Model selection and prediction with lasso

Suppose that

$$X = (X_1, X_2, X_3, X_4, X_5) \sim N(\mu, \Sigma)$$

with $\mu = (1, ..., 1)$ and correlation matrix $\Sigma = I_5$.

The outcome Y is generated as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_3 + \beta_4 X_1 X_4 + \epsilon$$

with $\beta = (1, 1, -0.25, 0.75, 0.4)'$, $\epsilon \ \sim N(0, \sigma^2)$, $\sigma = 1.5$, and ϵ independent of X.

Competing methods

We generate an estimation (training) sample $\{(X_i, Y_i)\}$ of size N_{tr} and a test sample $\{(X_i, Y_i)\}$ of size N_{test}. We want to compare the prediction performance of the following models:

1. OLS regression of Y on $b_2(X)$
2. OLS regression of Y on $b_2(X)$ with backward selection
3. Lasso with λ chosen by 5-fold CV
4. Lasso with λ chosen by 5-fold CV and the 1 SE rule

Note: $b_2(X)$ contains 21 variables (including the constant term)

The model selection and prediction exercise

We conduct the following Monte Carlo exercise:

- Draw a training and test sample from the data generating process.
- Execute each method over the training sample.
- Apply the selected models (estimated over the training sample) to form predictions $\hat{Y}_i = b_2(X_i)'\hat{\beta}_{tr}$ for each observation i in the test sample.
- Compute the MSPE for the predictions given by each method:

$$\frac{1}{N^{test}} \sum_{i \in \mathcal{S}^{test}} (Y_i - \hat{Y}_i)^2$$

- Repeat the exercise many times (generating new data in each cycle) and report the average MSPE for each method.

Results: prediction performance

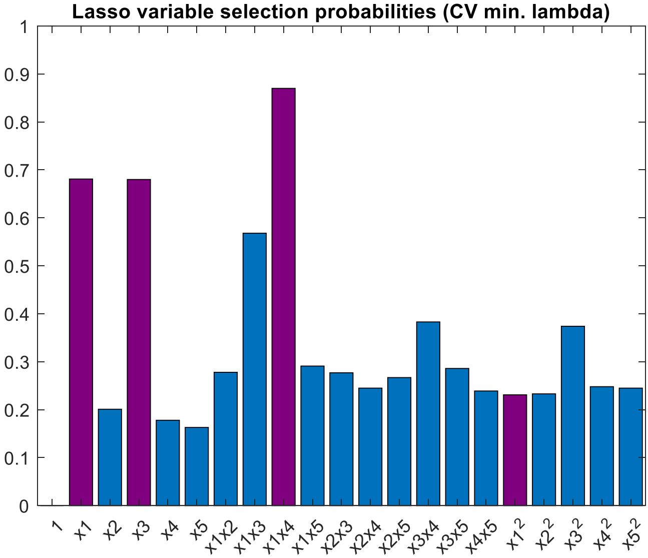
$dim(b_2(X)) = 21$; $N_{test} = 500$; averages over 1000 Monte Carlo repetitions

	$N_{tr} = 50$ MSPE	$N_{tr} = 100$ MSPE	$N_{tr} = 500$ MSPE
OLS	5.05	3.03	2.35
OLS+bw. sel.	3.99	2.81	2.31
Lasso (CV min)	2.97	2.63	2.31
Lasso (CV 1SE)	3.31	2.92	2.44

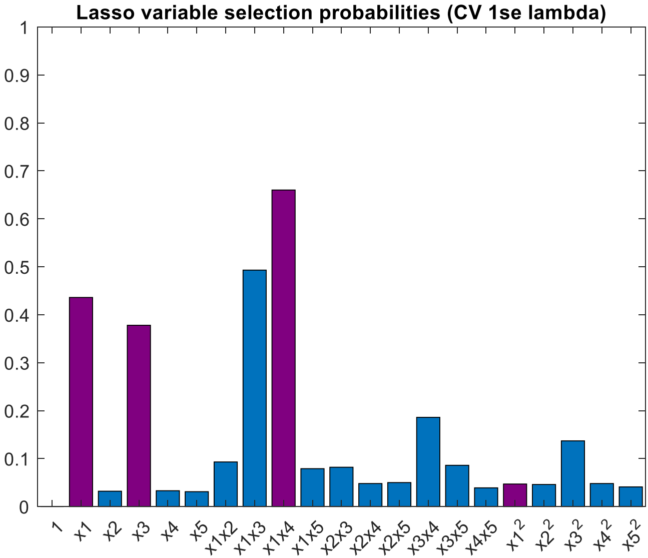
Results: a deeper look at the selected models

