STATISTICS IN TRANSITION new series, August 2020
Special Issue, pp. 73–88, DOI 10.21307/stattrans-2020-031
Submitted – 31.01.2020; paper accepted for published – 30.06.2020

Effective transformation-based variable selection under two-fold subarea models in small area estimation

Song Cai[1], **J. N. K. Rao**[2], **Laura Dumitrescu**[3], **Golshid Chatrchi**[4]

ABSTRACT

We present a simple yet effective variable selection method for the two-fold nested subarea model, which generalizes the widely-used Fay-Herriot area model. The two-fold subarea model consists of a sampling model and a linking model, which has a nested-error model structure but with unobserved responses. To select variables under the two-fold subarea model, we first transform the linking model into a model with the structure of a regular regression model and unobserved responses. We then estimate an information criterion based on the transformed linking model and use the estimated information criterion for variable selection. The proposed method is motivated by the variable selection method of Lahiri and Suntornchost (2015) for the Fay-Herriot model and the variable selection method of Li and Lahiri (2019) for the unit-level nested-error regression model. Simulation results show that the proposed variable selection method performs significantly better than some naive competitors, especially when the variance of the area-level random effect in the linking model is large.

Key words: bias correction, conditional AIC, Fay-Herriot model, information criterion.

1. Introduction

Small area estimation (SAE) aims to provide reliable estimates of some parameters of interest, such as means or totals, of subpopulations (areas). Sample surveys are usually carried out in some or all areas to collect unit-level data and design-based direct estimators of the parameters are obtained. A common practical issue in SAE is that the design-based direct estimators are usually unreliable because the sampled areas typically have small sample sizes. It is advantageous to use model-based approaches, which can incorporate auxiliary information through linking models to provide reliable estimates of small area parameters (Rao and Molina, 2015). In general, there are two types of small area models, unit-level models and area-level models. We focus on area-level models.

The celebrated Fay-Herriot (FH) area model (Fay and Herriot, 1979) combines direct estimators and auxiliary variables using a linking model to obtain accurate estimates of small area parameters. Let θ_i be the parameter of interest of a sampled area $i = 1, \ldots, m$

[1] Carleton University, Ottawa, ON, Canada. E-mail: scai@math.carleton.ca. ORCID: https://orcid.org/0000-0003-1368-394X .

[2] Carleton University, Ottawa, ON, Canada. E-mail: jrao@math.carleton.ca. ORCID: https://orcid.org/0000-0003-1103-5500.

[3] Victoria University of Wellington, Wellington, New Zealand. E-mail: laura.dumitrescu@vuw.ac.nz. ORCID: https://orcid.org/0000-0002-9205-9151.

[4] Statistics Canada, Ottawa, Ontario, Canada. E-mail: golshid.chatrchi@canada.ca

and x_i be an associated covariate vector. Let y_i be a direct estimator of θ_i, obtained using unit-level data. The FH model assumes that

$$y_i = \theta_i + e_i, \tag{1}$$

$$\theta_i = x_i^\mathsf{T}\beta + u_i, \tag{2}$$

where β is a parameter vector, u_i's are independent and identically distributed (iid) random effects following $\mathrm{N}(0, \sigma_u^2)$ with unknown σ_u^2, e_i's are independent (ind) sampling errors following $\mathrm{N}(0, \Psi_i)$ with known sampling variance Ψ_i, and u_i's are independent of e_i's. In practice, Ψ_i is obtained by smoothing the direct estimates of the sampling variances, based on the unit level data, and then treating the smoothed estimates as the true sampling variances. Model (1) is known as the "sampling model" and model (2) is called the "linking model". The empirical best linear unbiased prediction (EBLUP) estimator of θ_i for a sampled area is given by $\hat{\theta}_i = \hat{\gamma}_i y_i + (1-\hat{\gamma}_i)x_i^\mathsf{T}\hat{\beta}$, where $\hat{\gamma}_i = \hat{\sigma}_u^2/(\Psi_i + \hat{\sigma}_u^2)$, $\hat{\beta}$ is the best linear unbiased estimator of β and $\hat{\sigma}_u^2$ is the maximum likelihood (ML) estimator or the restricted ML (REML) estiamtor or a method of moments (MM) estimator of σ_u^2 (Rao and Molina, 2015, Chapter 6). The EBLUP estimator of θ_i is a weighted sum of the direct estimator y_i and the so-called "synthetic estimator" $x_i^\mathsf{T}\hat{\beta}$.

When multiple auxiliary variables are available, selecting a parsimonious model that fits the data well is especially important for attaining high estimation accuracy for small area parameters. Han (2013) used a conditional Akaike information criterion (cAIC) to select variables under the FH model. Lahiri and Suntornchost (2015) proposed a variable selection method for the FH model by estimating information criteria under the linking model (2). For variable selection under the unit-level nested-error regression (NER) model (Rao and Molina, 2015, Section 4.3), Meza and Lahiri (2005) proposed a method based on the Fuller-Battese transformation (Fuller and Battese, 1973), which requires estimated values of the variance parameters. Li and Lahiri (2019) used a parameter-free transformation method to avoid estimating the variance parameters.

In many applications, data for the subpopulations of interest are collected using a two-fold setup. First, some areas, e.g. states, are sampled. Then, a sample of subareas, e.g. counties, is further selected from each sampled area. Unit-level data then are collected from the sampled subareas. The goal is to estimate a subarea parameter θ_{ij} where i denotes an area and j denotes a subarea. An example of this nested two-fold setup is given by Mohadjer et al. (2012). In the two-fold case, subareas within an area are likely to share some common features and hence the variables of interest are correlated among those subareas. Naively applying the FH model to the subarea-level data will not capture the correlation.

The two-fold subarea model generalizes the FH model and is tailored for the two-fold setup. Suppose that m areas, labelled as $i = 1, \ldots, m$, are sampled from M areas, and for the ith sampled area, n_i subareas, labelled as $j = 1, \ldots, n_i$, are further sampled from N_i subareas. Let y_{ij}, $i = 1, \ldots, m$ and $j = 1, \ldots, n_i$, be design-unbiased direct estimators of θ_{ij}, and x_{ij} be associated covariate vectors. The two-fold subarea model consists of

$$\text{Sampling model: } y_{ij} = \theta_{ij} + e_{ij}, \tag{3}$$

$$\text{Linking model: } \theta_{ij} = x_{ij}^{\intercal}\beta + v_i + u_{ij}, \tag{4}$$

where $e_{ij} \overset{ind}{\sim} \mathrm{N}(0, \Psi_{ij})$ with known sampling variances Ψ_{ij}, β is a regression parameter vector, $v_i \overset{iid}{\sim} \mathrm{N}(0, \sigma_v^2)$ with unknown σ_v^2, and $u_{ij} \overset{iid}{\sim} \mathrm{N}(0, \sigma_u^2)$ with unknown σ_u^2. The random errors e_{ij}, v_i and u_{ij} are assumed to be independent. Different from the FH model, the linking model (4) under the two-fold subarea model has an area-level random effect v_i, which pools information across subareas within an area. Torabi and Rao (2014) developed the theory of EBLUP estimators under the two-fold subarea model.

