Lecture 7
Linear Model Selection and Regularization

CHUNG-MING KUAN

Department of Finance & CRETA

National Taiwan University

April 12, 2020

Lecture Outline

Introduction

- Despite that linear models are often questioned if they can properly characterize the behavior of the variable of interest, they remain a most popular choice in applications because they are easy to interpret and capable of reasonable predictions.
- On the other hand, finding a proper nonlinear model is always a challenging task; this difficulty becomes more severe when many potential regressors are available. Moreover, a specific nonlinear model may perform well for one data set, but such performance does not generalize to other data sets (or a validation set).
- If we want to confine ourselves with linear models with many potential regressors, it is important to determine a subset of most relevant regressors, based on some model selection criteria or regularization methods.

Model Selection Criteria

Given the model with p potential regressors, let $\hat{u}_{p,i}$ denote the i th OLS residual and $\tilde{\sigma}_p^2 = \sum_{i=1}^n \hat{u}_{p,i}^2/n$. Then, for the model with k out of these p regressors, the C_p measure defined by JWHT is:

$$C_p = \frac{1}{n}\left(\text{RSS}_k + 2k\tilde{\sigma}_p^2\right) = \tilde{\sigma}_p^2\left(\frac{\tilde{\sigma}_k^2}{\tilde{\sigma}_p^2} + \frac{2k}{n}\right),$$

where $\text{RSS}_k = \sum_{i=1}^n \hat{u}_{k,i}^2$, with $\hat{u}_{k,i}$ the i th OLS residual from the model with k regressors, and $\tilde{\sigma}_k^2 = \text{RSS}_k/n$. As $\tilde{\sigma}_p^2$ does not depend on k, this measure is in effect determined by the terms in the parentheses, which represent the trade-off between model complexity (in terms of the number of regressors, k) and model fitness (in terms of $\tilde{\sigma}_k^2$). We select the model with the smallest C_p in a collection of candidate models.

Note that Mallows' original C_p is slightly different:

$$\text{Mallows' } C_p = \frac{\text{RSS}_k}{\tilde{\sigma}_p^2} - n + 2k = n\left(\frac{\tilde{\sigma}_k^2}{\tilde{\sigma}_p^2} + \frac{2k}{n} - 1\right),$$

but it is also virtually determined by the first two terms in the parentheses; note that multiplying the constant n and subtracting the constant 1 do not affect the selection results.

Information Criteria

In the context of maximum likelihood, there are various information criteria for model selection. Let $L_n(\boldsymbol{\theta}_k)$ denote the log-likelihood function of n observations with k parameters $\boldsymbol{\theta}_k$ and $L_n(\tilde{\boldsymbol{\theta}}_k)$ its value based on the MLE $\tilde{\boldsymbol{\theta}}_k$. The Akaike information criterion (AIC) is

$$\text{AIC} = -2L_n(\tilde{\boldsymbol{\theta}}_k) + 2k,$$

and the Bayesian (Schwarz) information criterion (BIC or SIC) is:

$$\text{BIC} = -2L_n(\tilde{\boldsymbol{\theta}}_k) + \ln(n)k.$$

These criteria consider a balance between model complexity and likelihood value but use different penalty terms. We select the model with the smallest AIC/BIC from a collection of candidate models.

Recall that for the linear model $y = \boldsymbol{x}'\boldsymbol{\beta}_k + u$ with Gaussian error u,

$$L_n(\boldsymbol{\theta}_k) \propto -\frac{n}{2}\ln(\sigma^2) - \sum_{i=1}^{n}\frac{(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2}{2\sigma^2}.$$

Plugging the MLE $\tilde{\boldsymbol{\theta}}_k = (\tilde{\boldsymbol{\beta}}_k', \tilde{\sigma}_k^2)'$, where $\tilde{\sigma}_k^2 = \sum_{i=1}^{n}(y_i - \boldsymbol{x}_i'\tilde{\boldsymbol{\theta}}_k)^2/n$, into $L_n(\boldsymbol{\theta}_k)$, we have

$$L_n(\tilde{\boldsymbol{\theta}}_k) = -\frac{n}{2}\ln(\tilde{\sigma}_k^2) - \frac{n}{n}\sum_{i=1}^{n}\frac{(y_i - \boldsymbol{x}_i'\tilde{\boldsymbol{\beta}}_k)^2}{2\tilde{\sigma}_k^2} = -\frac{n}{2}\ln(\tilde{\sigma}_k^2) - \frac{n}{2}.$$

It follows that

$$\text{AIC} = -2L_n(\tilde{\boldsymbol{\theta}}_k) + 2k = n[\ln(\tilde{\sigma}_k^2) + 1 + 2k/n].$$

As such, model selection may be based on $\ln(\tilde{\sigma}_k^2) + 2k/n$. Similarly, the BIC may be determined as $\ln(\tilde{\sigma}_k^2) + \ln(n)k/n$.

Note that if we consider the log-likelihood function with a known σ_o^2,

$$L_n(\boldsymbol{\beta}_k) \propto -\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}_k)^2/(2\sigma_o^2),$$

and $L_n(\tilde{\boldsymbol{\beta}}_k) = -n\tilde{\sigma}_k^2/(2\sigma_o^2)$, so that

$$\text{AIC} = -2L_n(\tilde{\boldsymbol{\beta}}_k) + 2k = n\left(\frac{\tilde{\sigma}_k^2}{\sigma_o^2} + \frac{2k}{n}\right).$$

Replacing σ_o^2 with $\tilde{\sigma}_p^2$, we obtain the AIC in JWHT (2013, p. 212). Similarly, BIC may be determined as $\tilde{\sigma}_k^2/\tilde{\sigma}_p^2 + \ln(n)k/n$.

Note: BIC gives a higher penalty on model complexity than AIC (when $n > 7$) and hence usually results in a simpler model.

Model Selection Criteria or Cross Validation?

In model selection procedures, model with the "best" performance may be determined by a selection criteria (C_p, AIC, or BIC) or by cross validation. As discussed in the previous lecture, cross validation is computationally more demanding, but it (unlike C_p, AIC and BIC) requires less assumptions on the true model and data. JWHT (2013) suggest that, when computation is no longer a concern, "cross-validation is a very attractive approach for selecting from among a number of models under consideration" (p. 213).

Best Subset Selection

Let $\mathcal{M}_k$ denote the linear model with k out of p potential regressors and $\mathcal{M}_0$ the model with the constant term only. Best Subset Selection is a procedure searches among 2^p candidate models and finds the one with the smallest value of some model selection criterion.

Best Subset Selection Procedure

1. Starting from $k = 1$, fit all $\binom{p}{k}$ $\mathcal{M}_k$ models and denote the one with the highest R^2 as $\mathcal{M}_k^*$.
2. Repeat the first step with $k = 2, 3, \ldots, p$, and find $\mathcal{M}_2^*, \ldots, \mathcal{M}_p^*$.
3. Among $\mathcal{M}_0^*$, $\mathcal{M}_1^*, \ldots, \mathcal{M}_p^*$, select the one with the smallest C_p, AIC, BIC, or Cross Validation error.

- Best subset selection is computationally costly because it has to search among a very large number of candidate models. For $p = 10$, there are about 1,000 candidate models; when $p = 20$, the number of candidate models grows over a million. This approach is practically infeasible when $p > 40$.
- Another drawback: It is more likely to find an over-fitting model (the number of regressors is more than the model truly needs) when searching in a large space of models. An over-fitting model tends to perform well on training data but very poorly on test data.

Forward Stepwise Selection

A computationally simpler approach to selecting a model is Forward Stepwise Selection which starts from the simplest model and at each step searches among those models that add one regressor to the best model in the previous step. This procedure in effect searches in a much smaller space of candidate models.

Forward Stepwise Selection Procesure

1. Starting from $k = 0$, fit all $p - k$ models that add one regressor to the model $\mathcal{M}_k^*$ and denote the one with the highest R^2 as $\mathcal{M}_{k+1}^*$.
2. Repeat the first step for $k = 1, \ldots, p-1$ and find $\mathcal{M}_2^*, \ldots, \mathcal{M}_p^*$.
3. Among $\mathcal{M}_0^*$, $\mathcal{M}_1^*, \ldots, \mathcal{M}_p^*$, select the one with the smallest C_p, AIC, BIC, or Cross Validation error.

- Forward Stepwise Selection fits only $1 + p(p+1)/2$ models and is more feasible (than Best Subset Selection) when p is large. For example, when $p = 30$, the Best Subset Selection searches among $1,073,741,842$ models, while Forward Stepwise Selection fits only 466 models.
- Note that Forward Stepwise Selection need not find the best possible model even with infinitely many data are available. For instance, it may be the case that the best possible one-variable model contains x_1 and the best possible two-variable model contains x_2 and x_3. In this case, Forward Stepwise Selection would never find the best two-variable model, because it would find the model with x_1 and some other regressor.

