IMPERIAL

Imperial College London
Department of Mathematics

Selective Inference: The effect of estimating the variance of the error term in linear regression model

Zichun Wang

CID: 02405051

Supervised by Alastair Young

24 August 2024

Submitted in partial fulfilment of the requirements for the MSc in Statistics at Imperial College London

The work contained in this thesis is my own work unless otherwise stated.

Signed: Zichun Wang Date: 24 August 2024

Acknowledgements

I would like to thank my supervisor, Professor Alastair Young, for his continual guidance and encouragement throughout this process. I also want to express my gratitude to everyone who has generously supported me during my master's journey.

Abstract

We present the assumptions of the classical inferential theory and highlight the fact that these assumptions are often difficult to satisfy in practice. Selective inference is proposed to provide valid inferences after variable selection. One of the three main approaches to selective inference is data splitting. Rasines and Young (2023) introduced a new method called the (U, V) decomposition as an improvement to the data splitting technique. They were the first to suggest generating two independent components, U and V, from the dataset Y, with U used for variable selection and V used for making inferences for the selected variables. We aim to make inferences for the coefficients in the linear regression model using the (U, V) decomposition. One critical parameter is the variance of the error term, σ^2. If σ^2 is known, U and V are indeed independent of each other. However, when σ^2 is unknown, using its estimator introduces dependence between U and V. We are interested in finding what exactly happens to the inferential results if the estimated σ^2 is used.

Suppose n is the sample size and p is the dimensionality of the true parameter β. The method for estimating σ^2 varies depending on the relationship between n and p, which leads us to divide our study into two main sections. In the low-dimensional setting ($n > p$), the residual sum of squares is used to estimate σ^2, while in the high-dimensional setting ($n < p$), we use the refitted cross-validation estimator. Our further evaluations contain the effects of dimensionality and the sparsity level of the true parameter vector at the inference stage using V.

After lots of large simulations, we find that in the low-dimensional setting, the good performance of an estimated σ^2 results in the marginal difference in the lengths and coverage probability of the confidence interval compared to using the true value of σ^2. However, it is more challenging to obtain an accurate estimator in a high-dimensional setting. Using a biased estimate of σ^2 leads to wider confidence intervals and over-coverage. Both dimensionality and sparsity level influence the inferential results, with the impact of the sparsity level being more pronounced.

1 Introduction

In statistical inference, we assume the dataset is the observed value of a random variable Y from a known data-generating mechanism with unknown parameters. We are interested in doing parameter estimation and hypothesis testing to draw conclusions about the data. Fisher's likelihood theory is a core component of classical inference theory. An important assumption of Fisher's likelihood theory is that the true model is known before performing the parameter estimation and the precision assessment. In other words, the observational data cannot be used to identify the true model from the candidate set (Zhang et al., 2022). However, this assumption is usually violated in practice. People always use the same data for model selection and to provide inference for the selected variables, which means the final selected model is not fixed and classical inference results obtained based on this stochastic model become unreliable. Numerous contributions have been made to investigate the effects of this violation on inferential results. For example, Berk et al. (2010) found that parameter estimates were biased and their sampling distributions were distorted.

Hence, selective inference is proposed to provide valid inference after variable selection. There are three principal techniques in selective inference. The oldest and most straightforward one is data splitting (Cox, 1975; Faraway, 1998). Its basic idea is to split the whole dataset into two independent subsets: training and test sets. The training data is used to select variables and the test set is used to make inferences for those chosen variables. The second method is simultaneous inference, which was proposed by Berk et al. (2013). This approach results in valid but conservative simultaneous confidence sets, regardless of selection procedures and selected model. The third approach is conditional selective inference, which constructs confidence intervals using the distribution conditioned on the selected submodel (Lee et al., 2016).

One major disadvantage of the data splitting method is that results can be affected by randomness, which means different splits potentially result in different selected models and hence various parameter estimates. The randomness also causes problems in the

interpretation (Kuchibhotla et al., 2022). To address selection bias, Tian and Taylor (2018) proposed randomisation at the selection stage. The general idea is to use a randomised version of the data to select variables, followed by making inferences based on the distribution of the data conditional on the randomised form. Motivated by the work of Tian and Taylor (2018), Rasines and Young (2023) were the first to suggest splitting the original dataset Y into independent components (U, V) by introducing a noise variable, with U used for selection and V for the inference stage. This (U, V) decomposition method offers improved selection stability and higher inferential power compared to the data splitting method. However, this approach is only strictly valid when the variance of the error term, σ^2, is known, ensuring that U and V are independent of each other. Therefore, the objective here is to investigate the effect of the estimation of σ^2 at the inference stage using V.

In this report, we mainly focus on the evaluation of the (U, V) decomposition. In particular, we want to find the effect of estimating the variance of the error term on the inferential results. The report is structured as follows. In Section 2, we demonstrate details of the (U, V) decomposition based on a linear regression model. We define the partial regression coefficients as our inferential objectives, as suggested by Berk et al. (2013). Additionally, we provide a detailed description of the methods for estimating σ^2 in the low and high dimensional setting respectively. In Section 3, we clearly state the objectives of this report and our further evaluations. Section 4 includes the dimensionality analysis in the low-dimensional setting and corresponding results. The analyses in the high-dimensional setting can be found in Section 5, where we consider the effects of sparsity level and dimensionality, respectively. Finally, we conclude our results in Section 6 and present some limitations.

2 Methods

Linear regression model We are interested in making valid inferences for coefficients in the linear regression model. Let Y be the response variable containing n independent samples. The linear regression model can be defined as follows:

$$Y = \mu + \epsilon, \tag{1}$$

where $\mu = X\beta$ is a linear function of the covariates and corresponding coefficients. Here $X = (\boldsymbol{x}_1, ..., \boldsymbol{x}_n) \in \mathbb{R}^{n \times p}$ is an $n \times p$ design matrix with $\boldsymbol{x}_i = (x_{i1}, ..., x_{ip})$, $i = 1, ..., n$, and $\beta = (\beta_1, ..., \beta_p) \in \mathbb{R}^p$ is a vector of p unknown parameters. We assume $\epsilon = (\epsilon_1, ..., \epsilon_n) \in \mathbb{R}^n$ as noise vector, which independently and identically (i.i.d) follows the normal distribution with mean zero and variance, σ^2. $\epsilon \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0_n, \sigma^2 I_n)$, where I_n is an $n \times n$ identity matrix.

Selcetion method — LASSO Given the data above, one commonly used method for selecting potentially interesting variables from p covariates is the LASSO algorithm. Suppose the index set of the selected submodel is denoted as $M = \{j_1, j_2, \ldots, j_m\} \subseteq \{1, \ldots, p\}$. The submodel contains all the covariates with indices in set M, whose size is represented as $m = |M| \leq p$. In general, we normally apply LASSO to the dataset (Y, X) by solving:

$$\hat{\beta}^{\text{LASSO}} = \arg\min_{\beta \in \mathbb{R}^p} \left(\frac{1}{2} \|Y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right), \tag{2}$$

where λ represents the regularisation parameter and $\hat{\beta}^{\text{LASSO}}$ are the LASSO estimates, based on which we choose the interesting variables. The outcome of the selection procedure is the set of non-zero selected variables $\hat{M} = \{j : \hat{\beta}_j^{\text{LASSO}} \neq 0\}$, where $j \subseteq \{1, \ldots, p\}$. Because of this selection stage, we have to be careful when making inferences for those selected covariates.

Partial regression coefficients Before moving onto the inference stage, we should identify the target objectives for which we make inferences. A common choice is to provide inference for the full regression coefficients:

$$\beta = (X^T X)^{-1} X^T \mu. \tag{3}$$

However, in the high-dimensional setup, where $n < p$, it is not possible to compute the full regression coefficients, since $X^T X$ is not invertible. A good solution to this problem is proposed by Berk et al. (2013). Given the selected submodel M, let $X_M = (X_{j_1}, \ldots, X_{j_m}) \in \mathbb{R}^{n \times m}$ denotes the $n \times m$ submatrix of X containing the columns indexed by M. The partial regression coefficients are given by:

$$\beta_M = (X_M^T X_M)^{-1} X_M^T \mu, \tag{4}$$

where $\mu = X\beta$. We aim to provide valid inference for the partial regression coefficients in the following analyses.

2.1 (U, V) decomposition

Similar to the randomisation schemes proposed by Tian and Taylor (2018), Rasines and Young (2023) introduced an $i.i.d$ randomised noise vector $W \in \mathbb{R}^n$, which follows the normal distribution $\mathcal{N}(0_n, \sigma^2 I_n)$ with mean zero and variance σ^2. Here the variance of W is equal to that of the error term ϵ in Equation (1). Then U and V can be defined as follows:

$$U = Y + \gamma W \tag{5a}$$

$$V = Y - \frac{1}{\gamma} W, \tag{5b}$$

Model selection is based on U and inference is based on V. This strategy is similar to the

data splitting method, but it is more accurately described as an information splitting method. It improves the data splitting approach by allowing access to all data during both the selection and inferential stages. If the splitting fraction for the data splitting method is $f = \frac{|r|}{n}$, Rasines and Young (2023) proposed $\gamma = \sqrt{\frac{1-f}{f}}$. In this situation, the (U, V) decomposition can be seen as a method of averaging information over data splits of the size $|r|$.