Despite the fact that the two-fold subarea model is gaining popularity, little research has been conducted on variable selection under the model. In this paper, we propose a simple yet effective variable selection method for the two-fold subarea model, which combines and extends the variable selection method of Lahiri and Suntornchost (2015) for the FH model and the variable selection method of Li and Lahiri (2019) for the unit-level NER model.

The paper is organized as follows. In Section 2, we give a detailed review of some variable section methods for the FH model. In Section 3, we describe the proposed variable selection method for the two-fold subarea model. Simulation results for assessing the performance of the proposed method are provided in Section 4. Concluding remarks are given in Section 5. Proofs and additional simulation results are included in the Appendix.

2. Variable selection under the FH model

2.1. The Lahiri-Suntornchost method

Lahiri and Suntornchost (2015) developed a simple bias-correction method that can activate multiple information criteria for regular linear regression, including Akaike information criterion (AIC), Bayesian information criterion (BIC), Mallows' C_p and adjusted R^2, to be used for variable selection under the FH model. Note that the linking model (2) takes the form of a regular regression model although the response values θ_i are unobserved. A simple idea is to estimate an information criterion, for example BIC, for the linking model (2) and then use the estimated information criterion to carry out variable selection under the FH model.

To achieve this, Lahiri and Suntornchost (2015) proposed to estimate $\mathrm{MSE}_\theta := \frac{1}{m-p}\theta^{\intercal}(I_m - P)\theta$, where I_m is the m by m identity matrix, $\theta = (\theta_1 \ \dots \ \theta_m)^{\intercal}$, $P = X(X^{\intercal}X)^{-1}X^{\intercal}$, $X = (x_1 \ \dots \ x_m)^{\intercal}$, and p is the length of β under the FH model. The estimator of MSE_θ is given by

$$\widehat{\mathrm{MSE}}_\theta = \mathrm{MSE}_y - \overline{\Psi}_w,$$

where $\mathrm{MSE}_y = \frac{1}{m-p}y^{\intercal}(I_m - P)y$, $y = (y_1 \ \dots \ y_m)^{\intercal}$, $\overline{\Psi}_w = \frac{1}{m-p}\sum_{i=1}^m (1 - h_{ii})\Psi_i$, and $h_{ii} =$

$x_i^\mathsf{T}(X^\mathsf{T}X)^{-1}x_i$. Observing that the BIC for the linking model (2) is a continuous function of MSE_θ, i.e. $\mathrm{BIC} = m\log\{(m-p)\,\mathrm{MSE}_\theta\,/m\} + p\log m$, one can estimate the BIC by plugging in $\widehat{\mathrm{MSE}}_\theta$,

$$\widehat{\mathrm{BIC}} = m\log\Big\{(m-p)\widehat{\mathrm{MSE}}_\theta/m\Big\} + p\log m.$$

Other information criteria, including AIC, Mallows' C_p and adjusted R^2 for the linking model (2), can be estimated similarly. Lahiri and Suntornchost (2015) also proposed a modification to $\widehat{\mathrm{MSE}}_\theta$ that leads to a strictly positive estimator of MSE_θ.

Lahiri and Suntornchost (2015) commented that the goal of their method is to make simple adjustments to the regression packages available to data users, and their objective is not to decide on the best possible regression model selection criterion, but to suggest ways to adjust a data user's favourite model selection criterion. Indeed, given the conceptual and computational simplicity of the method and wide availability of software packages for the regular regression model, this is a method that is likely to be adopted by users.

2.2. The cAIC method

Han (2013) adapted the cAIC method for linear mixed-effects models (Vaida and Blanchard, 2005) to select variables under the FH model. Han (2013) showed that the cAIC for the FH model is given by

$$\mathrm{cAIC} = -2\log f_c(y|\hat{\theta}) + 2\Phi_0,$$

where $\hat{\theta} = (\hat{\theta}_1 \ \ldots \ \hat{\theta}_m)^\mathsf{T}$, $\hat{\theta}_i$ is the EBLUP of θ_i, $f_c(y|\hat{\theta})$ is the conditional density of y given $\hat{\theta}$, and $\Phi_0 = \sum_{i=1}^m(\partial\hat{\theta}_i/\partial y_i)$. When comparing submodels, the submodel with the smallest cAIC value is chosen.

In the expression of the EBLUP $\hat{\theta}_i$, estimated model parameters β and σ_u^2 are required. As a consequence, different estimators of model parameters lead to different expressions for the penalty term Φ_0. Han (2013) derived the analytical expressions of Φ_0 for three frequently used estimators of model parameters: the unbiased quadratic (UQ) estimator, the REML estimator, and the ML estimator. In all three cases, the penalty term Φ_0 has complicated expressions. Compared to the cAIC method, the Lahiri-Suntornchost (2015) method would be more attractive to data users because of its simplicity.

3. Variable selection under two-fold subarea model

We now turn to variable selection under the two-fold subarea model. The two-fold subarea model defined by (3) and (4) can be rewritten in vector form as

$$\text{Sampling model: } y_i = \theta_i + e_i, \tag{5}$$

$$\text{Linking model: } \theta_i = X_i\beta + \tau_i \tag{6}$$

for $i = 1, \ldots, m$, where $y_i = (y_{i1} \ \ldots \ y_{in_i})^\mathsf{T}$, $X_i = (x_{i1} \ \ldots \ x_{in_i})^\mathsf{T}$, $\theta_i = (\theta_{i1} \ \ldots \ \theta_{in_i})^\mathsf{T}$, $e_i = (e_{i1} \ \ldots \ e_{in_i})^\mathsf{T}$, and $\tau_i = v_i \mathbb{1}_{n_i} + u_i$ with $\mathbb{1}_k$ denoting a k-vector of 1s and $u_i = (u_{i1} \ \ldots \ u_{in_i})^\mathsf{T}$. We have $\tau_i \sim \mathrm{N}(0, \Sigma_i)$, where

$$\Sigma_i = \sigma_v^2 \mathbb{1}_{n_i} \mathbb{1}_{n_i}^\mathsf{T} + \sigma_u^2 I_{n_i}. \tag{7}$$

The key difference between the linking model (6) and the linking model (2) under the FH model is that the random effect τ_i in (6) does not have a diagonal structure. If the covariance matrix Σ_i of τ_i can be transformed to have a diagonal structure with equal diagonal entries, then the Lahiri-Suntornchost method for the FH model can be applied. Our proposed method is based on this simple idea and is outlined in two steps below.