Backward Stepwise Selection

Starting from the most complex model with p regressors, Backward Stepwise Selection at each step searches among the models that take one regressor out of the best model in the previous step. Similar to Forward Stepwise Selection, this procedure searches among $1 + p(p+1)/2$ models.

Backward Stepwise Selection Procedure

1. Starting from $k = p$, fit all k models that take out one regressor from the model $\mathcal{M}_k^*$ and denote the one with the highest R^2 as $\mathcal{M}_{k-1}^*$.
2. Repeat the first step for $k = p-1, \ldots, 1$ and find $\mathcal{M}_{p-2}^*, \ldots, \mathcal{M}_0^*$.
3. Among $\mathcal{M}_0^*, \mathcal{M}_1^*, \ldots, \mathcal{M}_p^*$, select the one with the smallest C_p, AIC, BIC, or Cross Validation error.

Example: Taiwan Criminal Data (Subset Selection)

We consider Taiwan's criminal and employment data from March 2001 through July 2018. These monthly data are taken from Department of Statistics, Ministry of Interior of Taiwan.

- `Theft`: This is the variable of interest y, which is the monthly theft incidents occurred in Taiwan.
- `Clearance`: Clearance rate of criminal cases, which is calculated by:

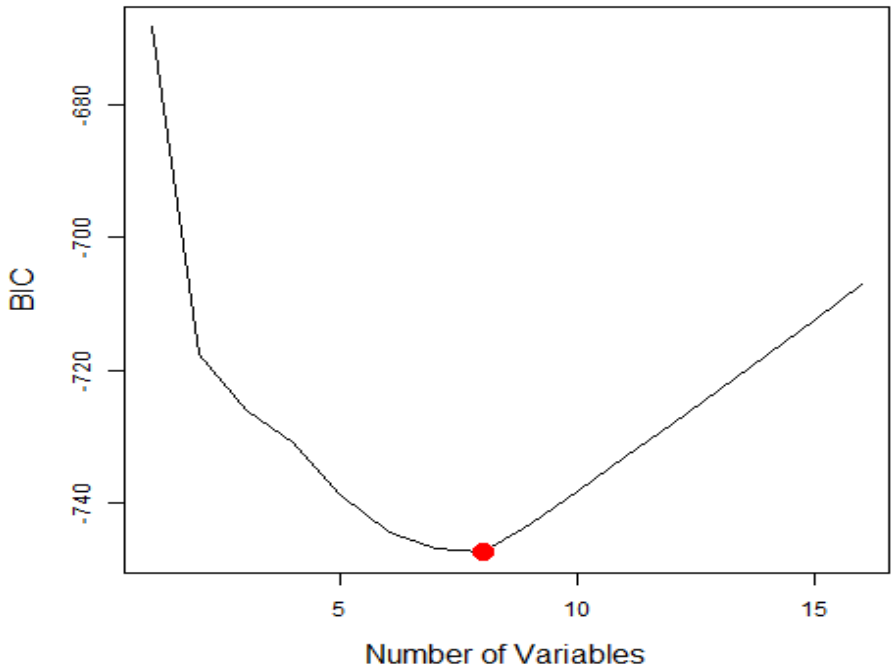
$$100 \times \frac{\text{(Monthly clearance cases)}}{\text{(Total criminal cases in the same month)}}.$$

Note that this variable may be larger than 100.

- `Recession`: A dummy variable, 1 for recession period.
- `Time`: Linear time trend.
- `Unemploy_ppl`: Number of unemployed persons in 1,000.
- `Workforce`: Number of workforce persons in 1,000.
- `Job_partici`: Job participation rate.
- `Drug_lag1`: Drug-related criminal cases in the previous month.
- `Violence_lag1`: Violence-related criminal cases in the previous month.

The other variables include: `Non_workforce`, `Clearance_lag1`, `Clearance_lag2`, `Recession_lag1`, `Recession_lag2`, `Unemploy_rate`, `Drug_lag2`, `Violence_lag2`. Note that we have a variable of interest, `Theft`, and 16 regressors in the data set.

We apply 3 model selection procedures to this sample. While Best Subset Selection and Backward Stepwise Selection both select a 8-variable model, Forward Stepwise Selection finds a 9-variable model. Below is the plot of the BIC values of the models found by Best Subset Selection.



- In this example, the models selected by Best Subset Selection and Backward Stepwise Selection are identical with the following variables: `Recession`, `Time`, `Clearance_lag1`, `Unemploy_rate`, `Job_partici`, `Workforce`, `Non_workforce`, `Violence_lag1`.
- Compared with the model selected above, the 9-variable model selected by Forward Stepwise Selection has 7 variables in common. Yet, it replaces `Unemploy_rate` with `Unemploy_ppl` and includes an additional variable Drug_lag2. Note that Forward Stepwise Selection identifies `Unemploy_ppl` in the 5^{th} round, which is not selected by other methods.

Selection results:

```
> Best Subset Selection: Select 8 Variables
(Intercept)      Recession           Time     Clearance_lag1
  -1.677e+06     -6.218e+02      1.489e+02      -5.053e+01
Unemploy_rate   Job_partici      Workforce
  -6.149e+02      3.461e+04      -9.286e+01
Non_workforce   Violence_lag1
   8.704e+01      4.557e+00
> Forward Stepwise Selection: Select 9 Variables
(Intercept)      Recession           Time     Clearance_lag1
  -1.618e+06     -6.3766e+02     1.442e+02      -5.10e+01
Unemploy_ppl    Job_partici      Workforce
  -5.559e+00      3.346e+04      -8.988e+01
Non_workforce     Drug_lag2   Violence_lag1
   8.385e+01      8.260e-02      4.476e+00
```

Shrinkage Methods

- Instead of searching among a collection of candidate models to find a best subset of regressors, shrinkage methods regularize the coefficient estimates in the complete model with p regressors by "forcing" some of them to shrink towards zero or to be exactly zero. This is done by imposing a constraint on parameters when minimizing the sum of squared errors.
- This approach finds more influential regressors (with coefficients away from zero) and may significantly reduce the variance of coefficient estimates.
- There are two major shrinkage methods, namely, ridge regression and LASSO (Least Absolute Shrinkage and Selection Operator), which are based on different parameter constraints.

Ridge Regression

The ridge regression considers the following minimization problem:

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \quad \text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 \leq s.$$

That is, we want to find solutions that minimize the sum of squared errors, subject to the constraints that there is a "budget," s, for $\sum_{j=1}^{p} \beta_j^2$. This is equivalent to minimizing the constrained objective function:

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2,$$

where λ denotes the tuning parameter.

- $\lambda \sum_{j=1}^{p} \beta_j^2$ is known as the shrinkage penalty, which applies to all coefficients but the intercept β_0.
- When $\lambda = 0$, this is just the usual least-squares problem, so that the ridge estimates are the OLS estimates.
- For a given $\lambda > 0$, the ridge estimates seek to minimize the sum of squared errors while decreasing the shrinkage penalty. When λ increases, the penalty is larger, and some coefficients are "forced" to shrink towards zero. The resulting ridge estimates would be pushed away from the unbiased estimates. Note that the ridge estimates may be very close to zero for some λ but not exactly zero.
- When λ approaches infinity, all coefficients would have to be arbitrarily small.

- While the OLS estimates are scale equivariant (when a regressor is multiplied by a constant c, the corresponding OLS estimate will be scaled by $1/c$), the ridge estimates are not. Indeed, the scaling of regressors has a strong effect on the ridge estimates. As such, it is typical to apply the ridge regression to standardized regressors:

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}}, \quad j = 1, \ldots, p,$$

  where $\bar{x}_j$ is the sample average of the j th variable: $\sum_{i=1}^{n} x_{ij}/n$.
- It is typical to consider a grid of λ values and compute the associated ridge estimates $\hat{\beta}_j^R(\lambda)$. The optimal λ is the one that minimizes the objective function and can be found using cross validation.

