Selection of Regression Models under Linear Restrictions for Fixed and Random Designs

Sen Tian* Clifford M. Hurvich Jeffrey S. Simonoff

Department of Technology, Operations, and Statistics,
Stern School of Business, New York University.

Abstract

Many important modeling tasks in linear regression, including variable selection (in which slopes of some predictors are set equal to zero) and simplified models based on sums or differences of predictors (in which slopes of those predictors are set equal to each other, or the negative of each other, respectively), can be viewed as being based on imposing linear restrictions on regression parameters. In this paper, we discuss how such models can be compared using information criteria designed to estimate predictive measures like squared error and Kullback-Leibler (KL) discrepancy, in the presence of either deterministic predictors (fixed-X) or random predictors (random-X). We extend the justifications for existing fixed-X criteria C_p, FPE and AICc, and random-X criteria S_p and RC_p, to general linear restrictions. We further propose and justify a KL-based criterion, RAICc, under random-X for variable selection and general linear restrictions. We show in simulations that the use of the KL-based criteria AICc and RAICc results in better predictive performance and sparser solutions than the use of squared error-based criteria, including cross-validation. Supplemental material containing the technical details of the theorems is attached at the end of the main document. The computer code to reproduce the results for this article, and the complete set of simulation results, are available online[1].

Keywords: AICc; C_p; Information criteria; Optimism; Random-X

1 Introduction

1.1 Model selection under linear restrictions

Consider a linear regression problem with an $n \times 1$ response vector y and an $n \times p$ design matrix X. The true model is generated from

$$y = X\beta_0 + \epsilon, \tag{1}$$

where β_0 is a $p \times 1$ true coefficient vector, and the $n \times 1$ vector ϵ is independent of X, with $\{\epsilon_i\}_{i=1}^n \overset{iid}{\sim} N(0, \sigma_0^2)$. Note that β_0 represents the true parameters, not an intercept term. We consider an approximating model

$$y = X\beta + u,$$

where β is $p \times 1$ and the $n \times 1$ vector u is independent of X, with $\{u_i\}_{i=1}^n \overset{iid}{\sim} N(0, \sigma^2)$. For this approximating model, we further impose m linear restrictions on the coefficient vectors β that are given by

$$R\beta = r, \tag{2}$$

*E-mail: stian@stern.nyu.edu

[1]https://github.com/sentian/RAICc.

where R is an $m \times p$ matrix with linearly independent rows ($\text{rank}(R) = m$) and r is an $m \times 1$ vector. Both R and r are nonrandom. Examples of such restrictions include setting some slopes equal to 0 (which corresponds to variable selection), setting slopes equal to each other (which corresponds to using the sum of predictors in a model), and setting sums of slopes to 0 (which for pairs of predictors corresponds to using the difference of the predictors in a model).

Suppose first that X is deterministic; we refer to this as the fixed-X design. Denote $f(y_i|x_i, \beta, \sigma^2)$ as the density for y conditional on the i-th row of x_i. We have the log-likelihood function (multiplied by -2)

$$-2\log f(y|X, \beta, \sigma^2) = -2\sum_{i=1}^{n} \log f(y_i|x_i, \beta, \sigma^2) = n\log(2\pi\sigma^2) + \frac{1}{\sigma^2}||y - X\beta||_2^2. \tag{3}$$

By minimizing (3) subject to (2), we obtain the restricted maximum likelihood estimator (MLE)

$$\begin{aligned} \hat{\beta} &= \hat{\beta}^f + (X^TX)^{-1}R^T(R(X^TX)^{-1}R^T)^{-1}(r - R\hat{\beta}^f), \\ \hat{\sigma}^2 &= \frac{1}{n}||y - X\hat{\beta}||^2, \end{aligned} \tag{4}$$

where $\hat{\beta}^f = (X^TX)^{-1}X^Ty$ is the unrestricted least squares estimator. Since the errors are assumed to be Gaussian, $\hat{\beta}$ is also the restricted least squares estimator.

In practice, a sequence of estimators $\hat{\beta}(R_i, r_i|X, y)$, each based on a different set of restrictions, is often generated, and the goal is to choose the one with the best predictive performance. This can be done on the basis of information criteria, which are designed to estimate the predictive accuracy for each considered model. Note that the notion of predictive accuracy can be as simple as distance of a predicted value from a future value, as is the case in squared-error prediction measures, but also can encompass the more general idea that the log-likelihood is a measure of the accuracy of a fitted distribution as a prediction for the distribution of a future observation. This idea can be traced back to Akaike (1973), as noted in an interview with Akaike (Findley and Parzen, 1995); see also Efron (1986).

1.2 Variable selection under fixed-X

An important example of comparing models with different linear restrictions on β is variable selection. We consider fitting the ordinary least squares (OLS) estimator on a predetermined subset of predictors with size k, and without loss of generality, the subset includes the first k predictors of X, i.e. $\hat{\beta}^f(X_1, \cdots, X_k, y)$. By letting $R_k = \begin{pmatrix} 0 & I_{p-k} \end{pmatrix}_{(p-k)\times p}$ and $r_k = \begin{pmatrix} 0 \end{pmatrix}_{(p-k)\times 1}$, it is easy to verify that $\hat{\beta}(R_k, r_k|X, y) = \hat{\beta}^f(X_1, \cdots, X_k, y)$. Therefore, comparing OLS fits on different subsets of predictors falls into the framework of comparing estimators with different linear restrictions on β.

Information criteria are designed to provide an unbiased estimate of the test error. We simplify the notation by denoting $\hat{\beta}(k) = \hat{\beta}(R_k, r_k|X, y)$. We also denote errF as the in-sample training error and ErrF as the out-of-sample test error. errF measures how well the estimated model fits on the training data (X, y), while ErrF measures how well the estimated model predicts the new test data $(X, \tilde{y})$, where $\tilde{y}$ is an independent copy of the original response y, i.e. $\tilde{y}$ is drawn from the conditional distribution of $y|X$. The notations of errF and ErrF are based on those in Efron (2004), and the notation F here indicates that we have a fixed-X design. Efron (1986) defined the optimism of a fitting procedure as the difference between the test error and the training error, i.e.

$$\text{optF} = \text{ErrF} - \text{errF},$$

and introduced the optimism theorem,

$$E_y(\text{optF}) = E_y(\text{ErrF}) - E_y(\text{errF}),$$

where E_y represents the expectation taken under the true model with respect to the random variable y. The optimism theorem provides an elegant framework to obtain an unbiased estimator of $E(\text{ErrF})$, that is

$$\widehat{\text{ErrF}} = \text{errF} + E_y(\text{optF}),$$

where the notation $\widehat{\text{ErrF}}$ follows from Efron (2004). It turns out that many existing information criteria can be derived using the concept of optimism.

A typical measure of the discrepancy between the true model and an approximating model is the squared error (SE), i.e.

$$\text{ErrF}_{\text{SE}} = E_{\tilde{y}}\left(\|\tilde{y} - X\hat{\beta}\|_2^2\right).$$

The training error is $\text{errF}_{\text{SE}} = \|y - X\hat{\beta}\|_2^2$. Ye (1998) and Efron (2004) showed that for any general fitting procedure $\hat{\mu}$ and any model distribution (not necessarily Gaussian)

$$E_y(\text{optF}_{\text{SE}}) = 2\sum_{i=1}^{n} \text{Cov}_y(\hat{\mu}_i, y_i), \tag{5}$$

which is often referred to as the covariance penalty. For the OLS estimator $\hat{\mu}(k) = X\hat{\beta}(k)$ it is easy to verify that $E_y(\text{optF}_{\text{SE}}(k)) = 2\sigma_0^2 k$. We denote $\text{RSS}(k)$ as the residual sum of squares for the OLS estimator, i.e. $\text{RSS}(k) = \|y - X\hat{\beta}(k)\|_2^2$. Hence,

$$\widehat{\text{ErrF}}_{\text{SE}}(k) = \text{RSS}(k) + 2\sigma_0^2 k$$

is an unbiased estimator of $E_y(\text{ErrF}_{\text{SE}})$. As suggested by Mallows (1973), typically σ_0^2 is estimated using the OLS fit on all the predictors, i.e. $\hat{\sigma}_0^2 = \text{RSS}(p)/(n-p)$. We then obtain the Mallows' C_p criterion (Mallows, 1973)

$$\text{C}_p(k) = \text{RSS}(k) + \frac{\text{RSS}(p)}{n-p} 2k. \tag{6}$$

An alternative is to use the OLS fit based on the k predictors in the subset to estimate σ_0^2. i.e. $\hat{\sigma}_0^2 = \text{RSS}(k)/(n-k)$, which yields the final prediction error (Akaike, 1969, 1970)

$$\text{FPE}(k) = \text{RSS}(k)\frac{n+k}{n-k}. \tag{7}$$

Another commonly-used error measure is (twice) the Kullback-Leibler (KL) divergence (see, e.g., Konishi and Kitagawa, 2008, Section 3)

$$\text{KLF} = E_{\tilde{y}}\left[2\log f(\tilde{y}|X, \beta_0, \sigma_0^2) - 2\log f(\tilde{y}|X, \hat{\beta}, \hat{\sigma}^2)\right]. \tag{8}$$

The right-hand side of (8) evaluates the predictive accuracy of the fitted model, by measuring the closeness of the distribution of $\tilde{y}$ based on the fitted model and the distribution of $\tilde{y}$ based on the true model. The term $E_{\tilde{y}}\left[2\log f(\tilde{y}|X, \beta_0, \sigma_0^2)\right]$ is the same for every fitted model. Therefore, an equivalent error measure is the expected likelihood

$$\text{ErrF}_{\text{KL}} = E_{\tilde{y}}\left[-2\log f(\tilde{y}|X, \hat{\beta}, \hat{\sigma}^2)\right].$$

The training error is

$$\text{errF}_{\text{KL}} = -2\log f(y|X, \hat{\beta}, \hat{\sigma}^2).$$

For the OLS estimator $\hat{\beta}(k)$, Sugiura (1978) and Hurvich and Tsai (1989) showed that under the Gaussian error (1)

$$E_y(\text{optF}_{\text{KL}}(k)) = n\frac{n+k}{n-k-2} - n,$$

and hence

$$\widehat{\mathrm{ErrF}}_{\mathrm{KL}}(k) = n\log\left(\frac{\mathrm{RSS}(k)}{n}\right) + n\frac{n+k}{n-k-2} + n\log(2\pi)$$

is an unbiased estimator of $E_y(\mathrm{ErrF}_{\mathrm{KL}})$. Since the term $n\log(2\pi)$ appears in all of the models being compared, and thus is irrelevant when comparing criteria for the models, the authors dropped it and introduced the corrected AIC

$$\mathrm{AICc}(k) = n\log\left(\frac{\mathrm{RSS}(k)}{n}\right) + n\frac{n+k}{n-k-2}. \tag{9}$$

Hurvich and Tsai (1991) showed that AICc has superior finite-sample predictive performance compared to AIC (Akaike, 1973)

$$\mathrm{AIC}(k) = n\log\left(\frac{\mathrm{RSS}(k)}{n}\right) + n + 2(k+1),$$

which does not require a Gaussian error assumption but relies on asymptotic results. The derivations of AICc and AIC require the assumption that the true model is included in the approximating models. Neither AICc nor AIC involve σ_0^2, a clear advantage over C_p. Note that the second term of AICc can be rewritten as $n[1+(2k+2)/(n-k-2)]$, which approximately equals the sum of the second and third terms of AIC when n is large relative to k, demonstrating their asymptotic equivalence when $n \to \infty$ and p is fixed.

1.3 From fixed-X to random-X

The assumption that X is fixed holds in many applications, for example in a designed experiment where categorical predictors are represented using indicator variables or effect codings. However, in many other cases where the data are observational and the experiment is conducted in an uncontrolled manner, fixed-X is not valid and it is more appropriate to treat $(x_i, y_i)_{i=1}^n$ as *iid* random draws from the joint distribution of X and y. We refer to this as the random-X design.

As noted by Breiman and Spector (1992), the choice between fixed-X and random-X is conceptual, and is normally determined based on the nature of the study. The extra source of randomness from X results in larger test error compared to the fixed-X situation, and therefore the information criteria designed under fixed-X can be biased estimates of the random-X test error. Furthermore, when applied as selection rules, the authors showed in simulations that C_p leads to significant overfitting under the random-X design. This motivates the derivation of information criteria for the random-X situation.