In the selection stage, instead of getting LASSO estimates based on (Y, X) in Equation (2), here we apply LASSO to data pair (U, X) by solving:

$$\hat{\beta}^{\text{LASSO}} = \arg\min_{\beta \in \mathbb{R}^p} \left(\frac{1}{2} \|U - X\beta\|_2^2 + \lambda \|\beta\|_1 \right), \tag{6}$$

from which we obtain the set of non-zero selected variables $\hat{M} = \{j : \hat{\beta}_j^{\text{LASSO}} \neq 0\}$. Then, we can provide inference for the partial regression coefficients defined in Equation (4) by fitting a linear regression on data pair $(V, X_{\hat{M}})$, resulting in the estimates $\hat{\beta}_{\hat{M}}$:

$$\hat{\beta}_{\hat{M}} = (X_{\hat{M}}^T X_{\hat{M}})^{-1} X_{\hat{M}}^T V. \tag{7}$$

However, a crucial parameter that determines whether the above procedure provides reliable inferences is σ^2. If σ^2 is known, U and V are independent. We assume that Y and W are i.i.d normally distributed, thus their linear combinations, U and V, both follow the normal distributions:

$$U \sim \mathcal{N}\left(\mu, (1 + \gamma^2)\sigma^2 I_n\right), \tag{8a}$$

$$V \sim \mathcal{N}\left(\mu, (1 + \gamma^{-2})\sigma^2 I_n\right). \tag{8b}$$

According to the definition of bivariate normal distribution, the joint distribution of U and V is also normally distributed:

$$\begin{pmatrix} U \\ V \end{pmatrix} \sim \mathcal{N}\left(\mu, \begin{pmatrix} \Sigma_U & \Sigma_{UV} \\ \Sigma_{VU} & \Sigma_V \end{pmatrix}\right), \tag{9}$$

where $\Sigma_U = (1+\gamma^2)\sigma^2 I_n$, $\Sigma_V = (1+\gamma^{-2})\sigma^2 I_n$ and $\Sigma_{UV} = \Sigma_{VU}^T \in \mathbb{R}^{n\times n}$.

To prove that U and V are independent is equivalent to showing that the elements of the covariance matrix between U and V are zeros, i.e., $\Sigma_{UV} = \Sigma_{VU}^T = 0_{n\times n}$. The derivations can be seen as follows:

$$\begin{aligned}
\Sigma_{UV} &\equiv \mathrm{Cov}(U, V) \\
&= \mathrm{E}[(U - \mathrm{E}[U])\,(V - \mathrm{E}[V])] \\
&= \mathrm{E}[(U-\mu)\,(V-\mu)] \\
&= \mathrm{E}[UV^T] - \mu\,(\mathrm{E}[U] + \mathrm{E}[V]) + \mu^T\mu \\
&= \mathrm{E}[UV^T] - 2\mu\mu^T + \mu\mu^T \\
&= \mathrm{E}[(Y+\gamma W)\left(Y - \gamma^{-1}W\right)] - \mu\mu^T \\
&= \mathrm{E}[YY^T - \gamma^{-1}YW^T + \gamma WY^T - WW^T] - \mu\mu^T \\
&= \mathrm{E}[Y]\mathrm{E}[Y]^T - \gamma^{-1}\mathrm{E}[Y]\mathrm{E}[W]^T + \gamma\mathrm{E}[W]\mathrm{E}[Y]^T - \mathrm{E}[W]\mathrm{E}[W]^T - \mu\mu^T \\
&= \mu\mu^T - 0 + 0 - 0 - \mu\mu^T \\
&= 0
\end{aligned}$$

This is consistent with the situation in the data splitting method, where the data used for selection and inference are independent. However, if σ^2 is unknown, we need to use appropriate methods to estimate it. After substituting the estimated σ^2 into Equations (8a) and (8b), U and V are no longer independent. Although Rasines and Young (2023) have proved the asymptotic validity of (U, V) decomposition in this situation, we are interested in what exactly happens to the inferential results when we use an estimate of σ^2, especially in a high-dimensional setting.

2.2 Methods for estimating σ^2

2.2.1 In the low-dimensional setting

If p is smaller than n, the residual sum of squares can be used for estimating σ^2 :

$$\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n}\left(Y_i - \boldsymbol{x}_i^T\hat{\beta}\right)^2, \tag{10}$$

where $\hat{\beta} = (X^TX)^{-1}X^TY$ is the ordinary least squares estimator in the full model containing all p covariates.

2.2.2 In the high-dimensional setting

When p is bigger than n, the method given in Equation (10) can not be used, since the X^TX in $\hat{\beta}$ is not invertible in this case. The refitted cross-validation (RCV) method, which was proposed by Fan et al. (2012), has been used in this report. The basic idea is to randomly split the whole data set (Y, X) into two even subsets $(Y^{(1)}, X^{(1)})$ and $(Y^{(2)}, X^{(2)})$. Firstly, we perform a variable selection algorithm such as LASSO on $(Y^{(1)}, X^{(1)})$ and get the set of selected variables $\hat{M}_1$. Then we use $(Y^{(2)}, X^{(2)}_{\hat{M}_1})$ to estimate σ_1^2 :

$$\hat{\sigma}_1^2 = \frac{Y^{(2)T}(I_{n/2} - P^{(2)}_{\hat{M}_1})Y^{(2)}}{n/2 - |\hat{M}_1|}, \tag{11}$$

where $P^{(2)}_{\hat{M}_1} = X^{(2)}_{\hat{M}_1}(X^{(2)T}_{\hat{M}_1}X^{(2)}_{\hat{M}_1})^{-1}X^{(2)T}_{\hat{M}_1}$.

Similarly, we repeat the above procedure, but this time $(Y^{(2)}, X^{(2)})$ is used to get the set $\hat{M}_2$ and $(Y^{(1)}, X^{(1)}_{\hat{M}_2})$ is then used for estimation of σ_2^2 :

$$\hat{\sigma}_2^2 = \frac{Y^{(1)T}(I_{n/2} - P^{(1)}_{\hat{M}_2})Y^{(1)}}{n/2 - |\hat{M}_2|}. \tag{12}$$

Then the final estimator is the weighted average of $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$:

$$\hat{\sigma}^2_{WRCV} = \frac{\hat{\sigma}_1^2 \times (n/2 - |\hat{M}_1|) + \hat{\sigma}_2^2 \times (n/2 - |\hat{M}_2|)}{n - |\hat{M}_1| - |\hat{M}_2|}. \tag{13}$$

There are two important assumptions when using the RCV estimator. The first one is the sparsity condition, which makes the high-dimensional problems solvable. Suppose we use $M_0 = \{j : \beta_j \neq 0\}$ to represent the set of variables in the true model. Then we assume that the number of non-zero parameters $s = |M_0|$ is small compared with the sample size n. Secondly, the model selection algorithm should exhibit a sure screening property, which means that all the important variables are chosen by the selection procedure, with a probability approaching 1 (Fan and Lv, 2008). Fan et al. (2012) found that the RCV estimator is unbiased if the sure screening property holds.

In this paper, we apply LASSO in the variable selection procedure when estimating the variance using the RCV method. We tune the regularisation parameter λ by monitoring the satisfaction of the sure screening property, which is represented by true positive rate (TPR).

	True 0	True 1
Selected 0	True negative	False negative
Selected 1	False positive	True positive

Table 1: Confusion Matrix for the state of the variables in the true and selected model, where 'True positive' refers to the number of variables correctly identified by LASSO as being part of the true model.

Table 1 is the confusion matrix for the state of the variables in the true and selected variables. Then true positive rate is defined as:

$$\text{TPR} = \frac{\text{True positive}}{\text{True positive} + \text{False negative}}. \tag{14}$$

TPR measures the proportion of important variables correctly chosen by LASSO. A higher TPR indicates that the selected model includes as many relevant variables as

possible, thus satisfying the sure screening property.

2.3 Assessing metrics

At the inferential stage of the (U, V) decomposition, we consider the problem of constructing confidence intervals for the partial regression coefficients β_M defined in Equation (4). This would be based on the coefficient estimates $\hat{\beta}_{\hat{M}}$ defined in Equation (7).

When σ^2 is known, $\hat{\beta}_{\hat{M}} \sim \mathcal{N}(\beta_M, \sigma^2(1+\gamma^{-2})(X_{\hat{M}}^T X_{\hat{M}})^{-1})$, then the $100(1-\alpha)\%$ confidence intervals (CIs) for the jth element of β_M are formed as:

$$\hat{\beta}^j_{\hat{M}} \ \pm \ z_{1-\frac{\alpha}{2}}\sigma\sqrt{(1+\gamma^{-2})\left[(X_{\hat{M}}^T X_{\hat{M}})^{-1}\right]_{jj}}, \tag{15}$$

where $z_{1-\frac{\alpha}{2}}$ is the $1-\frac{\alpha}{2}$ quantile of the standard normal distribution. When σ^2 is assumed unknown, we can plug in an estimate of it by using the methods described in Section 2.2. Then the CIs in Equation (15) become $\hat{\beta}^j_{\hat{M}} \ \pm \ z_{1-\frac{\alpha}{2}}\hat{\sigma}\sqrt{(1+\gamma^{-2})\left[(X_{\hat{M}}^T X_{\hat{M}})^{-1}\right]_{jj}}$.

We report the length of the confidence intervals and the actual coverage probability in the following sections to compare the differences in the inferential results when using the true value of σ^2 versus its estimator.

3 Aims

This report aims to examine the effect of estimating σ^2 on the inferential results obtained by using the (U, V) decomposition. We mainly focus on the following situations:

- In the low-dimensional setup, we use the residual sum of squares to estimate σ^2. We further investigate whether the results are affected by the dimensionality of the parameters.

