Review of *On the Cross-validation Bias due to Unsupervised Preprocessing*

Chandler Nielsen

1 Introduction

In this document, we review the paper *On the cross-validation bias due to unsupervised preprocessing* by Amit Moscovich and Saharon Rosset[MR22]. The aim of this work is to identify a common error committed by statisticians at all levels of skill and training. In particular, there is a widely held erroneous belief among statisticians that unsupervised transformations of feature vectors prior to data splitting will not result in biased error estimates. An unsupervised transformation $T : \mathcal{X} \to \mathcal{X}$ is a transformation constructed only from one's feature vectors $\mathbf{x}_1, \ldots, \mathbf{x}_N$. Some examples of **unsupervised transformations** include mean-centering, scaling, dimensionality reduction (like PCA), and grouping of categorical variables. In contrast, a **supervised transformation** is a transformation T depending on both the feature vectors $\mathbf{x}_1, \ldots, \mathbf{x}_N$ and the labels $y_1, \ldots, y_N$. It is well-known that involving the data labels in the construction of this transformation T will result in data leakage that will result in bias to cross-validation estimates of model performance. On the other hand, noted sources suggest that unsupervised preprocessing is safe from such bias. As indicated in [MR22], *The Elements of Statistical Learning* claims the following:

> *In general, with a multistep modeling procedure, cross-validation must be applied to the entire sequence of modeling steps. In particular, samples must be "left out" before any selection or filtering steps are applied. There is one qualification: initial unsupervised screening steps can be done before samples are left out. For example, we could select the 1000 predictors with highest variance across all 50 samples, before starting cross-validation. Since this filtering does not involve the class labels, it does not give the predictors an unfair advantage.*
>
> *While this point may seem obvious to the reader, we have seen this blunder committed many times in published papers in top rank journals. With the large number of predictors that are so common in genomic and other areas, the potential consequences of this error have also increased dramatically . . . [THF09]*

Note that the paper under review only includes the first paragraph; we have included the second for two reasons. First, it is pretty ironic that, while identifying the blunder associated with supervised preprocessing, the single most well-known work in statistical learning is advocating a blunder with regard to unsupervised preprocessing. The second paragraph is also instructive for another important reason - there is reason to suspect that those practicing machine learning on genomics data are making the same blunder that this most important work advocates. With the importance of analyzing genomics data correctly, the errors associated with this unsupervised preprocessing can be extremely costly. Indeed, the current work conducted a review of research articles published in *Science Magazine* over a period of one and a half years. During this period, the authors of the current work identified 20 publications that apply cross-validated predictive modeling. Moreover, the authors identified that seven of those papers (35%) performed some kind of unsupervised preprocessing on the entire dataset prior to cross-validation/data splitting. Thus, as was already stated, there is reason to suspect that a number of researchers are conducting cross-validation incorrectly, and their analyses of the predictive performance of their models suffer as a consequence. Moreover, as the current work also identifies, in both academia and industry, when data has been received by a statistician it has very often already undergone various stages of preprocessing. Thus, the literature review conducted by the authors likely underestimates the scope of the error in work by statisticians and practitioners.

So what do the authors recommend? What should a statistician conscious of this pitfall do to prevent themselves from committing the same mistake? The authors' summary of the correct approach to unsupervised preprocessing is provided below. Let S be the entire dataset of both feature vectors and class labels. Let S_{tr} denote the training set, while S_{val} denotes the validation set.

1. *Preprocessing.* Fit a transformation $\hat{T} : \mathcal{X} \to \tilde{\mathcal{X}}$ using just the feature vectors of the training set $\{\mathbf{x} : (\mathbf{x}, y) \in S_{tr}\}$
2. *Training.* Transform the feature vectors of S_{tr} using $\hat{T}$ and then learn a predictor $\hat{f}_{S_{tr}}$ from the transformed training set $\{(\hat{T}(\mathbf{x}), y) : (\mathbf{x}, y) \in S_{tr}\}$.
3. *Validation.* For every observation $(\mathbf{x}, y)$ in S_{val}, compute a prediction for the transformed feature vector $\hat{y} = \hat{f}_{S_{tr}}(\hat{T}(\mathbf{x}))$ and evaluate some loss function $\ell(y, \hat{y})$.

Critically, notice that the transformation T that was learned *only* on the training set is applied on the validation dataset, too. For example, suppose we compute $\hat{\mu}$ and $\hat{\sigma}$, the empirical mean and standard deviation for each of the features on our training data only, then we construct a standardizing transformation

$$\hat{T}(\mathbf{x}) = \frac{\mathbf{x} - \hat{\mu}}{\hat{\sigma}},$$

a transformation that will be applied to both the training and validation sets. If the validation set had been used to construct this T, then there would be data leakage when the predictor is constructed from the training set.