Simulations: Ridge Regression

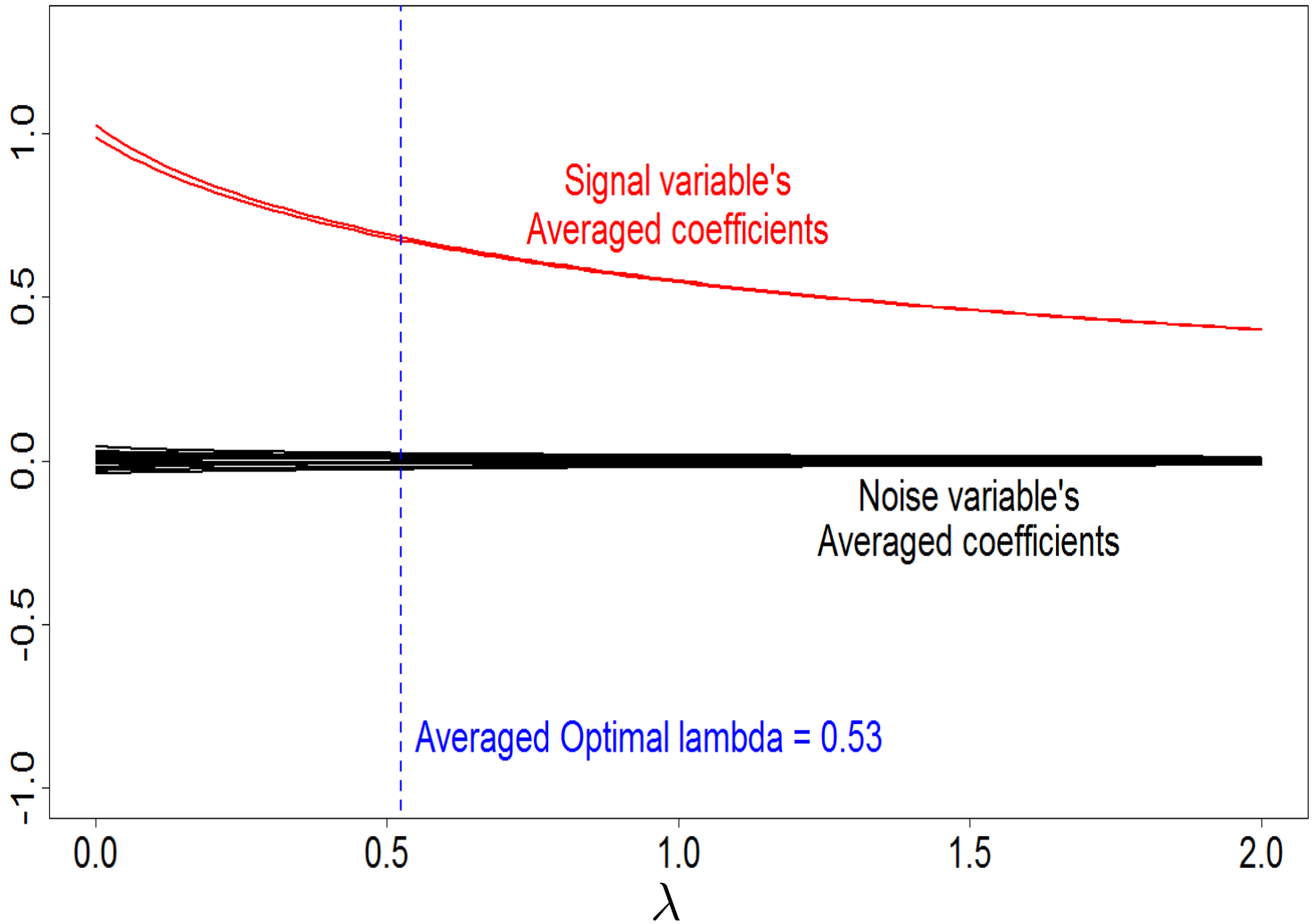
We conduct simulations to assess the performance of the ridge regression estimates. The data generating processes (DGPs) are

$$\text{DGP 1}: \; y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i, \quad \beta_0 = 2, \; \beta_1 = \beta_2 = 1,$$

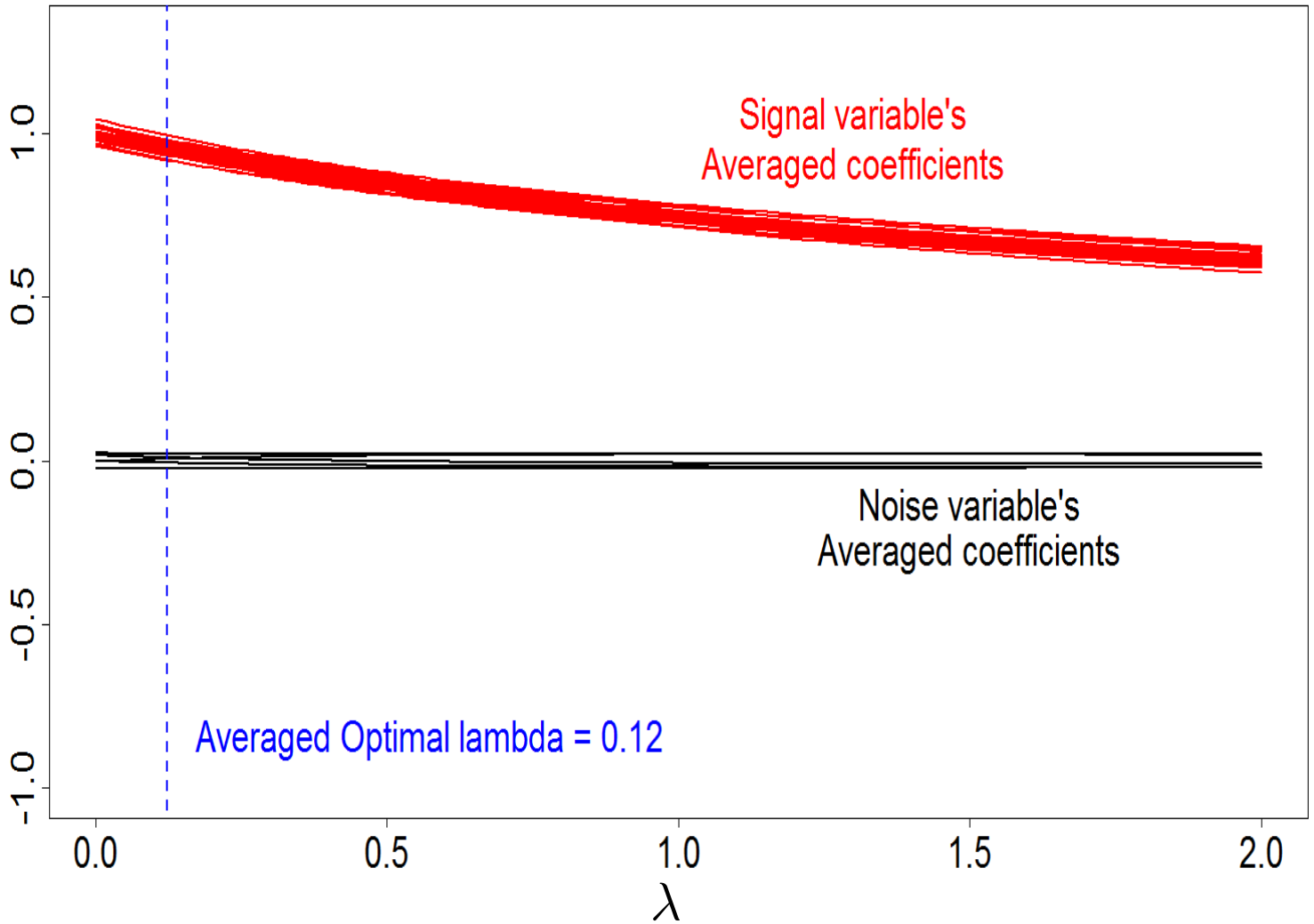
$$\text{DGP 2}: \; y_i = \beta_0 + \sum_{j=1}^{20} \beta_j x_{ij} + \epsilon_i, \quad \beta_0 = 2, \; \beta_1 = \cdots = \beta_{20} = 1,$$

where all x_{ij} and ϵ_i are i.i.d. $\mathcal{N}(0,1)$, and the sample size is $n = 55$. For each DGP, we estimate two models with 25 and 50 regressors, including the "signal" variables in the DGP and other "noise" variables (generated as i.i.d. $\mathcal{N}(0,1)$). We consider 100 λ values in $[0,2]$, with the optimal λ determined by 10-fold CV. All the lines in the following figures are the ridge estimates averaged over 100 replications.

