced by $\frac{5k}{K_{\text{true}}}1_p$ instead of $\frac{k}{K_{\text{true}}}1_p$ for each k.

Other scenarios considered are "MoE" and "io." Unlike the aforementioned scenarios, in the former, the explanatory variables are generated independently and identically from a multivariate normal distribution, and

cluster membership is generated from a multinomial logistic model conditional on the explanatory variables. The response variable is generated in an analogous manner. This scenario is more closely aligned with the MoE model and implies that the clusters cannot be recovered solely from the distribution of the explanatory variables, unlike in the other scenarios. The latter scenario simulates data under an intercept-only model, so the explanatory variables do not affect the response variable and are useful only for assigning cluster membership. In this scenario, criteria based on the intercept-only model or CCMP are expected to perform well. More details are provided in Supplementary Material.

Conditional on cluster membership $\mathcal{C}_i \in \{1, \ldots, K_{\text{true}}\}$, we then generate the explanatory variable as $x_i \sim N(\mu_{x,\mathcal{C}_i}, \Sigma_{x,\mathcal{C}_i})$ for $i = 1, \ldots, n$. We consider simulation scenarios with sample sizes $n = 500, 1000$, dimensions $p = 5, 10$, numbers of clusters $K_{\text{true}} = 1, 3, 9$.

3.1 Gaussian regression model with K-means clustering

First, the true regression coefficients are set as follows. Let the 9×2 matrix $B_{(-1,0,1)}$ consist of all pairwise combinations of three values, $-1, 0, 1$,

$$B^T_{(-1,0,1)} = \begin{pmatrix} -1 & 0 & 1 & -1 & 0 & 1 & -1 & 0 & 1 \\ -1 & -1 & -1 & 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix}.$$

Let the 9×10 matrix B_0 be $B_0 = (B_{(-1,0,1)}, B_{(-1,0,1)}, B_{(-1,0,1)}, B_{(-1,0,1)}, B_{(-1,0,1)})$. We construct the p-vector of regression coefficients $\beta_{p,K_{\text{true}}}$ from B_0 as $\beta_{p,K_{\text{true}},m_1} = D_{m_1} B_{0,[1:K_{\text{true}},1:p]}$, so that the coefficients differ from each other and redundancy is avoided, where $B_{0,[1:K_{\text{true}},1:p]}$ is the sub-matrix of B_0 and $D_{m_1} = \text{diag}(1, \frac{1}{m_1}, \ldots, \frac{1}{m_1})$ is the K_{true}-dimensional diagonal matrix with the first element equal to 1 and the remaining elements equal to $1/m_1$. The constant m_1 is introduced to increase the occurrence probability of clusters other than the first one. We consider $m_1 = 1$ and 2; the latter setting makes it more difficult to distinguish the clusters other than the first one described in Remark 3. The response variable y_i for $i = 1, \ldots, n$ is generated as $y_i = \mu_{i,\mathcal{C}_i} + \epsilon_i, \quad \epsilon_i \sim N(0, 3^2)$, where $\mu_{i,k} = \beta^T_{p,K_{\text{true}},m_1,k} \,\text{diag}(\Sigma_{x,k})^{-1/2}(x_i - \mu_{x,k})$.

We repeat the simulations 100 times and use three evaluation metrics: (i) the proportion that the estimated number of clusters $\widehat{K}$ equals to the true number of clusters K_{true}, $P(\widehat{K} = K_{\text{true}})$, across 100 simulations, as well as the proportions of overfitting and underfitting, $P(\widehat{K} > K_{\text{true}})$ and $P(\widehat{K} < K_{\text{true}})$, respectively; (ii) the adjusted Rand index (ARI; Rand, 1971), which measures the agreement between the true and estimated cluster memberships. In addition, we compute the ARI after binarizing membership in the first cluster, i.e., comparing whether each subject belongs

to the first cluster or not, and contrasting this with the estimated cluster assignment most similar to the true first cluster; and (iii) the mean squared error (MSE) between the true regression function $\mu_{i,\mathcal{C}_i}$ and the estimated linear predictor $\widehat{\mu}_i$ for $i = 1, \ldots, n$, $\frac{1}{n}\sum_{i=1}^{n}(\mu_{i,\mathcal{C}_i} - \widehat{\mu}_i)^2$. The results for the three metrics under scenarios "I" and "W" are presented in Table 2 and Supplementary Tables S2 and S3. The corresponding results for scenarios "W", "k1", "Dis", "MoE", and "io", using "I" as the reference, are provided in Supplementary Tables S4–S18.