Somewhat shockingly, it appears that the full theoretical analysis provided in this paper is the first to address the biases due to unsupervised preprocessing directly. To the best of the authors' knowledge, the only other work that considers this bias is an empirical study focused on microarray analysis [RH15]. In this work, the authors investigate empirically the bias that results from data preparation before training/test set prediction error is estimated via cross-validation. They refer to this approach as incomplete cross-validation. To assess the severity of the bias due to this incomplete cross-validation, the authors define the ”Cross-validation Incompleteness Impact Measure” (CVIIM), defined as follows[RH15]:

$$\text{CVIIM} = \begin{cases} 1 - \frac{\mathbb{E}[{}_{\text{incompl, K}}(\mathbf{S})]}{\mathbb{E}[e_{\text{full, K}}(\mathbf{S})]} & \text{if } \mathbb{E}[{}_{\text{incompl, K}}(\mathbf{S})] < \mathbb{E}[e_{\text{full, K}}(\mathbf{S})] \\ 0 & \text{otherwise} \end{cases}$$

Ultimately, the authors conclude that

> *Performing normalization on the entire dataset before CV did not result in a noteworthy optimistic bias in any of the investigated cases. In contrast, when performing PCA before CV, medium to strong underestimates of the prediction error were observed in multiple settings.*

The authors of the paper reviewed here produce theoretical results that shed light on these empirical findings. These are discussed in the next section.

2 Theory and Methods

2.1 Experimental Setup

The purpose of the paper was to explore the quote from *The Elements of Statistical Learning* described in the Introduction section above. To this end, for the sake of both simulation and theory, the authors consider variance-based feature selection on the entire dataset prior to cross-validated linear regression. In this document, we only consider the artificial simulation, thereby showing that a substantial bias in the validation error can arise with respect to the model risk.

We first describe the model that is considered in the section 3 below, which serves as the basis for the primary theoretical result. The simulation experiment is built from the following:

1. *Sampling Distribution* Generate a random vector of coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$, where each of the $\beta_i \sim \mathcal{N}(0,1)$. Moreover, each observation will be produced by

$$\mathbf{x} = (Cx_1, \ldots, Cx_M, x_{M+1}, \ldots, x_p) \quad y = \mathbf{x}\boldsymbol{\beta} + \epsilon,$$

   where $x_1, \ldots, x_p$ are iid draws from a zero-mean distribution. Moreover, the constant $C > 1$ to ensure that the first M features will have a larger magnitude and thus more influence on the response. By construction, the variance of $\mathbf{x}\boldsymbol{\beta}$ is proportional to $(p-M) + C^2M$, so given a noise level $\eta > 0$, the noise term $\epsilon \sim \mathcal{N}(0, \sigma^2)$, where $\sigma^2 = \eta \cdot [(p-M) + C^2M]$

2. *Preprocessing* As stated before, we perform variance-based feature selection. The unsupervised transformation is $\hat{T}(\mathbf{x}) = (x_{j_1}, \ldots, x_{j_K})$, where $j_1, \ldots, j_K$ are the K covariates with highest empirical variance. As stated before, the empirical variances will be computed using the entire dataset.

3. *Predictor* We perform ordinary least squares with no intercept: $\hat{f}\{\hat{T}(\mathbf{x})\} = (x_{j_1}, \ldots, x_{j_k})\hat{\boldsymbol{\beta}}$, where

$$\hat{\boldsymbol{\beta}} \coloneqq \operatorname*{arg\,min}_{\boldsymbol{\beta} \in \mathbb{R}^K} \sum_{i=1}^{n} (\beta_1 x_{j_1} + \cdots + \beta_K x_{j_K} - y_i)^2$$

After describing some numerical results (described below), the authors seek to determine the source of the bias. To this end, the authors consider the noiseless setting ($\eta = 0$) with independent and identically distributed $\mathcal{N}(0,1)$ random variables. Moreover, we set $K = 1$; in other words, the number of selected variables with highest variance is equal to one. Moreover, to simplify analysis, the authors consider the single variable with the largest squared norm instead of the largest variance. Asymptotically, these are the same since the covariates are drawn from a zero-mean distribution.

2.2 The Main Theorem

Let $X \in \mathbb{R}^{(n+m)\times p}$ be the matrix of feature vectors, where the first n rows correspond to the training set. Let $X_{1:n,j}$ be the vector containing all observations of the j^{th} feature in the training set, and let $X_{n+1:n+m,j}$ be the corresponding vector for the validation set. Recall that p is the total number of features. Finally, denote the normalized dot product between the j^{th} and the k^{th} features on the training set by

$$\hat{\rho}_{jk} = \frac{X_{1:n,j}^\top X_{1:n,k}}{\|X_{1:n,j}\| \|X_{1:n,k}\|}$$

The authors prove the following theorem, which is an expression for the bias due to the preprocessing step. Moreover, the authors analyze the asymptotic behavior of this bias.