For the random-X design, we assume that the row vectors of X, $\{x_i\}_{i=1}^n$, are *iid* multivariate normal with mean $E(x_i) = 0$ and covariance matrix $E(x_i x_i^T) = \Sigma_0$. Let $f(y_i, x_i|\beta, \sigma^2, \Sigma)$ denote the joint multivariate normal density for y_i and x_i. Let $g(x_i|\Sigma)$ denote the multivariate normal density for x_i. By partitioning the joint density of (y, X) into the product of the conditional and marginal densities, and by separating the parameters of interest, we have the log-likelihood function (multiplied by -2)

$$\begin{aligned} -2\log f(y, X|\beta, \sigma^2, \Sigma) &= \sum_{i=1}^n -2\log f(y_i, x_i|\beta, \sigma^2, \Sigma) = -2\sum_{i=1}^n [\log f(y_i|x_i, \beta, \sigma^2) + \log g(x_i|\Sigma)] \\ &= \left[n\log(2\pi\sigma^2) + \frac{1}{\sigma^2}||y - X\beta||_2^2\right] + \left[np\log(2\pi) + n\log|\Sigma| + \sum_{i=1}^n x_i^T\Sigma^{-1}x_i\right]. \end{aligned} \tag{10}$$

Minimizing (10) subject to (2), we find that the MLE $(\hat{\beta}, \hat{\sigma}^2)$ of (β, σ^2) remains the same as in the fixed-X design, i.e. (4). The MLE of Σ is given by

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^n x_i\, x_i^T = \frac{1}{n}X^T X.$$

Since $\hat{\beta}$ is unchanged when we move from fixed-X to random-X, variable selection as an example of linear restrictions on β is based on the same parameter estimates as in Section 1.2. Denote errR, ErrR and optR as the training error, test error and the optimism under random-X, respectively. We generate $X^{(n)}$ as an independent copy of X, where the rows $\{x_i^{(n)}\}_{i=1}^n$ are *iid* multivariate normal $\mathcal{N}(0, \Sigma_0)$. The new copy of the response $y^{(n)}$ is generated from the conditional distribution $y|X^{(n)}$. The optimism for random-X can be defined in the same way as for fixed-X, i.e. optR = ErrR − errR. Rosset and Tibshirani (2020) discussed the optimism for general fitting procedures, when the discrepancy between the true and approximating models is measured by the squared error (SE), i.e.

$$\mathrm{ErrR}_{\mathrm{SE}} = E_{X^{(n)},y^{(n)}}\left(\|y^{(n)} - X^{(n)}\hat{\beta}\|_2^2\right).$$

The training error is $\mathrm{errR}_{\mathrm{SE}} = \|y - X\hat{\beta}\|_2^2$. For the OLS estimator, the authors showed that

$$E_{X,y}(\mathrm{optR}_{\mathrm{SE}}(k)) = \sigma_0^2 k\left(2 + \frac{k+1}{n-k-1}\right),$$

and hence

$$\widehat{\mathrm{ErrR}}_{\mathrm{SE}}(k) = \mathrm{RSS}(k) + \sigma_0^2 k\left(2 + \frac{k+1}{n-k-1}\right)$$

is an unbiased estimator of $E_{X,y}(\mathrm{ErrR}_{\mathrm{SE}})$. The result holds for arbitrary joint distributions of (x_y, y_i) , and it only requires x_i being marginally normal. As in the fixed-X case, if we use the unbiased estimate of σ_0^2 based on the full OLS fit, we have the analog of the C_p rule for random-X,

$$\mathrm{RC}_p(k) = \mathrm{RSS}(k) + \frac{\mathrm{RSS}(p)}{n-p}k\left(2 + \frac{k+1}{n-k-1}\right). \tag{11}$$

If we use the alternative estimate of σ_0^2 based on the OLS fit on the k predictors in the subset, i.e. $\hat{\sigma}_0^2 = \mathrm{RSS}(k)/(n-k)$, we have the analog of the FPE rule for random-X,

$$\mathrm{S}_p(k) = \mathrm{RSS}(k)\frac{n(n-1)}{(n-k)(n-k-1)}. \tag{12}$$

Hocking (1976) refers to (12) as the S_p criterion of Sclove (1969); see also Thompson (1978a,b). Note that the notation used here is slightly different from that in Rosset and Tibshirani (2020), where the authors used RC_p to denote the infeasible criterion involving σ_0^2 and used $\widehat{\mathrm{RC}}_p$ to denote the feasible criterion S_p. The RC_p criterion in our notation was not studied in their paper.

Another class of selection rules is cross-validation (CV), which does not impose parametric assumptions on the model. A commonly used type of CV is the so-called K-fold CV. The data are randomly split into K equal folds. For each fold, the model is fitted using data in the remaining folds and is evaluated on the current fold. The process is repeated for all K folds, and an average squared error is obtained. In particular, the n-fold CV or leave-one-out (LOO) CV provides an approximately unbiased estimator of the test error under the random-X design, i.e. $E_{X,y}(\mathrm{ErrR}_{\mathrm{SE}})$. Burman (1989) showed that for OLS, LOOCV has the smallest bias and variance in estimating the squared error-based test error, among all K-fold CV estimators. LOOCV is generally not preferred due to its large computational cost, but for OLS, the LOOCV error estimate has an analytical expression: the predicted residual sum of squares (PRESS) statistic (Allen, 1974)

$$\mathrm{PRESS}(k) = \sum_{i=1}^{n}\left(\frac{y_i - x_i^T\hat{\beta}(k)}{1 - H_{ii}(k)}\right)^2,$$

where $H(k) = X(k)(X(k)^TX(k))^{-1}X(k)^T$ and $X(k)$ contains the first k columns of X.

1.4 General linear restrictions

Variable selection is a special case of linear restrictions on β, where certain entries of β are restricted to be zero. In practice, we may restrict predictors to have the same coefficient (e.g. $\beta_1 = \beta_2 = \beta_3$), or we may restrict the sum of their effects (e.g. $\beta_1 + \beta_2 + \beta_3 = 1$). Using the structure in (2), we formulate a sequence of models, each of which imposes a set of general restrictions on β, where the goal is to select the model with best predictive performance. The previously defined information criteria and PRESS cannot be applied to this problem, although Tarpey (2000) derived the PRESS statistic for the estimator under general restrictions as

$$\text{PRESS}(R, r) = \sum_{i=1}^{n} \left(\frac{y_i - x_i^T \hat{\beta}}{1 - H_{ii} + H_{Q\,ii}} \right)^2,$$

where $H = X(X^TX)^{-1}X^T$ and $H_Q = X(X^TX)^{-1}R^T \left[R(X^TX)^{-1}R^T\right]^{-1} R(X^TX)^{-1}X^T$.

1.5 The contribution of this paper

The information criteria introduced in Section 1.2 have been studied primarily in the context of variable selection problems under fixed-X. In this paper we discuss how such criteria can be generalized to model comparison under general linear restrictions under either fixed-X or random-X (in both cases including the special case of variable selection). Note that a selection rule is preferred if it chooses the models that lead to the best predictive performance. This is related to, but not the same as, providing the best estimate of the test error. These two goals are fundamentally different (see, e.g., Hastie et al., 2009, Section 7), and we focus on the predictive performance of the selected model.

In Section 2, we consider the fixed-X situation and derive general versions of AICc, C_p and FPE for arbitrary linear restrictions on β. Random-X is assumed in Section 3 and a version of RC_p and S_p for general linear restrictions is obtained. Furthermore, we propose and justify a novel criterion, RAICc, for general linear restrictions and discuss its connections with AICc. We further show that expressions for the information criteria for variable selection problems can be recovered as special cases of their expressions derived under general restrictions. In Section 4, we show via simulations that AICc and RAICc provide consistently strong predictive performance for both variable selection and general restrictions. Lastly, in Section 5, we provide conclusions and discussions of potential future work.

2 Information criteria for fixed-X

2.1 KL-based information criterion

Using the likelihood function (3) and the MLE (4), the expected log-likelihood can be derived as

$$\begin{aligned} \text{ErrF}_{\text{KL}} &= E_{\tilde{y}}[-2\log f(\tilde{y}|X, \hat{\beta}, \hat{\sigma}^2)] = n\log(2\pi\hat{\sigma}^2) + \frac{1}{\hat{\sigma}^2} E_{\tilde{y}}||\tilde{y} - X\hat{\beta}||_2^2 \\ &= n\log(2\pi\hat{\sigma}^2) + \frac{1}{\hat{\sigma}^2}(\hat{\beta} - \beta_0)^T X^T X(\hat{\beta} - \beta_0) + \frac{n\sigma_0^2}{\hat{\sigma}^2}, \end{aligned}$$

and the training error is

$$\text{errF}_{\text{KL}} = -2\log f(y|X, \hat{\beta}, \hat{\sigma}^2) = n\log(2\pi\hat{\sigma}^2) + n.$$

In the context of variable selection, the assumption that the approximating model includes the true model is used in the derivations of AIC (Linhart and Zucchini, 1986) and AICc (Hurvich and Tsai, 1989). This assumption can be generalized to the context of general restrictions.

Assumption 1. *If the approximating model satisfies the restrictions $R\beta = r$, then the true model satisfies the analogous restrictions $R\beta_0 = r$; that is, the true model is at least as restrictive as the approximating model.*

Under this assumption, we have the following lemma. The proofs for all of the lemmas and theorems in this paper are given in the Supplemental Material.

Lemma 1. *Under Assumption 1, $\hat{\sigma}^2$ and the quadratic form $(\hat{\beta} - \beta_0)^T X^T X(\hat{\beta} - \beta_0)$ are independent, and*

$$n\sigma_0^2 E_y\left[\frac{1}{\hat{\sigma}^2}\right] = n\frac{n}{n-p+m-2},$$
$$E_y\left[(\hat{\beta} - \beta_0)^T X^T X(\hat{\beta} - \beta_0)\right] = \sigma_0^2(p-m).$$

Lemma 1 provides the fundamentals for calculating the expected optimism.

Theorem 1. *Under Assumption 1,*

$$E_y(optF_{KL}) = n\frac{n+p-m}{n-p+m-2} - n.$$

Consequently,

$$\widehat{\mathrm{ErrF}}_{\mathrm{KL}} = \mathrm{errF}_{\mathrm{KL}} + E_y(\mathrm{optF}_{\mathrm{KL}}) = n\log(\hat{\sigma}^2) + n\frac{n+p-m}{n-p+m-2} + n\log(2\pi)$$

is an unbiased estimator of the test error $E_y\left[\mathrm{ErrF}_{\mathrm{KL}}\right]$. We follow the same tradition as in the derivations of AIC and AICc that since the term $n\log(2\pi)$ appears in $\widehat{\mathrm{ErrF}}_{\mathrm{KL}}$ for every model being compared, it is irrelevant for purposes of model selection. We therefore ignore this term and define

$$\mathrm{AICc}(R,r) = n\log\left(\frac{\mathrm{RSS}(R,r)}{n}\right) + n\frac{n+p-m}{n-p+m-2},$$

where $\mathrm{RSS}(R,r) = \|y - X\hat{\beta}\|_2^2$. For the variable selection problem, e.g. regressing on a subset of predictors with size k, we are restricting $p-k$ slope coefficients to be zero. By plugging $\hat{\beta} = \hat{\beta}(k)$ and $m = p-k$ into the expressions of $\mathrm{AICc}(R,r)$, we obtain $\mathrm{AICc}(k)$ given in (9).

2.2 Squared error-based information criterion

The covariance penalty (5) is defined for any general fitting procedure. By explicitly calculating the covariance term for $\hat{\mu} = X\hat{\beta}$, we can obtain the expected optimism.

Theorem 2.

$$E_y(optF_{SE}) = 2\sigma_0^2(p-m).$$

An immediate consequence of this is that

$$\widehat{\mathrm{ErrF}}_{\mathrm{SE}} = \mathrm{errF}_{\mathrm{SE}} + E_y(\mathrm{optF}_{\mathrm{SE}}) = \mathrm{RSS}(R,r) + 2\sigma_0^2(p-m)$$

is an unbiased estimator of $E_y(\mathrm{ErrF}_{\mathrm{SE}})$. Using the unbiased estimator of σ_0^2 given by the OLS fit based on all of the predictors, i.e. $\hat{\sigma}_0^2 = \mathrm{RSS}(p)/(n-p)$, we define

$$\mathrm{C}_p(R,r) = \mathrm{RSS}(R,r) + \frac{\mathrm{RSS}(p)}{n-p}2(p-m).$$

An alternative estimate of σ_0^2 is $\mathrm{RSS}(R,r)/(n-p+m)$, which yields

$$\mathrm{FPE}(R,r) = \mathrm{RSS}(R,r)\frac{n+p-m}{n-p+m}.$$

For the variable selection problem, by substituting $m = p-k$ into the expressions of C_p and FPE, we obtain the previously-noted definitions of them, i.e. C_p(k) and FPE(k) given in (6) and (7), respectively.

3 Information criteria for random-X

3.1 KL-based information criterion, RAICc

We replace the unknown parameters by their MLE, and have the fitted model $f(\cdot|\hat{\beta},\hat{\sigma}^2,\hat{\Sigma})$. The KL information measures how well the fitted model predicts the new set of data $(X^{(n)},y^{(n)})$, in terms of the closeness of the distributions of $(X^{(n)},y^{(n)})$ based on the fitted model and the true model, i.e.

$$\mathrm{KLR} = E_{X^{(n)},y^{(n)}}\left[2\log f(X^{(n)},y^{(n)}|\beta_0,\sigma_0^2,\Sigma_0) - 2\log f(X^{(n)},y^{(n)}|\hat{\beta},\hat{\sigma}^2,\hat{\Sigma})\right]. \tag{13}$$

An equivalent form for model comparisons is the expected log-likelihood

$$\begin{aligned}
&\mathrm{ErrR}_{\mathrm{KL}} = E_{X^{(n)},y^{(n)}}\left[-2\log f(X^{(n)},y^{(n)}|\hat{\beta},\hat{\sigma}^2,\hat{\Sigma})\right] \\
&= \left[n\log(2\pi\hat{\sigma}^2) + \frac{1}{\hat{\sigma}^2}E_{X^{(n)},y^{(n)}}||y^{(n)}-X^{(n)}\hat{\beta}||_2^2\right] + \left[np\log(2\pi) + n\log|\hat{\Sigma}| + E_{X^{(n)}}\left(\sum_{i=1}^{n} {x_i^{(n)}}^T\hat{\Sigma}^{-1}x_i^{(n)}\right)\right] \\
&= \left[n\log(2\pi\hat{\sigma}^2) + \frac{n}{\hat{\sigma}^2}(\hat{\beta}-\beta_0)^T\Sigma_0(\hat{\beta}-\beta_0) + \frac{n\sigma_0^2}{\hat{\sigma}^2}\right] + \left[np\log(2\pi) + n\log|\hat{\Sigma}| + n\mathrm{Tr}(\hat{\Sigma}^{-1}\Sigma_0)\right],
\end{aligned}$$

and the training error is

$$\mathrm{errR}_{\mathrm{KL}} = -2\log f(X,y|\hat{\beta},\hat{\sigma}^2,\hat{\Sigma}) = \left[n\log(2\pi\hat{\sigma}^2)+n\right] + \left[np\log(2\pi) + n\log|\hat{\Sigma}| + np\right].$$

As in the fixed-X case, we assume that the true model satisfies the restrictions, i.e. $R\beta_0 = r$, and we obtain the following lemma.