	$N_{tr} = 50$		
	Lasso λ-min	Lasso λ-1se	OLS bw. sel.
% of times true model found	0.1	0	0
% of times all relev. vars. found	12.1	2.7	5.0
Av. number of variables	6.94	3.05	10.6
% of times model beats OLS	99.4	90.9	90.9

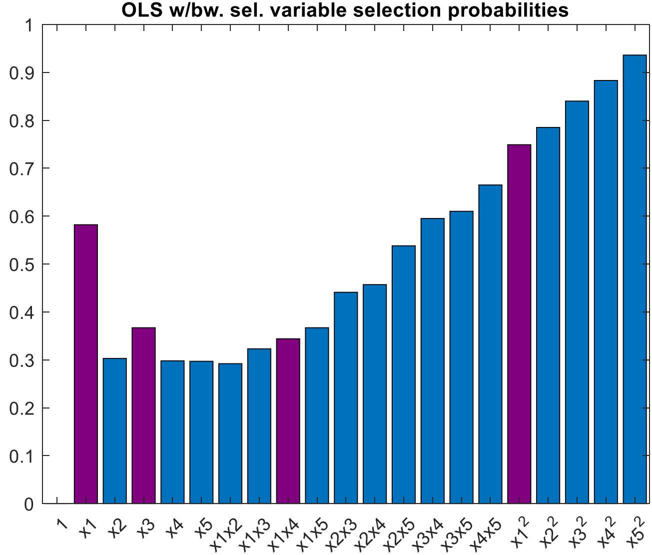
Lasso (λ-min) variable selection probabilities for $N_{tr} = 50$



Lasso (λ-1se) variable selection probabilities for $N_{tr} = 50$



OLS w/bw. sel. variable selection probabilities for $N_{tr} = 50$



Selection map: lasso min-λ, $N_{tr} = 50$

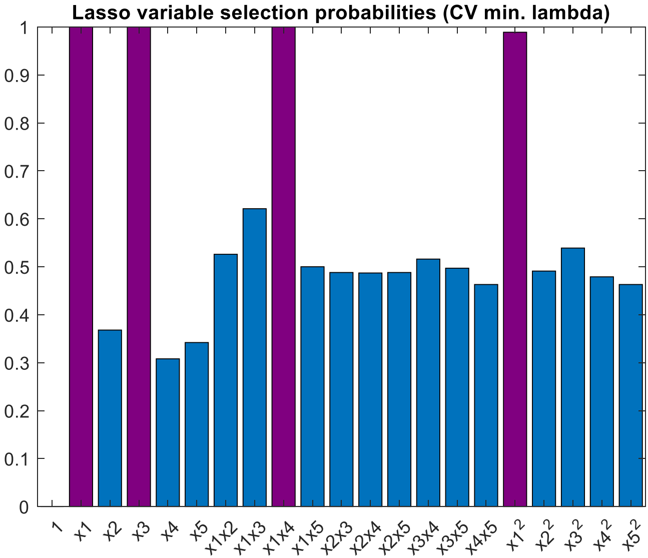
Selection map: lasso 1se-λ, $N_{tr} = 50$

Selection map: OLS w/bw. sel. $N_{tr} = 50$

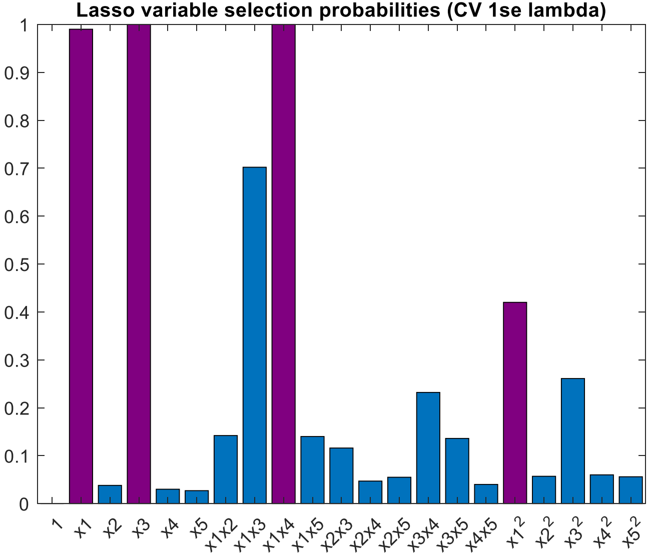
Results: a deeper look at the selected models