- In the high-dimensional setup, we use the refitted cross-validation method to estimate σ^2. Further evaluations include examining the effects of the sparsity level and dimensionality of the parameters.

4 The dimensionality analysis in the low-dimensional setting

4.1 Simulation setup

In this simulation study, we generated $B = 25,000$ replications. For each simulation, we set $n = 100$. Each row of the design matrix was independently sampled from the distribution $\mathcal{N}(0_p, \Gamma)$ with $\Gamma \in \mathbb{R}^{p \times p}$ following a Toeplitz covariance structure, where the (i, j) element of Γ is $\rho^{|i-j|}$, $\rho > 0$. We set $\rho = 0.5$ in our study. The true parameter $\beta = \{1, -1, 0.5, -0.5, 0.2, -0.2, 0, ..., 0\} \in \mathbb{R}^p$, with $s = 6$ non-zero elements and $(p - 6)$ zeros. In terms of the parameters in the (U, V) decomposition, we fixed $f = 0.5$ to compute U and V. We adopted the LASSO method in the selection procedure with regularisation parameter $\lambda = \sqrt{\frac{2 \log p}{n}}$, as suggested by Liu et al. (2018). The true value of the variance σ^2 was fixed at $\sigma^2 = 1$. The method used to estimate it has been formulated in Section 2.2.1. We set $\alpha = 0.05$, resulting in a 95% nominal coverage probability for the CIs. We gradually increased the dimensionality of the coefficients $p = \{10, 20, 40, 80\}$ to investigate its effect.

4.2 Results

4.2.1 Estimated σ^2

The results of the estimated σ^2 across different dimensionality p are shown in Figure 1. Generally speaking, using the residual sum of squares provides reasonable estimators. The median of $\hat{\sigma}^2$ is nearly equal to the true value of σ^2 for all dimensionalities, except

when $p = 80$, where $\hat{\sigma}^2$ is slightly below the red line. As the value of p increases, the range of $\hat{\sigma}^2$ expands, reaching its widest range when $p = 80$, which is nearly approaching the number of samples n. This indicates that we are more confident about the results when the true parameter has a relatively smaller dimensionality p.

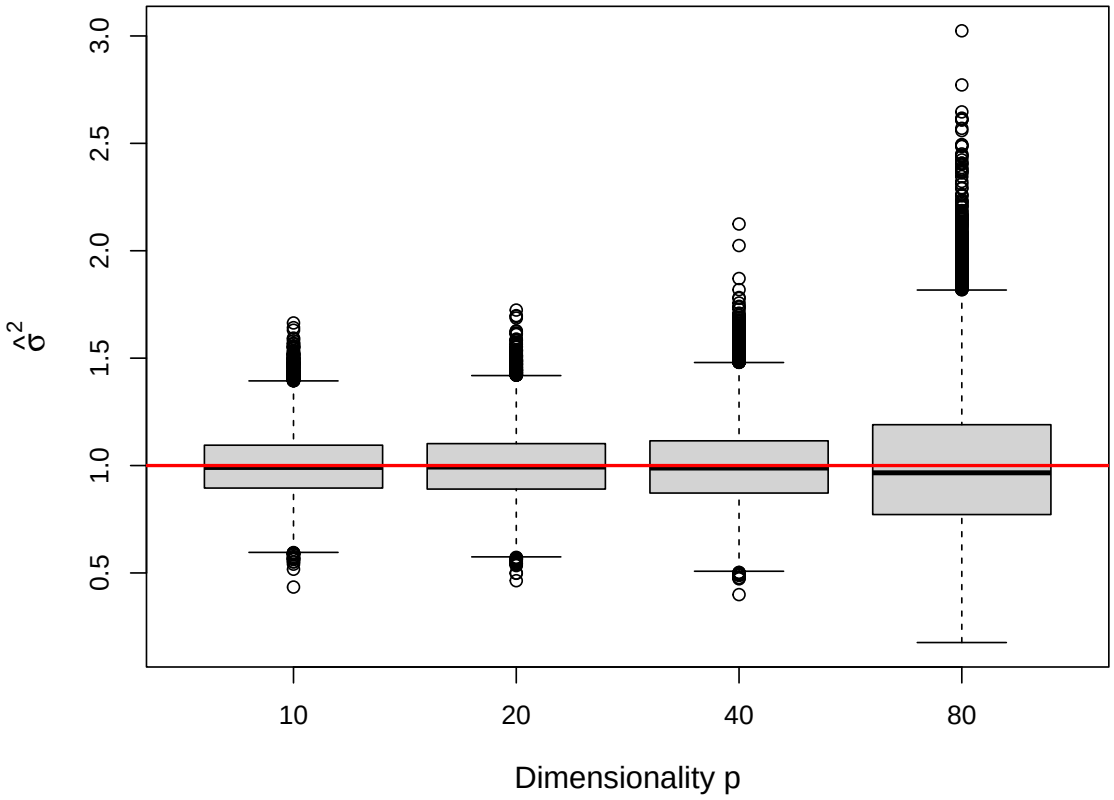

Figure 1: Boxplots of $\hat{\sigma}^2$ obtained using the residual sum of squares across different dimensionality of parameters. The red line represents the $\sigma^2_{\text{true}} = 1$.

4.2.2 Inference by using the (U, V) decompostion

Length of confidence interval When comparing the inferential results obtained using the true value of σ^2 against the estimators, we first examine the length of confidence intervals for the partial regression coefficients, as formulated in Section 2.3. The results can be seen in Figure 2. Here, the lengths are compared across the corresponding

absolute value of the full-model coefficients β_i chosen during the selection stage of the (U, V) decomposition.

Intuitively, from plot A we can see that, when $p = 10$, using $\hat{\sigma}^2$ results in wider confidence intervals for all the coefficients than using true σ^2. To be specific, the median and mean of the confidence interval lengths corresponding to $|\beta| = 0$ are marginally larger when using the estimator of σ^2. Applying the estimated σ^2 does not cause much difference to the median and mean for the remaining coefficients. However, we can find more variations in the confidence interval lengths when employing $\hat{\sigma}^2$ regardless of the absolute value of coefficients. Similar results can be found in plot B when $p = 80$, but it seems that the gap in confidence interval lengths between using the estimated σ^2 and the true value is larger in a relatively higher dimensionality.

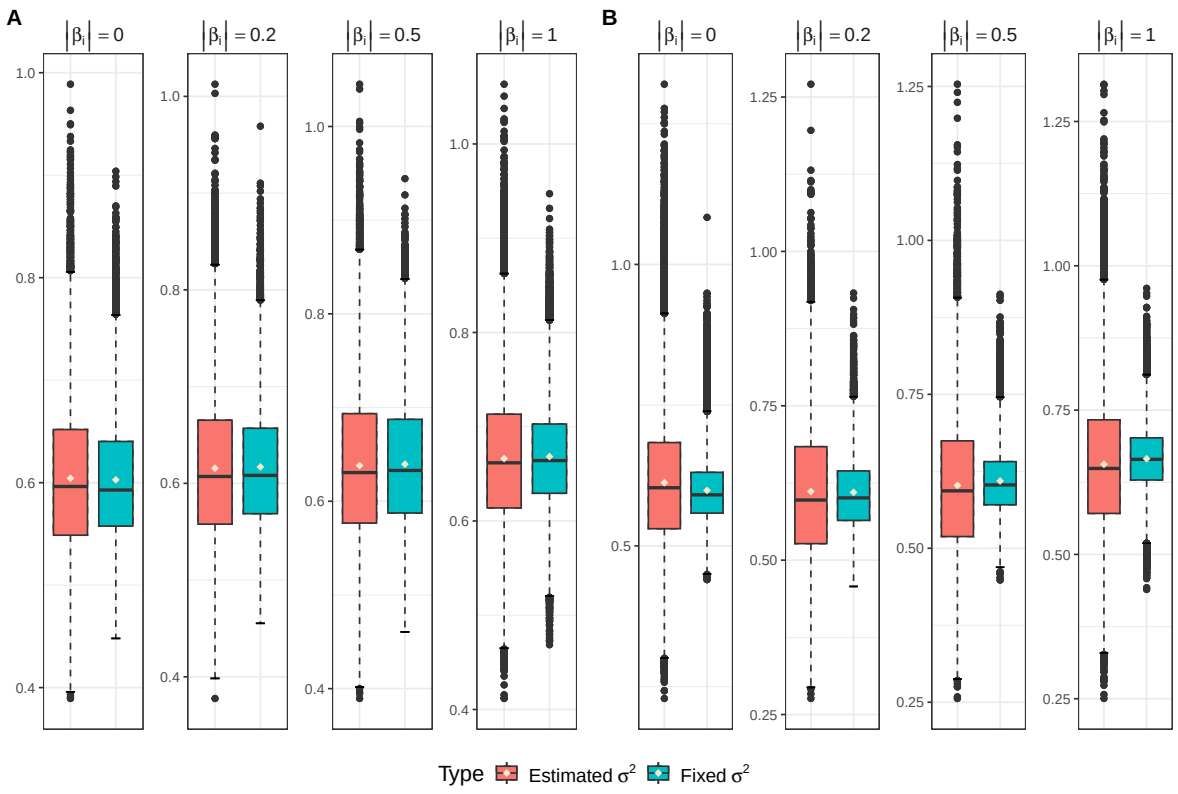

Figure 2: Boxplots of lengths of confidence intervals for the partial regression coefficients using $\hat{\sigma}^2$ (Estimated σ^2) and $\sigma^2_{\text{true}} = 1$ (Fixed σ^2). Plot A is obtained when $p = 10$, while plot B is obtained when $p = 80$. Lengths are compared across the absolute values of the full regression coefficients, $|\beta_i| \in \{0, 0.2, 0.5, 1\}$. The diamond symbol in each box represents the mean of the interval length.