Lemma 2. *Under Assumption 1, $\hat{\sigma}^2$ and $(\hat{\beta}-\beta_0)^T\Sigma_0(\hat{\beta}-\beta_0)$ are independent conditionally on X, and*

$$\begin{aligned}
E\left[Tr(\hat{\Sigma}^{-1}\Sigma_0)\right] &= \frac{np}{n-p-1}, \\
E_{X,y}\left[(\hat{\beta}-\beta_0)^T\Sigma_0(\hat{\beta}-\beta_0)\right] &= \sigma_0^2\frac{p-m}{n-p+m-1}.
\end{aligned}$$

Lemma 2 provides the components for calculating the expected optimism.

Theorem 3. *Under Assumption 1,*

$$E_{X,y}(optR_{KL}) = n\frac{n(n-1)}{(n-p+m-2)(n-p+m-1)} + n\frac{np}{n-p-1} - n(p+1).$$

Consequently,

$$\begin{aligned}\widehat{\mathrm{ErrR}}_{\mathrm{KL}} &= \mathrm{errR}_{\mathrm{KL}} + E_{X,y}(\mathrm{optR}_{\mathrm{KL}}) \\ &= n\log\left(\hat{\sigma}^2\right) + n\frac{n(n-1)}{(n-p+m-2)(n-p+m-1)} + n\log(2\pi)(p+1) + n\frac{np}{n-p-1} + n\log|\hat{\Sigma}|\end{aligned}$$

is an unbiased estimator of the test error $E_{X,y}(\mathrm{ErrR}_{\mathrm{KL}})$. Note that the last three terms are free of the restrictions and only depend on n, p and X. They are the same when we compare two models with different restrictions on β, and are thus irrelevant when comparing criteria for any two such models. Therefore, for the purpose of model selection, we define

$$\mathrm{RAICc}(R,r) = n\log\left(\frac{\mathrm{RSS}(R,r)}{n}\right) + n\frac{n(n-1)}{(n-p+m-2)(n-p+m-1)}.$$

An equivalent form is

$$\mathrm{RAICc}(R,r) = \mathrm{AICc}(R,r) + \frac{n(p-m)(p-m+1)}{(n-p+m-1)(n-p+m-2)}.$$

For linear regression on a subset of predictors with size k, we are restricting $p-k$ coefficients to be zero. By substituting $m = p-k$ and $\hat{\beta} = \hat{\beta}(k)$ into the expression of $\mathrm{RAICc}(R,r)$, we obtain the RAICc criterion for the variable selection problem, i.e.

$$\mathrm{RAICc}(k) = n\log\left(\frac{\mathrm{RSS}(k)}{n}\right) + n\frac{n(n-1)}{(n-k-2)(n-k-1)}.$$

3.2 Squared error-based information criteria

We note from Theorem 5 that $E_y(\mathrm{optF}_{\mathrm{SE}})$ is independent of X. According to Rosset and Tibshirani (2020, formula 6 and proposition 1), $E_{X,y}(\mathrm{optR}_{\mathrm{SE}})$ can be decomposed into $E_y(\mathrm{optF}_{\mathrm{SE}})$ plus an excess bias term and an excess variance term. We calculate both terms for our estimator $\hat{\beta}$ and obtain the following theorem.

Theorem 4.

$$E_{X,y}(optR_{SE}) = \sigma_0^2(p-m)\left(2 + \frac{p-m+1}{n-p+m-1}\right).$$

An immediate consequence is that

$$\widehat{\mathrm{ErrR}}_{\mathrm{SE}} = \mathrm{errR}_{\mathrm{SE}} + E_{X,y}(\mathrm{optR}_{\mathrm{SE}}) = \mathrm{RSS}(R,r) + \sigma_0^2(p-m)\left(2 + \frac{p-m+1}{n-p+m-1}\right)$$

is an unbiased estimator of $E_{X,y}(\mathrm{ErrR}_{\mathrm{SE}})$. Using the OLS fit on all of the predictors to estimate σ_0^2, we have

$$\mathrm{RC}_p(R,r) = \mathrm{RSS}(R,r) + \frac{\mathrm{RSS(p)}}{n-p}(p-m)\left(2 + \frac{p-m+1}{n-p+m-1}\right).$$

An alternative estimate of σ_0^2 is $\mathrm{RSS}(R,r)/(n-p+m)$, which yields

$$\mathrm{S}_p(R,r) = \mathrm{RSS}(R,r)\frac{n(n-1)}{(n-p+m)(n-p+m-1)}.$$

For the variable selection problem, by substituting $m = p-k$ into the expressions of RC_p and S_p, we obtain the previously-noted definitions of them, i.e. $\mathrm{RC}_p(\mathrm{k})$ and $\mathrm{S}_p(\mathrm{k})$ given in (11) and (12), respectively.

4 Performance of the selectors

4.1 Some other selectors

In this section we use computer simulations to explore the behavior of different criteria when used for model selection under linear restrictions (variable selection and general linear restrictions). In addition to the criteria already discussed, we also consider two other well-known criteria:

$$\mathrm{BIC}(k) = n \log\left(\frac{\mathrm{RSS}(k)}{n}\right) + \log(n)k$$

(Schwarz, 1978), and generalized cross-validation (GCV)

$$\mathrm{GCV}(k) = \mathrm{RSS}(k)\frac{n^2}{(n-k)^2}. \tag{14}$$

BIC is a consistent criterion, in the sense that under some conditions, if the true model is among the candidate models, the probability of selecting the true model approaches one, as the sample size becomes infinite. GCV, derived by Craven and Wahba (1978) in the context of smoothing, is equivalent to the mean square over degrees of freedom criterion proposed by Tukey (1967). By comparing the expressions of (14) and (12), GCV and S_p only differ by a multiplicative factor of $1+k/[(n-k)(n-1)]$.

By analogy to the criteria discussed in Section 2 and 3, if we substitute $k = p - m$ into the expressions of $\mathrm{BIC}(k)$ and $\mathrm{GCV}(k)$, we obtain their corresponding expressions for general linear restrictions $\mathrm{BIC}(R, r)$ and $\mathrm{GCV}(R, r)$, respectively. We also consider two types of the cross-validation (CV): 10-fold CV (denoted as 10FCV) and leave-one-out CV (LOOCV). The LOOCV is based on the PRESS(k) statistic for the variable selection problem and PRESS(R,r) for the general restriction problem.

4.2 Random-X

We first consider the variable selection problem. The candidate models include the predictors of X in a nested fashion, i.e. the candidate model of size k includes the first k columns of X $(X_1, \cdots, X_k)$. We describe the simulation settings reported here; description of and results from all other settings (243 configurations in total) can be found in the Online Supplemental Material[2], where we also provide the code to reproduce all of the simulation results in this paper. The sample sizes considered are $n \in \{40, 1000\}$, with number of predictors $p = n - 1$, being close to the sample size. Predictors exhibit moderate correlation with each other of an AR(1) type (see the Online Supplemental material for further details), and the strength of the overall regression is characterized as either low (average R^2 on the set of true predictors roughly 20%) or high (average R^2 on the set of true predictors roughly 90%). The true model is either sparse (with six nonzero slopes) or dense (with p nonzero slopes exhibiting diminishing strengths of coefficients; Taddy, 2017). The design matrix X is random. In each replication, we generate a matrix X such that the rows x_i $(i = 1, \cdots, n)$ are drawn from a p-dimensional multivariate normal distribution with mean zero and covariance matrix Σ_0, and we draw the response y from the conditional distribution of $y|X$ based on (1). The entire process is repeated 1000 times.

We consider the following metrics to evaluate the fit. The values of each criterion over all of the simulation runs are plotted using side-by-side boxplots, with the average value over the simulation runs given below the boxplot for the corresponding criterion.

- Root mean squared error for random-X:

$$\mathrm{RMSER} = \sqrt{E_{X^n}\|X^n\hat{\beta} - X^n\beta_0\|_2^2} = \sqrt{(\hat{\beta} - \beta_0)^T\Sigma_0(\hat{\beta} - \beta_0)}.$$

[2] `https://github.com/sentian/RAICc`

- KL discrepancy for random-X (13) in the log scale (denoted as logKLR).
- Size of the subset selected.

The results are presented in Figures 1 and 2. We find that RAICc provides the best predictive performance and the sparsest subset while rarely underfitting, compared to other information criteria designed for random-X, including RC_p and S_p. The underperformance of RC_p and S_p is due to overfitting. S_p, as an estimate of the squared prediction error, is a product of an unlogged residual sum of squares and a penalty term that is increasing in k. This results in higher variability in S_p for models that overfit, thereby potentially increasing the chances for spurious minima of these criteria at models that drastically overfit. In RAICc, on the other hand, the residual sum of squares is logged, thereby stabilizing the variance and avoiding the problem. RC_p drastically overfits in all scenarios, reflecting the price of estimating σ_0^2 using the full model, especially when p is close to n. S_p, on the other hand, estimates σ_0^2 using the candidate model, which mitigates the problem. Nevertheless, S_p also can sometimes strongly overfit, but only when n is small. Even for large n, S_p selects slightly larger subsets on average than does RAICc.

We also note that information criteria designed for random-X generally perform better than their counterparts for the fixed-X case. Both C_p and FPE are largely outperformed by RC_p and S_p, respectively. The advantage of RAICc over AICc is statistically significant in most scenarios, based on the Wilcoxon signed-rank test (the p-value for the test comparing the criteria for RAICc and AICc is given above the first two boxes in the first two columns of the table), but is not obvious in a practical sense. The only place that we see an advantage of AICc is for {Dense model, $n = 40$ and high signal}. In this scenario, a model with many predictors with nonzero slopes can predict well, but that advantage disappears when there is a relatively weak signal, as in that situation the added noise from including predictors with small slopes cannot be overcome by a small error variance.

We further note that choosing the appropriate family of information criteria (the KL-based AICc and RAICc) is more important than choosing the information criteria designed for the underlying designs of X. AICc, despite being designed for fixed-X, outperforms RC_p and S_p, which are designed for random-X in all of the scenarios, in terms of both predictive performance and providing sparse results. The KL-based criteria have a clear advantage compared to the squared-error based criteria.

Finally, we note some other findings that have been discussed previously in the literature. Despite its apparently strong penalty, BIC often chooses the model using all predictors when n is close to p, as discussed in Hurvich and Tsai (1989) and Baraud et al. (2009). We also see that even though GCV has a similar penalty term as S_p, it is more likely to suffer from overfitting. Unlike S_p, GCV can sometimes drastically overfit even when n is large. The overfitting problem of GCV was also observed in the context of smoothing by Hurvich et al. (1998). We further find that 10-fold CV performs better than LOOCV, the latter of which sometimes drastically overfits. The tendency of LOOCV to strongly overfit was noted by Scott and Terrell (1987) and Hall and Marron (1991) in the context of smoothing. Zhang and Yang (2015) showed that when applied as selection rules, the larger validation set used by 10-fold CV can better distinguish the candidate models than can LOOCV, and this results in a model with smaller predictive error. RAICc performs better than 10-fold CV for small n, and performs similarly for large n. Computationally, 10-fold CV is ten times more expensive compared to RAICc, and since the split of validation samples is random, 10-fold CV can select different subsets if applied multiple times on the same dataset. The fact that LOOCV provides better estimate of the test error while being outperformed by 10-fold CV, further emphasizes the difference between the goal of providing the best estimate of the test error, and the goal of selecting the models with the best predictive performance. Clearly, KL-based criteria (AICc and RAICc) bridge the gap between the two goals more effectively than the squared-error based criteria (including cross-validation).

In a related study by Leeb (2008), the author found that S_p and GCV outperform AICc under random-X, but those results are not directly comparable to ours. That paper did not consider the case where p is extremely close to n, which is the scenario that most separates the performances of the

different criteria.