First, we linearly transform the linking model (6) into a model with iid random errors. Specifically, for each $i = 1, \ldots, m$, we find a non-random matrix A_i such that $\tau_i^* := A_i \tau_i$ has a diagonal covariance matrix with all diagonal entries equalling some positive constant c across i, and then transform the linking model (6) into

$$\theta_i^* = X_i^* \beta + \tau_i^*, \tag{8}$$

where $\theta_i^* = A_i \theta_i$ and $X_i^* = A_i X_i$. Model (8) takes the form of a regular regression model but with unknown θ_i^*, which is similar to the linking model (2) under the FH model. Second, we estimate information criteria for the transformed linking model (8) using a method similar to the Lahiri-Suntornchost (2015) method for the FH model. The estimated information criteria then can be used for model selection.

In what follows, we give two transformation methods in subsection 3.1, and then describe the proposed method of estimating information criteria in subsection 3.2.

3.1. Transformation

3.1.1 The parameter-free Lahiri-Li transformation

The purpose of the linear transformation A_i is to make $\mathrm{Var}(\tau_i^*) = A_i \Sigma_i A_i^\mathsf{T}$ a diagonal matrix with constant diagonal entries. Ideally, the transformation matrix A_i should not depend on any unknown parameters. Lahiri and Li (2009) proposed a parameter-free transformation method, which can achieve this purpose, and Li and Lahiri (2019) used that transformation method for variable selection under the unit-level NER model. The idea of the transformation is as follows. By (7),

$$\mathrm{Var}(\tau_i^*) = A_i \Sigma_i A_i^\mathsf{T} = \sigma_v^2 (A_i \mathbb{1}_{n_i})(A_i \mathbb{1}_{n_i})^\mathsf{T} + \sigma_u^2 A_i A_i^\mathsf{T}.$$

Hence, to make a constant-diagonal structure for $\mathrm{Var}(\tau_i^*)$, it suffices to find an A_i such that (a) $A_i \mathbb{1}_{n_i} = 0$, and (b) $A_i A_i^\mathsf{T}$ is a diagonal matrix with diagonal entries being constant across $i = 1, \ldots, m$. The conditions (a) and (b) do not involve any parameter, so any matrix A_i satisfying them can be parameter free. Note that the rank of such an A_i is at most $n_i - 1$ because of the linear constraint (a).

Particular examples of parameter-free A_i that satisfy the conditions (a) and (b) were given by Lahiri and Li (2009) and Li and Lahiri (2019), but no general method for finding parameter-free A_i was suggested. Here, we complement their examples by giving a general method to construct a desired A_i as follows.

Step 1: Fix a set of $n_i - 1$ linearly independent vectors of length n_i, denoted $b_1, \ldots, b_{n_i-1}$, which satisfies $b_k^\mathsf{T} \mathbb{1}_{n_i} = 0$ for $k = 1, \ldots, n_i - 1$. For example, one can take b_k to be the vector with kth entry being 1, the last entry being -1 and all the other entries being 0, or, the vector with kth entry being 1, the $(k+1)$th entry being -1 and all the other entries being 0.

Step 2: Apply the Gram-Schmidt process to $b_1, \ldots, b_{n_i-1}$ to obtain a set of orthogonal vectors $a_1, \ldots, a_{n_i-1}$ with $a_1 = b_1$ and $a_k = b_k - \sum_{l=1}^{k-1} \mathrm{Proj}_{a_l}(b_k)$ for $k = 2, \ldots, n_i - 1$, where $\mathrm{Proj}_y(x) := \frac{x^\mathsf{T} y}{y^\mathsf{T} y} y$ is the projection of vector x onto the line spanned by y. Take $A_i = \left(\frac{a_1}{\|a_1\|} \ \cdots \ \frac{a_{n_i-1}}{\|a_{n_i-1}\|} \right)^\mathsf{T}$, where $\| \cdot \|$ is the Euclidean norm.

The A_i constructed this way is parameter free and satisfies the requirements (a) and (b), and correspondingly $A_i A_i^\mathsf{T} = I_{n_i-1}$.

In spite of being parameter free, this transformation has two drawbacks: (i) Since the rank of A_i is $n_i - 1$ instead of n_i, each area i loses one degree of freedom after transformation, which is undesirable when the number of sampled areas, m, is large. (ii) After transformation, the intercept term, if included in the original model, will be removed because of the requirement (a). Hence, this transformation method cannot be used if the intercept is to be selected. In practice, this is not an issue because the intercept is usually included in the model and only the other variables are to be selected. Moreover, transformation matrix that satisfies (a) and (b) is not unique, although we do not find that using different parameter-free transformation matrices affects variable selection results significantly. Overall, being simple and parameter-free is of practical importance and hence the Lahiri-Li transformation method is likely to be favoured by most data users.

3.1.2 The Fuller-Battese transformation

If not restricted to a parameter-free transformation, a straightforward idea to make $\mathrm{Var}(\tau_i^*) = A_i \Sigma_i A_i^\mathsf{T}$ a diagonal matrix with constant diagonal entries is to take $A_i = d\Sigma_i^{-1/2}$, where $\Sigma_i^{-1/2}$ is the positive definite square-root matrix of Σ_i^{-1} and d is a non-zero constant. Choosing $d = \sigma_u$ and working out $\Sigma_i^{-1/2}$, we get

$$A_i = I_{n_i} - \frac{1}{n_i} \left(1 - \sqrt{\frac{1-\rho}{1+(n_i-1)\rho}} \right) \mathbb{1}_{n_i} \mathbb{1}_{n_i}^\mathsf{T},$$

where $\rho = \sigma_v^2 / (\sigma_v^2 + \sigma_u^2)$, which depends on the model parameters σ_v^2 and σ_u^2. This is the same as the transformation used by Fuller and Battese (1973). Under the transformation, $\mathrm{Var}(\tau_i^*) = \sigma_u^2 I_{n_i}$.

In practice, ρ has to be estimated, which is undesirable. One can use the estimating equation (EE) method by Torabi and Rao (2014) or the ML method to estimate ρ

under the two-fold subarea model. Meza and Lahiri (2005) used the Fuller-Battese transformation for variable selection under the unit-level NER model.

3.2. Estimating information criteria

The transformed linking model (8) is a regular regression model with unobserved responses θ_i^*. We now adapt the Lahiri-Suntornchost (2015) method to estimate AIC, BIC and Mallows' C_p under model (8).