Theorem 1 (Moscovich and Rosset). *Let $\hat{\jmath}$ be the maximizer of $\sum_{i=1}^{n+m} X_{i,j}^2$ and let j_o be any other column (in our setup, they are exchangeable). For the model described above with $K = 1$ and $M = \eta = 0$, the bias of the MSE due to the preliminary feature selection is*

$$\mathit{bias} = \mathbb{E}\left[\left\{(p-1)\hat{\rho}_{\hat{\jmath}j_o}^2 \frac{A_{j_o}}{A_{\hat{\jmath}}} - 1\right\}\left(X_{n+1,\hat{\jmath}}^2 - 1\right)\right]$$

where $A_j \coloneqq \|X_{1:n,j}\|^2$. From this, we can infer the following asymptotic results:

a. If $n \to \infty$ and $p < n^\alpha$ for some $\alpha < 3/2$, then bias $\to 0$.

b. If we fix n and take $p \to \infty$ then $\frac{bias}{p/n} \to 1$ and in particular bias $\to \infty$.

We prove the expression for the bias and the first asymptotic result for the sake of brevity. The reader is encouraged to review the entire proof, as it is quite illustrative.

Proof of Main Expression. First, recall that the columns are interchangeable and assume without loss of generality that $\hat{\jmath} = 1$. In other words, the first variable was the one with the highest empirical variance. In this case, our preprocessed design matrix is given by $\tilde{X} = X_{1:n,1} \in \mathbb{R}^n$. Let $Y = (y_1, \dots, y_n)^\top$ be the responses recorded in the training set. Conditioned on $\hat{\jmath} = 1$, the estimated regression coefficient is

$$\hat{\beta}_1 = \left(\tilde{X}^\top \tilde{X}\right)^{-1} \tilde{X}^\top Y = \frac{X_{1:n,1}^\top Y}{\|X_{1:n,1}\|^2}$$

For the noiseless model, $Y = \sum\limits_{j=1}^{p} \beta_j X_{1:n,j}$. Plugging this into the foregoing equation, we have

$$\hat{\beta}_1 = \frac{X_{1:n,1}^\top \sum_{j=1}^{p} \beta_j X_{1:n,j}}{\|X_{1:n,1}\|^2} = \beta_1 + \sum_{j=2}^{p} \beta_j Z_j, \text{ where } Z_j = \frac{X_{1:n,1}^\top X_{1:n,1}}{\|X_{1:n,1}\|^2}$$

It follows that the prediction for the first observation in the dataset, conditioned on $\hat{\jmath} = 1$, is

$$\hat{y}_{n+1} = \hat{\beta}_1 X_{n+1,1} = \left(\beta_1 + \sum_{j=2}^{p} \beta_j Z_j\right) X_{n+1,1}$$

Meanwhile, the noiseless model satisfies $y_{n+1} = \sum_{j=1}^{p} \beta_j X_{n+1,j}$. Therefore, we have the following:

$$\begin{aligned}
\text{MSE} &= \mathbb{E}\left[(\hat{y}_{n+1} - y_{n+1})^2 | \hat{\jmath} = 1\right] \\
&= \mathbb{E}\left[\left\{\sum_{j=2}^{p} \beta_j (Z_j X_{n+1,1} - X_{n+1,j})\right\}^2 \middle| \hat{\jmath} = 1\right] \\
&= \sum_{j=2}^{p} \sum_{\ell=2}^{p} \mathbb{E}\left[\beta_j \beta_\ell (Z_j X_{n+1,1} - X_{n+1,j})(Z_\ell X_{n+1,1} - X_{n+1,\ell}) | \hat{\jmath} = 1\right] \\
&= \sum_{j=2}^{p} \sum_{\ell=2}^{p} \mathbb{E}\left[\beta_j \beta_\ell\right] \mathbb{E}\left[(Z_j X_{n+1,1} - X_{n+1,j})(Z_\ell X_{n+1,1} - X_{n+1,\ell}) | \hat{\jmath} = 1\right]
\end{aligned}$$

where the last line follows from the independence of our construction of β_j and X. Moreover, since $\beta_1, \dots, \beta_p \overset{iid}{\sim} \mathcal{N}(0,1)$, it follows that $\mathbb{E}\beta_j \beta_\ell = \delta_{j\ell}$. In that case, the foregoing becomes

$$\begin{aligned}
\text{MSE} &= \sum_{j=2}^{p} \mathbb{E}\left[(Z_j X_{n+1,1} - X_{n+1,j})^2 | \, \hat{\jmath} = 1\right] \\
&= \sum_{j=2}^{p} \left(\underbrace{\mathbb{E}\left[Z_j^2 X_{n+1,1}^2 | \, \hat{\jmath} = 1\right]}_{\text{I}} - 2 \underbrace{\mathbb{E}\left[Z_j X_{n+1,1} X_{n+1,j} | \, \hat{\jmath} = 1\right]}_{\text{II}} + \underbrace{\mathbb{E}\left[X_{n+1,j}^2 | \, \hat{\jmath} = 1\right]}_{\text{III}} \right)
\end{aligned}$$