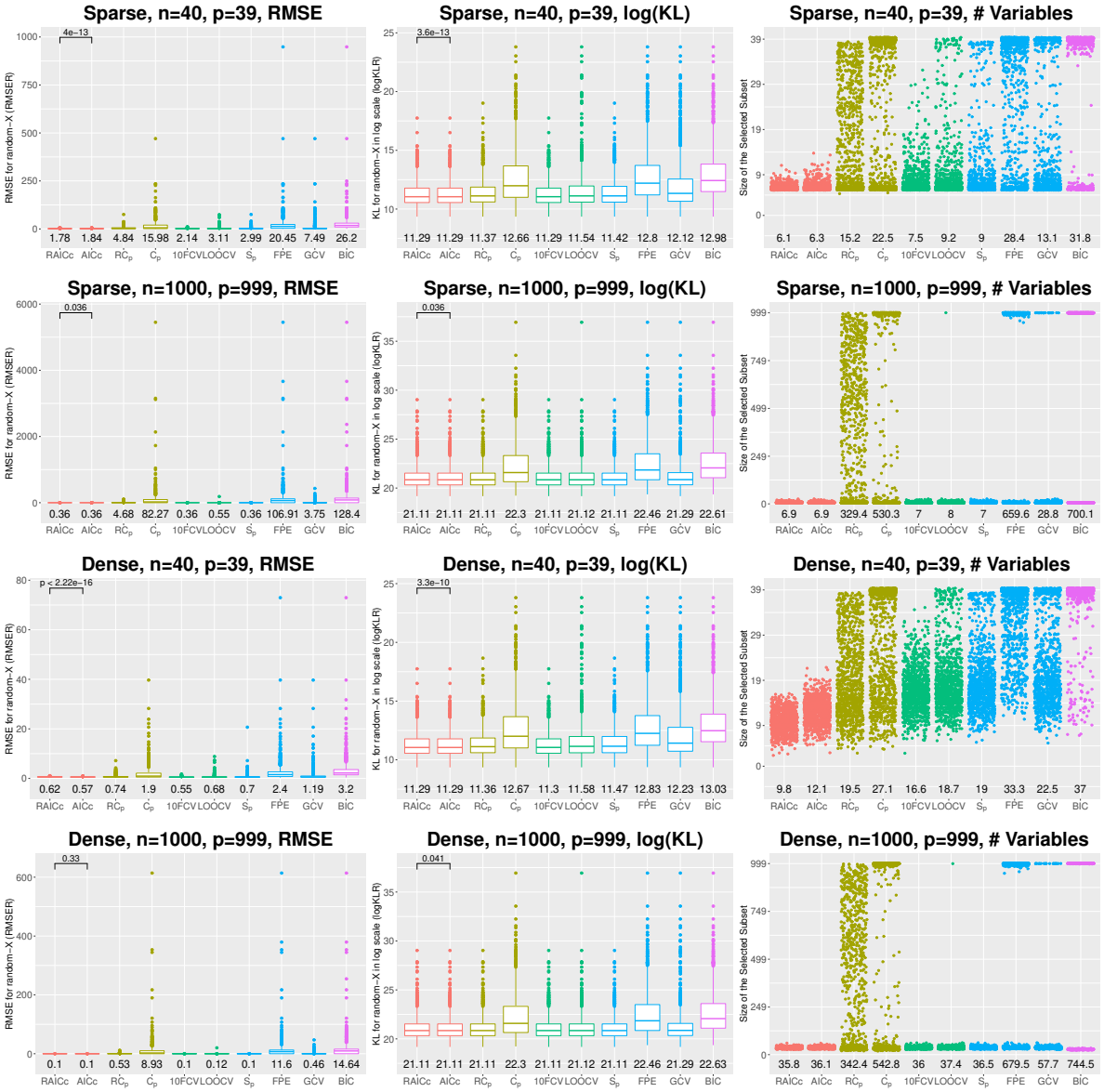

Figure 1: Results of simulations for variable selection. Random-X, high signal and $\rho = 0.5$. The Sparse and Dense models correspond to the VS-Ex2 and VS-Ex3 configurations (details are given in the Online Supplemental Material). The first column refers to RMSE, the second column corresponds to KL discrepancy (in log scale), and the third column gives the number of variables in the selected model with nonzero slopes, jittered horizontally and vertically, so the number of models with that number of nonzero slopes can be ascertained more easily. The mean values of the evaluation metrics for each criterion are presented at the bottom of each graph. The p-values of the Wilcoxon signed-rank test (paired and two-sided) for comparing RAICc and AICc are also presented.

We next consider the general restriction problem. We take $\beta_0 = [2, 2, 2, 1, 1, 1]^T$, $n \in \{10, 40\}$, moderate correlations between the predictors, and either high or low signal levels. The candidate

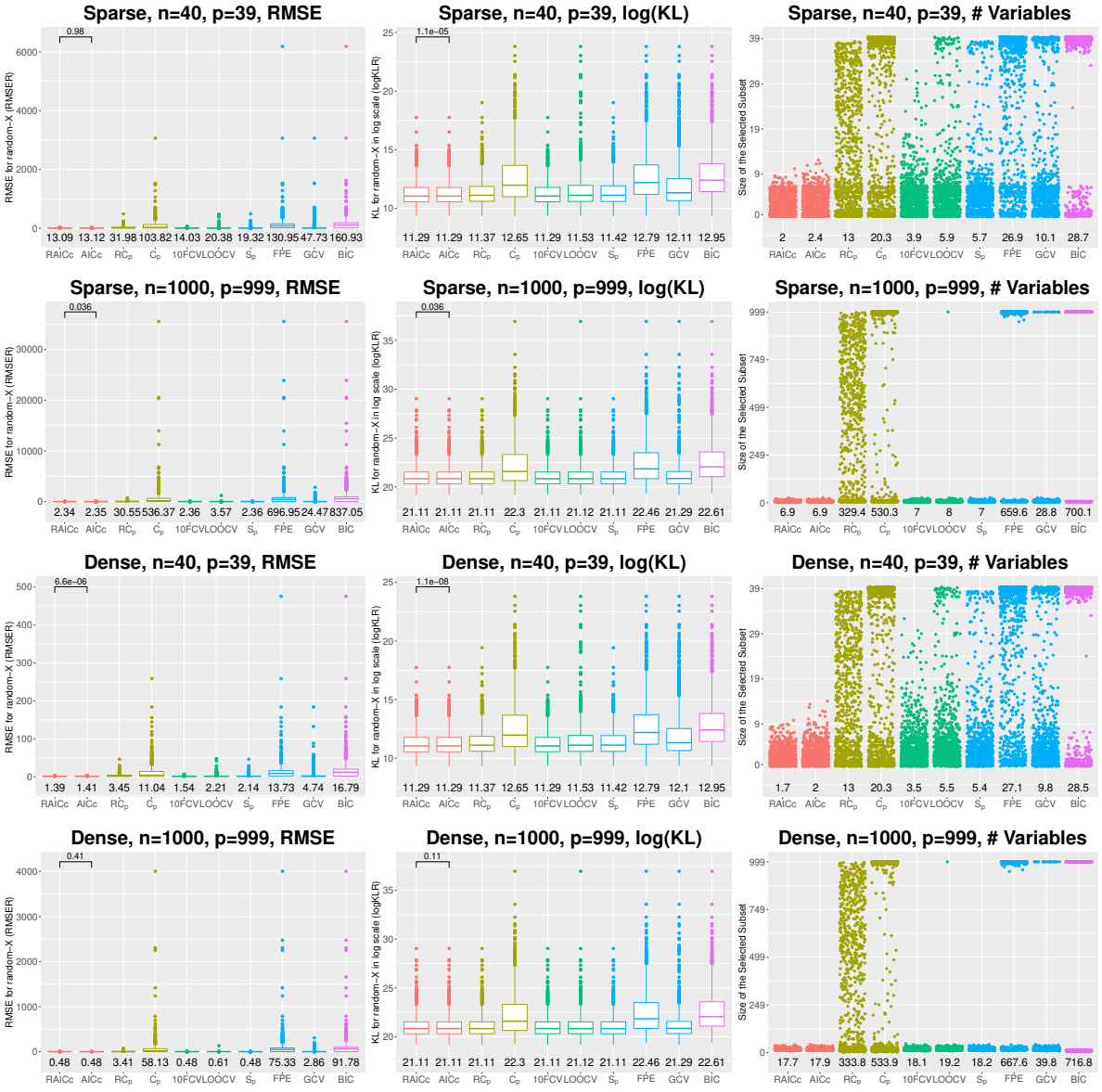

Figure 2: Results of simulations for variable selection. Random-X, low signal and $\rho = 0.5$.

models are constructed in the following way. We consider a set of restrictions: $\beta_1 = \beta_4$, $\beta_1 = 2\beta_2$, $\beta_1 = \beta_2$, $\beta_2 = \beta_3$, $\beta_4 = \beta_5$, $\beta_5 = \beta_6$, where the last four restrictions hold for our choice of β_0. We then consider all of the possible subsets of the six restrictions, resulting in 64 candidate models in total. The detailed configurations and complete results for this and other examples of the general restriction problem (54 scenarios in total) are given in the Online Supplemental Material.

We see from Figure 3 that differences in performance between the criteria are less dramatic. This is not surprising, since for these models the number of parameters never approaches the sample size. Still, RAICc is consistently the best selection rule for small sample size n, and it is second-best for large n, where it is outperformed by BIC (note that BIC has a strong tendency to select too few restrictions when the sample is small, which corresponds to overfitting in the variable selection context). We

also note an advantage of RAICc over AICc, with AICc having a stronger tendency to select too few restrictions.

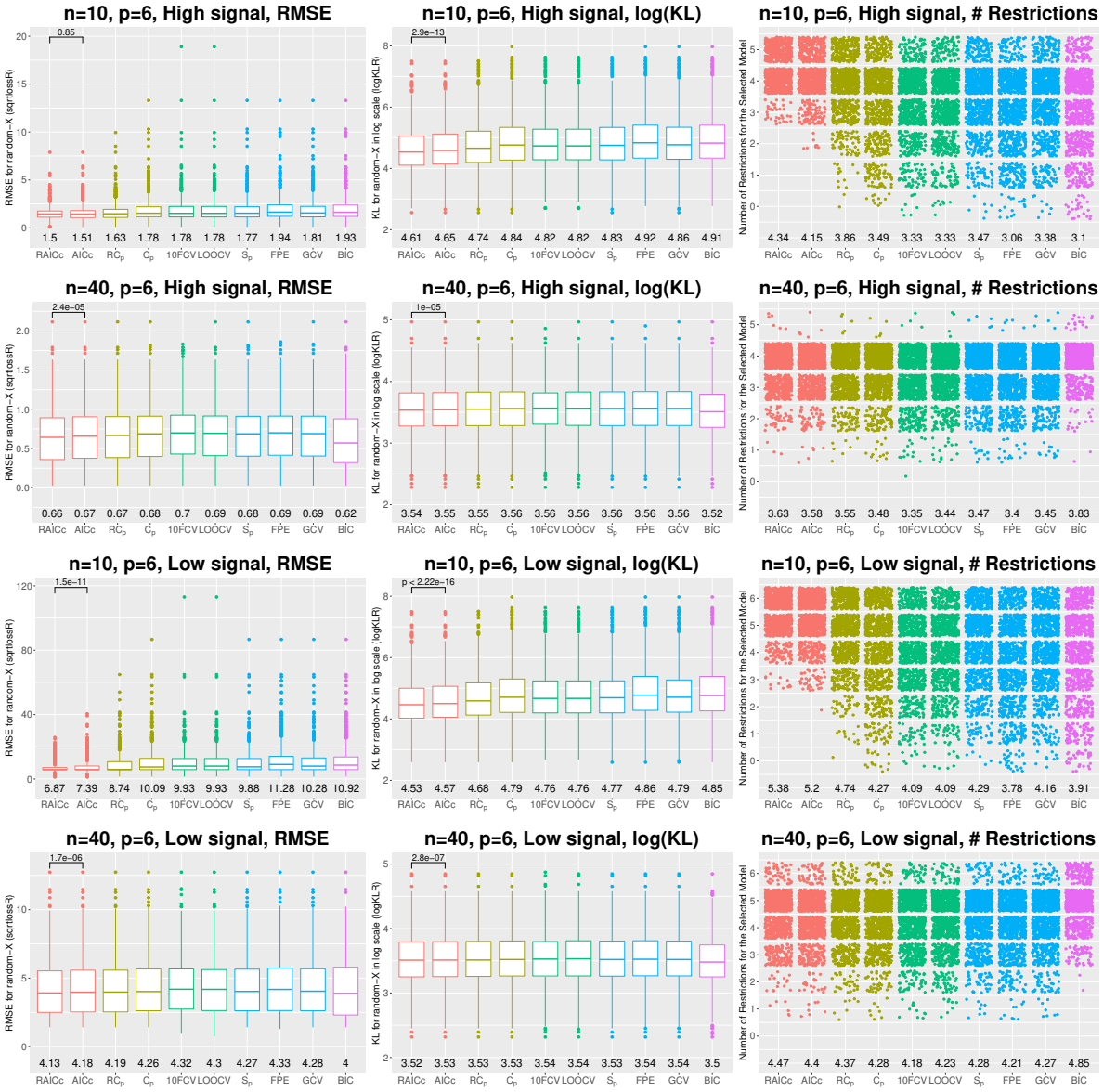

Figure 3: Results of simulations for general restrictions. Random-X, $\rho = 0.5$. The configuration of the model is GR-Ex1 (details can be found in the Online Supplemental Material). Third column gives the number of variables in the number of restrictions in the selected models, jittered horizontally and vertically, so the number of models with that number of imposed restrictions can be ascertained more easily. The mean values of the evaluation metrics for each criterion are presented at the bottom of each graph. The p-values of the Wilcoxon signed-rank test (paired and two-sided) for comparing RAICc and AICc are also presented.