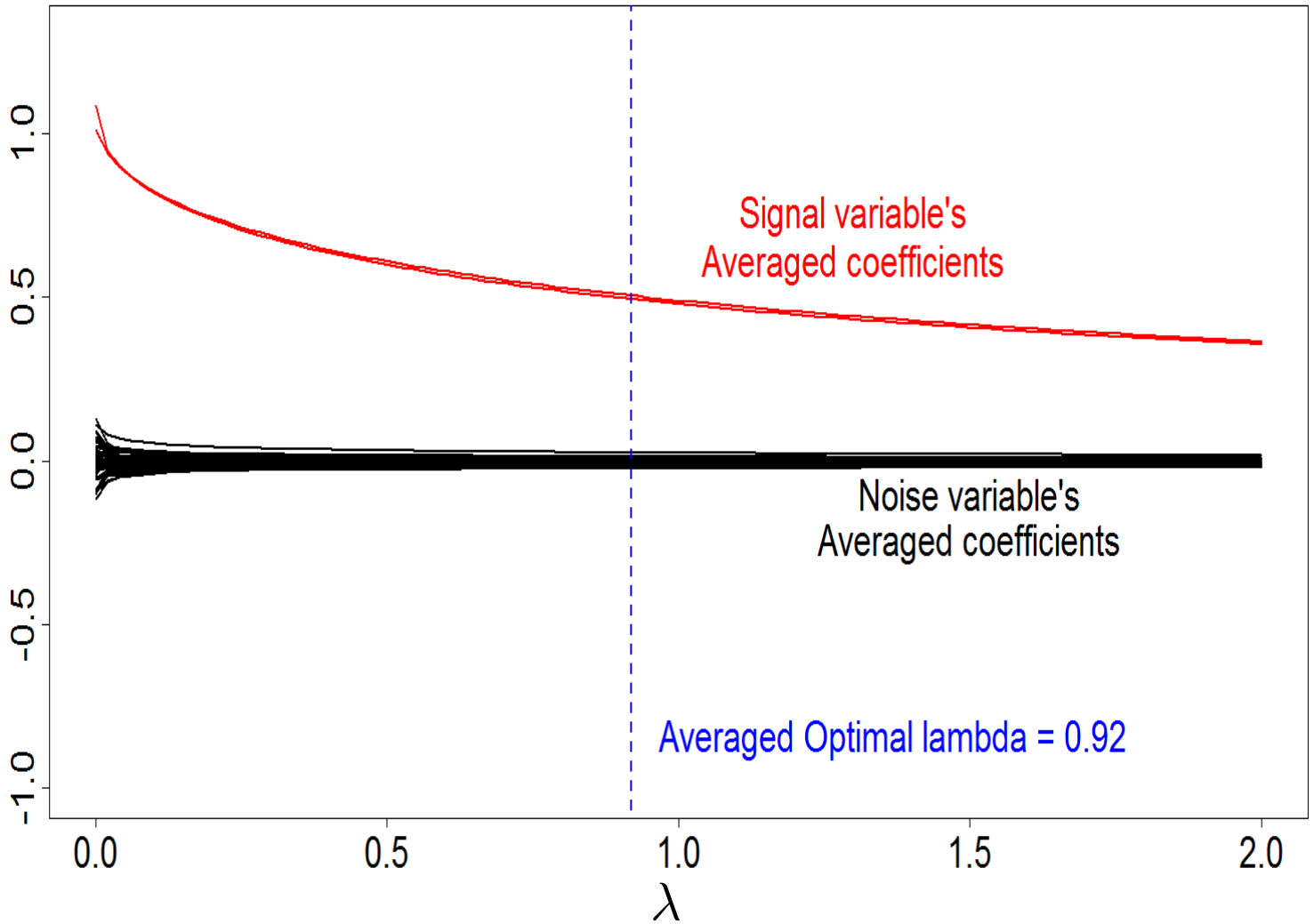
DGP 1: 2 signal variables; ridge regression with 25 regressors.



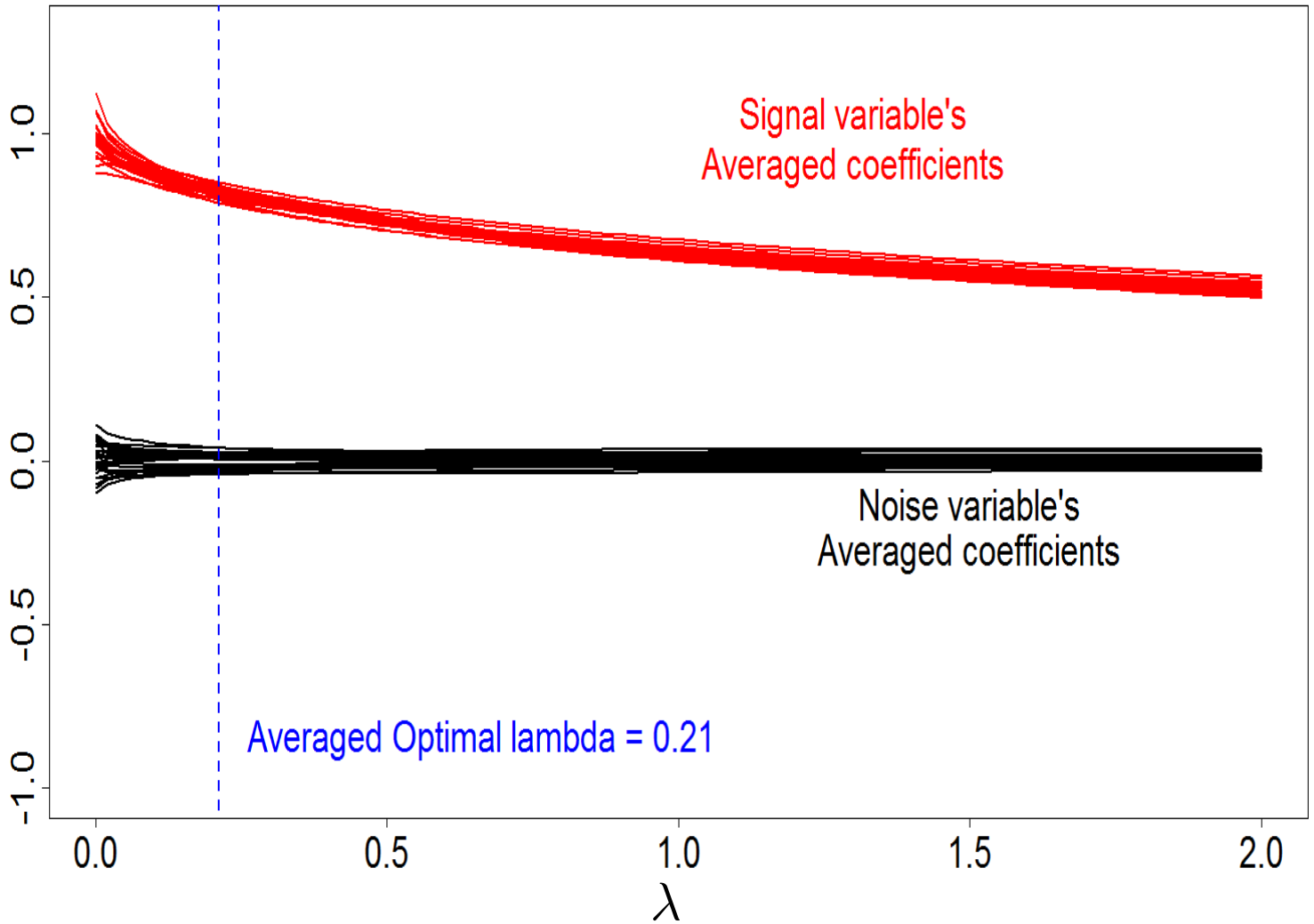
DGP 2: 20 signal variables; ridge regression with 25 regressors.



DGP 1: 2 signal variables; ridge regression with 50 regressors.



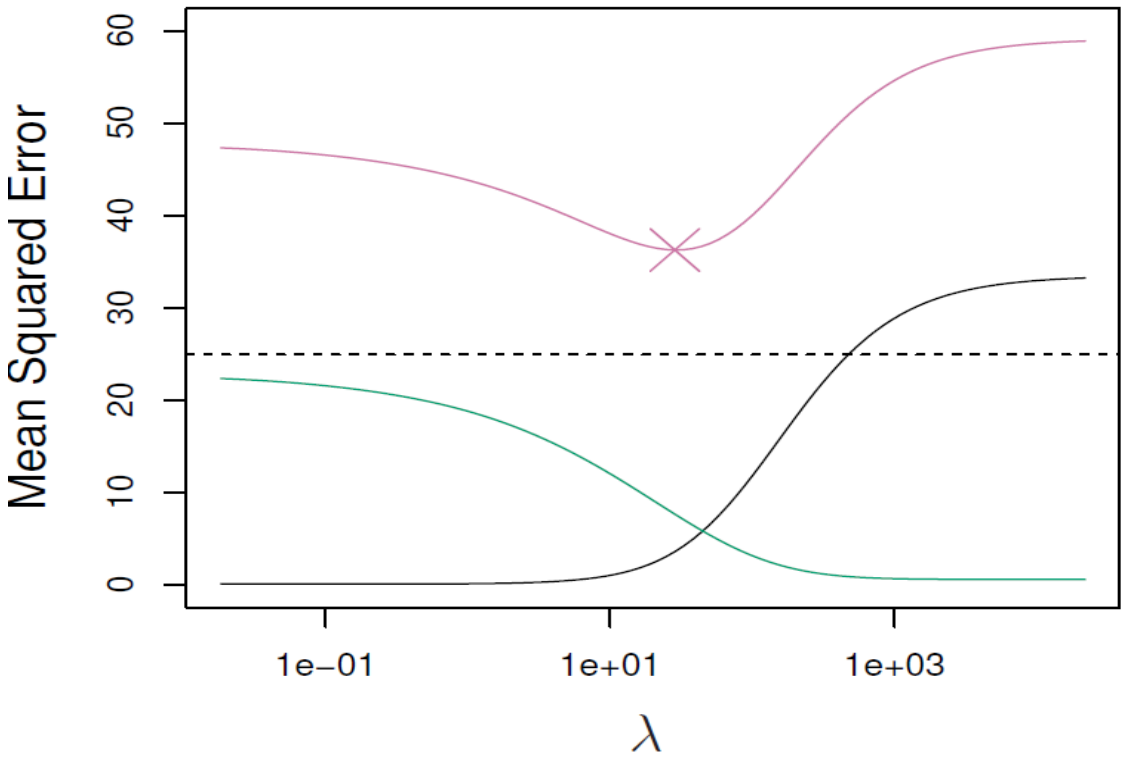
DGP 2: 20 signal variables; ridge regression with 50 regressors.