The term $\text{II} = 0$ by the following symmetry argument. Let $X' \in \mathbb{R}^{(n+m) \times p}$ be the same matrix with the sign of $X_{n+1,1}$ flipped. Since we assumed the covariates are iid $\mathcal{N}(0,1)$, we know that $X \overset{d}{=} X'$. Meanwhile, Z_j depends only on the first n observations, so it isn't affected by the sign flip. Therefore, $\mathbb{E}[Z_j X_{n+1,1} X_{n+1,j} | \, \hat{\jmath} = 1] = \mathbb{E}[Z_j(-X_{n+1,1}) X_{n+1,j} | \, \hat{\jmath} = 1]$. The term III can be reduced as follows:

$$\sum_{j=2}^{p} \mathbb{E}[X_{n+1,j}^2 | \, \hat{\jmath} = 1] = \sum_{j=1}^{p} \mathbb{E} X_{n+1,j}^2 - \mathbb{E}[X_{n+1,1}^2 | \, \hat{\jmath} = 1] = p - \mathbb{E}[X_{n+1,1}^2 | \, \hat{\jmath} = 1]$$

Finally, conditioned on $\hat{\jmath} = 1$, the remaining columns are identically distributed, so we have

$$\sum_{j=2}^{p} \mathbb{E}[Z_j^2 X_{n+1,1}^2 | \, \hat{\jmath} = 1] = (p-1)\mathbb{E}[Z_2^2 X_{n+1,1}^2 | \, \hat{\jmath} = 1] = (p-1)\mathbb{E}\left[\hat{\rho}_{1,2}^2 \frac{A_2}{A_1} X_{n+1,1}^2 \middle| \, \hat{\jmath} = 1\right]$$

Where this follows from the definition of $\hat{\rho}_{1,2}^2$. Thus, the expected validation error is given by

$$\mathbb{E}e_{\text{val}} = \mathbb{E}(\hat{y}_{n+1} - y_{n+1})^2 = (p-1)\mathbb{E}\left[\hat{\rho}_{1,2}^2 \frac{A_2}{A_1} X_{n+1,1}^2 \middle| \, \hat{\jmath} = 1\right] + p - \mathbb{E}[X_{n+1,1}^2 | \, \hat{\jmath} = 1]$$

Now, for a fresh observation $(\mathbf{x}, y) \in \mathbb{R}^p \times \mathbb{R}$, the derivation is exactly the same, except the new observation is independent of the training and validation data. This independence is also applied below. In particular, the expected value of the generalization error is given by

$$\begin{aligned}
\mathbb{E}e_{\text{gen}} &= \mathbb{E}[(\hat{y} - y)^2] = \mathbb{E}[(\hat{y} - y)^2 | \, \hat{\jmath} = 1] \\
&= (p-1)\mathbb{E}\left[\hat{\rho}_{1,2}^2 \frac{A_2}{A_1} \middle| \, \hat{\jmath} = 1\right] \mathbb{E}[x_1^2 | \hat{\jmath} = 1] + p - \mathbb{E}[x_1^2 | \hat{\jmath} = 1] \\
&= (p-1)\mathbb{E}\left[\hat{\rho}_{1,2}^2 \frac{A_2}{A_1} \middle| \, \hat{\jmath} = 1\right] + p - 1
\end{aligned}$$

Where we have applied the fact that $x_i \sim \mathcal{N}(0,1)$. Thus, we have derived the expression

$$\text{bias} = (p-1)\mathbb{E}\left[\hat{\rho}_{1,2}^2 \frac{A_2}{A_1}(X_{n+1,1}^2 - 1) \middle| \hat{\jmath} = 1\right] - \mathbb{E}\left[X_{n+1,1}^2 | \hat{\jmath} = 1\right] + 1$$

Taking the expectation of both sides and applying the tower property, we obtain the desired result. □

We now turn to a proof of the first asymptotic property. As before, we assume that $\hat{\jmath} = 1$ without loss of generality; moreover, we assume that $j_o = 2$ without loss of generality. Since the X_{ij} are iid $\mathcal{N}(0,1)$, the distribution of column vectors is spherically symmetric; therefore, we can write $X_{1:N,j} = \|X_{1:N,j}\| U_j$, where $U_j = X_{1:N,j} / \|X_{1:N,j}\|$. Note that $\|X_{1:N,j}\|^2 \sim \chi^2(N)$ and that U_j is uniformly distributed on the unit sphere. Moreover, since information about the magnitudes gives the statistician no information about the directions on the sphere, $\|X_{1:N,1}\|, \ldots, \|X_{1:N,p}\|$ are $U_1, \ldots, U_p$ are all independent. Moreover, $U_1, \ldots, U_p$ are uniformly distributed on the unit sphere and remain so even after conditioning on $\hat{\jmath} = 1$. Finally, to prove the first asymptotic result, we need the following lemmas. These lemmas are provided in the paper but are omitted here for the sake of brevity.