Finally, we extend the general restriction example by including restrictions that force additional predictors to have zero coefficients (as in the variable selection problem). Besides the six restrictions

specified, we also consider $\beta_i = 0$ for $i = 7, \cdots, p$ resulting in p possible restrictions in total. The candidate models are formulated by excluding the restrictions in a nested fashion. We start from the model including all p restrictions (corresponding to the null model), and the next model includes the $p-1$ restrictions except the first one $\beta_1 = \beta_4$. The process is repeated until all restrictions are excluded (the full model including all predictors with arbitrary slopes) resulting in $p+1$ candidate models in total. The true coefficient vector is the same as that used in Figure 3, implying that the correct number of restrictions is $p-2$. We present the detailed configurations and complete results for this and other examples (243 scenarios in total) in the Online Supplemental Material.

We see from Figure 4 that our findings for the variable selection problem also hold in this case. This is not surprising, since variable selection is just a special example of general restrictions, and in this scenario the set of candidate models includes ones where the number of parameters is close to the sample size. Thus, overall, RAICc and AICc are the best performers among all of the selectors. RAICc tends to provide the sparsest subset (or select more restrictions), while rarely underfitting, having a slight advantage over AICc in terms of predictive performance.

4.3 Fixed-X

The simulation structure for random-X can also be applied to fixed-X. We only generate the design matrix X once and draw 1000 replications of the response vector y from the conditional distribution of $y|X$ based on (1). The evaluation metrics for fixed-X are as follows. The complete simulation results are given in the Online Supplemental Material.

- Root mean squared error for fixed-X:

$$\text{RMSEF} = \sqrt{\frac{1}{n}\|X\hat{\beta} - X\beta_0\|_2^2}.$$

- KL discrepancy for fixed-X (8) in the log scale (denoted as logKLF).
- Size of the subset selected.

The patterns for the fixed-X scenario are similar to those for random-X, as can be seen in Figures 5, 6, 7 and 8. In some ways this is surprising, in that the random-X versions of the criteria still seem to outperform the fixed-X versions, even though that is not the scenario for which they are designed. This seems to be related to the tendency for the fixed-X versions to overfit (or choose too few restrictions) compared to their random-X counterparts, which apparently works against the goal of selecting the candidate with best predictive performance. Otherwise, the KL-based criteria (RAICc and AICc) noticeably outperform the other criteria in general, especially C_p and FPE, particularly for small samples.

5 Conclusion and future work

In this paper, the use of information criteria to compare regression models under general linear restrictions for both fixed and random predictors is discussed. It is shown that general versions for KL-based discrepancy (AICc and RAICc, respectively) and squared error-based discrepancy (C_p, FPE, RC_p and S_p, respectively) can be formulated as effectively unbiased estimators of the test error (up to some terms that are free of the linear restrictions and hence are irrelevant when comparing criteria for different models). Model comparison based on the KL-based discrepancy measures is shown via simulations to be better-behaved than squared error-based discrepancies (including cross-validation) in selecting models with low predictive error and sparse subset.

The study of RAICc for variable selection in this paper focuses on OLS fits on pre-fixed predictors (e.g. nested predictors based on their physical orders in X). The discussion can be extended to

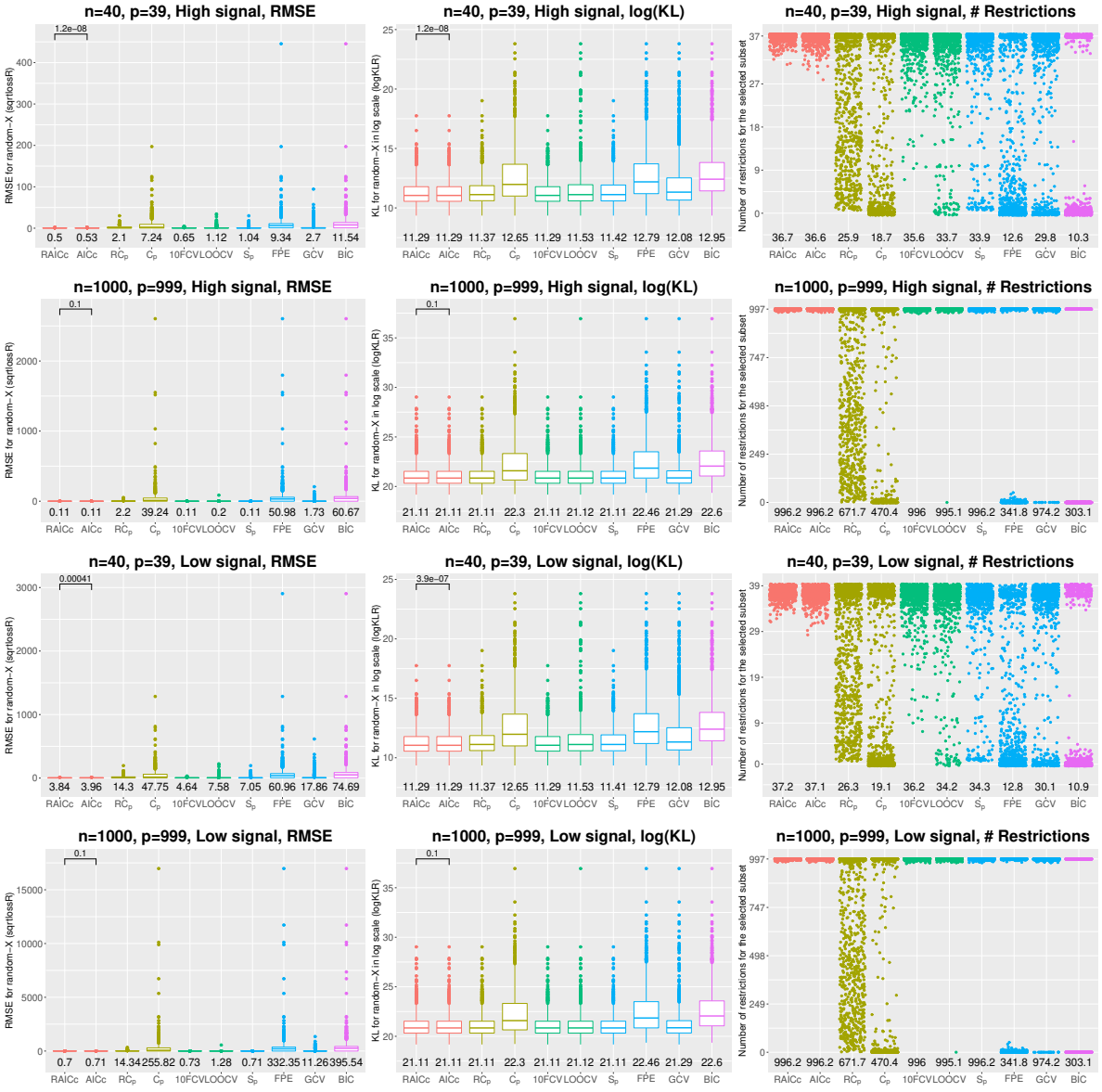

Figure 4: Results of simulations for general restrictions. Random-X, $\rho = 0.5$. The configuration of the model is GR-Ex4 (details can be found in the Online Supplemental Material).

other fitting procedures where the predictors in each subset are decided in a data-dependent way. For instance, Tian et al. (2019) discussed using AICc for least-squares based subset selection methods, and extending those results to the random-X scenario is a topic for future work.

Note also that only restrictions on the regression coefficients are considered here, corresponding to restrictions on the regression portion of the model. It is also possible that the data analyst could be interested in restrictions on the distributional parameters of the predictors (restricting the variances of some predictors to be equal to each other, for example, or restricting covariances to follow a specified pattern such as autoregressive of order 1 or compound symmetry), and it would be interesting to try to generalize the criteria discussed here to that situation.

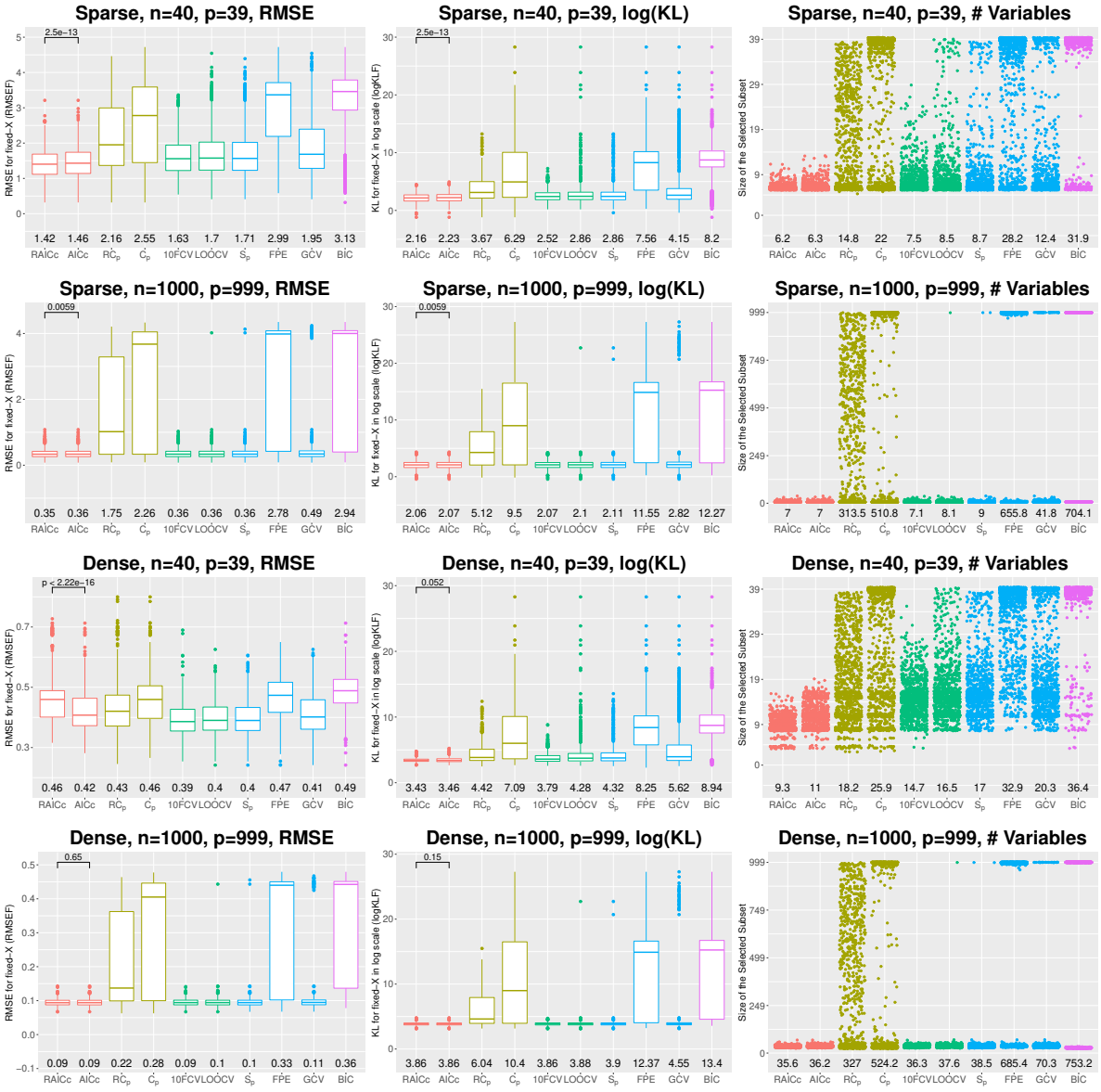

Figure 5: Results of simulations for variable selection. Fixed-X, high signal. The configurations are the same as in Figure 1.