	$N_{tr} = 500$		
	Lasso λ-min	Lasso λ-1se	OLS bw. sel.
% of times true model found	0	1.7	0
% of times all relev. vars. found	98.9	42.0	13.0
Av. number of variables	11.6	5.5	10.5
% of times model beats OLS	84.8	22.7	92.5

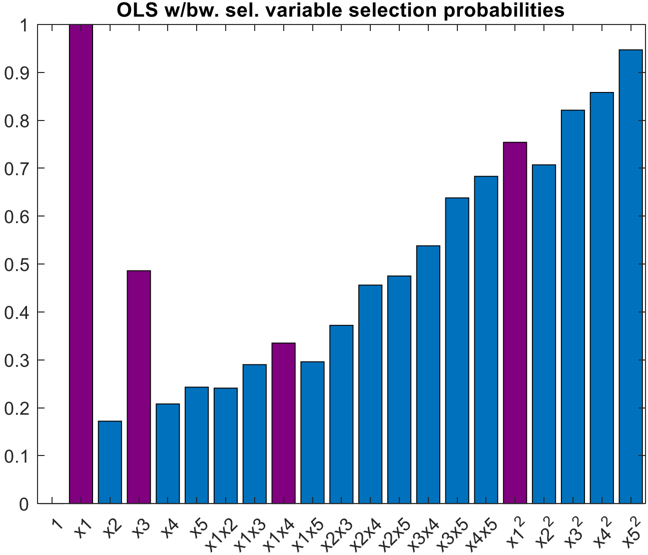
Lasso (λ-min) variable selection probs for $N_{tr} = 500$



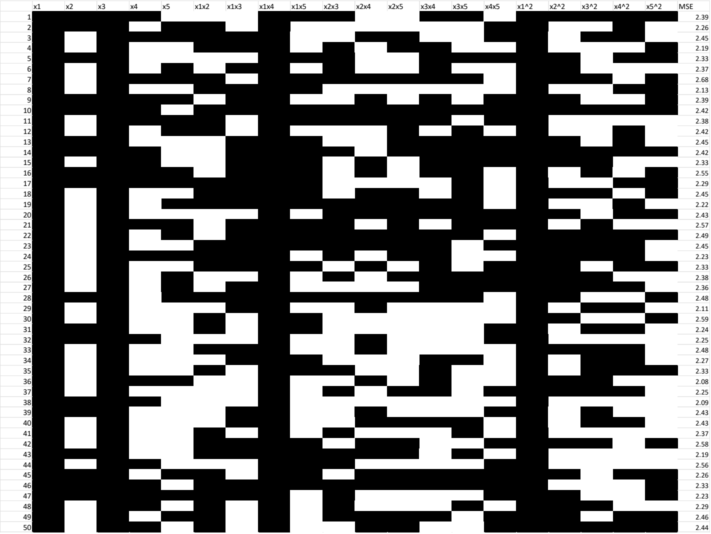
Lasso (λ-1se) variable selection probs for $N_{tr} = 500$



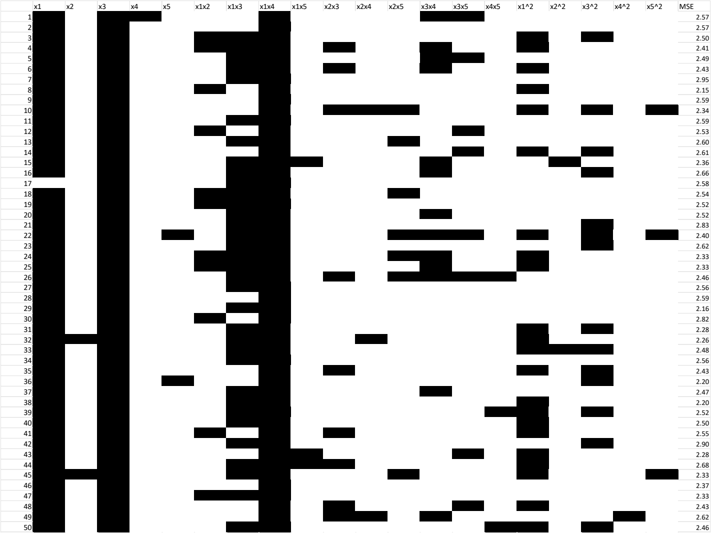
OLS w/bw. sel. variable selection probs for $N_{tr} = 500$