Define the mean sum of squares of errors (MSE) of (8) as

$$\mathrm{MSE}_{\theta^*} = \frac{1}{n^* - p}\theta^{*\mathsf{T}}(I_{n^*} - P^*)\theta^*,$$

where $\theta^* = \left(\theta_1^{*\mathsf{T}} \ldots \theta_m^{*\mathsf{T}}\right)^{\mathsf{T}}$, $P^* = X^*\left(X^{*\mathsf{T}}X^*\right)^{-1}X^{*\mathsf{T}}$ with $X^* = \left(X_1^{*\mathsf{T}} \ldots X_m^{*\mathsf{T}}\right)^{\mathsf{T}}$, n^* is the length of θ^*, and p is the length of β. For a submodel of (8) with p_s covariates, the AIC, BIC and Mallows' C_p are given, respectively, by

$$\begin{aligned}\mathrm{AIC}^{(s)} &= n^*\log\left\{\left(n^* - p_s\right)\mathrm{MSE}_{\theta^*}^{(s)}/n^*\right\} + 2p_s,\\ \mathrm{BIC}^{(s)} &= n^*\log\left\{\left(n^* - p_s\right)\mathrm{MSE}_{\theta^*}^{(s)}/n^*\right\} + p_s\log(n^*),\\ C_p^{(s)} &= \left(n^* - p_s\right)\mathrm{MSE}_{\theta^*}^{(s)}/\mathrm{MSE}_{\theta^*} + 2p_s - n^*,\end{aligned}$$

where $\mathrm{MSE}_{\theta^*}^{(s)}$ is the MSE from the submodel. Since θ^* is unknown, the above information criteria cannot be calculated. To estimate them, we first propose an estimator of MSE_{θ^*}.

Transform the direct estimator vector y_i using the same transformation matrix A_i by letting $y_i^* = A_i y_i$ and $y^* = \left(y_1^{*\mathsf{T}} \ldots y_m^{*\mathsf{T}}\right)^{\mathsf{T}}$. Define $\mathrm{MSE}_{y^*} = \frac{1}{n^*-p}y^{*\mathsf{T}}(I_{n^*} - P^*)y^*$. We propose to estimate MSE_{θ^*} by

$$\widehat{\mathrm{MSE}}_{\theta^*} = \mathrm{MSE}_{y^*} - \frac{1}{n^* - p}\mathrm{tr}\left\{(I_{n^*} - P^*)AV_eA^{\mathsf{T}}\right\}, \tag{9}$$

where $A = \mathrm{diag}(A_1, \ldots, A_m)$ and $V_e = \mathrm{diag}(\Psi_{11}, \ldots, \Psi_{mn_m})$. The second term on the right hand side of the above equation can be viewed as a bias-correction term. A simple modification to the MSE estimator as used by Lahiri and Suntornchost (2015) can be applied to $\widehat{\mathrm{MSE}}_{\theta^*}$ to ensure a strictly positive estimator of MSE_{θ^*}.

Theorem 1. *Suppose that the sampling variances Ψ_{ij} are bounded for all i and j, and $n_i \geq 2$ for all i. Then, as the number of areas $m \to \infty$,*

$$\widehat{\mathrm{MSE}}_{\theta^*} = \mathrm{MSE}_{\theta^*} + o_p(1).$$

The proof of Theorem 1 is given in Appendix A. Estimators of AIC, BIC and Mallows' C_p are obtained by plugging $\widehat{\mathrm{MSE}}_{\theta^*}$ into their corresponding expressions. Since all these information criteria are continuous functions of MSE_{θ^*}, by the continuous mapping theorem (van der Vaart, 1998, Theorem 2.3), the errors of the estimated information criteria are also of $o_p(1)$.

To carry out variable selection, one can choose one of the above information criteria and estimate its values for a set of submodels under consideration. The submodel with the smallest estimated information criterion value is selected as the final model.

4. Simulation study

We conducted a simulation study to assess the performance of the proposed variable selection method under the two-fold subarea model. In the simulation, the number of sampled areas m is set to 30. The number of sampled subareas is set to 8 for the first 10 sampled areas, 5 for the next 15 sampled areas, and 10 for the last 5 sampled areas. The sampling standard deviation $\sqrt{\Psi_{ij}}$ is generated from $\text{Unif}(0.5, 1.5)$. We set $\sigma_u = 2$ and consider a few settings for the standard deviation of the area-level random effect with $\sigma_v = 2$, 3.5, 5, 6.5 and 8. In the linking model, we consider an intercept and eight covariates with

$$\begin{aligned}
&x_{ij,1} \sim \text{Log-normal}(0.3, 0.5), \quad x_{ij,2} \sim \text{Gamma}(1.5, 2), \quad x_{ij,3} \sim \text{N}(0, 0.8),\\
&x_{ij,4} \sim \text{N}(1, 1.5), \quad x_{ij,5} \sim \text{Gamma}(0.6, 10), \quad x_{ij,6} \sim \text{Beta}(0.5, 0.5),\\
&x_{ij,7} \sim \text{Unif}(1, 3), \quad x_{ij,8} \sim \text{Poisson}(1.5),
\end{aligned}$$

where $x_{ij,k}$ represents the value of the kth covariate for the ith area and jth subarea, Log-normal(μ, σ) denotes the log-normal distribution with mean μ and standard deviation σ on the log-scale, $\text{Gamma}(\alpha, \beta)$ denotes the gamma distribution with shape parameter α and rate parameter β, $\text{Beta}(\kappa, \gamma)$ denotes the beta distribution with shape parameters κ and γ, $\text{Unif}(a, b)$ denotes the uniform distribution on the interval (a, b), and $\text{Poisson}(\lambda)$ denotes the Poisson distribution with mean parameter λ.

We consider two settings for the true underlying model. In the first setting (Setting I), the true regression parameter value is fixed to $\beta = (2, 0, 0, 4, 0, 8, 0, 0, 0)^\intercal$. The corresponding true model is the submodel with an intercept and covariates $x_{ij,3}$ and $x_{ij,5}$. In the second setting (Setting II), the true regression parameter value is set to $\beta = (2, 3, 0, 4, 0, 8, 0, 1, 0)^\intercal$, which corresponds to the true model with an intercept and covariates $x_{ij,1}$, $x_{ij,3}$, $x_{ij,5}$, and $x_{ij,7}$. When selecting variables, the intercept term is always included in the model, and we compare all submodels defined by inclusion/exclusion of $x_{ij,k}$, $k = 1, \ldots, 8$.

When generating data, the covariates are generated first and fixed throughout all simulation replications. Then in each simulation replication, y_i, $i = 1, \ldots, m$, are generated from the two-fold subarea model using the above settings. The total number of simulation replications is set to 10000.