To further examine the impact of dimensionality p on the results, we generate Table 2. This table presents the observed median and standard deviation of the confidence interval lengths for the partial regression coefficients across various dimensionalities. We compare the results among the absolute values of the selected coefficients $|\beta_i| \in \{0, 0.2, 0.5, 1\}$. We can see that, the empirical medians are similar in all cases. In contrast, estimating σ^2 increases the standard deviation of confidence interval lengths especially when p is large.

$\lvert\beta_i\rvert$	$p = 10$		$p = 40$		$p = 80$	
	$\hat{\sigma}^2$	σ^2_{true}	$\hat{\sigma}^2$	σ^2_{true}	$\hat{\sigma}^2$	σ^2_{true}
0.0	0.60 (0.08)	0.59 (0.06)	0.59 (0.08)	0.59 (0.06)	0.60 (0.12)	0.59 (0.06)
0.2	0.61 (0.08)	0.61 (0.07)	0.60 (0.09)	0.60 (0.06)	0.60 (0.12)	0.60 (0.06)
0.5	0.63 (0.09)	0.63 (0.07)	0.60 (0.08)	0.61 (0.06)	0.59 (0.12)	0.60 (0.06)
1.0	0.66 (0.07)	0.66 (0.06)	0.66 (0.08)	0.66 (0.06)	0.65 (0.12)	0.66 (0.06)

Table 2: Empirical median and standard deviation of confidence interval lengths for the partial regression coefficients using $\hat{\sigma}^2$ and σ^2_{true}. Results for different dimensionalities $p \in \{10, 40, 80\}$ are included. Lengths are compared across the absolute values of the full regression coefficients, $|\beta_i| \in \{0, 0.2, 0.5, 1\}$. The standard deviation is in the brackets.

Coverage probability Secondly, Figure 3 demonstrates the results of the coverage probability of the confidence intervals for the partial regression coefficients across various dimensionalities p. Separate plots are generated according to whether the corresponding coefficients are active in the true model.

Using estimated σ^2 decreases the average coverage probability of the confidence intervals compared to using the true value of σ^2 for all dimensionalities p. This does not matter whether we are considering a 'null' parameter or an 'active' one. The differences in the coverages between employing $\hat{\sigma}^2$ and σ^2_{true} become more pronounced as p increases, especially when $\beta = 0$.

Suppose we use C to represent the total number of confidence intervals containing the true partial regression coefficients. We know that C follows the Binomial distribution $Bin(B, p)$, where $B = 25,000$ is the number of replications and p is the coverage probability, which is expected to be close to 95%. Since the variance of C is $B \times p \times (1-p)$, the

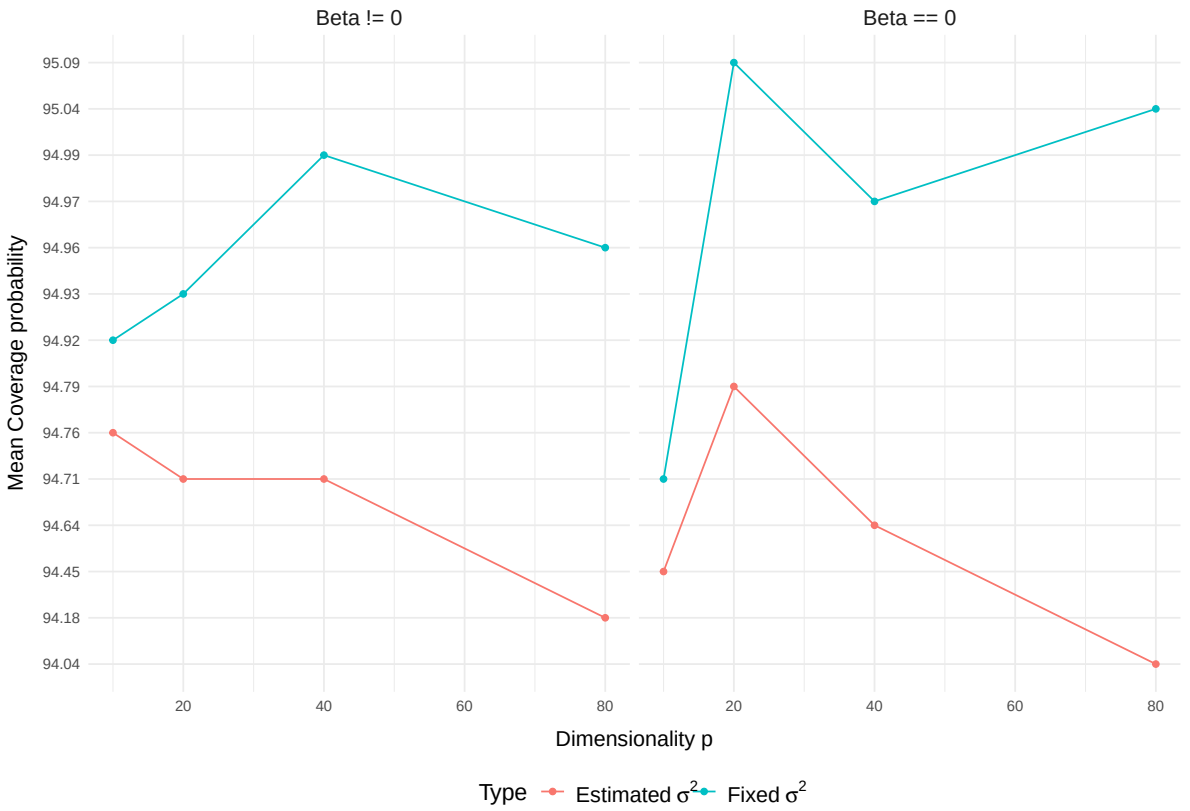

Figure 3: Mean coverage probability of the confidence intervals using $\hat{\sigma}^2$ (Estimated σ^2) and σ^2_{true} (Fixed σ^2) across different Dimensionalities p. The left plot refers to the active coefficients in the true model. The right plot corresponds to the coefficients that are zero in the true model.

variance of the estimated coverage probability can be approximated by $\frac{\hat{p}\times(1-\hat{p})}{B}$, where $\hat{p}$ is the observed coverage probability. Table 3 presents the results of the estimated variance of the coverage probability. All the values are very small given the large number of replications. Hence, we can deduce that the marginal loss in the coverage accuracy when estimating σ^2 is not caused by chance.

	$p = 10$	$p = 20$	$p = 40$	$p = 80$
Estimated σ^2	2×10^{-6}	1.99×10^{-6}	2.02×10^{-6}	2.22×10^{-6}
Fixed σ^2	1.94×10^{-6}	1.91×10^{-6}	1.91×10^{-6}	1.90×10^{-6}

Table 3: Estimated variance of the coverage probability obtained using $\hat{\sigma}^2$ (Estimated σ^2) and σ^2_{true} (Fixed σ^2) across various dimensionalities p.

5 Analyses in the high-dimensional setting

5.1 The sparsity level analysis

5.1.1 Simulation setup

Similar to Section 4.1, in this simulation study, we also generated $B = 25,000$ replications with a sample size of $n = 100$ for each simulation. In terms of the design matrix, we kept using the Toeplitz covariance structure for the variance of its distribution, with $\rho = 0.5$. We also fixed $f = 0.5$ and the regularisation parameter of LASSO with $\lambda = \sqrt{\frac{2 \log p}{n}}$. Again, we assumed $\sigma^2_{\text{true}} = 1$. However, in this section, the refitted cross-validation method was used to estimate σ^2. We continued using $\alpha = 0.05$.

We fixed $p = 150$ and gradually decreased the sparsity level of the coefficients, which was reflected by increasing the number of non-zero elements, s, in the coefficient vector, to examine its effect in this section. The value of s and corresponding β are defined as follows:

- $s = 6$
  $\beta = \{\pm 1, \pm 0.5, \pm 0.2, 0, ..., 0\}$
- $s = 10$
  $\beta = \{\pm 1, \pm 0.8, \pm 0.6, \pm 0.4, \pm 0.2, 0, ..., 0\}$
- $s = 14$
  $\beta = \{\pm 1, \pm 0.8, \pm 0.7, \pm 0.6, \pm 0.5, \pm 0.4, \pm 0.2, 0, ..., 0\}$
- $s = 20$
  $\beta = \{\pm 1, \pm 0.9, \pm 0.8, \pm 0.7, \pm 0.6, \pm 0.5, \pm 0.4, \pm 0.3, \pm 0.2, \pm 0.1, 0, ..., 0\}$
- $s = 50$
  $\beta = \{\pm 1, \pm 0.9, \pm 0.8, \pm 0.7, \pm 0.6, \pm 0.5, \pm 0.4, \pm 0.3, \pm 0.2, \pm 0.1, runif(30, 0, 0.1), 0, ..., 0\}$

- $s = 100$

  $\beta = \{\pm 1, \pm 0.9, \pm 0.8, \pm 0.7, \pm 0.6, \pm 0.5, \pm 0.4, \pm 0.3, \pm 0.2, \pm 0.1, runif(80, 0, 0.1), 0, ..., 0\}$

- $s = 150$

  $\beta = \{\pm 1, \pm 0.9, \pm 0.8, \pm 0.7, \pm 0.6, \pm 0.5, \pm 0.4, \pm 0.3, \pm 0.2, \pm 0.1, runif(130, 0, 0.1)\}$

where $runif(30, 0, 0.1)$ means we randomly generated 30 values from the Uniform distribution $\mathbb{U}(0, 0.1)$.