Summary of Simulation Results

- The larger the λ, the closer the ridge estimates $\hat{\beta}_j^R(\lambda)$ to 0.
- Model with 25 regressors
  - DGP 1: Optimal $\lambda^* = 0.53$, $\bar{\hat{\beta}}_1^R(\lambda^*) = 0.6739$ and $\bar{\hat{\beta}}_2^R(\lambda^*) = 0.6834$.
  - DGP 2: Optimal $\lambda^* = 0.12$, the average ridge estimates at λ^* for the signal variables range from 0.9191 to 0.9947.
- Model with 50 regressors
  - DGP 1: Optimal $\lambda^* = 0.92$, $\bar{\hat{\beta}}_1^R(\lambda^*) = 0.5067$ and $\bar{\hat{\beta}}_2^R(\lambda^*) = 0.4962$.
  - DGP 2: Optimal $\lambda^* = 0.21$, the average ridge estimates at λ^* for the signal variables range from 0.7844 to 0.8475.
- The ridge estimates perform well when the number of signal variables is large relative to the number of variables included in the model. The ridge estimates have large biases when the number of signal variables is relatively small.

The bias-variance trade-off: The graph below, taken from JWHT (2013), shows the $squared\ bias$ (black), $variance$ (green), and the $testing\ MSE$ (purple) for a ridge regression with different λ. The cross indicates the λ value that yields the smallest MSE.



Source: Figure 6.4 of JWHT (2013)

LASSO

The ridge regression does not possess the ability of variable (model) selection because all p regressors are included in the final model. This may not be a problem for prediction accuracy, but it may affect model interpretability, especially when p is large. The LASSO circumvents this problem by considering the following minimization problem:

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq s.$$

That is, the LASSO estimates minimize the sum of squared errors, subject to the constraints that there is a "budget," s, for $\sum_{j=1}^{p} |\beta_j|$. Note that the budget is determined by the absolute values of the parameters (except the intercept term).

The LASSO problem is equivalent to minimizing the objective function:

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j|,$$

where every regressor x_j is standardized.

- As in the ridge regression, the LASSO estimates $\hat{\beta}_j^L(\lambda)$ are the OLS estimates when $\lambda=0$, and they are zero when $\lambda\to\infty$. In practice, we may also apply cross validation to determine the optimal λ from a grid of λ values.
- In contrast with the ridge regression, the LASSO is capable of variable selection because the LASSO penalty, $\lambda\sum_{j=1}^{p}|\beta_j|$, has the effect of "forcing" the estimates $\hat{\beta}_j^L(\lambda)$ to be exactly zero when λ is sufficiently large.

Simulations: LASSO

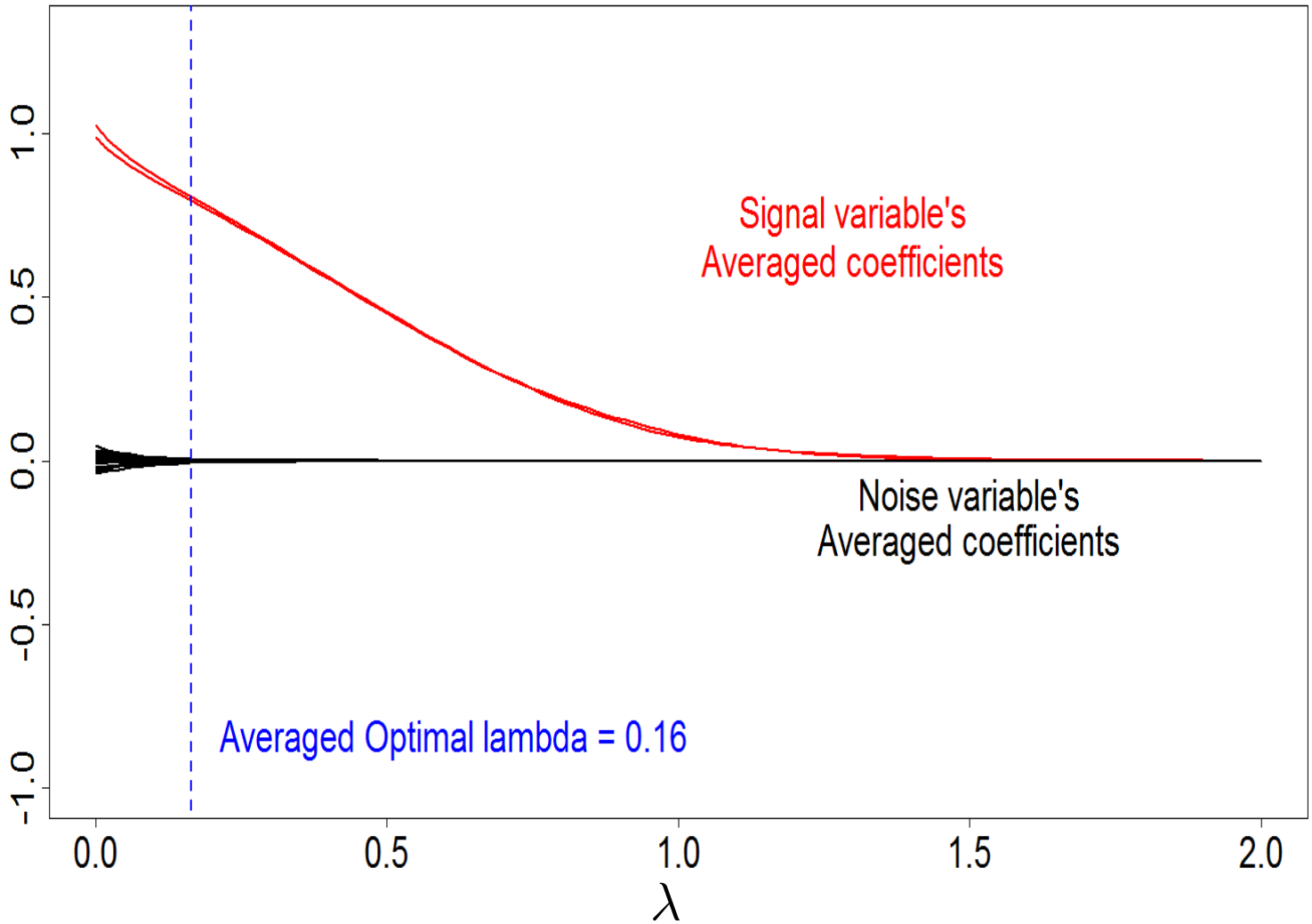
We conduct simulations to assess the performance of the LASSO estimates. The DGPs are as in the previous simulations:

$$\text{DGP 1}: \; y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i, \quad \beta_0 = 2, \; \beta_1 = \beta_2 = 1,$$