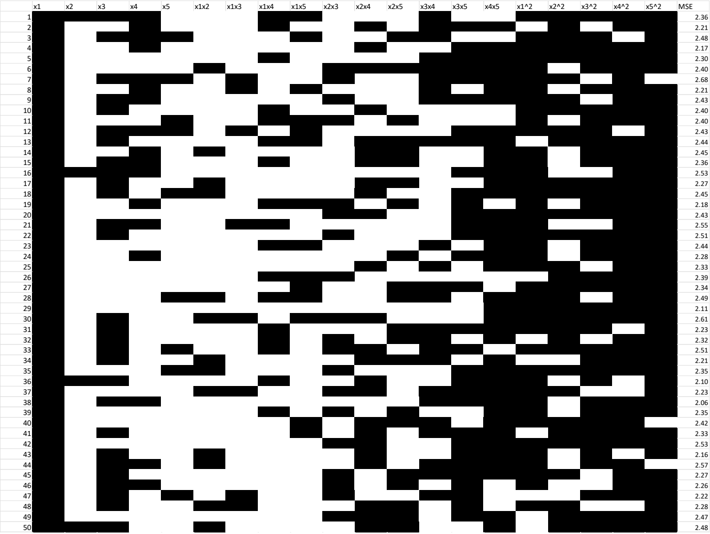
Selection map: lasso min-λ, $N_{tr} = 500$



Selection map: lasso 1se-λ, $N_{tr} = 500$



Selection map: OLS w/bw. sel. $N_{tr} = 500$



OLS versus PCA regression

Let $X = (X_1, X_2, \ldots, X_{50}) \sim N(0, \Sigma)$.
The correlations between the components of X (the elements of Σ) are randomly chosen.

The outcome Y is generated as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_{50} X_{50} + \epsilon$$

with $\epsilon \ \sim N(0, \sigma^2)$, $\sigma = 2$, and ϵ independent of X.

Case 1 ('sparse model'): $\beta_0 = 1$, $\beta_1 = 1$, $\beta_2 = 1$, $\beta_3 = 1$, $\beta_4 = 0$,..., $\beta_{50} = 0$.
Case 2 ('dense model'): $\beta_0 = 1$, all slope coefficients between 0 and 1, randomly picked from uniform distribution

Competing methods

We generate an estimation (training) sample $\{(X_i, Y_i)\}$ of size N_{tr} and a test sample $\{(X_i, Y_i)\}$ of size N_{test}. We want to compare the prediction performance of the following models:

1. OLS regression of Y on X
2. PCA regression of Y on $Z_1^*, \ldots, Z_k^*$

The model selection and prediction exercise

We conduct the following Monte Carlo exercise:

- Draw a training and test sample from the data generating process.
- Execute each method over the training sample.
- Use the estimated models to compute predictions for each observation i in the test sample.
- Compute the MSPE for the predictions given by each method:

$$\frac{1}{N^{test}} \sum_{i \in \mathcal{S}^{test}} (Y_i - \hat{Y}_i)^2$$

- Repeat the exercise many times (generating new data in each cycle) and report the average MSPE for each method.

Results: prediction performance

$N_{test} = 500$; averages over 1000 Monte Carlo repetitions

SPARSE DGP	$N_{tr} = 75$ MSPE	$N_{tr} = 150$ MSPE	$N_{tr} = 500$ MSPE
OLS	12.8	6.0	4.5
PCA (k=1)	8.7	8.7	8.7
PCA (k=5)	5.8	5.5	5.4
PCA (k=10)	5.4	4.9	4.7

DENSE DGP	$N_{tr} = 75$ MSPE	$N_{tr} = 150$ MSPE	$N_{tr} = 500$ MSPE
OLS	12.9	6.0	4.5
PCA (k=1)	14.9	14.7	14.6
PCA (k=5)	13.6	13.0	12.7
PCA (k=10)	9.3	8.5	8.0