In Section 2.2.2 we mentioned that we need to tune the regularisation parameter λ in the RCV method. We tried $\lambda = \{0.2, 0.1, 0.05\}$ and chose the best one in the following analyses.

5.1.2 Results

Choosing suitable λ for the RCV Table 4 contains the mean squared error (MSE) of $\hat{\sigma}^2$ when using different regularisation parameter λ and sparsity level s. We always prefer a λ with a small MSE. As can be seen, using $\lambda = 0.1$ consistently results in the smallest MSE, regardless of the value of s.

λ \ s	6	10	14	20	50	100	150
0.2	0.1808	0.9926	2.2724	4.2061	4.8981	6.3620	8.7164
0.1	0.1589	0.7598	1.8574	3.5562	4.0959	5.3370	6.7712
0.05	0.2174	0.8187	2.1951	3.9866	9.6467	6.5937	98.509

Table 4: Mean squared error (MSE) of the estimated σ^2 for different combinations of regularisation parameter λ and sparsity level s.

Furthermore, Figure 4 displays the performance of the true positive rate and the estimated σ^2 across different values of λ when $s = 20$. From the left boxplot, we can see that as λ decreases, the TPR increases, indicating that the assumption of the sure screening property mentioned in Section 2.2.2 is more easily satisfied with a smaller λ. However, when $\lambda = 0.05$, more errors occur in the right subplot, which is not good for the inferential stage. Looking again at the left plot, the improvement of the TPR with $\lambda = 0.05$

compared to $\lambda = 0.1$ is not significant. Therefore, to balance the performance of TPR and $\hat{\sigma}^2$, we prefer to use $\lambda = 0.1$, which aligns with the result shown in Table 4.

After analysing the TPR and $\hat{\sigma}^2$ for other values of s, we can conclude that, in general, $\lambda = 0.1$ gives sensible $\hat{\sigma}^2$. Thus, we keep using $\lambda = 0.1$ for all the values of s in the subsequent analyses.

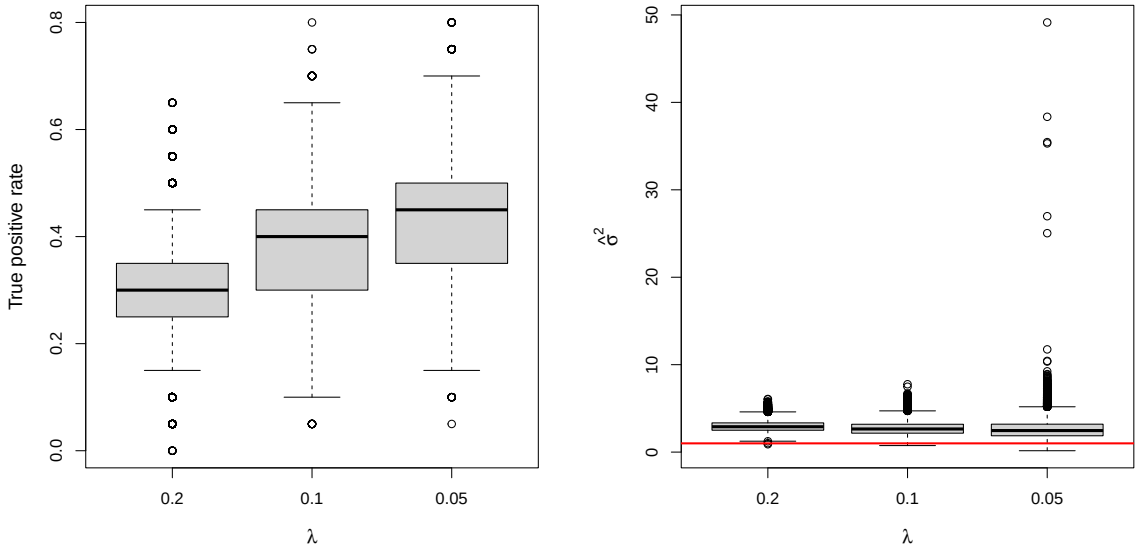

Figure 4: Boxplots of true positive rate (left) and estimated σ^2 (right) for $\lambda \in \{0.2, 0.1, 0.05\}$ when $s = 20$.

Estimated σ^2 Boxplots of the estimated σ^2 across different sparsity levels s when using $\lambda = 0.1$ can be seen in Figure 5. The median of $\hat{\sigma}^2$ at all sparsity levels exceeds the true value of σ^2. Moreover, as the sparsity level decreases, the estimator of σ^2 becomes more and more biased. The quality of estimation deteriorates rapidly as the number of non-zero elements in β grows, especially from $s = 6$ to $s = 20$. In contrast, after $s = 20$, the rate at which the bias increases slows down. This is probably because although we add more non-zero elements in β, the absolute value of most of the added elements is extremely small, ranging from 0 to 0.1, which reduces their effect on the estimate of σ^2. These results coincide with what we have said in Section 2.2.2. Since one of the

core assumptions for using the refitted cross-validation method to estimate σ^2 in a high-dimensional setting is the sparsity condition, it is unsurprising that the RCV provides better estimators when s is small.

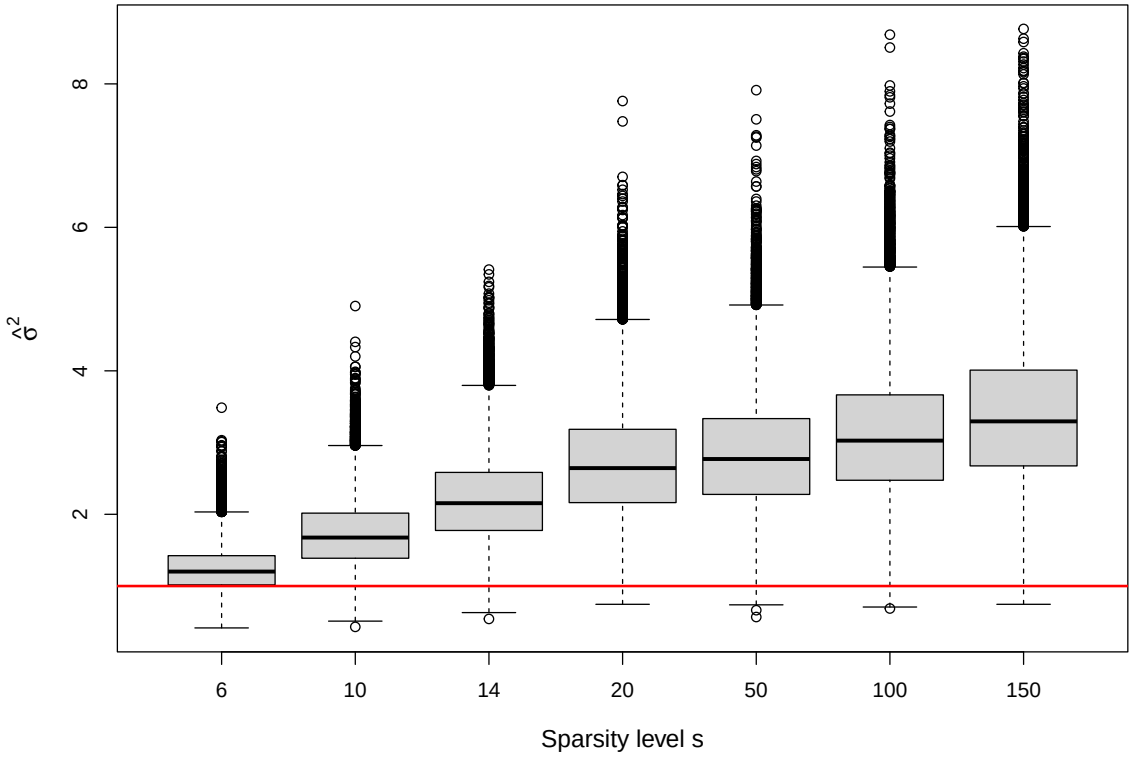

Figure 5: Boxplots of $\hat{\sigma}^2$ obtained using the refitted cross-validation method across different sparsity levels s when $\lambda = 0.1$. The red line represents the $\sigma^2_{\text{true}} = 1$.

Inference by using the (U, V) decompostion Figure 6 shows the boxplots of the confidence interval lengths obtained by using estimated σ^2 and fixed σ^2 in the case of $s = 6$. It is obvious that if we do not know the true value of σ^2 and use an estimate instead, we may get wider confidence intervals for all the coefficients. And there are more variations in the results.