From Table 2, the probabilities that AIC selects the true model are away from 1, even when $n = 1000$. In contrast, BIC achieves probabilities closer to 1 in most scenarios, and its performance generally improves as the sample size increases. This finding is consistent with theoretical results: AIC tends to select an overfitted model (Theorem 1), whereas BIC tends to select the correct model (Theorem 3). The proposed exclusion methods adjust AIC, ensuring that its selection probability approaches 1. As expected, AICex1 is more aggressive, whereas AICex2 yields more balanced performance in practice. An increase in the number of dimensions p reduces the probability of selecting the true model, reflecting the well-known effect of dimensionality on BIC (Chen and Chen, 2008). Similarly, increasing the number of clusters K_{true} reduces the probability of selecting the true model.

Supplementary Table S2 supplements Table 2 by presenting the estimated cluster memberships, whereas Table 2 only reports the number of clusters and therefore does not guarantee recovery of the true memberships. The ARI evaluates the agreement between the estimated and true memberships. Larger ARI values generally correspond to higher selection probabilities of the true number of clusters, although a high ARI can occasionally occur even when the selection probability is small. Overall, AICex2 yields better ARI values than AICex1, AIC, and BIC. The ARI for the first cluster highlights the impact of m_1 when comparing BIC with AICex1 or AICex2. The ARIs for AICex1 and AICex2 are close to 1 in the scenarios where BIC fails, suggesting that the first cluster is more reliably detected by the proposed exclusion procedure, consistent with Remark 3.

Supplementary Table S3 further reports estimation results for the regression function. It is well known that AIC performs better in regression estimation, whereas BIC is suboptimal because of its emphasis on selection consistency (Yang, 2005). From this perspective, the smaller MSE obtained by AIC compared with other methods is expected. Although AICex1 and AICex2 achieve selection consistency, they also yield the best or comparable MSEs relative to AIC, whereas BIC generally produces larger MSEs.

In most scenarios, the MoEs exhibit inferior performance to the pro-

posed AICex1 and AICex2, especially when $K_{\text{true}} = 9$. We show in Supplementary Material that simulation model "I" can be regarded as an MoE model with multinomial logistic regression for cluster assignment given the explanatory variables, whereas simulation model "W" cannot. Thus, the performance of the MoEs is worse in "W" because the MoE misspecifies the underlying model. In addition, by construction of the simulation design, the explanatory variables help identify the true clusters. As a consequence, MoEk, which uses K-means on the explanatory variables for the initial clustering, performs better than MoE with random initial clusters. In particular, in simulation "Dis", K-means on the explanatory variables can perfectly reconstruct the true clusters, and thus it is unnecessary, especially for MoEk, to update the initial clustering. Nevertheless, MoEk (and also MoE) fails to reconstruct the true clusters. The failure of MoEk is not caused by an optimization failure, since the initial cluster membership is the true one. To understand this phenomenon, in "Behavior of MoE model" in Supplementary Material, we further conduct additional numerical experiments by varying the magnitude of the regression coefficients, and find that increasing the magnitude improves the accuracy of cluster reconstruction (see Supplementary Figure S1). Thus, we conclude that MoE can perform poorly because of an inadequate effect size in the response variable, and

that unsupervised clustering based on the explanatory variables alone is a better method in the simulations considered.

From the "k1" simulation shown in Supplementary Tables S4–S6, the effect of m_1 indicates that BIC performs worse when $m_1 = 2$ than when $m_1 = 1$, corresponding to scenarios in which BIC's selection probability is far from 1. When $m_1 = 2$, BIC has greater difficulty detecting the first cluster, as discussed in Remark 3. AICex1 and AICex2 demonstrate superior performance in selecting the true number of clusters, even in scenarios where BIC performs poorly.

To examine the effect of cluster imbalance, we consider the "Imb" simulation shown in Supplementary Tables S7–S9. The performance of all methods is generally reduced, but their relative ordering remains the same.