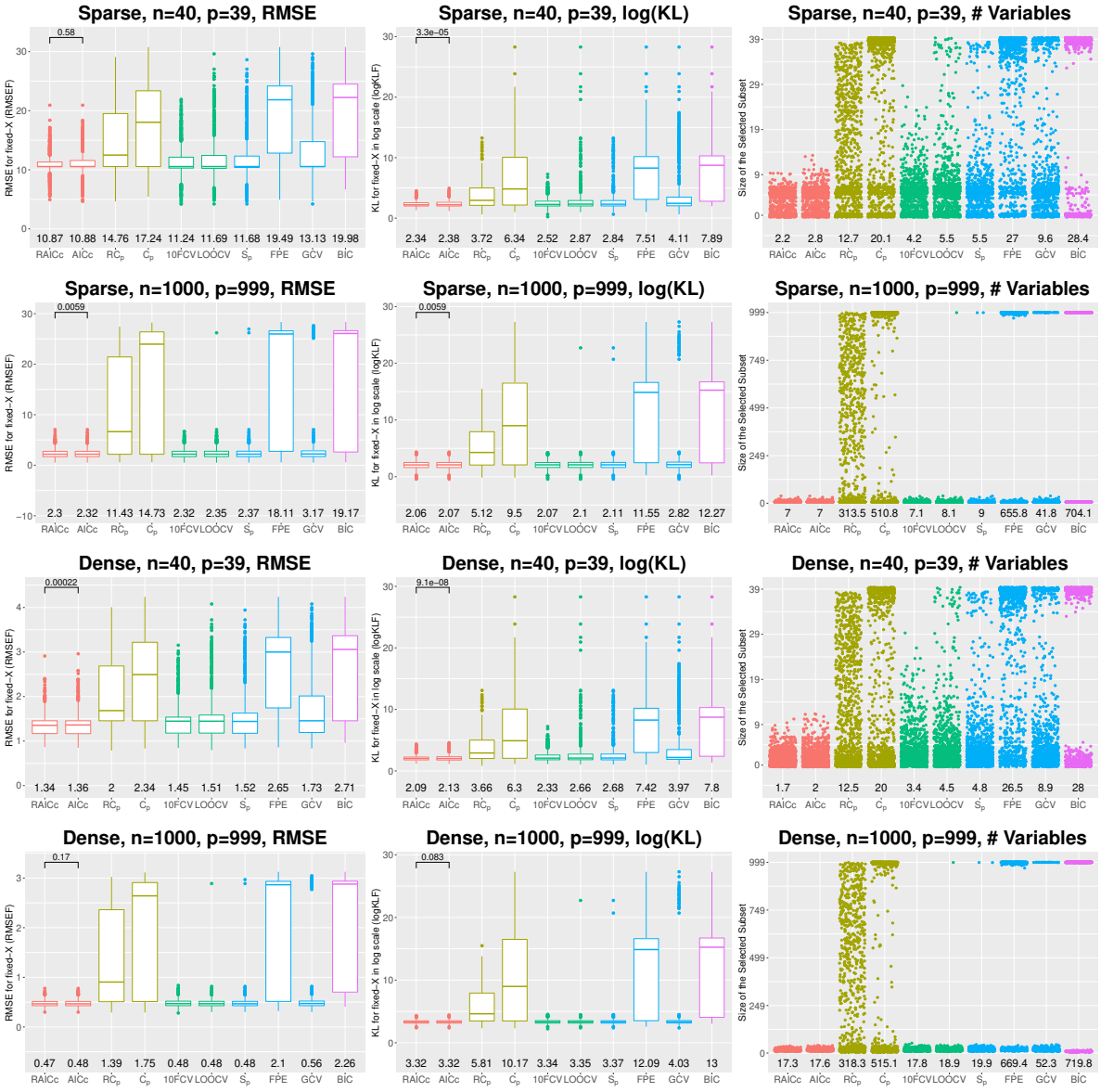

Figure 6: Results of simulations for variable selection. Fixed-X, low signal. The configurations are the same as in Figure 2.

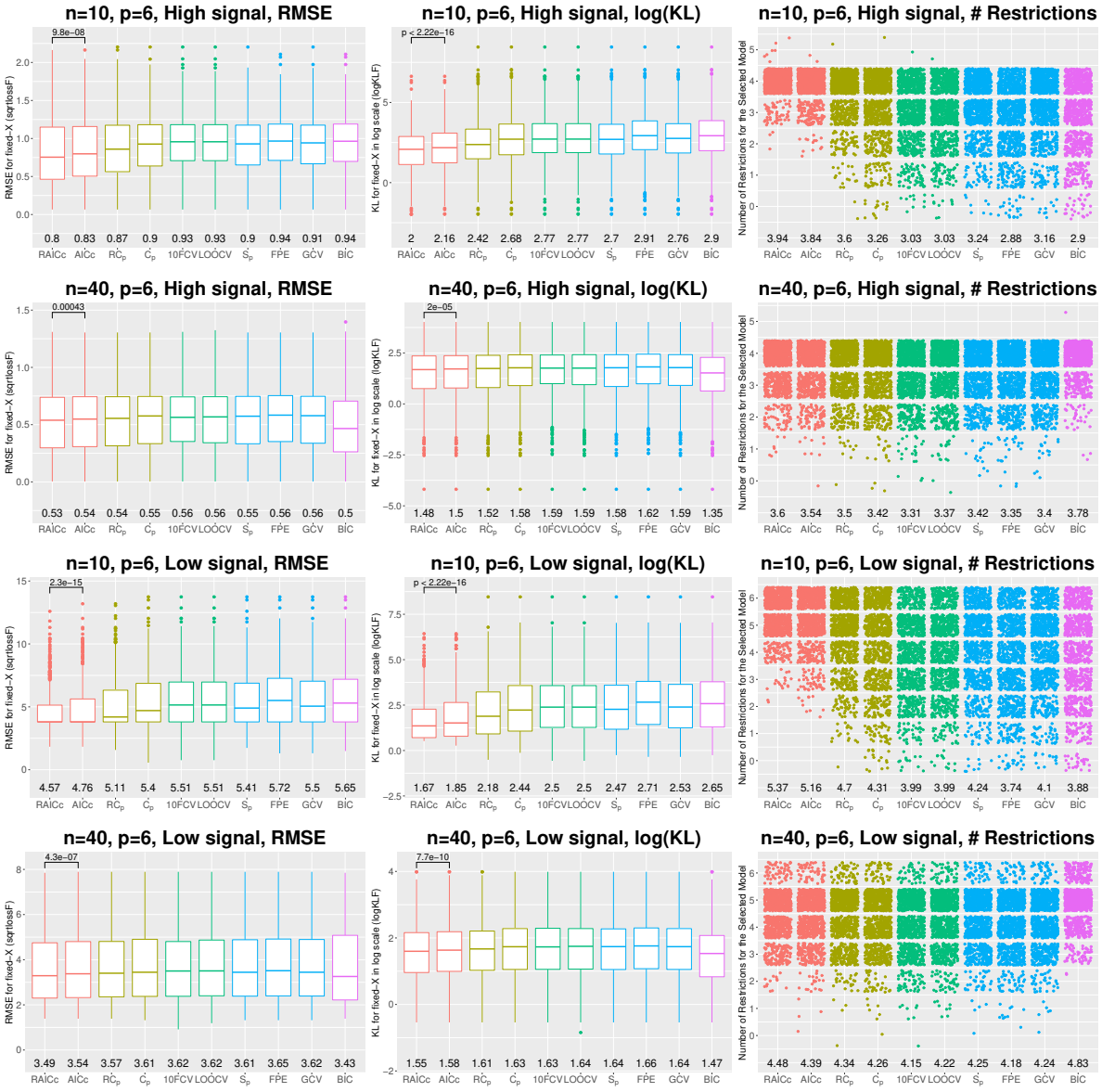

Figure 7: Results of simulations for general restrictions. Fixed-X, GR-Ex1, $\rho = 0.5$. The configurations are the same as in Figure 3.

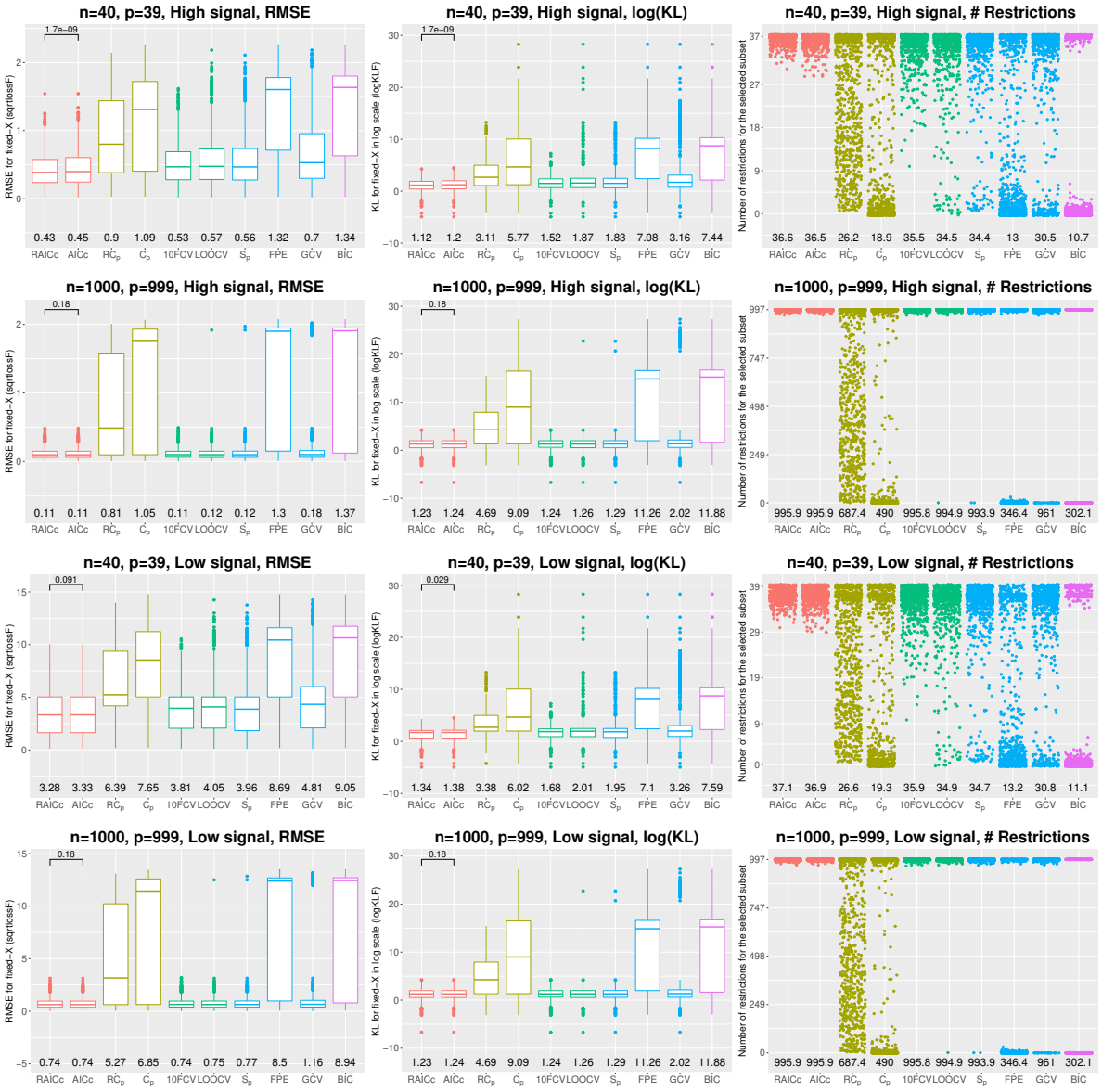

Figure 8: Results of simulations for general restrictions. Fixed-X, GR-Ex4, $\rho = 0.5$. The configurations are the same as in Figure 4.

References

Akaike, H. (1969). Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics 21*, 243–247.

Akaike, H. (1970). Statistical predictor identification. *Annals of the Institute of Statistical Mathematics 22*, 203–217.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. P. F. Csaki (Ed.), *2nd International Symposium on Information Theory*, Budapest, Hungary, 267-281. Akadémiai Kiadó, Budapest.

Allen, D. M. (1974). The relationship between variable selection and data agumentation and a method for prediction. *Technometrics 16*, 125–127.

Baraud, Y., C. Giraud, and S. Huet (2009). Gaussian model selection with an unknown variance. *The Annals of Statistics 37*, 630–672.

Breiman, L. and P. Spector (1992). Submodel selection and evaluation in regression. The X-random case. *International Statistical Review 6*, 291–319.

Burman, P. (1989). A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. *Biometrika 76*, 503–514.

Craven, P. and G. Wahba (1978). Smoothing noisy data with spline functions. *Numerische Mathematik 31*, 377–403.

Efron, B. (1986). How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association 81*, 461–470.

Efron, B. (2004). The estimation of prediction error: covariance penalties and cross-validation. *Journal of the American Statistical Association 99*, 619–632.

Findley, D. F. and E. Parzen (1995). A conversation with Hirotugu Akaike. *Statistical Science 10*, 104–117.

Hall, P. and J. S. Marron (1991). Local minima in cross-validation functions. *Journal of the Royal Statistical Society: Series B (Methodological) 53*, 245–252.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer Science & Business Media.

Hocking, R. R. (1976). The analysis and selection of variables in linear regression. *Biometrics 32*, 1–49.

Hurvich, C. M., J. S. Simonoff, and C.-L. Tsai (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 60*, 271–293.

Hurvich, C. M. and C.-L. Tsai (1989). Regression and time series model selection in small samples. *Biometrika 76*, 297–307.

Hurvich, C. M. and C.-L. Tsai (1991). Bias of the corrected AIC criterion for underfitted regression and time series models. *Biometrika 78*, 499–509.

Konishi, S. and G. Kitagawa (2008). *Information Criteria and Statistical Modeling*. New York: Springer Science & Business Media.

Leeb, H. (2008). Evaluation and selection of models for out-of-sample prediction when the sample size is small relative to the complexity of the data-generating process. *Bernoulli 14*, 661–690.

Linhart, H. and W. Zucchini (1986). *Model Selection*. New York: John Wiley & Sons.

Mallows, C. L. (1973). Some comments on Cp. *Technometrics 15*, 661–675.

Rosset, S. and R. J. Tibshirani (2020). From fixed-X to random-X regression: Bias-variance decompositions, covariance penalties, and prediction error estimation. *Journal of the American Statistical Association 115*, 138–151.

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

Sclove, S. L. (1969). On criteria for choosing a regression equation for prediction. Technical Report No. 28, Department of Statistics, Carnegie-Mellon University.

Scott, D. W. and G. R. Terrell (1987). Biased and unbiased cross-validation in density estimation. *Journal of the American Statistical Association 82*, 1131–1146.