Further investigation into the impact of sparsity level s on the results can be seen in Table 5. The increase in the number of non-zero elements in the parameters causes dramatic growth in both the empirical medians and standard deviations of the confidence interval

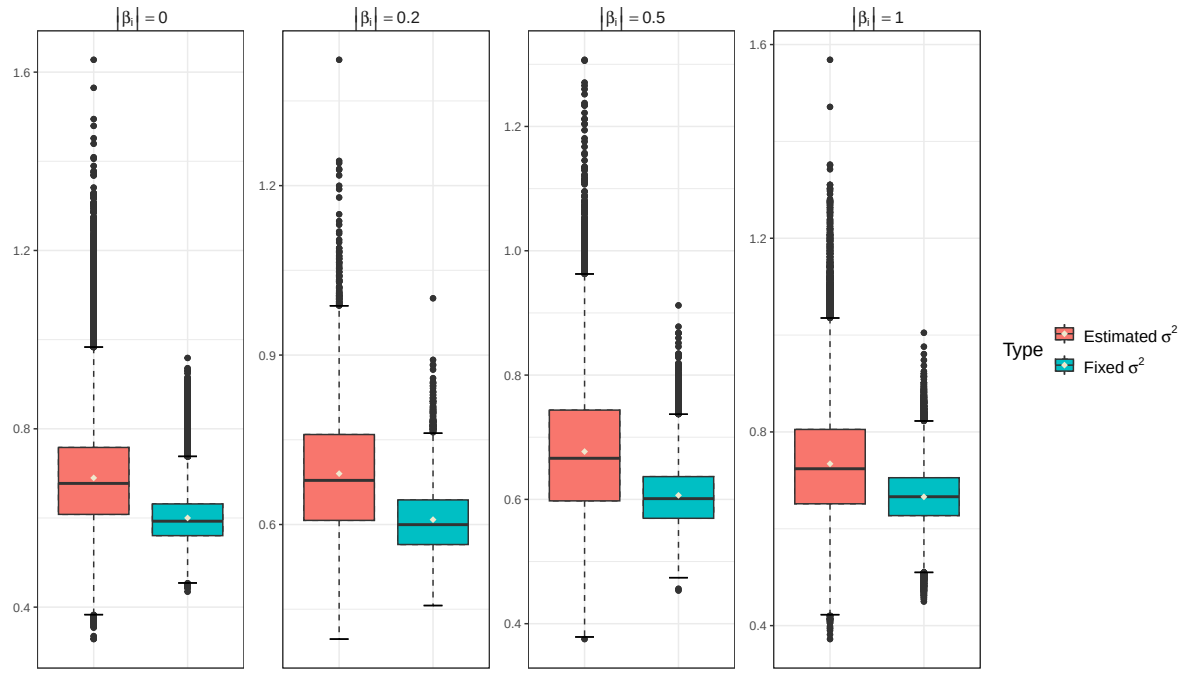

Figure 6: Boxplots of lengths of confidence intervals for the partial regression coefficients using $\hat{\sigma}^2$ (Estimated σ^2) and $\sigma^2_{\text{true}} = 1$ (Fixed σ^2). This is obtained when $s = 6$. Lengths are compared across the absolute values of the full regression coefficients, $|\beta_i| \in \{0, 0.2, 0.5, 1\}$. The diamond symbol in each box represents the mean of the interval length.

lengths for all the coefficients when using $\hat{\sigma}^2$. To illustrate, for $|\beta_i| = 0.2$, when we use the true value of σ^2, the medians of confidence interval lengths are 0.6, 0.62 and 0.65 for $s = 6$, $s = 20$ and $s = 150$, respectively. However, when we use $\hat{\sigma}^2$, the observed median rises significantly from 0.68 ($s = 6$) to 1.06 ($s = 20$). Finally, the median of the confidence interval lengths for $s = 150$ is nearly twice that for $s = 6$. Therefore, the differences in the confidence interval lengths become larger as the degree of sparsity level decreases.

In terms of the average coverage probability of the confidence intervals, the results can be seen in Figure 7. The coverage probabilities are satisfactory if we use the true σ^2, with the mean coverage probability fluctuating around the nominal value of 95%. However, when σ^2 is estimated, we observe significantly higher coverage probabilities. For instance, when $s = 6$, the mean coverage probability for the active coefficients is approximately 95.71%, which is acceptable. However, as s increases to the non-sparse

$\lvert\beta_i\rvert$	$s = 6$		$s = 20$		$s = 150$	
	$\hat{\sigma}^2$	σ^2_{true}	$\hat{\sigma}^2$	σ^2_{true}	$\hat{\sigma}^2$	σ^2_{true}
0.0	0.68 (0.12)	0.59 (0.06)	1.07 (0.21)	0.62 (0.06)	-	-
0-0.1	-	-	-	-	1.25 (0.26)	0.66 (0.08)
0.1	-	-	1.07 (0.21)	0.62 (0.06)	1.24 (0.25)	0.65 (0.07)
0.2	0.68 (0.12)	0.60 (0.06)	1.06 (0.21)	0.62 (0.07)	1.24 (0.25)	0.65 (0.07)
0.3	-	-	1.06 (0.21)	0.62 (0.06)	1.23 (0.26)	0.64 (0.07)
0.4	-	-	1.06 (0.20)	0.62 (0.06)	1.22 (0.25)	0.64 (0.07)
0.5	0.67 (0.11)	0.60 (0.05)	1.06 (0.21)	0.62 (0.06)	1.23 (0.25)	0.65 (0.07)
0.6	-	-	1.06 (0.21)	0.63 (0.07)	1.23 (0.26)	0.65 (0.07)
0.7	-	-	1.08 (0.22)	0.64 (0.07)	1.25 (0.26)	0.66 (0.07)
0.8	-	-	1.10 (0.22)	0.66 (0.08)	1.26 (0.27)	0.67 (0.08)
0.9	-	-	1.11 (0.22)	0.68 (0.08)	1.28 (0.27)	0.69 (0.08)
1.0	0.72 (0.12)	0.67 (0.06)	1.10 (0.22)	0.67 (0.08)	1.26 (0.26)	0.69 (0.08)

Table 5: Empirical median and standard deviation of confidence interval lengths for the partial regression coefficients using $\hat{\sigma}^2$ and σ^2_{true}. Results for different sparsity levels $s \in \{6, 20, 150\}$ are included. Lengths are compared across the absolute values of the full regression coefficients, $|\beta_i|$. The standard deviation is in the brackets.

case, the large coverage probabilities, such as 98%, indicate over-coverage. This is likely due to the more biased estimation of σ^2 at relatively lower sparsity levels. Additionally, the differences in the coverage probabilities between using estimated and fixed σ^2 become more notable as the number of non-zero elements s increases.

5.2 The dimensionality analysis

5.2.1 Simulation setup

This simulation study mainly focuses on the dimensionality analysis in high-dimensional settings. Similar to the setup in Section 4.1, we fixed $B = 25,000$, $n = 100$, $\rho = 0.5$, $f = 0.5$, $\alpha = 0.05$ and the true value of the variance $\sigma^2 = 1$. The true parameter vector $\beta = \{1, -1, 0.5, -0.5, 0.2, -0.2, 0, ..., 0\} \in \mathbb{R}^p$, with the first $s = 6$ elements are non-zero and the remaining $(p - 6)$ zeros. We gradually increased $p = \{150, 200, 300, 500\}$ in this case. The regularisation parameter λ in the (U, V) decomposition selection stage was fixed at $\lambda = \sqrt{\frac{2 \log p}{n}}$. Additionally, we simply used $\lambda = 0.1$ in the estimation of σ^2 by

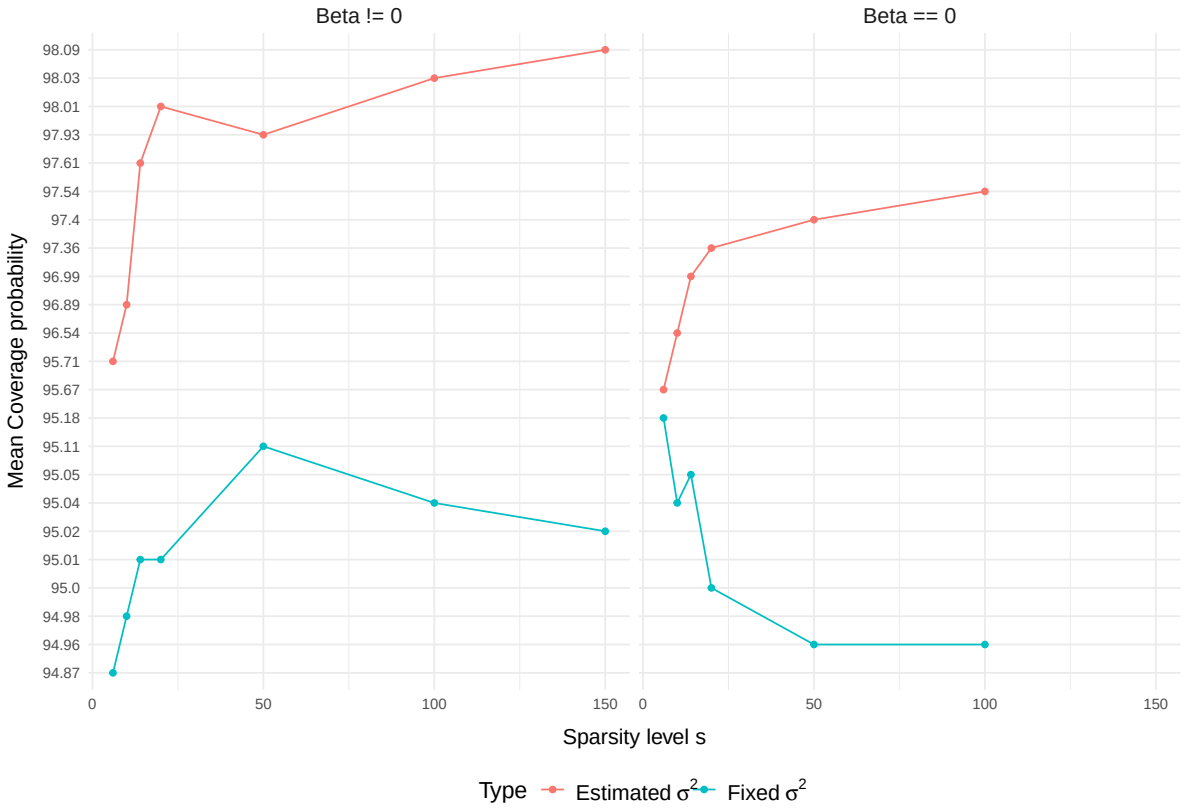

Figure 7: Mean coverage probability of the confidence intervals using $\hat{\sigma}^2$ (Estimated σ^2) and σ^2_{true} (Fixed σ^2) across different sparsity levels s. The left plot refers to the active coefficients in the true model. The right plot corresponds to the coefficients that are zero in the true model.

using the RCV method, since the investigation in Section 5.1.2 indicates $\lambda = 0.1$ always provides sensible estimators.