On the other hand, for well-separated clusters in the explanatory-variable space, the "Dis" simulation shown in Supplementary Tables S10–S12 indicates that the proposed approaches perform similarly well to those in simulation "I", although the performance of the MoEs is reduced, as mentioned above. In this simulation setting, because clustering based on the explanatory variables recovers the true cluster membership, separation in the covariate space does not itself drive the exclusion procedure. The proposed procedure still performs well because, conditional on correct cluster recov-

ery, large out-of-cluster prediction error arises from differences between the true cluster-specific regression coefficients, rather than from the distance between clusters in the covariate space itself.

Supplementary Appendix also provides simulations for logistic regression models, with results summarized in Supplementary Tables S19–S36. Overall, the findings are consistent with those for the Gaussian regression models.

3.2 Cox regression model with K-means clustering

Simulations for the Cox regression model are conducted in the same way as for linear and logistic regression, except that the response variables are generated from a survival model. Specifically, explanatory variables x_i are generated as in the linear regression simulations, conditional on cluster membership, and survival outcomes are then generated from a Cox proportional hazards model. The event times T_i are drawn from a Weibull distribution with proportional hazards. Let $\mu_{i,\mathcal{C}_i}$ denote the linear predictor for individual i, defined as $\mu_{i,k} = \beta^T_{p,K_{\text{true}},m_1,k}\,\text{diag}(\Sigma_{x,k})^{-1/2}(x_i - \mu_{x,k})$, where x_i is the explanatory variable vector, and $\beta_{p,K_{\text{true}},m_1,k}$ is the vector of regression coefficients given cluster membership k. The hazard function λ_i for individual i is then $\lambda_i = \lambda_0 e^{\mu_{i,\mathcal{C}_i}}$, with the baseline hazard fixed at

$\lambda_0 = 1$. The survival time T_i is generated from a Weibull distribution with scale parameter λ_i and shape parameter $\frac{1}{2}$. Censoring times C_i are drawn from Exp$\left(\frac{1}{r_i}\right)$, where $r_i = \frac{1}{2U_i\lambda_i^{-1/2}}$ and $U_i \sim U[1,3]$. For each individual, the observed time is $Y_i = \min(T_i, C_i)$, and the event indicator is $\delta_i = 1_{\{T_i \le C_i\}}$.

The evaluation metrics are (i) the probability of selecting the true number of clusters, (ii) the ARI, and (iii) the MSE for the linear predictor in the true regression function $\mu_{i,\mathcal{C}_i}$. The results for scenarios "I", "W", "k1", "Imb", "Dis" and "MoE" are reported in Table 3 and Supplementary Tables S37–S50.

AIC and BIC based on the standard partial log-likelihood exhibit severe overfitting, consistently selecting an excessively large number of clusters. This behavior arises from the divergence of the standard partial log-likelihood, as analyzed in Li et al. (2017) and discussed in the previous section. By contrast, AIC and BIC based on the normalized partial log-likelihood more often identify the correct number of clusters, suggesting that the proposed normalization is effective. Between the two, AIC shows a slight tendency to select larger models more frequently than BIC. The proposed exclusion procedures, AICex1 and AICex2, further mitigate AIC's tendency to overfit. Notably, when combined with AIC under the

standard partial log-likelihood, the exclusion procedures substantially reduce the overfitting observed without them. This phenomenon is explained by Lemma 3 and Remark 4, where, after the exclusion procedure is applied, the candidate models consist only of the true model and underfitted models. In this setting, the tendency of the standard partial log-likelihood to select the largest model leads to consistent selection of the true model. The ARI values reinforce this finding: selecting the true number of clusters corresponds to recovering the true cluster membership. Moreover, the advantage of the exclusion procedures over BIC persists under the normalized partial log-likelihood. For example, when comparing $m_1 = 1$ with $m_1 = 2$, the ARI values indicate that AICex1 and AICex2 can correctly recover the first cluster even when BIC fails. This pattern is consistent across both Gaussian and logistic regression simulations. Finally, the MSE results for the normalized partial log-likelihood show that AIC often outperforms the other criteria, particularly BIC, while AICex1 and AICex2 achieve MSE values comparable to that of AIC.

4. Real data application

We illustrate the proposed approaches using real data through three analyses corresponding to Gaussian, logistic, and Cox regression models. Here,

we present the Cox regression analysis, and the remaining two analyses are given in Supplementary Material.