Sugiura, N. (1978). Further analysts of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics-Theory and Methods 7*, 13–26.

Taddy, M. (2017). One-step estimator paths for concave regularization. *Journal of Computational and Graphical Statistics 26*, 525–536.

Tarpey, T. (2000). A note on the prediction sum of squares statistic for restricted least squares. *The American Statistician 54*, 116–118.

Thompson, M. L. (1978a). Selection of variables in multiple regression: Part I. A review and evaluation. *International Statistical Review 46*, 1–19.

Thompson, M. L. (1978b). Selection of variables in multiple regression: Part II. Chosen procedures, computations and examples. *International Statistical Review 46*, 129–146.

Tian, S., C. M. Hurvich, and J. S. Simonoff (2019). On the use of information criteria for subset selection in least squares regression. *arXiv preprint arXiv:1911.10191*.

Tukey, J. W. (1967). Discussion of 'Topics in the investigation of linear relations fitted by the method of least squares' by FJ Anscombe. *J. Roy. Statist. Soc. Ser. B 29*, 47–48.

Ye, J. (1998). On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association 93*, 120–131.

Zhang, Y. and Y. Yang (2015). Cross-validation for selecting a model selection procedure. *Journal of Econometrics 187*, 95–112.

Supplemental Material
Selection of Regression Models under Linear Restrictions for Fixed and Random Designs

Sen Tian, Clifford M. Hurvich, Jeffrey S. Simonoff

This document provides theoretical details of the theorems and lemmas in the paper. The complete simulation results and the computer code to reproduce the results can be viewed online[3].

A Proof of Lemma 1

Proof. As is well known (see, e.g., Greene, 2011, p. 122),

$$n\hat{\sigma}^2 = \|y - X\hat{\beta}\|_2^2 \sim \sigma_0^2\chi^2(n-p+m), \tag{S.1}$$

and by using Assumption 1

$$\begin{aligned} E_y\left(\hat{\beta}-\beta_0\right) &= 0, \\ \mathrm{Cov}_y\left(\hat{\beta}-\beta_0\right) &= E\left[\left(\hat{\beta}-\beta_0\right)\left(\hat{\beta}-\beta_0\right)^T\right] \\ &= \sigma_0^2\left\{(X^TX)^{-1} - (X^TX)^{-1}R^T\left[R(X^TX)^{-1}R^T\right]^{-1}R(X^TX)^{-1}\right\}. \end{aligned} \tag{S.2}$$

From (S.1), $1/\hat{\sigma}^2$ follows an inverse χ^2 distribution and we have

$$n\sigma_0^2 E_y\left[\frac{1}{\hat{\sigma}^2}\right] = n\frac{n}{n-p+m-2}.$$

From (S.2), we have

$$E_y\left[(\hat{\beta}-\beta_0)^TX^TX(\hat{\beta}-\beta_0)\right] = \mathrm{Tr}\left[X^TX\cdot\mathrm{Cov}_y\left(\hat{\beta}-\beta_0\right)\right] = \sigma_0^2(p-m).$$

We next show that $\hat{\sigma}^2$ and $(\hat{\beta}-\beta_0)^TX^TX(\hat{\beta}-\beta_0)$ are independent. Define the idempotent matrix $H_R = (X^TX)^{-1}R^T\left[R(X^TX)^{-1}R^T\right]^{-1}R$. Recall that two other idempotent matrices are defined as $H = X(X^TX)^{-1}X^T$ and $H_Q = XH_R(X^TX)^{-1}X^T$, respectively. We have

$$\begin{aligned} y - X\hat{\beta}^f &= (I-H)\epsilon, \\ X\hat{\beta}^f - X\hat{\beta} &= XH_R(\hat{\beta}^f-\beta_0) = H_Q\epsilon, \\ X\hat{\beta} - X\beta_0 &= X(I-H_R)(\hat{\beta}^f-\beta_0) = (H-H_Q)\epsilon, \end{aligned}$$

where we use the fact that $\hat{\beta}^f - \beta_0 = (X^TX)^{-1}X^T\epsilon$. Also since $HH_Q = H_QH = H_Q$, any two of the three idempotent symmetric matrices $I-H$, H_Q and $H-H_Q$ have product zero. Then by Craig's Theorem (Craig, 1943) on the independence of two quadratic forms in a normal vector,

$$n\hat{\sigma}^2 = \|y-X\hat{\beta}\|_2^2 = \|y-X\hat{\beta}^f\|_2^2 + \|X\hat{\beta}^f - X\hat{\beta}\|_2^2 = \epsilon^T(I-H)\epsilon + \epsilon^TH_Q\epsilon$$

and

$$\|X\hat{\beta} - X\beta_0\|_2^2 = \epsilon^T(H-H_Q)\epsilon$$

are independent. □

[3] `https://github.com/sentian/RAICc`

B Proof of Theorem 1

Proof. By using Lemma 1, the expected KL discrepancy can be derived as

$$\begin{aligned} E_y[\mathrm{ErrF_{KL}}] &= E_y\left\{ n\log(2\pi\hat{\sigma}^2) + \frac{1}{\hat{\sigma}^2}(\hat{\beta}-\beta_0)^T X^T X(\hat{\beta}-\beta_0) + \frac{n\sigma_0^2}{\hat{\sigma}^2} \right\} \\ &= E_y\left[n\log(2\pi\hat{\sigma}^2) \right] + (p-m)\frac{n}{n-p+m-2} + n\frac{n}{n-p+m-2} \\ &= E_y\left[n\log(2\pi\hat{\sigma}^2) \right] + n\frac{n+p-m}{n-p+m-2}. \end{aligned}$$

Recall that

$$\mathrm{errF_{KL}} = n\log(2\pi\hat{\sigma}^2) + n.$$

The expected optimism is then

$$E_y(\mathrm{opF_{KL}}) = E_y[\mathrm{ErrF_{KL}}] - E_y[\mathrm{errF_{KL}}] = n\frac{n+p-m}{n-p+m-2} - n.$$

□

C Proof of Theorem 2

Proof. Using the expression of $\hat{\beta}$ (4) and the definitions of H and H_Q, we have

$$\hat{\mu} = X\hat{\beta} = (H - H_Q)y + X(X^TX)^{-1}R^T\left[R(X^TX)^{-1}R^T\right]^{-1} r,$$

where the second term on the right-hand side is deterministic. Denote h_i and h_{Q_i} as the i-th rows of H and H_Q, respectively. We then have

$$\mathrm{Cov}_y\left(\hat{\mu}_i, y_i\right) = \mathrm{Cov}_y\left[(h_i - h_{Q_i})y, y_i\right] = \mathrm{Cov}_y\left[(H_{ii} - H_{Q_{ii}})y_i, y_i\right] = \sigma_0^2(H_{ii} - H_{Q_{ii}}).$$

Therefore, the covariance penalty (5) can be derived as

$$E_y(\mathrm{optF_{SE}}) = 2\sum_{i=1}^{n}\mathrm{Cov}_y(\hat{\mu}_i, y_i) = 2\sigma_0^2\mathrm{Tr}(H - H_Q) = 2\sigma_0^2(p-m).$$

□

D Proof of Lemma 2

Proof. Since x_i are iid $\mathcal{N}(0, \Sigma_0)$, $X^TX \sim \mathcal{W}(\Sigma_0, n)$ and $(X^TX)^{-1} \sim \mathcal{W}^{-1}(\Sigma_0^{-1}, n)$, where $\mathcal{W}$ and $\mathcal{W}^{-1}$ denotes a Wishart and an inverse Wishart distribution with n degrees of freedom, respectively. We have $E(X^TX) = n\Sigma_0$ and $E((X^TX)^{-1}) = \Sigma_0^{-1}/(n-p-1)$. Hence,

$$E\left[\mathrm{Tr}(\hat{\Sigma}^{-1}\Sigma_0)\right] = E\left[\mathrm{Tr}(n(X^TX)^{-1}\Sigma_0)\right] = n\mathrm{Tr}\left[E\left((X^TX)^{-1}\right)\Sigma_0\right] = \frac{np}{n-p-1}.$$

Define $H_S = X(X^TX)^{-1}(I-H_R)^T\Sigma_0(I-H_R)(X^TX)^{-1}X^T$. Conditionally on X, the random variable $\hat{\sigma}^2$ and

$$(\hat{\beta}-\beta_0)^T\Sigma_0(\hat{\beta}-\beta_0) = \epsilon^T H_S \epsilon$$

are independent by Craig's Theorem, since H_S is symmetric and $H_S(I - H + H_Q) = 0$. In order to calculate $E_{X,y}\left[(\hat{\beta} - \beta_0)^T \Sigma_0 (\hat{\beta} - \beta_0)\right]$, we transform the original basis of the problem. Denote $\tilde{R} = \binom{R}{R^c}$, a $(p \times p)$ matrix, where the rows of R^c span the orthogonal complement of the row space of R. Hence $\tilde{R}$ has full rank. The true model now becomes

$$y = X\beta_0 + \epsilon = \tilde{X}\tilde{\beta}_0 + \epsilon,$$

where $\tilde{X} = X\tilde{R}^T$, $\tilde{\beta}_0 = \tilde{R}^{T^{-1}}\beta_0$. Denote $\tilde{M} = R\tilde{R}^T$. Assumption 1 indicates that the true model in the new basis satisfies $\tilde{M}\tilde{\beta}_0 = r$. The approximating model is

$$y = X\beta + u = \tilde{X}\tilde{\beta} + u,$$

with restrictions $\tilde{M}\tilde{\beta} = r$ where $\tilde{\beta} = \tilde{R}^{T^{-1}}\beta$. Denote $\hat{\tilde{\beta}}^f = (\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T y$ as the OLS estimator in the regression of y on $\tilde{X}$. The restricted MLE is then

$$\hat{\tilde{\beta}} = \hat{\tilde{\beta}}^f - \left(\tilde{X}^T\tilde{X}\right)^{-1}\tilde{M}^T\left[\tilde{M}(\tilde{X}^T\tilde{X})^{-1}\tilde{M}^T\right]^{-1}\left(\tilde{M}\hat{\tilde{\beta}}^f - r\right),$$

and it can be easily verified that $\hat{\tilde{\beta}} = \tilde{R}^{T^{-1}}\hat{\beta}$. Denote $\tilde{X}_m$ and $\tilde{X}_{p-m}$ as the matrices containing the first m and last $p-m$ columns of $\tilde{X}$, respectively. Let $\hat{\tilde{\beta}}_m$ and $\hat{\tilde{\beta}}_{p-m}$ be column vectors consisting of the first m and last $p-m$ entries in $\hat{\tilde{\beta}}$, respectively. Also let $\tilde{\beta}_{0,m}$ and $\tilde{\beta}_{0,p-m}$ be column vectors consisting of the first m and last $p-m$ entries in $\tilde{\beta}_0$, respectively. By using the formula for the inverse of partitioned matrices and some algebra, it can be shown that (details are given in Supplemental Material Section G)

$$\begin{aligned}\hat{\tilde{\beta}}_m &= \tilde{r},\\ \hat{\tilde{\beta}}_{p-m} &= \left(\tilde{X}_{p-m}^T\tilde{X}_{p-m}\right)^{-1}\left(\tilde{X}_{p-m}^T y - \tilde{X}_{p-m}^T\tilde{X}_m\tilde{r}\right),\end{aligned} \tag{S.3}$$

where $\tilde{r} = (RR^T)^{-1}r$. The restrictions $\tilde{M}\tilde{\beta}_0 = r$ results in $\tilde{\beta}_{0,m} = \tilde{r}$. We then have

$$\hat{\tilde{\beta}}_{p-m} - \tilde{\beta}_{0,p-m} = \left(\tilde{X}_{p-m}^T\tilde{X}_{p-m}\right)^{-1}\tilde{X}_{p-m}^T\epsilon,$$

and therefore

$$\hat{\tilde{\beta}}_{p-m} - \tilde{\beta}_{0,p-m}|\tilde{X} \sim \mathcal{N}\left(0, \sigma_0^2\left(\tilde{X}_{p-m}^T\tilde{X}_{p-m}\right)^{-1}\right).$$

We also note that $\tilde{X}_{p-m} = XR^{cT}$, and hence the rows $\tilde{x}_{p-m,i}$ of $\tilde{X}_{p-m}$, are independent and satisfy $\tilde{x}_{p-m,i} \sim \mathcal{N}\left(0, R^c\Sigma_0 R^{cT}\right)$. We then have that $\left(\tilde{X}_{p-m}^T\tilde{X}_{p-m}\right)^{-1}$ follows the inverse Wishart distribution $W^{-1}(R^c\Sigma_0 R^{cT}, n)$. The expectation of the quadratic form can be derived as

$$\begin{aligned}E_{X,y}\left[(\hat{\beta} - \beta_0)^T\Sigma_0(\hat{\beta} - \beta_0)\right] &= E_{\tilde{X},y}\left[\left(\hat{\tilde{\beta}} - \tilde{\beta}_0\right)^T\tilde{R}\Sigma_0\tilde{R}^T\left(\hat{\tilde{\beta}} - \tilde{\beta}_0\right)\right]\\ &= E_{\tilde{X}}\left\{E\left[\left(\hat{\tilde{\beta}}_{p-m} - \tilde{\beta}_{0,p-m}\right)^T R^c\Sigma_0 R^{cT}\left(\hat{\tilde{\beta}}_{p-m} - \tilde{\beta}_{0,p-m}\right)\Big|\tilde{X}\right]\right\}\\ &= \sigma_0^2\mathrm{Tr}\left\{R^c\Sigma_0 R^{cT}E\left[\left(\tilde{X}_{p-m}^T\tilde{X}_{p-m}\right)^{-1}\right]\right\}\\ &= \sigma_0^2\frac{p-m}{n-p+m-1}.\end{aligned}$$