We use the proposed method to select covariates by comparing all submodels defined by the subsets of the eight covariates. We consider the proposed method using the parameter-free Lahiri-Li transformation (TWOF_{LL}), the Fuller-Battese transformation with the true ρ value ($\text{TWOF}_{\text{FB}}(\rho_0)$), that with the MLE of ρ ($\text{TWOF}_{\text{FB}}(\hat{\rho}_{mle})$), and that with the estimated ρ based on the estimating equation method of Torabi and

Rao (2014) ($\mathrm{TWOF}_{\mathsf{FB}}(\hat{\rho}_{ee})$). For comparison, we consider three naive competitors, the method of Lahiri and Suntornchost (2015) for the FH model fitted naively to the data (Naive 1), information criterion approach for the regular linear regression model fitted naively to the data (Naive 2), and the cAIC method of Han (2013) for the FH model fitted naively to the data (Naive cAIC). Note that different information criteria can be used with Naive 1 and Naive 2 methods, but Naive 3 uses cAIC only.

The simulation results using BIC for variable selection under Setting I of the underlying model are reported in Table 1. All versions of the proposed method have significantly higher percentages of selecting the true model in all cases. When the standard deviation σ_v of the area-level random effect increases, all versions of the proposed method exhibit stable rate of selecting the true model at approximately 77% level, while both naive methods show dramatic decay in performance from approximately 72% rate of selecting the true model when $\sigma_v = 2$ to nearly 12% when $\sigma_v = 8$. This suggests that when there is a strong area-level effect, as it commonly happens in practice, the proposed method is a clear choice over the naive ones. The proposed method based on the parameter-free Lahiri-Li transformation and that based on the Fuller-Battese transformation perform equally well. Moreover, using an estimated ρ instead of the true value of ρ in the Fuller-Battese transformation does not adversely affect the performance of variable selection in this case.

Table 1: Percentage (%) of selecting the true model using BIC;
True model, Setting I: $\beta = (2,0,0,4,0,8,0,0,0)^{\mathsf{T}}$

Method	σ_v				
	2	3.5	5	6.5	8
$\mathrm{TWOF}_{\mathsf{LL}}$	76.98	77.00	76.06	76.52	77.54
$\mathrm{TWOF}_{\mathsf{FB}}(\rho_0)$	78.92	78.58	77.74	77.94	78.46
$\mathrm{TWOF}_{\mathsf{FB}}(\hat{\rho}_{mle})$	78.12	78.16	77.02	77.62	78.46
$\mathrm{TWOF}_{\mathsf{FB}}(\hat{\rho}_{ee})$	78.56	78.34	77.22	76.52	78.70
Naive 1	71.42	49.92	29.48	18.80	11.92
Naive 2	73.90	49.52	29.08	18.34	11.66

The simulation results using AIC and Naive cAIC for variable selection under Setting I are given in Table 2. Compared to BIC, AIC yields lower percentage of selecting the true model for all the methods. However, the comparison between the proposed method and the naive methods is similar to the case using BIC. All versions of the proposed method perform similarly and give stable results for different values of σ_v. The naive methods, on the other hand, have poorer performance, and their performance drops considerably as σ_v increases. The Naive cAIC method performs worse than the Naive 1 and Naive 2 methods, likely because the cAIC has a complicated expression.

The simulation results using Mallows' C_p for variable selection under Setting I are reported in Table 3. These results are similar to those using AIC, and the same conclusion can be drawn: the proposed method has stable performance for different values of σ_v and it outperforms the Naive methods in all cases.

The simulation results for variable selection using BIC, AIC/cAIC and Mallows' C_p

Table 2: Percentage (%) of selecting the true model using AIC or Naive cAIC; True model, Setting I: $\beta = (2,0,0,4,0,8,0,0,0)^{\mathsf{T}}$

Method	σ_v				
	2	3.5	5	6.5	8
$\mathrm{TWOF}_{\mathsf{LL}}$	29.92	28.86	28.12	29.26	30.18
$\mathrm{TWOF}_{\mathsf{FB}}(\rho_0)$	30.30	28.52	28.52	29.32	29.94
$\mathrm{TWOF}_{\mathsf{FB}}(\hat{\rho}_{mle})$	29.94	28.16	28.30	29.14	29.90
$\mathrm{TWOF}_{\mathsf{FB}}(\hat{\rho}_{ee})$	30.02	28.52	28.74	29.56	30.46
Naive 1	27.40	24.94	19.88	15.36	12.82
Naive 2	29.92	26.20	20.04	15.52	12.92
Naive cAIC	22.90	19.18	16.50	12.51	11.44

Table 3: Percentage (%) of selecting the true model using Mallows' C_p; True model, Setting I: $\beta = (2,0,0,4,0,8,0,0,0)^{\mathsf{T}}$

Method	σ_v				
	2	3.5	5	6.5	8
$\mathrm{TWOF}_{\mathsf{LL}}$	31.18	29.98	29.52	30.68	31.48
$\mathrm{TWOF}_{\mathsf{FB}}(\rho_0)$	31.60	29.76	29.62	30.88	31.06
$\mathrm{TWOF}_{\mathsf{FB}}(\hat{\rho}_{mle})$	31.40	29.58	29.50	30.46	31.34
$\mathrm{TWOF}_{\mathsf{FB}}(\hat{\rho}_{ee})$	31.40	29.82	29.82	30.84	31.58
Naive 1	28.40	25.92	20.70	15.84	13.24
Naive 2	31.02	27.22	21.20	16.30	13.22

under Setting II of the underlying model are reported in Table 4, Table 5 and Table 6, respectively, in Appendix B. The comparison among different methods is similar to that under Setting I. It is worth noting that, compared to Setting I, although more covariates are included in the true model under Setting II, the performance gap between the proposed method and the naive methods is larger, and the performance of the naive methods drops quicker as σ_v increases when using AIC, cAIC or Mallows' C_p.

5. Concluding remarks

We proposed a simple transformation-based variable selection method for the two-fold subarea model. This method is a blend of the variable selection method of Lahiri and Suntornchost (2015) for the FH model and the variable selection method of Li and Lahiri (2019) for the unit-level NER model. The proposed method can be used with the parameter-free Lahiri-Li (Lahiri and Li, 2009) transformation or the Fuller-Battese transformation which requires estimating model parameters σ_v^2 and σ_u^2. The performance of the proposed method using two different transformations is found to be comparable and substantially better than some naive competitors, especially when the variance of the area-level random effect is large. In practice, using the proposed method with the parameter-free Lahiri-Li transformation is preferred because of the simplicity of the transformation.