We apply the method to non-alcoholic fatty liver disease (NAFLD) data, `nafld1`, from the `survival` package (Therneau, 2024) for R. We consider time to death or last follow-up (`futime`) as the survival time, and `status` (0 = alive at last follow-up, 1 = dead) as the event indicator. The explanatory variables are age at study entry (`age`), sex (`male`, coded as 1 for male and 0 for female), `weight`, `height`, and `bmi` (body mass index). A complete-case analysis reduces the sample size to 12,588.

The normalized partial log-likelihood AIC and BIC select $K = 2$ and $K = 1$, respectively. The normalized partial log-likelihood criteria AICex1 and AICex2 both select $K = 2$. In contrast, the partial log-likelihood AIC and BIC select $K = 12$ and $K = 1$, respectively, while the partial log-likelihood AICex1 and AICex2 select $K = 2$ and $K = 7$, respectively.

The Cox regression analysis for the entire sample indicates that `age` ($e^{\beta} = 1.1$, $P < 1.0 \times 10^{-15}$) and `male` ($e^{\beta} = 1.7$, $P = 1.8 \times 10^{-9}$) are positively associated with the event, whereas `weight` shows no evidence of association ($e^{\beta} = 1.01$, $P = 0.43$).

The sample sizes of clusters 1 and 2 are 6,869 and 5,719, respectively. For cluster 1, `age` ($e^{\beta} = 1.1$, $P < 1.0 \times 10^{-15}$) and `male` ($e^{\beta} = 2.0$, $P = 1.7 \times$

10^{-9}) are positively associated with the event, while `weight` is negatively associated ($e^{\beta} = 0.95$, $P = 0.04$). For cluster 2, `age` ($e^{\beta} = 1.1$, $P < 1.0 \times 10^{-15}$) and `male` ($e^{\beta} = 1.4$, $P = 0.04$) are positively associated with the event, while `weight` again shows no evidence of association ($e^{\beta} = 1.01$, $P = 0.47$).

Although an effect of `weight` emerges only in cluster 1, the scatterplot in Supplementary Figure S3 colored by cluster suggests that `weight` also contributes to distinguishing the two clusters. Cluster 1 mainly consists of samples with lower `weight`. Thus, an increase in `weight` appears to reduce the risk of the event only in cluster 1, which is characterized by lower weight in samples in cluster 1.

5. Discussion

We explored the utility of out-of-cluster prediction for model selection in regression with unsupervised clustering, in contrast to existing well-established methods that primarily rely on in-cluster prediction. Two exclusion procedures were considered. We showed that, as known from previous studies, AIC does not necessarily provide consistency in selecting the true number of clusters, whereas BIC does. We further demonstrated that combining AIC with the exclusion procedures yields model selection consis-

tency. While BIC also ensures consistency, the exclusion procedures have additional advantages: they enable more detailed analyses, such as identifying which clusters are redundant, and they improve the ability to detect cases where only a specific cluster has an effect, as described in Remark 3.

While we suggested a combined use of the exclusion procedures with AIC, Lemma 3 and Remark 4 indicate that other criteria that can differentiate the true model from the underfitted models, but do not require the ability to distinguish the true model from overfitted models, may also suffice. For example, the best model in terms of apparent error, such as the log-likelihood, may in principle be used instead of AIC for candidate models after applying the exclusion procedure. Indeed, although we proposed the normalized partial log-likelihood to correct the poor behavior of the standard partial log-likelihood, the exclusion procedures applied to the partial log-likelihood still provided substantially improved model selection performance, although somewhat inferior to the proposed normalized version. This suggests that the exclusion procedures are applicable to broader problems where AIC is unavailable. Furthermore, although our procedures were developed for generalized linear models and the Cox regression model, the idea of supplementing in-cluster prediction with out-of-cluster prediction is general and can be extended to clustering evaluation, aligning with

existing measures such as the silhouette score (Rousseeuw, 1987).

Katahira (2023) recently proposed a novel clustering strategy for prediction through the CCMP criterion. We showed that CCMP is approximately equivalent to AIC under an intercept-only model, namely, a constant model within each cluster. Therefore, it has limited ability to capture the true structure when the underlying model is more complex than a constant model, thereby hampering both the predictive performance and the interpretability of the selected model. In contrast, the proposed approaches performed better in the simulation studies under such scenarios.