5.2.2 Results

Estimated σ^2 Figure 8 shows the performance of the estimated σ^2 obtained using the refitted cross-validation method across different dimensionalities p. Overall, the medians of the $\hat{\sigma}^2$ are nearly the same for all the dimensionalities, with the value at $p = 500$ being slightly higher than the others. In additional, all the medians lie above the red line, which represents the true value. We also observe the increasing variations in the estimator as p increases. Comparing this plot to Figure 1 and Figure 5, we can conclude that it is

harder to obtain an accurate estimator of σ^2 in the high-dimensional setting than in the low-dimensional setting. The refitted cross-validation method tends to produce a more biased estimator compared to using the residual sum of squares. Moreover, increasing p may not have much effect on the median of the estimators in both the low and high-dimensional setups, but it may increase the variation. In contrast, increasing s in the high dimensional setting can lead to a considerable rise in the median.

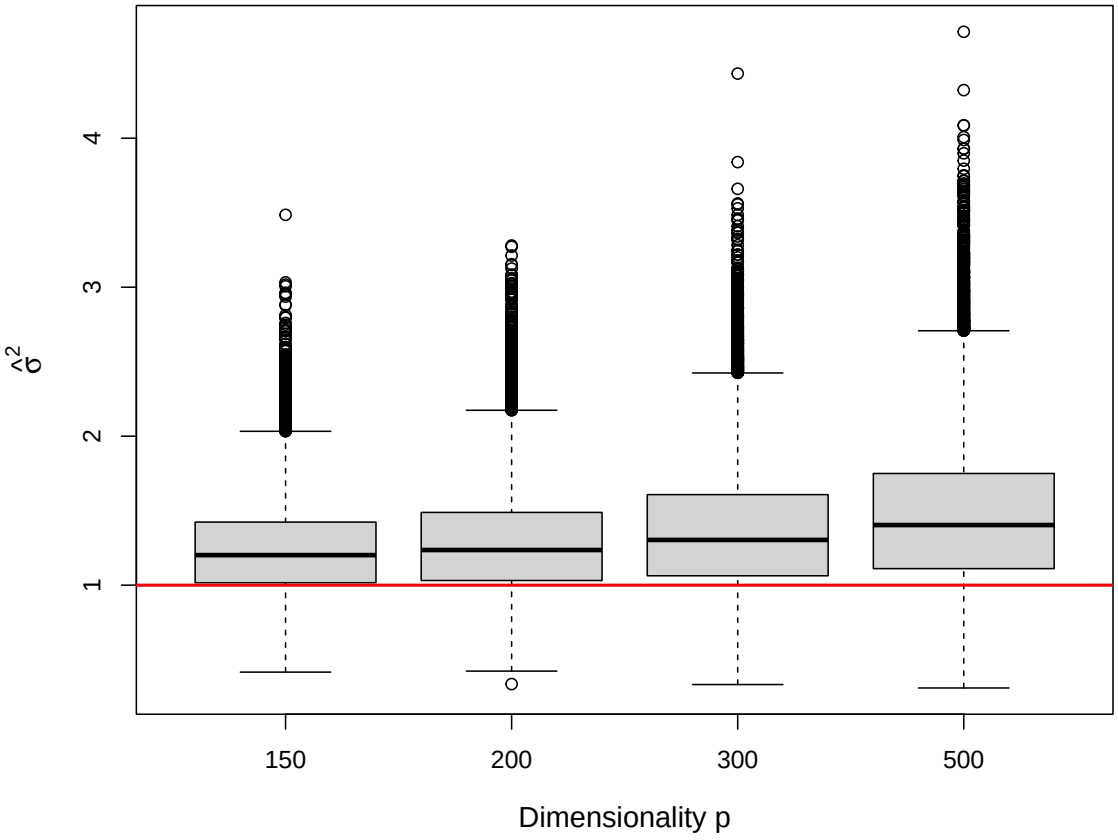

Figure 8: Boxplots of $\hat{\sigma}^2$ obtained using the refitted cross-validation method across different dimensionalities p when $\lambda = 0.1$. The red line represents the $\sigma^2_{\text{true}} = 1$.

Inference by using the (U, V) decompostion Figure 9 compares the lengths of the confidence intervals using true values and estimators of σ^2 for $p = 200$ (plot A) and $p = 500$ (plot B). To be specific, in plot A, we observe that estimating σ^2 gives

rise to an increase in the confidence interval lengths for all the absolute values of the coefficients. A similar pattern can be found in plot B. From an intuitive perspective, higher dimensionality, such as $p = 500$, tends to cause wider confidence intervals when using $\hat{\sigma}^2$.

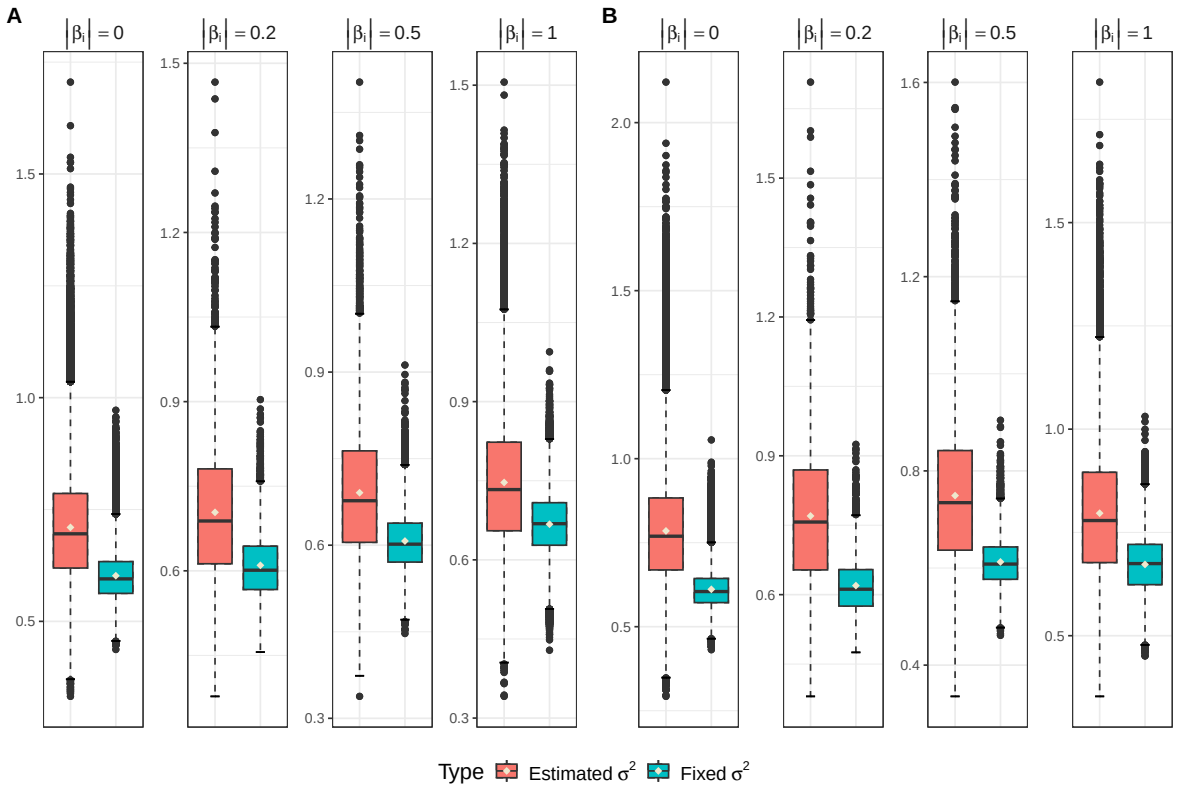

Figure 9: Boxplots of lengths of confidence intervals for the partial regression coefficients using $\hat{\sigma}^2$ (Estimated σ^2) and $\sigma^2_{\text{true}} = 1$ (Fixed σ^2). Plot A is obtained when $p = 200$, while plot B is obtained when $p = 500$. Lengths are compared across the absolute values of the full regression coefficients, $|\beta_i| \in \{0, 0.2, 0.5, 1\}$. The diamond symbol in each box represents the mean of the interval length.

More precise measurements of the confidence interval lengths can be seen in Table 6. For different values of p, using the true value of σ^2 results in similar empirical medians and standard deviations of confidence interval lengths. In comparison, estimating σ^2 brings about a growth in the medians and standard deviations compared to employing the σ^2_{true}. As p increases, the gap in the lengths of the confidence intervals between using estimated and fixed σ^2 becomes more pronounced. Another interesting finding is that among the different absolute values of β_i, $|\beta_i| = 1$ always has the largest median,

regardless of whether σ^2 is estimated or not.