Lemma 1. Let $Q_1, \ldots, Q_p \overset{iid}{\sim} \chi^2(N)$. Then for any $0 < \epsilon < 1$,

$$\mathbb{E}\left[\max_i Q_i\right] \leq \frac{N}{1-\epsilon} + \frac{2 \ln p}{\epsilon}$$

□

Lemma 2. Let $Z \in \mathbb{R}^N$ be a random vector uniformly distributed on the unit sphere. Then for any ℓ,

$$\mathbb{E}[Z_\ell^2] = 1/N$$

□

With all of this preparation, we can now prove asymptotic result a.

Proof of Asymptotic Result a. Denote

$$b_1 := \mathbb{E}\left[\hat{\rho}_{1,2}^2 \frac{A_2}{A_1}(X_{n+1,1}^2 - 1) \middle| \, \hat{\jmath} = 1\right], \quad b_2 := \mathbb{E}\left[X_{n+1,1}^2 - 1 | \hat{\jmath} = 1\right]$$

By the foregoing proof, we know we can write bias $= (p-1)b_1 - b_2$. Note that $\mathbb{E}[X_{n+1,1}^2] = 1$ without conditioning. Therefore, if the column with the maximum empirical variance is the first, then we must have $\mathbb{E}[X_{n+1,1}^2 | \hat{\jmath} = 1] \geq 1$. Thus, $b_2 \geq 0$. As was done previously, write $X_{1:N,1} = \|X_{1:N,1}\| Z$, where $Z \in \mathbb{R}^N$ is a unit vector. Moreover, since all entries of the vectors in X are iid standard normal draws (and thus Z is rotationally symmetric / is uniform on the unit sphere), we must have $Z \perp\!\!\!\perp \|X_{1:N,1}\|$. Therefore, applying independence and the second lemma above, we must have

$$b_2 = \mathbb{E}[\|X_{1:N,1}\|^2|\hat{\jmath}=1]\mathbb{E}[Z_{n+1}^2|\hat{\jmath}=1]-1 = \frac{1}{N}\mathbb{E}[\|X_{1:N,1}\|^2|\hat{\jmath}=1]-1$$

Combining this result with Lemma 1, we obtain, for all $\epsilon \in (0,1)$,

$$0 \leq b_2 \leq \frac{\epsilon}{1-\epsilon} + \frac{2\ln p}{\epsilon N}$$

We now turn to b_1. Let $V_j \coloneqq X_{1:n,j}/\|X_{1:n,j}\|$. Again, as stated above, V_1 and V_2 are rotationally symmetric, and the variable selection (conditioning on $\hat{\jmath}$) is invariant under such rotations. Thus,

$$b_1 = \mathbb{E}\left[\langle V_1, V_2\rangle^2 \cdot \frac{A_2}{A_1} \cdot (X_{n+1,1}^2 - 1) \mid \hat{\jmath}=1\right]$$

By rotational symmetry, we can assume without loss of generality that $V_1 = e_1 = (1,0,\ldots,0)$, thus $\mathbb{E}[\langle V_1, V_2\rangle^2] = \mathbb{E}[\langle V_1, e_1\rangle^2]$. Applying Lemma 2, this implies that $\mathbb{E}[\langle V_1, V_2\rangle^2] = 1/n$. Finally, note that since $\hat{\jmath}=1$, this must mean that A_2 is stochastic smaller than A_1. Thus, we finally have the following:

$$\begin{aligned}
b_1 &\leq \frac{1}{n}\mathbb{E}\left[1 \cdot \left(\|X_{1:N,1}\|^2 Z_{n+1}^2 - 1\right) \mid \hat{\jmath}=1\right] \\
&= \frac{1}{n}\mathbb{E}\left[\|X_{1:N,1}\|^2 \mid \hat{\jmath}=1\right]\mathbb{E}\left[Z_{n+1}^2\right] - \frac{1}{n} \\
&= \frac{1}{Nn}\mathbb{E}\left[\|X_{1:N,1}\|^2 \mid \hat{\jmath}=1\right] - \frac{1}{n} \\
&\leq \frac{\epsilon}{n(1-\epsilon)} + \frac{2\ln p}{\epsilon N n}
\end{aligned}$$

Thus, for any $0 < \epsilon < 1$,

$$|\text{bias}| = |(p-1)b_1 - b_2| \leq (p-1)|b_1| + |b_2| \leq \frac{n-p-1}{n}\left(\frac{\epsilon}{1-\epsilon} + \frac{2\ln p}{\epsilon N}\right)$$