The two-stage approach considered in this paper is intended for settings in which subgroup structure is reflected primarily in the distribution of the explanatory variables, rather than being defined solely through the distribution of the response conditional on explanatory variables. In such cases, unsupervised pre-clustering based on the explanatory variables can recover meaningful subgroup structure, after which cluster-specific regression models can be estimated and evaluated. By contrast, simultaneous approaches such as MoE incorporate the response variable into subgroup assignment. This can be advantageous when subgroup differences are expressed mainly through the outcome model, but it may be disadvantageous when the true subgroup structure is already well represented in the explanatory variables,

because incorporating the response may introduce instability or misclassification in subgroup identification, as shown in the simulation studies and Supplementary Figure S1. Thus, the proposed method is not intended to uniformly dominate simultaneous approaches, but rather to provide a more appropriate alternative when heterogeneity is primarily encoded in the distribution of explanatory variables.

In future work, it will be important to extend the methodology to more general regression models, including those without explicit likelihood functions. Machine learning methods provide typical examples. For such models, AIC and BIC are not well-defined, but we believe the basic idea can still be generalized. In particular, out-of-cluster prediction can be incorporated alongside in-cluster prediction. Moreover, incorporating variable selection will be valuable in practical applications, both for improving predictive performance and enhancing interpretability. This extension should be straightforward, since the proposed exclusion procedures are based on loss functions, which also underlie most variable selection methods. In addition, evaluating performance with variable selection is especially beneficial in high-dimensional settings.

Supplementary Material

Supplementary Material include Supplementary Appendix, Tables, and Figures.

Acknowledgements

This work was partially supported by JSPS KAKENHI Grant Numbers 23K11009 and 26K14742. During the preparation of this work the author used ChatGPT-5 in order to improve the readability and language of the manuscript. After using this service, the author reviewed and edited the content as needed and takes full responsibility for the content of the published article.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *2nd Inter. Symp. on Information Theory, Akademiai Kidao, Budapest, 1973*, 267–281.

Batool, F. and C. Hennig (2021). Clustering with the average silhouette width. *Computational Statistics & Data Analysis 158*, 107190.

Chen, J. and Z. Chen (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika 95*(3), 759–771.

Choi, M. Y., I. Chen, A. E. Clarke, M. J. Fritzler, K. A. Buhler, M. Urowitz, J. Hanly, Y. St-

Pierre, C. Gordon, S.-C. Bae, J. Romero-Diaz, J. Sanchez-Guerrero, S. Bernatsky, D. J. Wallace, D. A. Isenberg, A. Rahman, J. T. Merrill, P. R. Fortin, D. D. Gladman, I. N. Bruce, M. Petri, E. M. Ginzler, M. A. Dooley, R. Ramsey-Goldman, S. Manzi, A. Jönsen, G. S. Alarcón, R. F. van Vollenhoven, C. Aranow, M. Mackay, G. Ruiz-Irastorza, S. Lim, M. Inanc, K. Kalunian, S. Jacobsen, C. Peschken, D. L. Kamen, A. Askanase, J. P. Buyon, D. Sontag, and K. H. Costenbader (2023). Machine learning identifies clusters of longitudinal autoantibody profiles predictive of systemic lupus erythematosus disease outcomes. *Annals of the Rheumatic Diseases 82*(7), 927–936.

Grün, B. and F. Leisch (2007). Fitting finite mixtures of generalized linear regressions in r. *Computational Statistics & Data Analysis 51*(11), 5247–5252.

Katahira, K. (2023). Evaluating the predictive performance of subtyping: A criterion for cluster mean-based prediction. *Statistics in Medicine 42*(7), 1045–1065.

Leisch, F. (2004). Flexmix: A general framework for finite mixture models and latent class regression in r. *Journal of Statistical Software 11*(8).

Li, R., J.-J. Ren, G. Yang, and Y. Yu (2017). Asymptotic behavior of Cox's partial likelihood and its application to variable selection. *Statistica Sinica 28*(4), 2713–2731.

Mallows, C. L. (1973). Some comments on C_p. *Technometrics 15*(4), 661.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association 66*(336), 846–850.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of

cluster analysis. *Journal of Computational and Applied Mathematics 20*, 53–65.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics 6*(2), 461–464.

Shao, J. (1997). An asymptotic theory for linear model selection. *Statistica sinica 7*(2), 221–242.

Teng, H.-W., M.-H. Kang, I.-H. Lee, and L.-C. Bai (2024). Bridging accuracy and interpretability: A rescaled cluster-then-predict approach for enhanced credit scoring. *International Review of Financial Analysis 91*, 103005.