$\lvert\beta_i\rvert$	$p = 200$		$p = 300$		$p = 500$	
	$\hat{\sigma}^2$	σ^2_{true}	$\hat{\sigma}^2$	σ^2_{true}	$\hat{\sigma}^2$	σ^2_{true}
0.0	0.70 (0.13)	0.59 (0.06)	0.73 (0.15)	0.60 (0.06)	0.77 (0.16)	0.60 (0.06)
0.2	0.69 (0.13)	0.60 (0.06)	0.72 (0.15)	0.61 (0.06)	0.76 (0.16)	0.61 (0.06)
0.5	0.68 (0.12)	0.60 (0.05)	0.70 (0.14)	0.60 (0.05)	0.73 (0.16)	0.61 (0.05)
1.0	0.73 (0.13)	0.67 (0.06)	0.75 (0.15)	0.67 (0.07)	0.78 (0.17)	0.67 (0.07)

Table 6: Empirical median and standard deviation of confidence interval lengths for the partial regression coefficients using $\hat{\sigma}^2$ and σ^2_{true}. Results for different dimensionalities $p \in \{200, 300, 500\}$ are included. Lengths are compared across the absolute values of the full regression coefficients, $|\beta_i| \in \{0, 0.2, 0.5, 1\}$. The standard deviation is in the brackets.

Regarding the mean coverage probability, the results are shown in Figure 10. Generally, using a fixed σ^2 always yields the desired coverage for the confidence intervals, with probabilities around 95%. However, similar to the pattern observed in Figure 7, the red line consistently lies above the blue line, regardless of whether the coefficient is active or not. This suggests that in high-dimensional settings, estimating σ^2 leads to an increase in the average coverage probabilities of the confidence intervals. Furthermore, as the dimensionality p increases, the mean coverage probability also rises, reaching a higher value than the nominal coverage probability of 95%.

6 Discussion

In conclusion, our analyses were divided into two main parts based on the relationship between n and p: the low-dimensional setting ($n > p$) and the high-dimensional setting ($n < p$). In the low-dimensional setting, we used the residual sum of squares to estimate σ^2, while in the high-dimensional setting, we adopted the refitted cross-validation method. We find that using the residual sum of squares provides reasonable estimators, whereas the refitted cross-validation estimators are always biased. The performance of the estimators is influenced by both the dimensionality and the sparsity level of the true parameter vector. Although increasing dimensionality p raises the variation of the

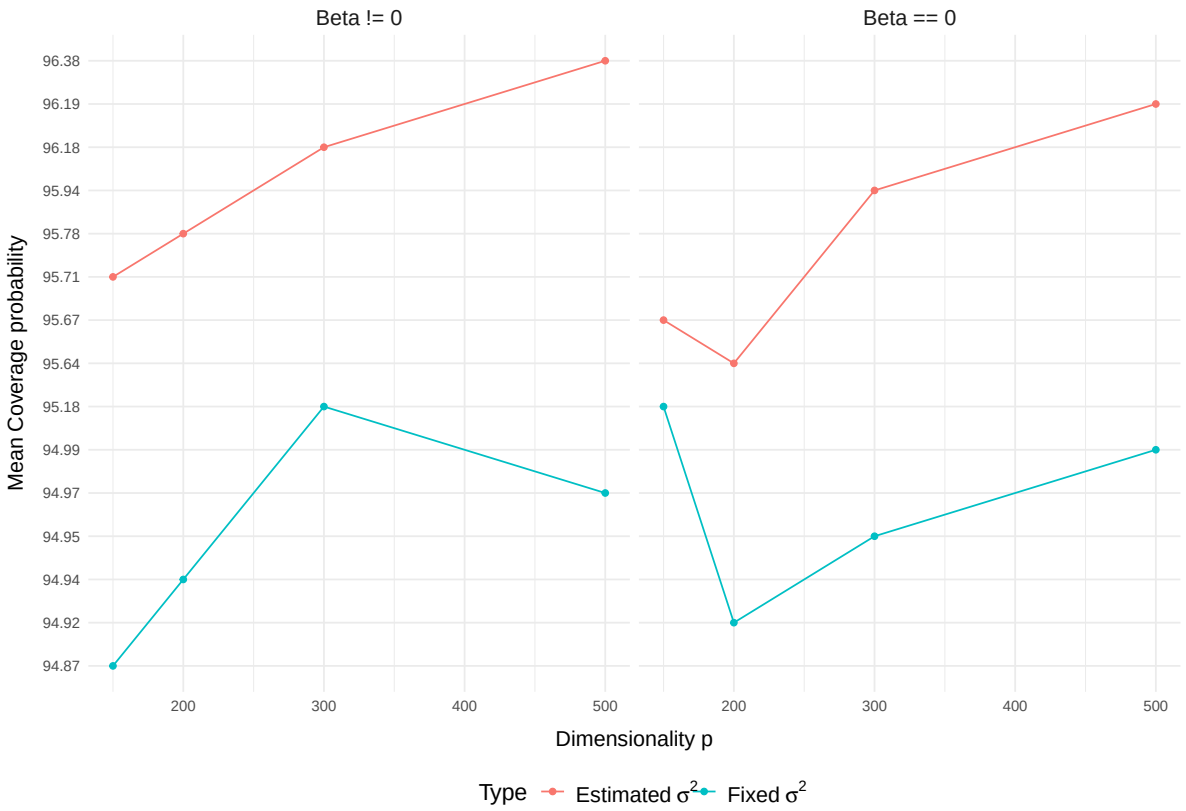

Figure 10: Mean coverage probability of the confidence intervals using $\hat{\sigma}^2$ (Estimated σ^2) and σ^2_{true} (Fixed σ^2) across different dimensionalities p. The left plot refers to the active coefficients in the true model. The right plot corresponds to the coefficients that are zero in the true model.

estimated σ^2, it does not significantly affect the median value in either low- or high-dimensional settings. In contrast, when $n < p$, a reduction in the degree of the sparsity level simultaneously increases both the median and the variation, with the impact of the sparsity level being greater than that of dimensionality.

Concerning the effect of estimating σ^2 on the inferential results based on the (U, V) decomposition, we observe that when $n > p$, using the estimated σ^2 introduces greater variation in the lengths of the confidence intervals. As p increases, the standard deviation increases as well. However, there is no such difference in the empirical median of the confidence interval lengths between using $\hat{\sigma}^2$ and σ^2_{true}. Additionally, estimating σ^2 can lead to lower coverage of the confidence intervals. In the high-dimensional setting, we notice wider confidence intervals with more variations and higher mean coverage

probabilities when using the refitted cross-validation estimator compared to using the true value of σ^2. These differences become more significant as p and s increase, with the impact of the sparsity level being more pronounced. The empirical average coverage probabilities obtained using $\hat{\sigma}^2$ are much higher than the nominal coverage probability of 95%, which indicates substantial over-coverage in the high-dimensional setting.

In the low-dimensional setting, when we have to estimate σ^2, the inferential results do not deviate considerably from those obtained when the true value of σ^2 is known. However, there are some significant differences in the lengths and coverage probabilities of the confidence intervals if we use the RCV estimator in the high-dimensional setting. These discrepancies primarily stem from the method chosen to estimate σ^2. We fixed the value of λ at 0.1 for simplicity when obtaining the RCV estimator. Exploring different values of λ in future studies may yield better $\hat{\sigma}^2$, potentially reducing the differences in the inferential results between using estimated and true value of σ^2. Although it is hard to estimate σ^2 when $n < p$, Fan et al. (2012) proposed alternative methods to RCV that could be explored for more accurate estimators, which might improve inferential results using the (U, V) decomposition. Furthermore, we defined the sparsity level as the number of non-zero elements in the true parameters. We gradually reduced the sparsity level by increasing the number of non-zero elements in true β. In our study, the absolute values of the elements added ranged between 0 and 1. Notable, when $s = 50$, $s = 100$ and $s = 150$, we introduced numerous small values. One possible choice in the future is to set the absolute values of the elements in the parameter to be relatively larger, particularly in non-sparse situations. Additionally, our simulation studies were limited to a fixed Toeplitz structure of the design matrix X. It would be valuable to use alternative covariance structures, as listed in (Zhang et al., 2022) to investigate their actual effects on the inferential results based on the (U, V) decomposition.

7 Endmatter

R code for simulation and reproducing results are available at:
https://github.com/Zichun0314/Selective-Inference

References

Berk, R., Brown, L., Buja, A., Zhang, K., and Zhao, L. (2013). Valid post-selection inference. *The Annals of Statistics*, pages 802–837.

Berk, R., Brown, L., and Zhao, L. (2010). Statistical inference after model selection. *Journal of Quantitative Criminology*, 26:217–236.

Cox, D. R. (1975). A note on data-splitting for the evaluation of significance levels. *Biometrika*, 62(2):441–444.

Fan, J., Guo, S., and Hao, N. (2012). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 74(1):37–65.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 70(5):849–911.

Faraway, J. J. (1998). Data splitting strategies for reducing the effect of model selection on inference. *Comput Sci Stat*, 30:332–41.

Kuchibhotla, A. K., Kolassa, J. E., and Kuffner, T. A. (2022). Post-selection inference. *Annual Review of Statistics and Its Application*, 9(1):505–527.

Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907 – 927.

Liu, K., Markovic, J., and Tibshirani, R. (2018). More powerful post-selection inference, with application to the lasso. *arXiv preprint arXiv:1801.09037*.

Rasines, D. G. and Young, G. A. (2023). Splitting strategies for post-selection inference. *Biometrika*, 110(3):597–614.

Tian, X. and Taylor, J. (2018). Selective inference with a randomized response. *The Annals of Statistics*, 46(2):679–710.

Zhang, D., Khalili, A., and Asgharian, M. (2022). Post-model-selection inference in linear regression models: An integrated review. *Statistics Surveys*, 16:86–136.