Finally, pick $\epsilon(n) = \sqrt{\ln p/n}$. Then the foregoing becomes

$$\begin{aligned}
\frac{n+p-1}{n}\left(\frac{\sqrt{\ln p/n}}{1-\sqrt{\ln p/n}} + \frac{2\ln p}{N\sqrt{\ln p/n}}\right) &\asymp \frac{n+p-1}{n}\left(\frac{\sqrt{\ln p}}{\sqrt{n}-\sqrt{\ln p}} + \frac{2\sqrt{\ln p}}{\sqrt{n}}\right) \\
&= \frac{n+p-1}{n}\underbrace{\left(\frac{3\sqrt{n\ln p} - 2\ln p}{n - \sqrt{n\ln p}}\right)}_{\asymp\sqrt{\frac{\ln p}{n}}}
\end{aligned}$$

Thus, the term with which we must concern ourselves involves p/n outside of the parentheses. Thus, we need

$$\frac{p}{n}\sqrt{\frac{\ln p}{n}} \sim \frac{p\sqrt{\ln p}}{n^{3/2}} \xrightarrow{n\to+\infty} 0$$

Thus, if $p \sim n^\alpha$ with $\alpha < 3/2$, we have the desired result. This concludes the proof. □

The proof of the second asymptotic result is arguably more interesting than the proof just provided, but it is omitted here due to a concern over space. We next illustrate how the bias behaves using the simulation procedure described at the beginning of this section.

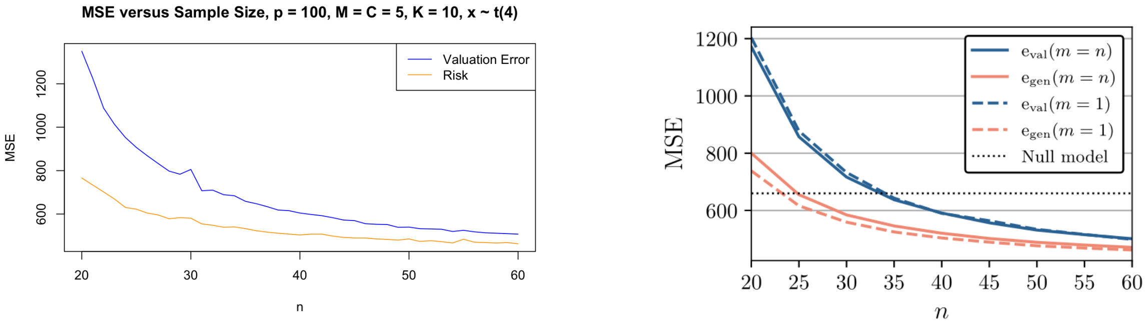
3 Numerical Results

We now conduct a numerical experiment wherein we run the procedure described in section 2.1. We first aim to reproduce the results in the paper, after which we alter the parameters a bit to see how this affects the results. We restricted these simulations only to the main example provided in the paper. All of the code used to produce these plots is provided in the R markdown file associated with this document. The reader will find code for both the incorrect and correct preprocessing procedures in addition to code used to produce the plots below.

As just mentioned, we have two data pipelines described above: one correct and the other incorrect. We input the following parameters into both data pipelines:

- num_runs corresponds to the number of times a dataset is generated for each sample size.
- K corresponds to the number of covariates that are selected for training and testing in both the correct and incorrect data pipelines.
- p is the number of parameters in our model before feature selection is performed.
- $C > 1$ is the constant that is multiplied with some of our features to give these features larger influence.
- M corresponds to the number of covariates that are multiplied by this constant C
- η is the noise variance. For all simulations performed on this dataset, $\eta = 1$.

Moreover, note that we perform leave-one-out cross-validation for all simulations. We compare our plots to those produced in our paper. The first such result is provided in Figure 1 below.



(a) Our Plot

(b) Paper Plot

Figure 1: These plots illustrate our reproduction of the simulation conducted in *On the cross-validation bias due to unsupervised preprocessing* for parameters $p = 100$, $M = C = 5$, $K = 10$, and the covariates $x \sim t(4)$, a t-distribution with 4 degrees of freedom. Note that our plot was produced for 1000 runs, while the plot produced by the paper is 100,000 runs.

Importantly, note that our results were produced for 1000 runs of leave-one-out cross-validation. As a consequence, there are irregularities that are present in our plot that are not present in the plot produced by the authors of our paper. Indeed, the curves on the right are the pointwise means for a large number of runs.

To illustrate this, we produced a plot for only 100 runs at this setting. We run our simulation when the settings are the same as above but $x \sim \mathcal{N}(0, 1)$. This is illustrated in Figure 2 above.