Acknowledgements

This work was supported by research grants to Song Cai and J. N. K. Rao from the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

FAY, R. E., HERRIOT, R. A., (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74(366), pp. 269–277.

FULLER, W. A., BATTESE, G. E., (1973). Transformations for estimation of linear models with nested-error structure. *Journal of the American Statistical Association*, 68(343), pp. 626–632.

HAN, B., (2013). Conditional Akaike information criterion in the Fay-Herriot model. *Statistical Methodology*, 11, pp. 53–67.

LAHIRI, P., LI, Y., (2009). A new alternative to the standard F test for clustered data. *Journal of Statistical Planning and Inference*, 139(10), pp. 3430–3441.

LAHIRI, P., SUNTORNCHOST, J., (2015). Variable selection for linear mixed models with applications in small area estimation. *Sankhyā B*, 77(2), pp. 312–320.

LI, Y., LAHIRI, P., (2019). A simple adaptation of variable selection software for regression models to select variables in nested error regression models. *Sankhyā B*, 81(2), pp. 302–371.

MAGNUS, J. R., NEUDECKER, H., (2019). *Matrix Differential Calculus with Applications in Statistics and Econometrics, 3rd Edition*. Hoboken: Wiley.

MEZA, J. L., LAHIRI, P., (2005). A note on the C_p statistic under the nested error regression model. *Survey Methodology*, 31(1), pp. 105–109.

MOHADJER, L., RAO, J. N. K., LIU, B., KRENZKE, T., VAN DE KERCKHOVE, W., (2012). Hierarchical Bayes small area estimates of adult literacy using unmatched sampling and linking models. *Journal of the Indian Society of Agricultural Statistics*, 66(1), pp. 55–63.

RAO, J. N. K., MOLINA, I., (2015). *Small Area Estimation, 2nd Edition*. Hoboken: Wiley.

TORABI, M., RAO, J. N. K., (2014). On small area estimation under a sub-area level model. *Journal of Multivariate Analysis*, 127, pp. 36–55.

VAIDA, F., BLANCHARD, S., (2005). Conditional Akaike information for mixed-effects models. *Biometrika*, 92(2), pp. 251–370.

VAN DER VAART, A. W., (1998). *Asymptotic Statistics*. Cambridge: Cambridge University Press.

APPENDICES

A. Proof of Theorem 1

The idea of the proof is to show that

$$\mathrm{E}\big(\widehat{\mathrm{MSE}}_{\theta^*}|\theta^*\big)=\mathrm{MSE}_{\theta^*} \tag{10}$$

and

$$\mathrm{E}\big\{\mathrm{Var}\big(\widehat{\mathrm{MSE}}_{\theta^*}|\theta^*\big)\big\}\to 0 \quad \text{as } m\to\infty. \tag{11}$$

Then, by (10) and Markov's inequality, for any given $\varepsilon>0$, we have

$$\begin{aligned}\Pr\Big(\big|\widehat{\mathrm{MSE}}_{\theta^*}-\mathrm{MSE}_{\theta^*}\big|\geq\varepsilon\,\Big|\,\theta^*\Big)&=\Pr\Big(\big|\widehat{\mathrm{MSE}}_{\theta^*}-\mathrm{E}\big(\widehat{\mathrm{MSE}}_{\theta^*}|\theta^*\big)\big|\geq\varepsilon\,\Big|\,\theta^*\Big)\\&\leq\frac{\mathrm{Var}\big(\widehat{\mathrm{MSE}}_{\theta^*}|\theta^*\big)}{\varepsilon^2}.\end{aligned}$$

Taking expectation on both sides of the above inequality and applying (11) gives

$$\Pr\Big(\big|\widehat{\mathrm{MSE}}_{\theta^*}-\mathrm{MSE}_{\theta^*}\big|\geq\varepsilon\Big)\leq\frac{\mathrm{E}\big\{\,\mathrm{Var}\big(\widehat{\mathrm{MSE}}_{\theta^*}|\theta^*\big)\big\}}{\varepsilon^2}\to 0$$

as $m\to\infty$, which proves the claimed. To complete the proof, we show (10) and (11) in the sequel.

Lemma 1. *Let C and D be real-valued matrices of the same order, then*

$$\mathrm{tr}\big\{\big(C^{\mathsf{T}}D\big)^2\big\}\leq\mathrm{tr}\big(C^{\mathsf{T}}CD^{\mathsf{T}}D\big)$$

and

$$\big\{\mathrm{tr}\big(C^{\mathsf{T}}D\big)\big\}^2\leq\mathrm{tr}\big(C^{\mathsf{T}}C\big)\,\mathrm{tr}\big(D^{\mathsf{T}}D\big).$$

Lemma 1 is Theorem 11.2 of Magnus and Neudecker (2019). See a proof therein.

We now prove (10). By the sampling model (5) and the definite of y^*, we have $y^*=\theta^*+e^*$, where $e^*=\big(e_1^{*\mathsf{T}}\dots e_m^{*\mathsf{T}}\big)^{\mathsf{T}}$ with $e_i^*=A_ie_i$ for $i=1,\dots,m$. This gives

$$\mathrm{MSE}_{y^*}=\frac{y^{*\mathsf{T}}(I_{n^*}-P^*)y^*}{n^*-p}\,\frac{\theta^{*\mathsf{T}}(I_{n^*}-P^*)\theta^*+2\theta^{*\mathsf{T}}(I_{n^*}-P^*)e^*+e^{*\mathsf{T}}(I_{n^*}-P^*)e^*}{n^*-p}.$$

Because e^* is independent of θ^*, we have

$$\begin{aligned}\mathrm{E}\big(\mathrm{MSE}_{y^*}\,|\theta^*\big) &= \frac{1}{n^*-p}\Big[\theta^{*\mathsf{T}}(I_{n^*}-P^*)\theta^* + 2\theta^{*\mathsf{T}}(I_{n^*}-P^*)\,\mathrm{E}(e^*) + \mathrm{E}\big\{e^{*\mathsf{T}}(I_{n^*}-P^*)e^*\big\}\Big]\\ &= \mathrm{MSE}_{\theta^*} + \frac{1}{n^*-p}\Big[2\theta^{*\mathsf{T}}(I_{n^*}-P^*)\,\mathrm{E}(e^*) + \mathrm{E}\big\{e^{*\mathsf{T}}(I_{n^*}-P^*)e^*\big\}\Big].\end{aligned}$$

Put $e = \big(e_1^{\mathsf{T}} \ldots e_m^{\mathsf{T}}\big)^{\mathsf{T}}$. Then $\mathrm{E}(e) = 0$, $\mathrm{Var}(e) = V_e$ and $e^* = Ae$, where A and V_e are defined just before Theorem 1. Hence, $\mathrm{E}(e^*) = 0$ and $\mathrm{Var}(e^*) = AV_eA^T$, which implies that $\mathrm{E}\big\{e^{*\mathsf{T}}(I_{n^*}-P^*)e^*\big\} = \mathrm{tr}\big\{(I_{n^*}-P^*)AV_eA^{\mathsf{T}}\big\}$ by a standard result in multivariate statistics. This leads to

$$\mathrm{E}\big(\mathrm{MSE}_{y^*}\,|\theta^*\big) = \mathrm{MSE}_{\theta^*} + \frac{1}{n^*-p}\,\mathrm{tr}\big\{(I_{n^*}-P^*)AV_eA^{\mathsf{T}}\big\}.$$

Then by the definition, (9), of $\widehat{\mathrm{MSE}}_{\theta^*}$, equation (10) is true.