Therneau, T. M. (2024). *survival: Survival Analysis*. R package version 3.8-3.

Tibshirani, R., G. Walther, and T. Hastie (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society Series B: Statistical Methodology 63*(2), 411–423.

Ueki, M. (2025). A deflation-adjusted bayesian information criterion for selecting the number of clusters in k-means clustering. *Computational Statistics & Data Analysis 209*, 108170.

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

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.

E-mail: uekimrsd@nifty.com

Table 2: 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 "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|W	91| 9| 0	99| 1| 0	97| 3| 0	100| 0| 0	0|100| 0	20| 80| 0	19| 81| 0	13| 87| 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|W	93| 7| 0	97| 3| 0	95| 5| 0	100| 0| 0	1| 99| 0	12| 88| 0	12| 88| 0	11| 89| 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|W	95| 5| 0	100| 0| 0	98| 2| 0	100| 0| 0	0|100| 0	15| 85| 0	14| 86| 0	6| 94| 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|W	98| 2| 0	100| 0| 0	100| 0| 0	100| 0| 0	0|100| 0	9| 91| 0	8| 92| 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|W	83| 8| 9	88| 0| 12	86| 4| 10	47| 0| 53	0| 97| 3	0| 1| 99	0| 1| 99	0| 58| 42	2| 0| 98	3| 0| 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|W	87| 11| 2	97| 0| 3	90| 8| 2	73| 0| 27	0|100| 0	0| 1| 99	0| 5| 95	0| 90| 10	17| 3| 80	19| 2| 79
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|W	93| 3| 4	96| 0| 4	94| 2| 4	50| 1| 49	0|100| 0	0| 2| 98	0| 4| 96	1| 65| 34	3| 0| 97	1| 0| 99
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|W	97| 2| 1	98| 1| 1	97| 2| 1	72| 0| 28	0|100| 0	0| 3| 97	0| 6| 94	1| 86| 13	11| 0| 89	9| 1| 90
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|W	45| 29| 26	59| 3| 38	54| 20| 26	0| 1| 99	0| 73| 27	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|W	73| 16| 11	72| 1| 27	78| 6| 16	5| 0| 95	0| 97| 3	0| 0|100	0| 0|100	0| 2| 98	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|W	21| 54| 25	58| 10| 32	37| 28| 35	0| 7| 93	0| 82| 18	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|W	68| 25| 7	74| 1| 25	68| 9| 23	0| 0|100	1| 98| 1	0| 0|100	0| 0|100	0| 2| 98	0| 0|100	0| 0|100

Table 3: 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 "W". 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|W	99| 1| 0	99| 1| 0	99| 1| 0	100| 0| 0	0|100| 0	99| 1| 0	89| 11| 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|W	100| 0| 0	100| 0| 0	100| 0| 0	100| 0| 0	0|100| 0	100| 0| 0	100| 0| 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|W	100| 0| 0	100| 0| 0	100| 0| 0	100| 0| 0	0|100| 0	100| 0| 0	88| 12| 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|W	100| 0| 0	100| 0| 0	100| 0| 0	100| 0| 0	0|100| 0	100| 0| 0	98| 2| 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|W	97| 2| 1	96| 1| 3	96| 2| 2	95| 0| 5	0|100| 0	95| 3| 2	74| 25| 1	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|W	98| 1| 1	98| 1| 1	98| 1| 1	98| 0| 2	0|100| 0	96| 4| 0	83| 17| 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|W	100| 0| 0	99| 0| 1	99| 0| 1	98| 0| 2	0|100| 0	96| 3| 1	64| 35| 1	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|W	99| 1| 0	100| 0| 0	99| 1| 0	98| 0| 2	0|100| 0	100| 0| 0	85| 15| 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|W	82| 16| 2	42| 2| 56	78| 9| 13	74| 2| 24	0|100| 0	40| 5| 55	28| 62| 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|W	89| 8| 3	74| 1| 25	91| 5| 4	87| 4| 9	0|100| 0	70| 5| 25	56| 41| 3	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|W	67| 31| 2	74| 9| 17	75| 17| 8	42| 0| 58	1| 99| 0	54| 32| 14	7| 87| 6	15| 85| 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|W	80| 19| 1	79| 4| 17	78| 12| 10	68| 7| 25	0|100| 0	78| 7| 15	34| 58| 8	0|100| 0