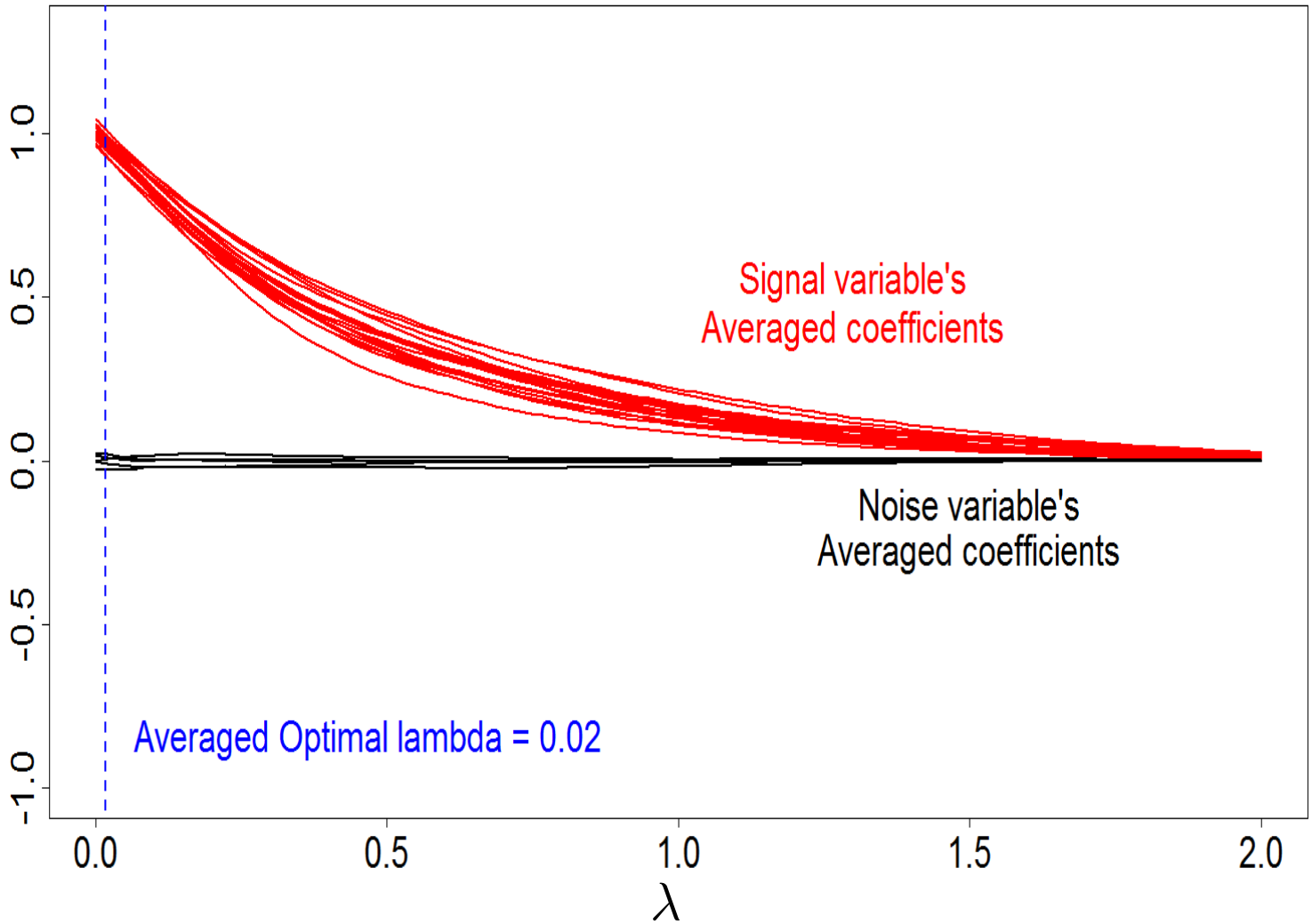
$$\text{DGP 2}: \; y_i = \beta_0 + \sum_{j=1}^{20} \beta_j x_{ij} + \epsilon_i, \quad \beta_0 = 2, \; \beta_1 = \cdots = \beta_{20} = 1,$$

where all x_{ij} and ϵ_i are i.i.d. $\mathcal{N}(0,1)$, and the sample size is $n = 55$. For each DGP, we estimate two models with 25 and 50 regressors, including the "signal" variables in the DGP and other "noise" variables (generated as i.i.d. $\mathcal{N}(0,1)$). We consider 100 λ values in $[0,2]$, with the optimal λ determined by 10-fold CV. All the lines in the following figures are the LASSO estimates averaged over 100 replications.

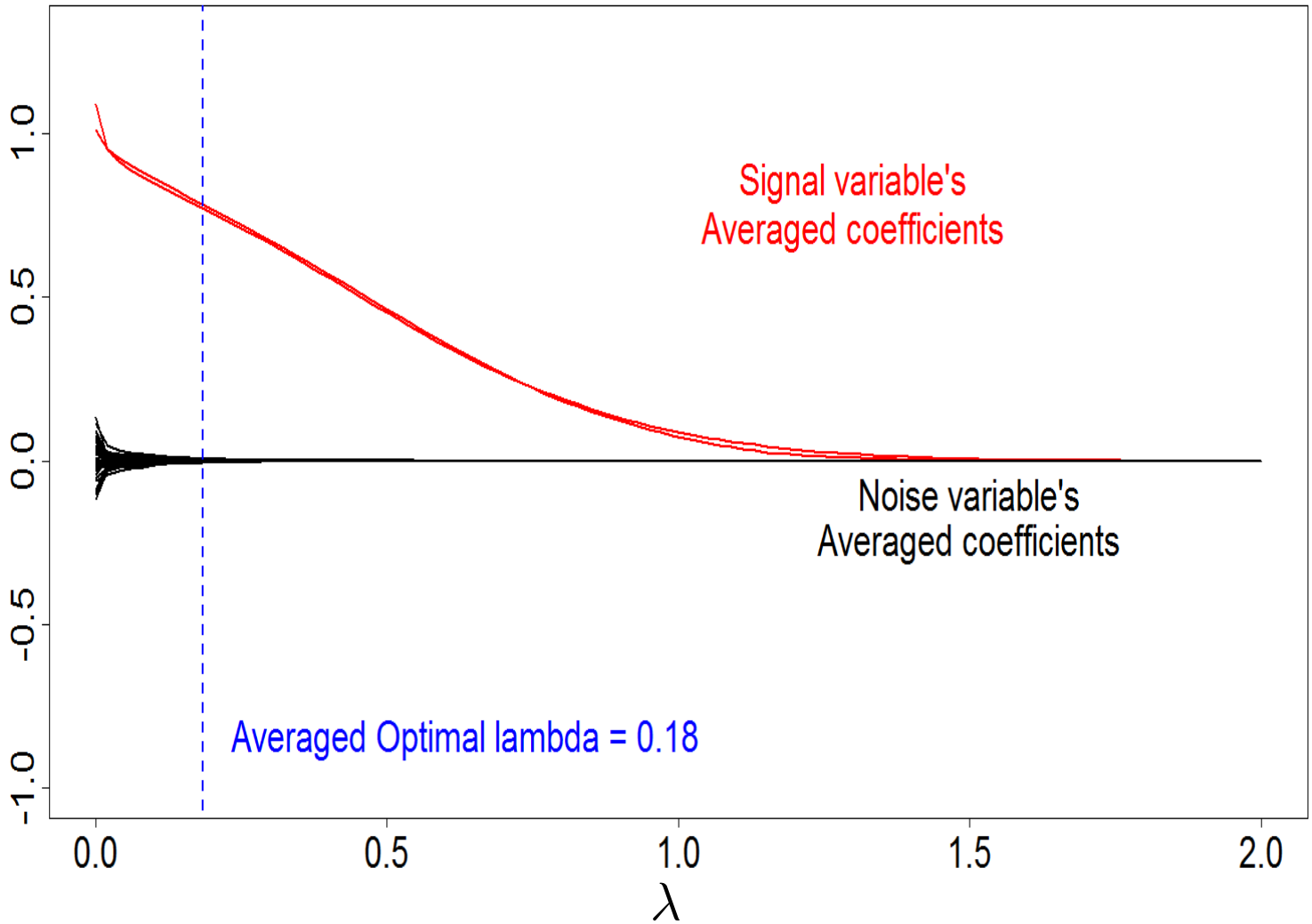
DGP 1: 2 signal variables; LASSO with 25 regressors.