Finally, we prove (11). With simple algebra, we obtain the following decomposition:

$$\mathrm{Var}\big(\widehat{\mathrm{MSE}}_{\theta^*}|\theta^*\big) = \frac{1}{(n^*-p)^2}(T_1+T_2+T_3), \tag{12}$$

where

$$\begin{aligned}T_1 &= \mathrm{Var}\big\{e^{*\mathsf{T}}(I_{n^*}-P^*)e^*\big\},\\ T_2 &= 4\theta^*(I_{n^*}-P^*)\,\mathrm{E}\big\{e^*e^{*\mathsf{T}}(I_{n^*}-P^*)e^*\big\},\\ T_3 &= 4\,\mathrm{E}\Big[\big\{\theta^{*\mathsf{T}}(I_{n^*}-P^*)e^*\big\}^2 \mid \theta^*\Big].\end{aligned}$$

Since $e \sim \mathrm{N}(0, V_e)$, we have $e^* \sim \mathrm{N}(0, AV_eA^{\mathsf{T}})$. Then by a standard result for multivariate normal distribution, we have $\mathrm{E}\big\{e^*e^{*\mathsf{T}}(I_{n^*}-P^*)e^*\big\} = 0$, which gives $T_2 = 0$. In what follows, we derive upper bounds for T_1 and T_3.

By normality of e^*, we have

$$T_1 = 2\,\mathrm{tr}\Big[\big\{(I_{n^*}-P^*)AV_eA^{\mathsf{T}}\big\}^2\Big].$$

Noting that $I_{n^*}-P^*$ is symmetric and idempotent, and AV_eA^{T} is symmetric, by Lemma 1, we have

$$T_1 \le 2\,\mathrm{tr}\Big\{(I_{n^*}-P^*)(AV_eA^{\mathsf{T}})^2\Big\} = 2\,\mathrm{tr}\Big\{(AV_eA^{\mathsf{T}})^2\Big\} - 2\,\mathrm{tr}\Big\{P^*(AV_eA^{\mathsf{T}})^2\Big\}$$

Since P^* is symmetric and idempotent, by the cyclic property of trace, we have

$$\mathrm{tr}\Big\{P^*(AV_eA^{\mathsf{T}})^2\Big\} = \mathrm{tr}\Big\{P^{*2}(AV_eA^{\mathsf{T}})^2\Big\} = \mathrm{tr}\Big\{P^*(AV_eA^{\mathsf{T}})^2P^*\Big\} = \mathrm{tr}\Big\{Q^{\mathsf{T}}Q\Big\} \ge 0,$$

where $Q = (AV_eA^T)P^*$. Therefore,

$$T_1 \le 2\,\mathrm{tr}\Big\{(AV_eA^\mathsf{T})^2\Big\}.$$

Noting that $AV_eA^\mathsf{T} = \mathrm{diag}\{(A_1V_{e_1}A_1^\mathsf{T}), \ldots, (A_mV_{e_m}A_m^\mathsf{T})\}$ where $V_{e_i} = \mathrm{diag}(\Psi_{i1}, \ldots, \Psi_{in_i})$ for $i = 1, \ldots, m$, we have

$$\mathrm{tr}\Big\{(AV_eA^\mathsf{T})^2\Big\} = \sum_{i=1}^m \mathrm{tr}\Big\{\big(A_iV_{e_i}A_i^\mathsf{T}\big)^2\Big\} = \sum_{i=1}^m \mathrm{tr}\Big\{\big(V_{e_i}A_i^\mathsf{T}A_i\big)^2\Big\}.$$

By Lemma 1, we further have

$$\mathrm{tr}\Big\{\big(V_{e_i}A_i^\mathsf{T}A_i\big)^2\Big\} \le \mathrm{tr}\Big\{V_{e_i}^2(A_i^\mathsf{T}A_i)^2\Big\} = \mathrm{tr}\Big\{V_{e_i}(A_i^\mathsf{T}A_i)^2V_{e_i}\Big\}.$$

Let λ_i be the largest eigenvalue of $(A_i^\mathsf{T}A_i)^2$. By an inequality about quadratic forms, we have that the jth diagonal entry of $V_{e_i}(A_i^\mathsf{T}A_i)^2V_{e_i}$ is bounded by $\lambda_i\Psi_{ij}^2$. For both the parameter-free Lahiri-Li transformation based on the proposed procedure using the Gram-Schmidt process and the Fuller-Battese transformation, it is easy to show that $\lambda_i = 1$. Then, since Ψ_{ij} is bounded by some constant Ψ_0 for all i and j, we have

$$\mathrm{tr}\Big\{\big(V_{e_i}A_i^\mathsf{T}A_i\big)^2\Big\} \le \sum_{j=1}^{n_i} \lambda_i\Psi_{ij}^2 \le n_i\Psi_0^2.$$

Therefore,

$$T_1 \le 2\,\mathrm{tr}\Big\{(AV_eA^\mathsf{T})^2\Big\} \le 2\sum_{i=1}^m n_i\Psi_0^2 = 2n\Psi_0^2. \tag{13}$$

We now turn to T_3. Because e^* is independent of θ^* and $\mathrm{E}(e^*) = 0$, we have

$$\begin{aligned} T_3 &= 4\,\mathrm{E}\Big[\big\{\theta^{*\mathsf{T}}(I_{n^*} - P^*)e^*\big\}^2 \mid \theta^*\Big]. \\ &= 4\theta^{*\mathsf{T}}(I_{n^*} - P^*)\,\mathrm{E}\big(e^*e^{*\mathsf{T}}\big)(I_{n^*} - P^*)\theta^* \\ &= 4\theta^{*\mathsf{T}}(I_{n^*} - P^*)\big(AV_eA\big)(I_{n^*} - P^*)\theta^*. \end{aligned}$$