When we use 1000 runs as before, we obtain results very similar to those produced by the paper. This is illustrated below:

Note critically that for all of the simulations considered here, validation error overestimates the risk; in other words, the bias is positive. It may seem counterintuitive that this should be so - indeed, the incorrect procedure sees more of the dataset than the correct procedure, so one might naively expect the MSE to be smaller for the validation error than for the risk. This is a consequence of the feature selection procedure. Because we select the highest-variance features on the entire dataset in

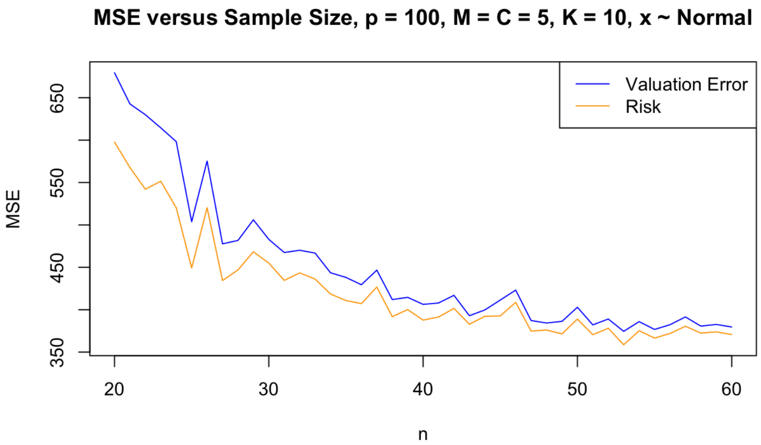
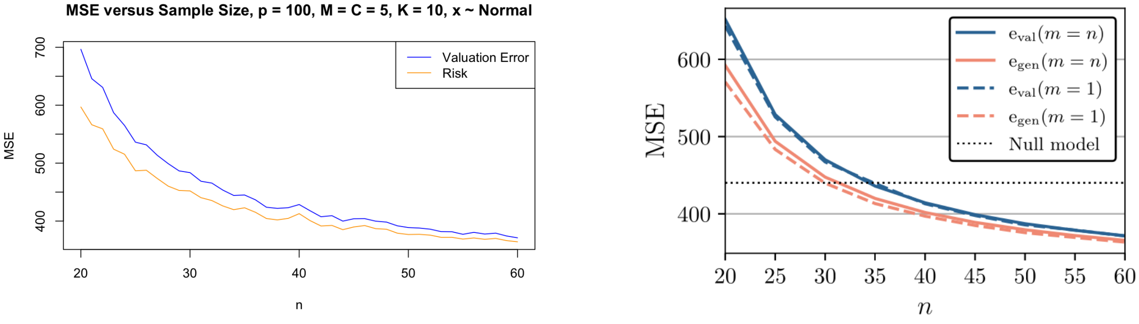


Figure 2: Mean Squared Error versus Sample Size with $p = 100$, $M = C = 5$, $K = 10$, and $x \sim \mathcal{N}(0, 1)$, and 100 total samples generated for each sample size.



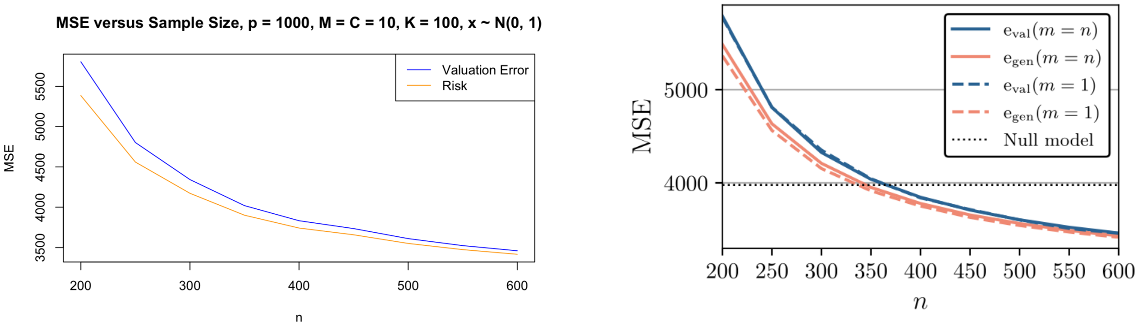
(a) Our Plot

(b) Paper Plot

Figure 3: Settings $p = 100$, $M = C = 5$, $K = 10$, and $x \sim \mathcal{N}(0, 1)$

the incorrect data pipeline, this variance propagates through the model and implies excess variance in the validation error, resulting in a positive bias.

We provide one more group of settings to produce the plots in the main example of the paper. In particular, we do 1000 runs for which $p = 1000$, $K = 100$, $M = 10$, $C = 10$, and $x \sim \mathcal{N}(0, 1)$. These results are provided below:



(a) Our Plot

(b) Paper Plot

Figure 4: Settings $p = 1000$, $M = C = 10$, $K = 100$, and $x \sim \mathcal{N}(0, 1)$

Note that the bias is not always positive. In section 5 of the paper, the authors consider the bias in the validation error due to improper preprocessing when the LASSO is performed. In particular, the authors consider a data generating procedure with $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ with $\beta_i \sim \mathcal{N}(0,1)$. Moreover, the observations $(\mathbf{x}, y)$ are drawn in the following manner:

$$\mathbf{x} \sim \mathcal{N}(0, I_{p\times p}), \qquad y = \mathbf{x}\boldsymbol{\beta} + \epsilon \quad \text{where } \epsilon \sim \mathcal{N}(0, \sigma^2)$$

