Empirical Finance: Study Questions Week 5
Imperial College London
Hormoz Ramian

Question 1 Suppose a time-series of variable y_t for

$$y_t = u_t + \phi u_{t-1} \tag{1}$$

where u_t is a white noise with variance term σ^2.

(1.1) Show that $\mathbb{E}[y_t] = 0$, $\text{var}[y_t] = (1+\phi^2)\sigma^2$ and $\rho_1 = \phi/(1+\phi^2)$ where ρ_1 is the first autocorrelation function of y_t.

(1.2) Suppose $|\phi| < 1$ and show that:

$$y_t = \sum_{j=1}^{m} \alpha_j y_{t-j} + u_t + (-1)^{m+2}\phi^{m+1} u_{t-m-1} \tag{2}$$

where $\alpha_j = (-1)^{j+1}\phi^j$ for any positive integer m.

(1.3) Given $|\phi| < 1$ then show that we can re-write the MA(1) model as an AR(∞):

$$y_t = \sum_{j=1}^{\infty} \alpha_j y_{t-j} + u_t \tag{3}$$

Question 2 Ridge regression shrinks the regression coefficients by imposing a penalty on their size. The ridge coefficients minimize a penalized residual sum of squares

$$\widehat{\boldsymbol{\beta}}^{\text{Ridge}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{K} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{K} \beta_j^2 \right\} \tag{4}$$

where $\lambda \geq 0$ is a complexity parameter that controls the amount of shrinkage: the larger the value of λ, the greater the amount of shrinkage which implies a size constraint on the parameters. Notice that the intercept β_0 can be left out of the penalty term.

(2.1) Derive the first order conditions for this problem.

(2.2) Derive an analytical solution for $\widehat{\boldsymbol{\beta}}^{\text{Ridge}}$ in terms of $\{x_i, y_i, \lambda\}$ and discuss what conditions are needed for this problem to be well-defined.

(2.3) Show the bias term $\widehat{\boldsymbol{\beta}}^{\text{Ridge}}$ in terms of λ.

(2.4) Provide an expression for $\mathbb{V}ar[\widehat{\boldsymbol{\beta}}^{\text{Ridge}}|\boldsymbol{X}]$ and discuss how choice of λ derives the trade off between bias and variance.

(2.5) Suppose $\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{I}$ and show that $\widehat{\boldsymbol{\beta}}^{\text{Ridge}} = \widehat{\beta}^{\text{OLS}}/(1+\lambda)$

Answers

Question 1

(1.1) Since u_t is white noise then $\mathbb{E}u_t = 0\ \forall_t$, and $\mathbb{E}u_\tau u_s = \text{cov}(u_\tau, u_s) = 0\ \forall_{\tau \neq s}$, then:

$$\begin{aligned}
\mathbb{E}[y_t] &= \mathbb{E}[u_t + \phi u_{t-1}] \ = \ \mathbb{E}[u_t] + \phi\mathbb{E}[u_{t-1}] \ = \ 0 \\
\text{var}[y_t] &= \mathbb{E}\left[(u_t + \phi u_{t-1})^2\right] \ = \ \mathbb{E}u_t^2 + \phi^2\mathbb{E}u_{t-1}^2 + 2\phi\mathbb{E}u_t u_{t-1} \ = \ (1+\phi^2)\sigma^2 \\
\text{cov}[y_t, y_{t-1}] &= \mathbb{E}[(u_t + \phi u_{t-1})(u_{t-1} + \phi u_{t-2})] \\
&= \mathbb{E}[u_t u_{t-1}] + \phi\mathbb{E}[u_t u_{t-2}] + \phi\mathbb{E}[u_{t-1}^2] + \phi^2\mathbb{E}[u_{t-1}u_{t-2}] \ = \ \phi\mathbb{E}[u_{t-1}^2] \ = \ \phi\sigma^2
\end{aligned}$$

given $\mathbb{E}[u_t u_{t-1}] = \mathbb{E}[u_t u_{t-2}] = \mathbb{E}[u_{t-1}u_{t-2}] = 0$. The first autocorrelation function is $\rho_1 = \text{cov}[y_t, y_{t-1}]/\,\text{var}[y_t] = \phi/(1+\phi^2)$.

(1.2) Re-write $u_{t-k} = y_{t-k} - \phi u_{t-k-1}$ for any k and iterate to obtain:

$$\begin{aligned}
y_t &= u_t + \phi(y_{t-1} - \phi u_{t-2}) \ = \ \phi y_{t-1} + u_t - \phi^2 u_{t-2} \\
&= \phi y_{t-1} + u_t - \phi^2(y_{t-2} - \phi u_{t-3}) \\
&= \phi y_{t-1} - \phi^2 y_{t-2} + u_t + \phi^3 u_{t-3} \\
&= \sum_{j=1} \alpha_j y_{t-j} + u_t + (-1)^{m+2}\phi^{m+1}u_{t-m-1}
\end{aligned}$$

with $\alpha_j = (-1)^{j+1}\phi^j$.