Observing that $T_3 = \mathrm{tr}(T_3)$, we further have

$$\begin{aligned} T_3 &= 4\,\mathrm{tr}\big\{\theta^{*\mathsf{T}}(I_{n^*} - P^*)\big(AV_eA\big)(I_{n^*} - P^*)\theta^*\big\} \\ &= 4\,\mathrm{tr}\big\{\big(AV_eA\big)(I_{n^*} - P^*)\theta^*\theta^{*\mathsf{T}}(I_{n^*} - P^*)\big\}. \end{aligned}$$

Then, by Lemma 1 and (13), we get

$$T_3 \le 4\sqrt{\mathrm{tr}\big\{(AV_eA^\mathsf{T})^2\big\}\,\mathrm{tr}\big\{(UU^\mathsf{T})^2\big\}} \le 4\sqrt{n}\Psi_0\sqrt{\mathrm{tr}\big\{(UU^\mathsf{T})^2\big\}},$$

where $U = (I_{n^*} - P^*)\theta^*$. In addition, by the cyclic property of trace, $\operatorname{tr}\{(UU^\intercal)^2\} = \operatorname{tr}\{(U^\intercal U)^2\} = (U^\intercal U)^2$. Hence

$$T_3 \le 4\sqrt{n}\Psi_0(U^\intercal U) = 4\sqrt{n}\Psi_0\{\theta^{*\intercal}(I_{n^*} - P^*)\theta^*\}. \tag{14}$$

Combining (12), (13), (14) and the fact that $T_2 = 0$, we get

$$\begin{aligned}\operatorname{Var}\big(\widehat{\mathrm{MSE}}_{\theta^*}|\theta^*\big) &= \frac{1}{(n^*-p)^2}(T_1+T_2+T_3)\\ &\le \frac{n}{(n^*-p)^2}2\Psi_0^2 + \frac{\sqrt{n}}{(n^*-p)^2}4\Psi_0\{\theta^{*\intercal}(I_{n^*}-P^*)\theta^*\}.\end{aligned}$$

Therefore,

$$\mathrm{E}\{\operatorname{Var}\big(\widehat{\mathrm{MSE}}_{\theta^*}|\theta^*\big)\} \le \frac{n}{(n^*-p)^2}2\Psi_0^2 + \frac{\sqrt{n}}{(n^*-p)^2}4\Psi_0\,\mathrm{E}\{\theta^{*\intercal}(I_{n^*}-P^*)\theta^*\}.$$

Since θ^* is normally distributed with covariance matrix $\sigma_u^2 I_{n^*}$, $I_{n^*} - P^*$ is a symmetric idempotent matrix and $\mathrm{E}\{(I_{n^*} - P^*)\theta^*\} = 0$, $\frac{1}{\sigma_u^2}\theta^{*\intercal}(I_{n^*} - P^*)\theta^*$ has a chi-square distribution with $n^* - p$ degrees of freedom, and so $\mathrm{E}\{\theta^{*\intercal}(I_{n^*} - P^*)\theta^*\} = (n^* - p)\sigma_u^2$. Further recall that $n^* = n - m$ for the Lahiri-Li transformation, $n^* = n$ for the Fuller-Battese transformation, and $n_i \ge 2$. Then, under both transformations, we have $\frac{n}{(n^*-p)^2} \to 0$ and $\frac{\sqrt{n}}{(n^*-p)} \to 0$ as $m \to \infty$. With the above results, we conclude that

$$\mathrm{E}\{\operatorname{Var}\big(\widehat{\mathrm{MSE}}_{\theta^*}|\theta^*\big)\} \le \frac{n}{(n^*-p)^2}2\Psi_0^2 + \frac{\sqrt{n}}{n^*-p}4\Psi_0\sigma_u^2 \to 0$$

as $m \to \infty$, and hence Theorem 1 is proved. □

B. Simulation results under Setting II of the underlying model

Table 4: Percentage (%) of selecting the true model using BIC;
True model, Setting II: $\beta = (2,3,0,4,0,8,0,1,0)^\intercal$

Method	σ_v				
	2	3.5	5	6.5	8
$\mathrm{TWOF}_{\mathrm{LL}}$	71.95	72.46	72.60	72.54	72.36
$\mathrm{TWOF}_{\mathrm{FB}}(\rho_0)$	73.67	73.22	73.29	73.40	72.76
$\mathrm{TWOF}_{\mathrm{FB}}(\hat\rho_{mle})$	73.02	72.88	73.24	73.18	72.75
$\mathrm{TWOF}_{\mathrm{FB}}(\hat\rho_{ee})$	73.15	73.02	73.28	73.12	72.66
Naive 1	53.53	23.20	9.75	4.04	2.06
Naive 2	50.64	21.52	8.91	3.86	1.95

Table 5: Percentage (%) of selecting the true model using AIC or Naive cAIC; True model, Setting II: $\beta = (2,3,0,4,0,8,0,1,0)^{\mathsf{T}}$

Method	σ_v				
	2	3.5	5	6.5	8
$\mathrm{TWOF}_{\mathsf{LL}}$	42.56	41.59	41.99	42.17	42.48
$\mathrm{TWOF}_{\mathsf{FB}}(\rho_0)$	42.34	42.27	42.26	42.49	42.05
$\mathrm{TWOF}_{\mathsf{FB}}(\hat{\rho}_{mle})$	42.04	41.96	42.05	42.12	42.04
$\mathrm{TWOF}_{\mathsf{FB}}(\hat{\rho}_{ee})$	42.25	42.16	41.37	42.37	42.43
Naive 1	39.20	27.72	18.88	12.18	8.59
Naive 2	41.37	28.32	19.11	12.17	8.64
Naive cAIC	37.11	22.77	14.46	7.71	8.57

Table 6: Percentage (%) of selecting the true model using Mallows' C_p; True model, Setting II: $\beta = (2,3,0,4,0,8,0,1,0)^{\mathsf{T}}$

Method	σ_v				
	2	3.5	5	6.5	8
$\mathrm{TWOF}_{\mathsf{LL}}$	43.83	43.01	43.35	43.49	43.78
$\mathrm{TWOF}_{\mathsf{FB}}(\rho_0)$	43.78	43.55	43.54	43.74	43.47
$\mathrm{TWOF}_{\mathsf{FB}}(\hat{\rho}_{mle})$	43.44	43.27	43.36	43.51	43.39
$\mathrm{TWOF}_{\mathsf{FB}}(\hat{\rho}_{ee})$	43.77	43.69	43.68	43.85	43.91
Naive 1	40.51	28.20	19.17	12.15	8.67
Naive 2	42.46	28.92	19.35	12.16	8.67