Instead of highest-variance feature selection as before, the authors scale each covariate by $\hat{\sigma}_j$, where

$$\hat{\sigma}_j \coloneqq \frac{1}{n+m} \sum_{i=1}^{n+m} x_{i,j}^2$$

Note that these are computed using the entire dataset. Finally, the authors perform the LASSO in the usual way:

$$\hat{\boldsymbol{\beta}}_{\text{LASSO}} \coloneqq \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \frac{1}{2}\|Y - \tilde{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_1,$$

where $\tilde{X} \in \mathbb{R}^{n\times p}$ is the design matrix with the j^{th} feature rescaled by $\hat{\sigma}_j$.

The following plots illustrate the bias with respect to the generalization error for this model setup. We have the following:

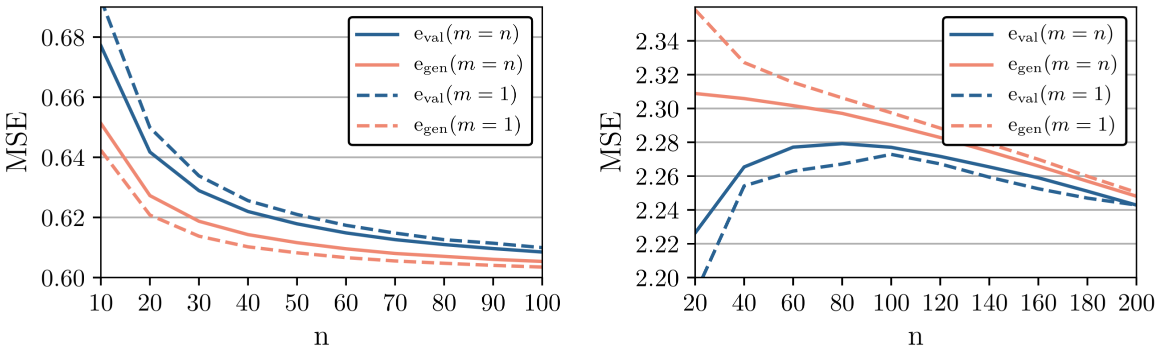


Figure 5: Validation and Generalization Errors for rescaled data prior to LASSO. The left plot corresponds to the setting $p = 5$, $\lambda = 0.5$, and $\sigma = 0.1$ averaged over 10 million runs. The right hand plot corresponds to a high-dimensional setting for which $p = 10000$, $\lambda = 0.1$, $\sigma = 1$ averaged over 1000000 runs. Note that the bias can be positive or negative, depending on the settings.

The settings are provided in the caption of the plot. The critical object of note is that, depending on the settings, the bias of the validation error with respect to the generalization error may be positive or negative depending on these settings. Thus, should a statistical practitioner perform cross-validation incorrectly, they cannot even be certain of the bias of their model error. In turn, this can result in the selection of a model which is inferior to the one produced if preprocessing were performed correctly. This is potentially very costly in high-dimensional settings like genomics. This concludes our study of this phenomenon via simulation.

4 Conclusion

It is important to reflect on the implications of this paper. There are a large number of statistical works, including those used by students, that make incorrect claims regarding methods as basic as cross-validation. This is a pretty serious problem, since there are a number of statistics-adjacent subjects like genomics for which incorrect analyses can have consequences for human health and well-being. Moreover, this paper's simplicity reinforces the importance of critical thinking and asking questions of the most basic methods and results. The method for performing cross-validation is taken for granted by a large number of statisticians and statistical practitioners, and this paper illustrates the extent to which one's analysis of a model and model selection can be inaccurate if preprocessing and cross-validation are performed incorrectly.

We saw that, should cross-validation be performed incorrectly, the bias can be positive or negative depending on the setup of the experiment. Moreover, we saw that if n is fixed and p is made large, the bias in the mean squared error can be made arbitrarily large. As already mentioned, this has serious implications for high-dimensional settings, such as in the analysis of genomics data.

We encourage the reader of this document to explore the paper for themselves and carefully consider its ideas. It was invaluable for our understanding of statistical practice, and the paper illustrates the danger of performing cross-validation incorrectly with a number of powerful examples that we did not have the space or time to detail here. It was excellent for illustrating the techniques learned in our Statistical Learning course and permitted the formation of connections between Statistical Learning and Advanced Regression analysis.

References

[MR22] Amit Moscovich and Saharon Rosset. On the cross-validation bias due to unsupervised preprocessing. *JRSSB*, 84(4):1474–1502, 2022.

[RH15] et al. Roman Hornung. A measure of the impact of cv incompleteness on prediction error estimation with application to pca and normalization. *BMC Medical Research Methodology*, 15(1):95, 2015.

[THF09] Robert Tibshirani Trevor Hastie and JH Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, 2 edition, 2009.