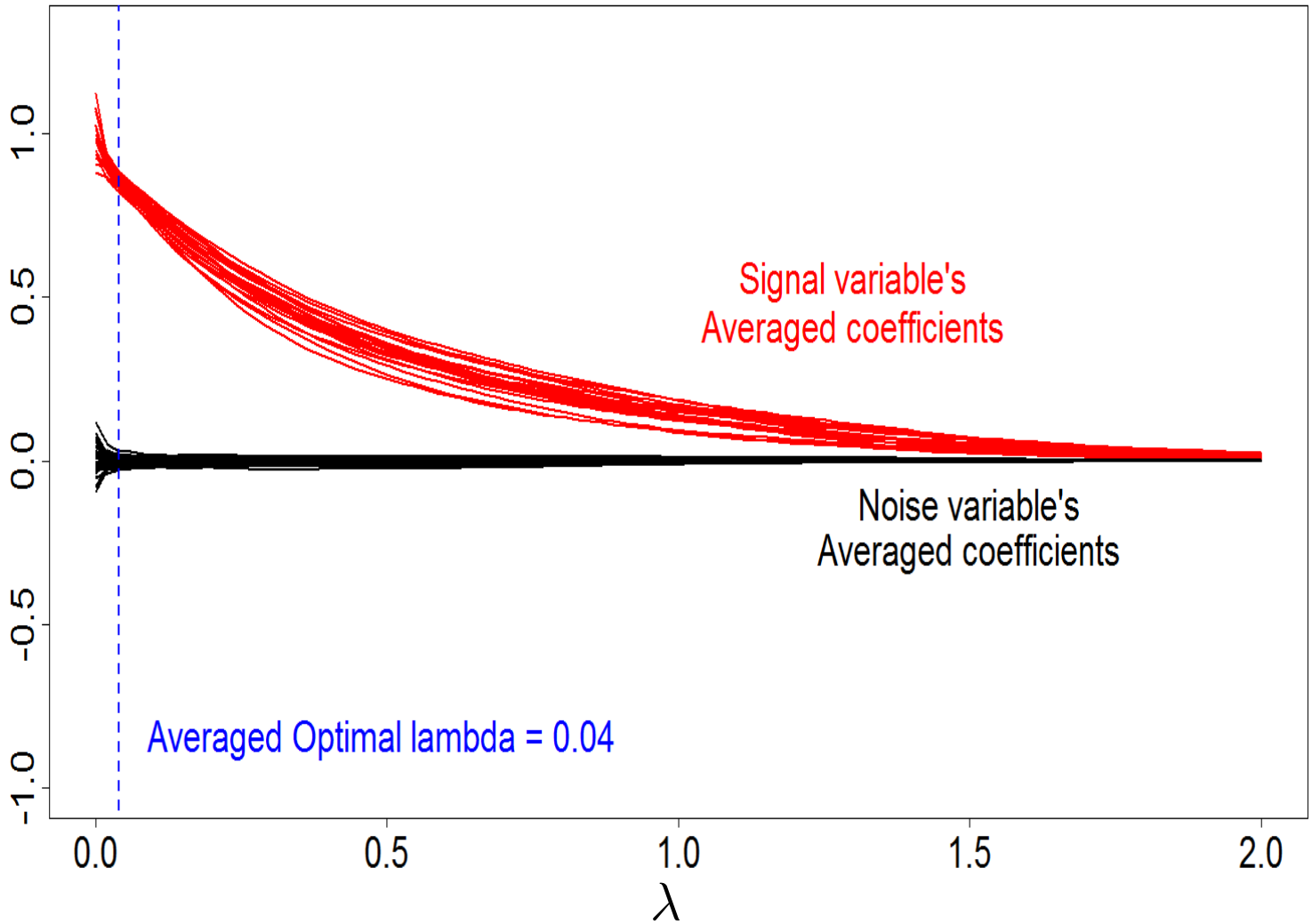
DGP 2: 20 signal variables; LASSO with 25 regressors.



DGP 1: 2 signal variables; LASSO with 50 regressors.



DGP 2: 20 signal variables; LASSO with 50 regressors.



Summary of Simulation Results

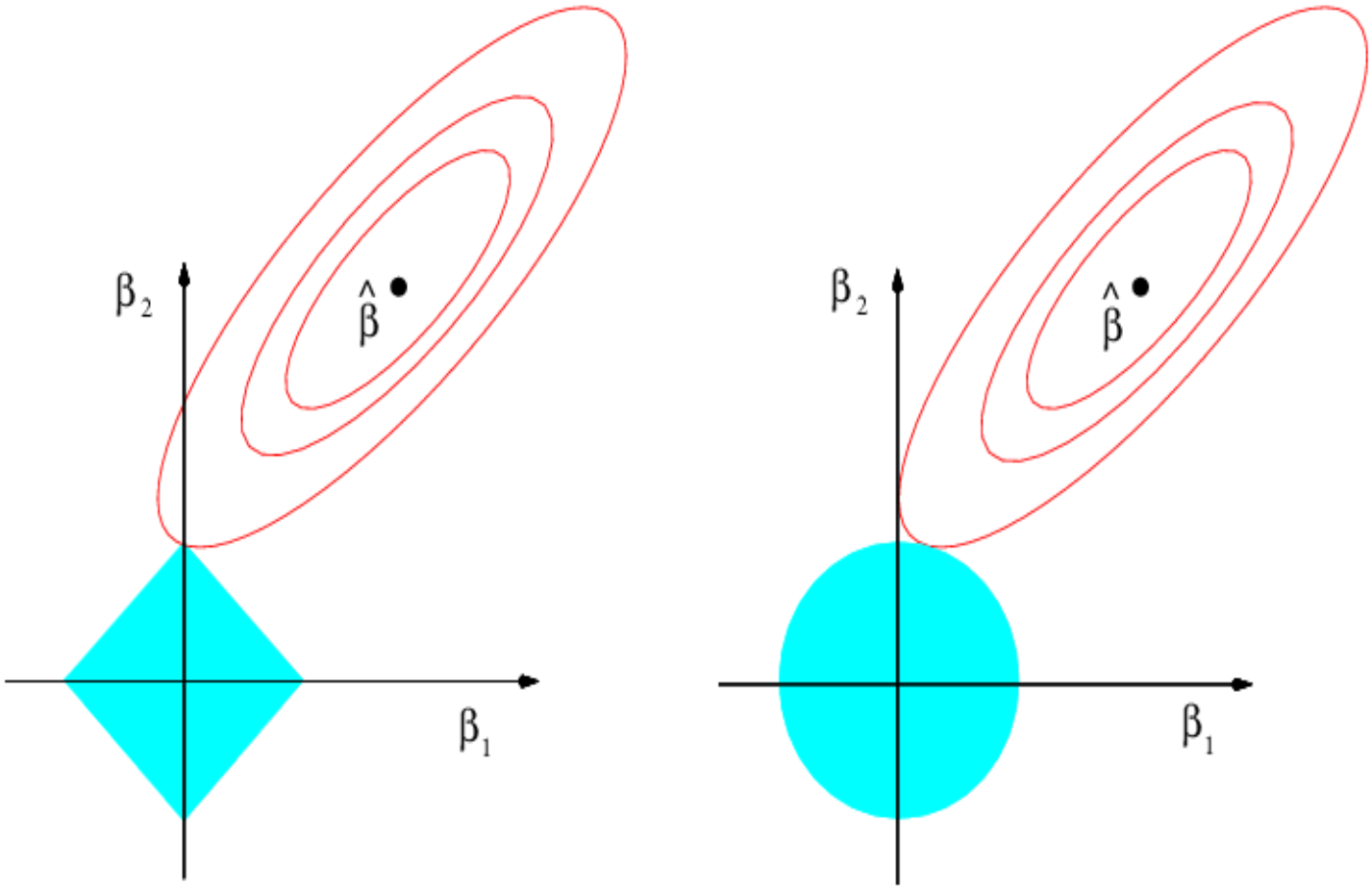
- The LASSO estimates $\hat{\beta}_j^L(\lambda)$ become 0 quickly as λ increases.
- Model with 25 regressors
  - DGP 1: Optimal $\lambda^* = 0.16$, $\bar{\hat{\beta}}_1^L(\lambda^*) = 0.7969$ and $\bar{\hat{\beta}}_2^L(\lambda^*) = 0.8082$.
  - DGP 2: Optimal $\lambda^* = 0.02$, the average ridge estimates at λ^* for the signal variables range from 0.9262 to 1.0077.
- Model with 50 regressors
  - DGP 1: Optimal $\lambda^* = 0.18$, $\bar{\hat{\beta}}_1^L(\lambda^*) = 0.7850$ and $\bar{\hat{\beta}}_2^L(\lambda^*) = 0.7729$.
  - DGP 2: Optimal $\lambda^* = 0.04$, the average ridge estimates at λ^* for the signal variables range from 0.8207 to 0.8834.
- Compared with the ridge estimates, the LASSO estimates tend to have a smaller bias when the number of signal variables is small. The LASSO also performs very well when there are many signal variables.

LASSO vs. Ridge Regression

To see how the LASSO differs from the ridge regression, consider the case of $p = 2$. In this case, the budget set for the ridge regression is the circle: $\beta_1^2 + \beta_2^2 \leq s$, and that for the LASSO is the diamond: $|\beta_1| + |\beta_2| \leq s$.

- Optimization with a smooth budget sets often leads to interior solutions, while optimization with a budget sets with non-differentiable points often yields corner solutions. See the graph in the next page.
- If s is so large that the OLS estimates fall within the budget set, $\sum_{j=1}^p \beta_j^2 \leq s$ and $\sum_{j=1}^p |\beta_j| \leq s$ are not binding, and the ridge regression and the LASSO would yield the OLS solutions.

In the graphs below, the red solid lines are the contours of the sum of squared errors, and the light blue regions are the budget sets.



