The LASSO

- Time series focuses on dynamic relationships within and across variables
  - We have a choice of how many lags to include, etc...
- The LASSO:
  - "**L**east **A**bsolute **S**hrinkage and **S**election **O**perator"
  - Regression-based model selection procedure popular in *data science*
- Suppose that we have N observations, P potential explanatory variables
- The Lasso Problem:

$$\max_{\beta_p} \sum_{i=1}^{N} \left(y_i - \sum_{p=1}^{P} \beta_p x_{ip} \right)^2 \tag{1}$$

$$s.t. \qquad \sum_{p=1}^{P} \left| \beta_p \right| < \lambda \tag{2}$$

- (1) is the OLS problem.
- (2) constrains the total absolute size of all coefficients

The LASSO (cont.)

- Within the context of an ARp model, the problem can be written as:

$$\max_{\beta_p} \sum_{i=1}^{N} \left(y_t - \sum_{p=1}^{P} \beta_p y_{t-p} \right)^2$$

$$s.t. \qquad \sum_{p=1}^{P} \left| \beta_p \right| < \lambda$$

- λ can be chosen by cross-validation. Let's first look at the procedure
- Load "lars" library and some SPY data.

```
library(lars)
getSymbols('SPY',from='2001-05-12',to='2015-11-12')
prices.spy <- (as.numeric(SPY$SPY.Open))
```

- Define a first difference and the length of the vector

```
dSPY0<-diff(prices.spy,lag=1)
len<-length(dSPY)
```

The LASSO (cont.)

- Define the current difference (dSPY) and five lags

```
dSPY<-dSPY0[6:(len)]
dSPY1<-dSPY0[5:(len-1)]
dSPY2<-dSPY0[4:(len-2)]
dSPY3<-dSPY0[3:(len-3)]
dSPY4<-dSPY0[2:(len-4)]
dSPY5<-dSPY0[1:(len-5)]
```

- Run a regression, a LASSO, and compare coefficients

```
lm.reg<-lm(dSPY~dSPY1+dSPY2+dSPY3+dSPY4+dSPY5-1)
d<-cbind(dSPY1,dSPY2,dSPY3,dSPY4,dSPY5)
lasso.reg<-lars(d,dSPY,type="lasso",normalize=FALSE)
coef(lm.reg)
coef(lasso.reg)
plot(lasso.reg)
```

The LASSO (cont.)

- The x-axis on the plot represents:

$$s = \frac{\sum_p |\beta_p|}{\sum_p |\beta_p^{ols}|}$$

- Choose the optimal λ via cross validation

```
CVlasso<-cv.lars(d,dSPY,type="lasso",normalize=FALSE)
str(CVlasso)
```

- Extract the optimal s using "which.min" and "index"

```
opt<-CVlasso$index[which.min(CVlasso$cv)]
predict(lasso.reg,s=opt,type="coef",mode="fraction")
```

- This particular prediction code gives the coefficient estimates under the CV-optimal solution.