□

E Proof of Theorem 3

Proof. The expected KL can be derived as

$$
\begin{aligned}
&E_{X,y}(\mathrm{ErrR_{KL}}) \\
&= E_{X,y}\left[n\log(2\pi\hat{\sigma}^2) + \frac{n}{\hat{\sigma}^2}(\hat{\beta}-\beta_0)^T\Sigma_0(\hat{\beta}-\beta_0) + \frac{n\sigma_0^2}{\hat{\sigma}^2}\right] + E\left[np\log(2\pi) + n\log|\hat{\Sigma}| + n\mathrm{Tr}(\hat{\Sigma}^{-1}\Sigma_0)\right] \\
&= E_{X,y}\left[n\log(2\pi\hat{\sigma}^2)\right] + E_X\left[E\left(\frac{n}{\hat{\sigma}^2}|X\right)E\left((\hat{\beta}-\beta_0)^T\Sigma_0(\hat{\beta}-\beta_0)|X\right) + E\left(\frac{n\sigma_0^2}{\hat{\sigma}^2}|X\right)\right] + \\
&\quad E\left[np\log(2\pi) + n\log|\hat{\Sigma}| + n\mathrm{Tr}(\hat{\Sigma}^{-1}\Sigma_0)\right] \\
&= E_{X,y}\left[n\log(2\pi\hat{\sigma}^2)\right] + n\frac{n(n-1)}{(n-p+m-2)(n-p+m-1)} + E\left[n\log|\hat{\Sigma}|\right] + np\log(2\pi) + n\frac{np}{n-p-1},
\end{aligned}
$$

where the second equality is based on Lemma 2 for the independence of $\hat{\sigma}^2$ and $(\hat{\beta}-\beta_0)^T\Sigma_0(\hat{\beta}-\beta_0)$ conditionally on X, and in the last equality we use results from Lemmas 1 and 2. Since the training error is

$$
\mathrm{errR_{KL}} = -2\log L(\hat{\beta},\hat{\sigma}^2,\hat{\Sigma}|X,y) = \left[n\log(2\pi\hat{\sigma}^2) + n\right] + \left[np\log(2\pi) + n\log|\hat{\Sigma}| + np\right],
$$

the expected optimism can be obtained as

$$
\begin{aligned}
E_{X,y}(\mathrm{optR_{KL}}) &= E_{X,y}[\mathrm{ErrR_{KL}}] - E_{X,y}[\mathrm{errR_{KL}}] \\
&= n\frac{n(n-1)}{(n-p+m-2)(n-p+m-1)} + n\frac{np}{n-p-1} - n(p+1).
\end{aligned}
$$

□

F Proof of Theorem 4

Proof. We first note from Theorem 2 that

$$
E_{X,y}(\mathrm{optF_{SE}}) = E\left[E(\mathrm{optF_{SE}}|X)\right] = 2\sigma_0^2(p-m).
$$

Based on formula 6 and proposition 1 in Rosset and Tibshirani (2020), the expected optimism can be decomposed into

$$
E_{X,y}(\mathrm{optR_{SE}}) = E_{X,y}(\mathrm{optF_{SE}}) + B^+ + V^+ = 2\sigma_0^2(p-m) + B^+ + V^+,
$$

where B^+ and V^+ are the excess bias and excess variance of the fit. In particular, the excess bias is defined as

$$
B^+ = E_{X,X^{(n)}}\left\|E(X^{(n)}\hat{\beta}|X,X^{(n)}) - X^{(n)}\beta_0\right\|_2^2 - E_X\left\|E(X\hat{\beta}|X) - X\beta_0\right\|_2^2.
$$

Because of our assumption that the true model satisfies the restrictions, it follows that $\hat{\beta}$ is unbiased, and hence it is easy to see that $B^+ = 0$. Next, V^+ is defined as

$$
V^+ = E_{X,X^{(n)}}\left\{\mathrm{Tr}\left[\mathrm{Cov}\left(X^{(n)}\hat{\beta}|X,X^{(n)}\right)\right]\right\} - E_X\left\{\mathrm{Tr}\left[\mathrm{Cov}\left(X\hat{\beta}|X\right)\right]\right\}.
$$

The second term on the right-hand side is

$$
E_X\left\{\mathrm{Tr}\left[\mathrm{Cov}\left(X\hat{\beta}|X\right)\right]\right\} = E_X\left\{\mathrm{Tr}\left[\mathrm{Cov}\left((H-H_Q)y|X\right)\right]\right\} = E\left\{\sigma_0^2\mathrm{Tr}\left(H-H_Q\right)\right\} = \sigma_0^2(p-m).
$$

The first term on the right-hand side is

$$\begin{aligned}
&E_{X,X^{(n)}}\mathrm{Tr}\left[\mathrm{Cov}\left(X^{(n)}\hat{\beta}|X,X^{(n)}\right)\right]\\
&=E_{X,X^{(n)}}\mathrm{Tr}\left[\mathrm{Cov}\left(X^{(n)}(\hat{\beta}-\beta_0)|X,X^{(n)}\right)\right]\\
&=\mathrm{Tr}\left\{E_X\left[\mathrm{Cov}\left(\hat{\beta}-\beta_0|X\right)\right]E\left({X^{(n)}}^TX^{(n)}\right)\right\}\\
&=nE_X\left\{\mathrm{Tr}\left[\Sigma_0\mathrm{Cov}\left(\hat{\beta}-\beta_0|X\right)\right]\right\}\\
&=nE_X\left\{E\left[\left(\hat{\beta}-\beta_0\right)^T\Sigma_0\left(\hat{\beta}-\beta_0\right)|X\right]\right\}\\
&=nE_{X,y}\left[\left(\hat{\beta}-\beta_0\right)^T\Sigma_0\left(\hat{\beta}-\beta_0\right)\right]\\
&=n\sigma_0^2\frac{p-m}{n-p+m-1},
\end{aligned}$$

where in the third equality we use the independence and identical distribution of X and X_0, and $E(X^TX)=n\Sigma_0$, while in the last equality we use the result in Lemma 2. Combining the results together, we have

$$V^+=n\sigma_0^2\frac{p-m}{n-p+m-1}-\sigma_0^2(p-m)=\sigma_0^2\frac{(p-m)(p-m+1)}{n-p+m-1},$$

and

$$E_{X,y}(\mathrm{optR_{SE}})=2\sigma_0^2(p-m)+\sigma_0^2\frac{(p-m)(p-m+1)}{n-p+m-1}.$$

□

G Derivation of the expression of $\hat{\tilde{\beta}}$ in (S.3)

Denote $\tilde{H}_m=\tilde{X}_m\left(\tilde{X}_m^T\tilde{X}_m\right)^{-1}\tilde{X}_m^T$ and $\tilde{H}_{p-m}=\tilde{X}_{p-m}\left(\tilde{X}_{p-m}^T\tilde{X}_{p-m}\right)^{-1}\tilde{X}_{p-m}^T$. Then the partitioned form of the matrix $\tilde{X}^T\tilde{X}$ is given by

$$\begin{aligned}
\left(\tilde{X}^T\tilde{X}\right)^{-1}&=\begin{bmatrix}\tilde{X}_m^T\tilde{X}_m & \tilde{X}_m^T\tilde{X}_{p-m}\\ \tilde{X}_{p-m}^T\tilde{X}_m & \tilde{X}_{p-m}^T\tilde{X}_{p-m}\end{bmatrix}^{-1}\\
&=\begin{bmatrix}\left[\tilde{X}_m^T(I-\tilde{H}_{p-m})\tilde{X}_m\right]^{-1} & -\left[\tilde{X}_m^T(I-\tilde{H}_{p-m})\tilde{X}_m\right]^{-1}\tilde{X}_m^T\tilde{X}_{p-m}\left(\tilde{X}_{p-m}^T\tilde{X}_{p-m}\right)^{-1}\\ -\left[\tilde{X}_{p-m}^T\left(I-\tilde{H}_m\right)\tilde{X}_{p-m}\right]^{-1}\tilde{X}_{p-m}^T\tilde{X}_m\left(\tilde{X}_m^T\tilde{X}_m\right)^{-1} & \left[\tilde{X}_{p-m}^T\left(I-\tilde{H}_m\right)\tilde{X}_{p-m}\right]^{-1}\end{bmatrix},
\end{aligned}$$

and the partitioned form of $\hat{\tilde{\beta}}^f$ is given by

$$\begin{aligned}
\hat{\tilde{\beta}}^f&=\left(\tilde{X}^T\tilde{X}\right)^{-1}\tilde{X}^Ty\\
&=\begin{bmatrix}\left[\tilde{X}_m^T\left(I-\tilde{H}_{p-m}\right)\tilde{X}_m\right]^{-1}\left(\tilde{X}_m^Ty-\tilde{X}_m^T\tilde{H}_{p-m}y\right)\\ \left[\tilde{X}_{p-m}^T\left(I-\tilde{H}_m\right)\tilde{X}_{p-m}\right]^{-1}\left(\tilde{X}_{p-m}^Ty-\tilde{X}_{p-m}^T\tilde{H}_my\right)\end{bmatrix}.
\end{aligned}$$

We also have

$$\begin{aligned}
&\left[\tilde{X}_{p-m}^T\left(I-\tilde{H}_m\right)\tilde{X}_{p-m}\right]^{-1}\tilde{X}_{p-m}^T\tilde{H}_m\left(I_n-\tilde{H}_{p-m}\right)\tilde{X}_m \\
&=\left[\tilde{X}_{p-m}^T\left(I-\tilde{H}_m\right)\tilde{X}_{p-m}\right]^{-1}\left(\tilde{X}_{p-m}^T\tilde{X}_m-\tilde{X}_{p-m}^T\tilde{H}_m\tilde{H}_{p-m}\tilde{X}_m\right) \\
&=\left[\tilde{X}_{p-m}^T\left(I-\tilde{H}_m\right)\tilde{X}_{p-m}\right]^{-1}\tilde{X}_{p-m}^T\left(I-\tilde{H}_m\right)\tilde{X}_{p-m}\left(\tilde{X}_{p-m}^T\tilde{X}_{p-m}\right)^{-1}\tilde{X}_{p-m}^T\tilde{X}_m \\
&=\left(\tilde{X}_{p-m}^T\tilde{X}_{p-m}\right)^{-1}\tilde{X}_{p-m}^T\tilde{X}_m.
\end{aligned}$$

Using this property and $\tilde{M}=R\tilde{R}^T=\begin{pmatrix} RR^T & 0\end{pmatrix}$, we have

$$\left(\tilde{X}^T\tilde{X}\right)^{-1}\tilde{M}^T\left[\tilde{M}\left(\tilde{X}^T\tilde{X}\right)^{-1}\tilde{M}^T\right]^{-1}=\begin{bmatrix}(RR^T)^{-1}\\ -\left(\tilde{X}_{p-m}^T\tilde{X}_{p-m}\right)^{-1}\tilde{X}_{p-m}^T\tilde{X}_m(RR^T)^{-1}\end{bmatrix},$$

$$I_p-\left(\tilde{X}^T\tilde{X}\right)^{-1}\tilde{M}^T\left[\tilde{M}\left(\tilde{X}^T\tilde{X}\right)^{-1}\tilde{M}^T\right]^{-1}\tilde{M}=\begin{bmatrix}0 & 0\\ \left(\tilde{X}_{p-m}^T\tilde{X}_{p-m}\right)^{-1}\tilde{X}_{p-m}^T\tilde{X}_m & I_{p-m}\end{bmatrix}.$$

Therefore, (S.3) can be derived as

$$\begin{aligned}
\hat{\tilde{\beta}}&=\left\{I_p-\left(\tilde{X}^T\tilde{X}\right)^{-1}\tilde{M}^T\left[\tilde{M}\left(\tilde{X}^T\tilde{X}\right)^{-1}\tilde{M}^T\right]^{-1}\tilde{M}\right\}\hat{\tilde{\beta}}^f+\left\{\left(\tilde{X}^T\tilde{X}\right)^{-1}\tilde{M}^T\left[\tilde{M}\left(\tilde{X}^T\tilde{X}\right)^{-1}\tilde{M}^T\right]^{-1}\right\}r \\
&=\begin{bmatrix}\tilde{r}\\ \left(\tilde{X}_{p-m}^T\tilde{X}_{p-m}\right)^{-1}\tilde{X}_{p-m}^T\left(y-\tilde{X}_m\tilde{r}\right)\end{bmatrix}.
\end{aligned}$$

References

Craig, A. T. (1943). Note on the independence of certain quadratic forms. *The Annals of Mathematical Statistics 14*, 195–197.

Greene, W. H. (2011). *Econometric Analysis* (Seventh ed.). Upper Saddle River, NJ: Prentice Hall.

Rosset, S. and R. J. Tibshirani (2020). From fixed-X to random-X regression: Bias-variance decompositions, covariance penalties, and prediction error estimation. *Journal of the American Statistical Association 115*, 138–151.