(1.3) Since $|\phi| < 1$ then $y_t = u_t + \phi u_{t-1} \ = \ (1+\phi L)u_t$ can be arranged as:

$$\begin{aligned}
u_t &= \frac{1}{1+\phi L} y_t \\
&= (1 - \phi L + \phi^2 L^2 - \phi^3 L^3 + \ldots)y_t \\
&= y_t + (-\phi L + \phi^2 L^2 - \phi^3 L^3 + \ldots)y_t \\
&= y_t + \sum_{j=1}^{\infty}(-1)^j \phi^j L^j y_t \ = \ y_t + \sum_{j=1}^{\infty}(-1)^j\phi^j y_{t-j}
\end{aligned}$$

hence we obtain:

$$y_t \ = \ u_t + \sum_{j=1}^{\infty}(-1)^{j+1}\phi^j y_{t-j}$$

Question 2

(2.1) Re-write the penalized sum of squared residuals (PRSS) as the following (all terms in the second line are scalars 1×1):

$$\begin{aligned}
\text{PRSS} &= (\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta}) + \lambda\boldsymbol{\beta}'\boldsymbol{\beta} \\
&= \boldsymbol{y}'\boldsymbol{y} - \boldsymbol{y}'\boldsymbol{X\beta} - \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X\beta} + \lambda\boldsymbol{\beta}'\boldsymbol{\beta} \\
&= \boldsymbol{y}'\boldsymbol{y} - 2\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X\beta} + \lambda\boldsymbol{\beta}'\boldsymbol{\beta}
\end{aligned}$$

differentiate w.r.t. to $\boldsymbol{\beta}$ to obtain the following vector:

$$\begin{aligned}
\boldsymbol{0}_{K\times 1} &= -2\boldsymbol{X}'\boldsymbol{y} + 2\boldsymbol{X}'\boldsymbol{X}\widehat{\boldsymbol{\beta}}^{\text{ridge}} + 2\lambda\widehat{\boldsymbol{\beta}}^{\text{ridge}} \\
&= -\boldsymbol{X}'\boldsymbol{y} + (\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})\widehat{\boldsymbol{\beta}}^{\text{ridge}}
\end{aligned}$$

(2.2) The solution requires $(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}$ to exist or to require $\text{rank}(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I}) = K$. Note that in this problem $K > N$ is not a requirement for invertibility because the linear combination of $\boldsymbol{X}'\boldsymbol{X}$ with the identity matrix is full column rank regardless of number of observations:

$$\widehat{\boldsymbol{\beta}}^{\text{ridge}} = (\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y} \tag{5}$$

(2.3)

$$\begin{aligned}\widehat{\boldsymbol{\beta}}^{\text{ridge}} &= (\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}) \\ &= (\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}\boldsymbol{u} \end{aligned} \tag{6}$$

then:

$$\begin{aligned}\mathbb{E}[\widehat{\boldsymbol{\beta}}^{\text{ridge}}|\boldsymbol{X}] &= (\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}\mathbb{E}[\boldsymbol{u}|\boldsymbol{X}] \\ &= (\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} \\ &= (\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I} - \lambda\boldsymbol{I})\boldsymbol{\beta} \\ &= (\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})\boldsymbol{\beta} - \lambda(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{\beta} \\ &= \left[\boldsymbol{I} - \lambda(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\right]\boldsymbol{\beta}\end{aligned}$$

(2.4) Take variance directly from equation (6) to obtain the following $K \times K$ matrix:

$$\begin{aligned}\mathbb{V}ar\left[\widehat{\boldsymbol{\beta}}^{\text{ridge}}|\boldsymbol{X}\right] &= \mathbf{0}_{K\times 1} + \mathbb{V}ar\left[(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}\boldsymbol{u}|\boldsymbol{X}\right] \\ &= (\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}\mathbb{E}\left[\boldsymbol{u}\boldsymbol{u}'|\boldsymbol{X}\right]\boldsymbol{X}'(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\end{aligned}$$

when $\mathbb{E}\left[\boldsymbol{u}\boldsymbol{u}'|\boldsymbol{X}\right] = \sigma^2\boldsymbol{I}$ then:

$$\mathbb{V}ar\left[\widehat{\boldsymbol{\beta}}^{\text{ridge}}|\boldsymbol{X}\right] = \sigma^2(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}$$

where we note that λ is inversely proportional to variance expression[1].

(2.5) use $\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{I}$ together with equation (5):

$$\begin{aligned}\widehat{\boldsymbol{\beta}}^{\text{ridge}} &= (\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y} \\ &= (1+\lambda)^{-1}\boldsymbol{X}'\boldsymbol{y}\end{aligned}$$

and that $\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{I}$ implies $\widehat{\boldsymbol{\beta}}^{\text{OLS}} = \boldsymbol{X}'\boldsymbol{Y}$ then,

$$\widehat{\boldsymbol{\beta}}^{\text{ridge}} = \widehat{\boldsymbol{\beta}}^{\text{OLS}}/(1+\lambda)$$

[1]Optional: Although the vector of ridge estimators incur a greater bias, it possesses a smaller variance than the vector of OLS estimators, and one may compare these two quantities by taking the trace of the variance matrices of the two methods.