Source: Figure 6.7 of JWHT (2013)

A Special Case

Consider now the special case that $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, with $\boldsymbol{X} = \boldsymbol{I}_p$ (so that $n = p$). In this case, the OLS estimates minimize:

$$\sum_{j=1}^{p} \left(y_j - \beta_j\right)^2,$$

and hence $\hat{\beta}_j = y_j$. The ridge regression and LASSO estimates minimize, respecitvely,

$$\sum_{j=1}^{p} \left(y_j - \beta_j\right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2,$$

$$\sum_{j=1}^{p} \left(y_j - \beta_j\right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|.$$

It is easy to verify that the ridge regression estimates are:

$$\hat{\beta}_j^R = y_j/(1+\lambda),$$

and the LASSO estimates are:

$$\hat{\beta}_j^L = \begin{cases} y_j - \lambda/2 & \text{if } y_j > \lambda/2, \\ y_j + \lambda/2 & \text{if } y_j < -\lambda/2, \\ 0 & \text{if } |y_j| \leq \lambda/2. \end{cases}$$

Thus, the ridge regression shrinks each OLS coefficient towards zero by the same proportion $1/(1+\lambda)$. Clearly, the ridge estimates approach zero when λ increases, but they will not be exactly zero. On the other hand, the LASSO shrinks the OLS coefficients by a given amount $\lambda/2$; the larger the λ values, the more the estimates shrink. See Figure 6.10 of JWHT (2013).

LASSO vs. Best Subset Selection

Note that the Best Subset Selection problem can be written as:

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \quad \text{subject to} \quad \sum_{j=1}^{p} \mathbf{1}(\beta_j \neq 0) \leq s,$$

where $\mathbf{1}(\cdot)$ denotes the indicator function. In other words, this problem finds parameter estimates to minimize the sum of squared errors, subject to the constraint that no more than s coefficients can be selected. As we learned earlier, this minimization problem is computationally prohibited when p is large. Hence, we may view both the ridge regression and LASSO as two computationally feasible alternatives to Best Subset Selection.

LASSO vs. OLS

To further examine the difference between the LASSO and OLS estimates, we conduct simulations with the DGPs considered earlier:

$$\text{DGP 1}: \; y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i, \quad \beta_0 = 2, \; \beta_1 = \beta_2 = 1,$$

$$\text{DGP 2}: \; y_i = \beta_0 + \sum_{j=1}^{20} \beta_j x_{ij} + \epsilon_i, \quad \beta_0 = 2, \; \beta_1 = \cdots = \beta_{20} = 1,$$

and the sample size is $n = 55$; another 55 observations are also generated and reserved as the testing sample. We estimate two models with 25 and 50 regressors, including the "signal" variables in the DGP and other "noise" variables. The optimal λ is chosen from 100 λ values in $[0, 2]$ using 10-fold CV. The number of replications is $10,000$.

- In the following tables, we consider three methods: OLS, LASSO, and LASSO-OLS, where for LASSO-OLS we first select the variables by the LASSO and then apply OLS with the LASSO-selected variables as regressors.
- To illustrate, we report only the results of the coefficient of the first signal variable $\hat{\beta}_1$ and those of the coefficient of the first noise variable $\hat{\beta}_3$ (for DGP 1) or $\hat{\beta}_{21}$ (for DGP 2). Their bias, variance, and MSE are the averages over $10,000$ replications.
- The model testing MSEs are the averages over $10,000$ testing samples.

DGP 1; $p = 25$		OLS	LASSO	LASSO-OLS
$\hat{\beta}_1$	Bias	0.0011	-0.1845	-0.0281
	Variance	0.0367	0.0262	0.0255
	MSE	0.0367	0.0603	0.0263
$\hat{\beta}_3$	Bias	0.0019	-0.0005	-0.0001
	Vriance	0.0353	0.0032	0.0117
	MSE	0.0353	0.0032	0.0117
Model Testing MSE		1.9323	1.2187	1.3539
DGP 1; $p = 50$				
$\hat{\beta}_1$	Bias	-0.0060	-0.2269	-0.0609
	Variance	0.3432	0.0281	0.0282
	MSE	0.3432	0.0796	0.0319
$\hat{\beta}_3$	Bias	-0.0004	0	0.0005
	Variance	0.3335	0.0020	0.0081
	MSE	0.3335	0.0020	0.0081
Model Testing MSE		18.1495	1.2812	1.4951

Summary of Simulation Results: DGP 1

- Bias-variance trade-off for $\hat{\beta}_1$ (coefficient of the first signal variable): Compared with OLS, the LASSO estimates have much larger biases, smaller variances, and smaller MSEs. When $p = 50$, OLS leads to a very large variance, but the LASSO estimate has a stable variance.
- $\hat{\beta}_3$ (coefficient of the first noise variable): Compared with OLS, the LASSO estimates have little bias, smaller variances, and smaller MSEs. Thus, there is no bias-variance trade-off for this coefficient.
- LASSO-OLS yields smaller MSE for $\hat{\beta}_1$ than both OLS and LASSO but larger MSE for $\hat{\beta}_3$ than LASSO.
- For both models, the LASSO yields the smallest testing MSEs, and OLS yields the largest testing MSEs. The advantage of LASSO is very significant when $p = 50$.

DGP 2; $p = 25$		OLS	LASSO	LASSO-OLS
$\hat{\beta}_1$	Bias	0.0011	-0.0292	-0.0001
	Variance	0.0367	0.0372	0.0365
	MSE	0.0367	0.0381	0.0365
$\hat{\beta}_{21}$	Bias	0.0002	0.0004	0.0005
	Vriance	0.0348	0.0279	0.0340
	MSE	0.0348	0.0279	0.0340
Model Testing MSE		1.9203	1.9301	1.9234
DGP 2; $p = 50$				
$\hat{\beta}_1$	Bias	-0.0060	-0.1411	-0.0636
	Variance	0.3433	0.0742	0.0888
	MSE	0.3433	0.0942	0.0928
$\hat{\beta}_{21}$	Bias	-0.0054	-0.0003	-0.0016
	Variance	0.3245	0.0309	0.0612
	MSE	0.3245	0.0309	0.0612
Model Testing MSE		18.0081	3.7543	4.5389

Summary of Simulation Results: DGP 2

- $\hat{\beta}_1$: We observe bias-variance trade-off for the LASSO estimate when $p = 50$ but no such trade-off when $p = 25$.
- $\hat{\beta}_{21}$ (the coefficient of the first noise variable): Compared with OLS, the LASSO estimate has a smaller bias, smaller variance and smaller MSE when $p = 50$, i.e., no bias-variance trade-off. For $p = 25$, the LASSO has slightly larger bias but smaller variance and MSE.
- When $p = 50$, the LASSO yields the smallest testing MSE, but OLS has very large testing MSE. When $p = 25$, the LASSO yields slightly larger testing MSE than does OLS.
- The LASSO has clear advantages when the number of regressors is much larger than the number of signal variables. Note that bias-variance trade-off does not always exist.

Example: Taiwan Criminal Data (Shrinkage Methods)

We apply the ridge regression and LASSO to Taiwan's criminal data. Recall that the variable of interest is `Theft` and that there are 16 independent variables. We report the ridge estimates at $\lambda^* = 0.5005$ and the LASSO estimates at $\lambda^* = 51.2512$. Note that the ridge estimates ought to be close to the OLS estimates because λ^* is very small.

The LASSO selects 10 out of 16 regressors, among them 5 are the same as those selected by Best Subset Selection (`Recession`, `Clearance_lag1`, `Workforce`, `Non_workforce`, `Violence_lag1`) and one is the same as that selected by Forward Stepwise Selection (`Unemploy_ppl`). In addition, the LASSO selects `Clearance_lag2`, `Recession_lag1`, `Drug_lag2`, and `Violence_lag2`.

Ridge Regression Estimates

Clearance	Recession	Time	Clearance_lag1
$7.118042e+00$	$-7.165727e+02$	$5.900018e+01$	$-4.455641e+01$
Clearance_lag2	Recession_lag1	Recession_lag2	Unemploy_ppl
$-1.562688e+01$	$-1.703746e+02$	$-3.224905e+01$	$4.666036e+01$
Unemploy_rate	Job_partici	Workforce	Non_workforce
$-5.629722e+03$	$3.285448e+03$	$-1.716189e+01$	$-6.595184e+00$
Drug_lag1	Drug_lag2	Violence_lag1	Violence_lag2
$-1.984927e-02$	$1.873150e-01$	$4.773189e+00$	$9.504042e-01$
(Intercept)			
$6.047442e+04$			

LASSO Estimates

Clearance	Recession	Time	Clearance_lag1
$7.118042e+00$	$-7.240257e+02$	0	$-4.061501e+01$
Clearance_lag2	Recession_lag1	Recession_lag2	Unemploy_ppl
$-5.711850e+00$	$-1.766387e+02$	0	$-6.140723e+00$
Unemploy_rate	Job_partici	Workforce	Non_workforce
0	0	$-2.628722e+00$	$-1.454844e+01$
Drug_lag1	Drug_lag2	Violence_lag1	Violence_lag2
0	$9.107362e-02$	$4.992001e+00$	$1.338615e+00$
(Intercept)			
$1.612056e+05$			

References and Acknowledgement

References

1. James, G., D. Witten, T. Hastie, and R. Tibshirani (2013). *An Introduction to Statistical Learning, with Applications in R*, New York: Springer. (JWHT (2013))
2. Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning*, Second Edition, New York: Springer.

Some of the figures in this presentation are taken from JWHT (2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani