Frequentist Properties of Bayes Estimators

STA721 Linear Models Duke University

Merlise Clyde

October 14, 2019

Recap: Bayesian Estimation with 2 Block g-prior (Normal-Jeffreys)

- Model in centered parameterization $\mathbf{X}_c = (\mathbf{I}_n - \mathbf{P}_1)\mathbf{X}$

$$
\begin{aligned}
\mathbf{Y} &= \mathbf{1}\beta_0 + (\mathbf{I}_n - \mathbf{P}_1)\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \\
\mathbf{Y} &= \mathbf{1}\beta_0 + (\mathbf{X}_c\boldsymbol{\beta} + \boldsymbol{\epsilon} \\
p(\beta_0, \phi) &\propto 1/\phi \\
\boldsymbol{\beta} \mid \beta_0, \phi &\sim \mathrm{N}(\mathbf{0}, \frac{g}{\phi}(\mathbf{X}_c^T\mathbf{X}_c)^{-1})
\end{aligned}
$$

Recap: Bayesian Estimation with 2 Block g-prior (Normal-Jeffreys)

- Model in centered parameterization $\mathbf{X}_c = (\mathbf{I}_n - \mathbf{P}_1)\mathbf{X}$

$$
\begin{aligned}
\mathbf{Y} &= \mathbf{1}\beta_0 + (\mathbf{I}_n - \mathbf{P}_1)\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \\
\mathbf{Y} &= \mathbf{1}\beta_0 + (\mathbf{X}_c\boldsymbol{\beta} + \boldsymbol{\epsilon} \\
p(\beta_0, \phi) &\propto 1/\phi \\
\boldsymbol{\beta} \mid \beta_0, \phi &\sim \mathrm{N}(\mathbf{0}, \frac{g}{\phi}(\mathbf{X}_c^T\mathbf{X}_c)^{-1})
\end{aligned}
$$

- Zellner-Siow prior (assume $\mathbf{X}$ is centered)

Recap: Bayesian Estimation with 2 Block g-prior (Normal-Jeffreys)

- Model in centered parameterization $\mathbf{X}_c = (\mathbf{I}_n - \mathbf{P}_1)\mathbf{X}$

$$
\begin{aligned}
\mathbf{Y} &= \mathbf{1}\beta_0 + (\mathbf{I}_n - \mathbf{P}_1)\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \\
\mathbf{Y} &= \mathbf{1}\beta_0 + (\mathbf{X}_c\boldsymbol{\beta} + \boldsymbol{\epsilon} \\
p(\beta_0, \phi) &\propto 1/\phi \\
\boldsymbol{\beta} \mid \beta_0, \phi &\sim \mathrm{N}(\mathbf{0}, \frac{g}{\phi}(\mathbf{X}_c^T\mathbf{X}_c)^{-1})
\end{aligned}
$$

- Zellner-Siow prior (assume $\mathbf{X}$ is centered) Zellner's g-prior $\boldsymbol{\beta} \mid \phi \sim \mathrm{N}(\mathbf{0}_p, g(\mathbf{X}_c^T\mathbf{X}_c)^{-1}/\phi)$

Recap: Bayesian Estimation with 2 Block g-prior (Normal-Jeffreys)

- Model in centered parameterization $\mathbf{X}_c = (\mathbf{I}_n - \mathbf{P}_1)\mathbf{X}$

$$
\begin{aligned}
\mathbf{Y} &= \mathbf{1}\beta_0 + (\mathbf{I}_n - \mathbf{P}_1)\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \\
\mathbf{Y} &= \mathbf{1}\beta_0 + (\mathbf{X}_c\boldsymbol{\beta} + \boldsymbol{\epsilon} \\
p(\beta_0, \phi) &\propto 1/\phi \\
\boldsymbol{\beta} \mid \beta_0, \phi &\sim \mathrm{N}(\mathbf{0}, \frac{g}{\phi}(\mathbf{X}_c^T\mathbf{X}_c)^{-1})
\end{aligned}
$$

- Zellner-Siow prior (assume $\mathbf{X}$ is centered) Zellner's g-prior $\boldsymbol{\beta} \mid \phi \sim \mathrm{N}(\mathbf{0}_p, g(\mathbf{X}_c^T\mathbf{X}_c)^{-1}/\phi)$
- Let $\tau = 1/g$ assign $\tau \sim G(1/2, n/2)$

Recap: Bayesian Estimation with 2 Block g-prior (Normal-Jeffreys)

- Model in centered parameterization $\mathbf{X}_c = (\mathbf{I}_n - \mathbf{P}_1)\mathbf{X}$

$$
\begin{aligned}
\mathbf{Y} &= \mathbf{1}\beta_0 + (\mathbf{I}_n - \mathbf{P}_1)\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \\
\mathbf{Y} &= \mathbf{1}\beta_0 + (\mathbf{X}_c\boldsymbol{\beta} + \boldsymbol{\epsilon} \\
p(\beta_0, \phi) &\propto 1/\phi \\
\boldsymbol{\beta} \mid \beta_0, \phi &\sim \mathrm{N}(\mathbf{0}, \frac{g}{\phi}(\mathbf{X}_c^T\mathbf{X}_c)^{-1})
\end{aligned}
$$

- Zellner-Siow prior (assume $\mathbf{X}$ is centered) Zellner's g-prior $\boldsymbol{\beta} \mid \phi \sim \mathrm{N}(\mathbf{0}_p, g(\mathbf{X}_c^T\mathbf{X}_c)^{-1}/\phi)$
- Let $\tau = 1/g$ assign $\tau \sim G(1/2, n/2)$
- Marginal prior on $\boldsymbol{\beta} \sim C(0, \phi^{-1}(\mathbf{X}_c^T\mathbf{X}_c/n)^{-1})$

Recap: Bayesian Estimation with 2 Block g-prior (Normal-Jeffreys)

- Model in centered parameterization $\mathbf{X}_c = (\mathbf{I}_n - \mathbf{P}_1)\mathbf{X}$

$$
\begin{aligned}
\mathbf{Y} &= \mathbf{1}\beta_0 + (\mathbf{I}_n - \mathbf{P}_1)\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \\
\mathbf{Y} &= \mathbf{1}\beta_0 + (\mathbf{X}_c\boldsymbol{\beta} + \boldsymbol{\epsilon} \\
p(\beta_0, \phi) &\propto 1/\phi \\
\boldsymbol{\beta} \mid \beta_0, \phi &\sim \mathrm{N}(\mathbf{0}, \frac{g}{\phi}(\mathbf{X}_c^T\mathbf{X}_c)^{-1})
\end{aligned}
$$

- Zellner-Siow prior (assume $\mathbf{X}$ is centered) Zellner's g-prior $\boldsymbol{\beta} \mid \phi \sim \mathrm{N}(\mathbf{0}_p, g(\mathbf{X}_c^T\mathbf{X}_c)^{-1}/\phi)$
- Let $\tau = 1/g$ assign $\tau \sim G(1/2, n/2)$
- Marginal prior on $\boldsymbol{\beta} \sim C(0, \phi^{-1}(\mathbf{X}_c^T\mathbf{X}_c/n)^{-1})$
- Use Gibbs sampling or MCMC

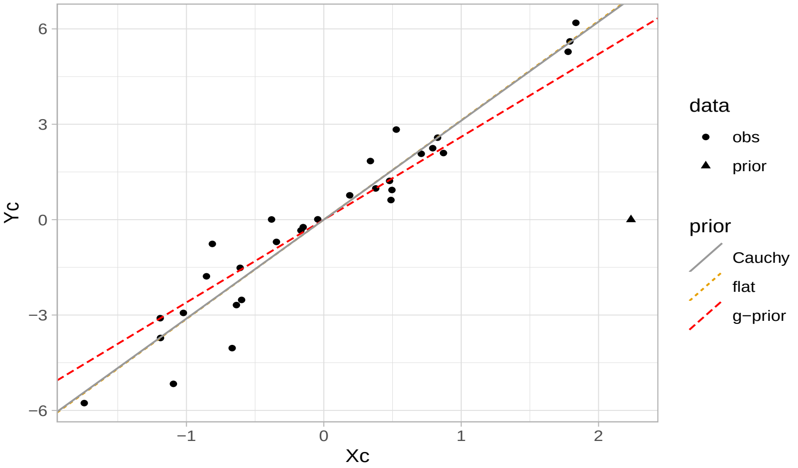
JAGS Code: library(R2jags)

```
model = function(){
  for (i in 1:n) {
      Y[i] ~ dnorm(beta0+ (X[i] -Xbar)*beta, phi)
  }
  beta0 ~ dnorm(0, .000001*phi) #precision is 2nd arg
  beta ~ dnorm(0, phi*tau*SSX)  #precision is 2nd arg
  phi ~ dgamma(.001, .001)
  tau ~ dgamma(.5, .5*n)
  g <- 1/tau
  sigma <- pow(phi, -.5)
}
data = list(Y=Y, X=X, n =length(Y), SSX=sum(Xc^2),
            Xbar=mean(X))
ZSout = jags(data, inits=NULL,
             parameters.to.save=c("beta0","beta", "g",
                                  "sigma"),
             model=model, n.iter=10000)
```

HPD intervals

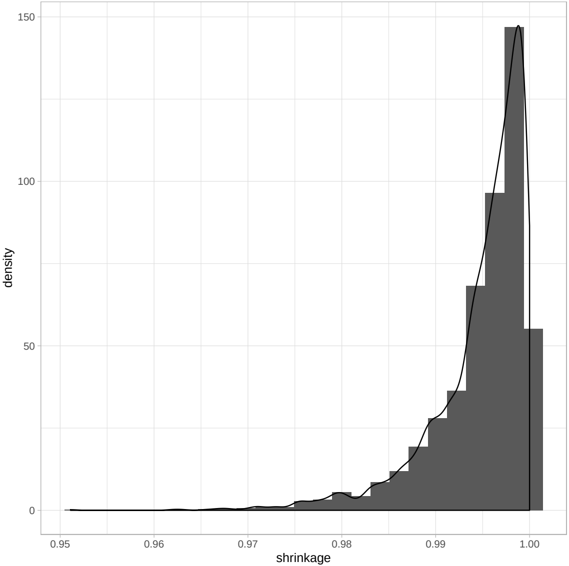
```r
confint(lm(Y ~ Xc))

##                  2.5 %    97.5 %
## (Intercept) -0.3985359 0.2048303
## Xc           2.7945824 3.4555162

HPDinterval(as.mcmc(ZSout$BUGSoutput$sims.matrix))

##                lower        upper
## beta       2.7823047    3.4453690
## beta0     -0.3764027    0.2095465
## deviance  70.2043917   78.4813041
## g         19.4503373 3782.7134974
## sigma      0.6171029    1.0504892
## attr(,"Probability")
## [1] 0.95
```

Compare



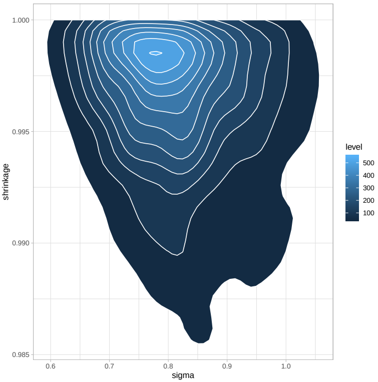
```
ZSout

## Inference for Bugs model at "/var/folders/n4/nj1122xj6bn5_xgbptv7bml40000gp/T//RtmpwDmY3Y/model17fcc743
##  3 chains, each with 10000 iterations (first 5000 discarded), n.thin = 5
##  n.sims = 3000 iterations saved
##           mu.vect   sd.vect   2.5%     25%     50%     75%    97.5%  Rhat
## beta        3.112     0.170  2.782   2.997   3.115   3.225    3.445 1.001
## beta0      -0.099     0.152 -0.384  -0.204  -0.099   0.001    0.204 1.002
## g        2263.147 38967.029 48.273 146.129 282.298 697.063 9018.709 1.001
## sigma       0.827     0.114  0.636   0.747   0.816   0.896    1.079 1.001
## deviance   73.347     2.563 70.390  71.458  72.680  74.500   79.882 1.002
##          n.eff
## beta      3000
## beta0     1200
## g         3000
## sigma     3000
## deviance  1600
##
## For each parameter, n.eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor (at convergence, Rhat=1).
##
## DIC info (using the rule, pD = var(deviance)/2)
## pD = 3.3 and DIC = 76.6
## DIC is an estimate of expected predictive error (lower deviance is better).
```

Posterior Distribution of shrinkage



Joint Distribution of σ and $g/(1+g)$

```
ggplot(postdf, aes(x=sigma, y=shrinkage) ) +
 stat_density_2d(aes(fill = ..level..),
                geom = "polygon", colour="white") +
  theme_light()
```



Cauchy Summary

- Cauchy rejects prior mean if it is an "outlier"
- robustness related to "bounded" influence (more later)
- requires numerical integration or Monte Carlo sampling (MCMC)

How Good are Bayes Estimators?

Quadratic loss for estimating $\boldsymbol{\beta}$ using estimator $\mathbf{a}$

$$L(\boldsymbol{\beta}, \mathbf{a}) = (\boldsymbol{\beta} - \mathbf{a})^T(\boldsymbol{\beta} - \mathbf{a})$$

How Good are Bayes Estimators?

Quadratic loss for estimating $\boldsymbol{\beta}$ using estimator $\mathbf{a}$

$$L(\boldsymbol{\beta}, \mathbf{a}) = (\boldsymbol{\beta} - \mathbf{a})^T(\boldsymbol{\beta} - \mathbf{a})$$

- Consider our expected loss (before we see the data) of taking an "action" $\mathbf{a}$

How Good are Bayes Estimators?

Quadratic loss for estimating $\boldsymbol{\beta}$ using estimator $\mathbf{a}$

$$L(\boldsymbol{\beta}, \mathbf{a}) = (\boldsymbol{\beta} - \mathbf{a})^T(\boldsymbol{\beta} - \mathbf{a})$$

- Consider our expected loss (before we see the data) of taking an "action" $\mathbf{a}$
- Under OLS or the Reference prior the Expected Mean Square Error

How Good are Bayes Estimators?

Quadratic loss for estimating $\boldsymbol{\beta}$ using estimator $\mathbf{a}$

$$L(\boldsymbol{\beta}, \mathbf{a}) = (\boldsymbol{\beta} - \mathbf{a})^T(\boldsymbol{\beta} - \mathbf{a})$$

- Consider our expected loss (before we see the data) of taking an "action" $\mathbf{a}$
- Under OLS or the Reference prior the Expected Mean Square Error

$$\mathsf{E}_{\mathbf{Y}}[(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \quad = \quad \sigma^2 \mathrm{tr}[(\mathbf{X}^T\mathbf{X})^{-1}]$$

How Good are Bayes Estimators?

Quadratic loss for estimating $\boldsymbol{\beta}$ using estimator $\mathbf{a}$

$$L(\boldsymbol{\beta}, \mathbf{a}) = (\boldsymbol{\beta} - \mathbf{a})^T(\boldsymbol{\beta} - \mathbf{a})$$

- Consider our expected loss (before we see the data) of taking an "action" $\mathbf{a}$
- Under OLS or the Reference prior the Expected Mean Square Error

$$\begin{aligned} \mathsf{E}_{\mathbf{Y}}[(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) &= \sigma^2 \mathrm{tr}[(\mathbf{X}^T\mathbf{X})^{-1}] \\ &= \sigma^2 \sum_{j=1}^{p} \lambda_j^{-1} \end{aligned}$$

where λ_j are eigenvalues of $\mathbf{X}^T\mathbf{X}$.

How Good are Bayes Estimators?

Quadratic loss for estimating $\boldsymbol{\beta}$ using estimator $\mathbf{a}$

$$L(\boldsymbol{\beta}, \mathbf{a}) = (\boldsymbol{\beta} - \mathbf{a})^T(\boldsymbol{\beta} - \mathbf{a})$$

- Consider our expected loss (before we see the data) of taking an "action" $\mathbf{a}$
- Under OLS or the Reference prior the Expected Mean Square Error

$$\begin{aligned} \mathsf{E}_{\mathbf{Y}}[(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) &= \sigma^2 \text{tr}[(\mathbf{X}^T\mathbf{X})^{-1}] \\ &= \sigma^2 \sum_{j=1}^{p} \lambda_j^{-1} \end{aligned}$$

  where λ_j are eigenvalues of $\mathbf{X}^T\mathbf{X}$.
- If smallest $\lambda_j \to 0$ then MSE $\to \infty$
- Note: estimate is unbiased!

Is the g-prior better?

-
- Explore Frequentist properties of using a Bayesian estimator

$$\mathsf{E}_{\mathbf{Y}}[(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_g)^T(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_g)$$

but now $\hat{\boldsymbol{\beta}}_g = g/(1+g)\hat{\boldsymbol{\beta}}$

Is the g-prior better?

-
- Explore Frequentist properties of using a Bayesian estimator

$$\mathsf{E}_{\mathbf{Y}}[(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_g)^T(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_g)$$

  but now $\hat{\boldsymbol{\beta}}_g = g/(1+g)\hat{\boldsymbol{\beta}}$
- Sampling distribution of $\hat{\boldsymbol{\beta}}_g = \frac{g}{1+g}(\mathbf{X}^{\mathbf{X}})^{-1}\mathbf{X}^T\mathbf{Y}$

Is the g-prior better?

-
- Explore Frequentist properties of using a Bayesian estimator

$$\mathsf{E}_{\mathbf{Y}}[(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_g)^T(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_g)$$

  but now $\hat{\boldsymbol{\beta}}_g = g/(1+g)\hat{\boldsymbol{\beta}}$
- Sampling distribution of $\hat{\boldsymbol{\beta}}_g = \frac{g}{1+g}(\mathbf{X}^{\mathbf{X}})^{-1}\mathbf{X}^T\mathbf{Y}$
- HW: show that there is a value of g prior such that the g-prior is always better than the Reference prior/OLS
- Potential problem: MSE also blows up if smallest eigenvalue goes to zero!

Estimator Properties

- Bias

Estimator Properties

- Bias
- Variance

Estimator Properties

- Bias
- Variance
- $\text{MSE} = \text{Bias}^2 + \text{Variance}$ (multivariate analogs)

Estimator Properties

- Bias
- Variance
- MSE $=$ Bias^2 $+$ Variance (multivariate analogs)
- Problems with OLS, g-priors & mixtures of g-priors with collinearity

Estimator Properties

- Bias
- Variance
- MSE $=$ Bias^2 $+$ Variance (multivariate analogs)
- Problems with OLS, g-priors & mixtures of g-priors with collinearity
- Solutions:

Estimator Properties

- Bias
- Variance
- MSE $=$ Bias^2 $+$ Variance (multivariate analogs)
- Problems with OLS, g-priors & mixtures of g-priors with collinearity
- Solutions:
  - removal of terms

Estimator Properties

- Bias
- Variance
- MSE $=$ Bias^2 $+$ Variance (multivariate analogs)
- Problems with OLS, g-priors & mixtures of g-priors with collinearity
- Solutions:
  - removal of terms
  - other shrinkage estimators

Canonical Representation & Ridge Regression

- Assume that $\mathbf{X}$ has been centered and standardized so that $\mathbf{X}^T\mathbf{X} = \mathrm{corr}(\mathbf{X})$

Canonical Representation & Ridge Regression

- Assume that $\mathbf{X}$ has been centered and standardized so that $\mathbf{X}^T\mathbf{X} = \text{corr}(\mathbf{X})$ (use `scale` or `sweep` functions in `R`)

Canonical Representation & Ridge Regression

- Assume that $\mathbf{X}$ has been centered and standardized so that $\mathbf{X}^T\mathbf{X} = \text{corr}(\mathbf{X})$ (use `scale` or `sweep` functions in `R`)
- Write $\mathbf{X} = \mathbf{U}_p L \mathbf{V}^T$ Singular Value Decomposition

Canonical Representation & Ridge Regression

- Assume that $\mathbf{X}$ has been centered and standardized so that $\mathbf{X}^T\mathbf{X} = \text{corr}(\mathbf{X})$ (use `scale` or `sweep` functions in `R`)
- Write $\mathbf{X} = \mathbf{U}_p L \mathbf{V}^T$ Singular Value Decomposition where $\mathbf{U}_p^T \mathbf{U}_p = \mathbf{I}_p$ and $\mathbf{V}$ is $p \times p$ orthogonal matrix, L is diagonal

Canonical Representation & Ridge Regression

- Assume that $\mathbf{X}$ has been centered and standardized so that $\mathbf{X}^T\mathbf{X} = \text{corr}(\mathbf{X})$ (use `scale` or `sweep` functions in `R`)
- Write $\mathbf{X} = \mathbf{U}_p L \mathbf{V}^T$ Singular Value Decomposition where $\mathbf{U}_p^T \mathbf{U}_p = \mathbf{I}_p$ and $\mathbf{V}$ is $p \times p$ orthogonal matrix, L is diagonal

$$\mathbf{Y} = \mathbf{1}\alpha + \mathbf{U}_p L \mathbf{V}^T \boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Canonical Representation & Ridge Regression

- Assume that $\mathbf{X}$ has been centered and standardized so that $\mathbf{X}^T\mathbf{X} = \text{corr}(\mathbf{X})$ (use `scale` or `sweep` functions in `R`)
- Write $\mathbf{X} = \mathbf{U}_p L \mathbf{V}^T$ Singular Value Decomposition where $\mathbf{U}_p^T \mathbf{U}_p = \mathbf{I}_p$ and $\mathbf{V}$ is $p \times p$ orthogonal matrix, L is diagonal

$$\mathbf{Y} = \mathbf{1}\alpha + \mathbf{U}_p L \mathbf{V}^T \boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- Let $\gamma = \mathbf{V}^T\boldsymbol{\beta}$ and create $\mathbf{U}$ an $n \times n$ orthogonal matrix

Canonical Representation & Ridge Regression

- Assume that $\mathbf{X}$ has been centered and standardized so that $\mathbf{X}^T\mathbf{X} = \text{corr}(\mathbf{X})$ (use `scale` or `sweep` functions in `R`)
- Write $\mathbf{X} = \mathbf{U}_p L \mathbf{V}^T$ Singular Value Decomposition where $\mathbf{U}_p^T \mathbf{U}_p = \mathbf{I}_p$ and $\mathbf{V}$ is $p \times p$ orthogonal matrix, L is diagonal

$$\mathbf{Y} = \mathbf{1}\alpha + \mathbf{U}_p L \mathbf{V}^T \boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- Let $\gamma = \mathbf{V}^T\boldsymbol{\beta}$ and create $\mathbf{U}$ an $n \times n$ orthogonal matrix

$$\mathbf{U} = [\mathbf{U}_0 \,|\, \mathbf{U}_p \,|\, \mathbf{U}_{n-p-1}]$$

where $\mathbf{U}_0 = \mathbf{1}/\sqrt{n}$

Canonical Representation & Ridge Regression

- Assume that $\mathbf{X}$ has been centered and standardized so that $\mathbf{X}^T\mathbf{X} = \text{corr}(\mathbf{X})$ (use `scale` or `sweep` functions in `R`)
- Write $\mathbf{X} = \mathbf{U}_p L \mathbf{V}^T$ Singular Value Decomposition where $\mathbf{U}_p^T \mathbf{U}_p = \mathbf{I}_p$ and $\mathbf{V}$ is $p \times p$ orthogonal matrix, L is diagonal

$$\mathbf{Y} = \mathbf{1}\alpha + \mathbf{U}_p L \mathbf{V}^T \boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- Let $\gamma = \mathbf{V}^T\boldsymbol{\beta}$ and create $\mathbf{U}$ an $n \times n$ orthogonal matrix

$$\mathbf{U} = [\mathbf{U}_0 \,|\, \mathbf{U}_p \,|\, \mathbf{U}_{n-p-1}]$$

  where $\mathbf{U}_0 = \mathbf{1}/\sqrt{n}$
- $\mathbf{U}_0^T \mathbf{U}_p = 0$, $\mathbf{U}_0^T \mathbf{U}_{n-p-1} = 0$ and $\mathbf{U}_p^T \mathbf{U}_{n-p-1} = 0$ (orthogonal columns)

Orthogonal Regression

Rotate by multiplyting by $\mathbf{U}^T$:

$$\mathbf{U}^T\mathbf{Y} = \mathbf{U}^T\mathbf{1}\alpha + \mathbf{U}^T\mathbf{U}_p L \mathbf{V}^T\boldsymbol{\beta} + \mathbf{U}^T\boldsymbol{\epsilon}$$

Orthogonal Regression

Rotate by multiplyting by $\mathbf{U}^T$:

$$\begin{aligned}
\mathbf{U}^T\mathbf{Y} &= \mathbf{U}^T\mathbf{1}\alpha + \mathbf{U}^T\mathbf{U}_p L\mathbf{V}^T\boldsymbol{\beta} + \mathbf{U}^T\boldsymbol{\epsilon} \\
\mathbf{Y}^* &= \begin{bmatrix} \sqrt{n} & \mathbf{0}_p^T \\ \mathbf{0}_p & L \\ \mathbf{0}_{n-p-1} & \mathbf{0}_{n-p-1\times p} \end{bmatrix} \begin{pmatrix} \alpha \\ \gamma \end{pmatrix} + \boldsymbol{\epsilon}^*
\end{aligned}$$

Orthogonal Regression

Rotate by multiplyting by $\mathbf{U}^T$:

$$\begin{aligned}
\mathbf{U}^T\mathbf{Y} &= \mathbf{U}^T\mathbf{1}\alpha + \mathbf{U}^T\mathbf{U}_p L \mathbf{V}^T\boldsymbol{\beta} + \mathbf{U}^T\boldsymbol{\epsilon} \\
\mathbf{Y}^* &= \begin{bmatrix} \sqrt{n} & \mathbf{0}_p^T \\ \mathbf{0}_p & L \\ \mathbf{0}_{n-p-1} & \mathbf{0}_{n-p-1\times p} \end{bmatrix} \begin{pmatrix} \alpha \\ \gamma \end{pmatrix} + \boldsymbol{\epsilon}^*
\end{aligned}$$

- $y_0^* \equiv \hat{\alpha} = \bar{y}$

Orthogonal Regression

Rotate by multiplyting by $\mathbf{U}^T$:

$$\begin{aligned}
\mathbf{U}^T\mathbf{Y} &= \mathbf{U}^T\mathbf{1}\alpha + \mathbf{U}^T\mathbf{U}_p L \mathbf{V}^T\boldsymbol{\beta} + \mathbf{U}^T\boldsymbol{\epsilon} \\
\mathbf{Y}^* &= \begin{bmatrix} \sqrt{n} & \mathbf{0}_p^T \\ \mathbf{0}_p & L \\ \mathbf{0}_{n-p-1} & \mathbf{0}_{n-p-1\times p} \end{bmatrix} \begin{pmatrix} \alpha \\ \gamma \end{pmatrix} + \boldsymbol{\epsilon}^*
\end{aligned}$$

- $y_0^* \equiv \hat{\alpha} = \bar{y}$
- $\hat{\gamma} = (L^T L)^{-1} L^T \mathbf{U}_p^T \mathbf{Y}$ or $\hat{\gamma}_i = y_i^*/l_i$ for $i = 1, \ldots, p$

Orthogonal Regression

Rotate by multiplyting by $\mathbf{U}^T$:

$$
\begin{aligned}
\mathbf{U}^T\mathbf{Y} &= \mathbf{U}^T\mathbf{1}\alpha + \mathbf{U}^T\mathbf{U}_p L \mathbf{V}^T\boldsymbol{\beta} + \mathbf{U}^T\boldsymbol{\epsilon} \\
\mathbf{Y}^* &= \begin{bmatrix} \sqrt{n} & \mathbf{0}_p^T \\ \mathbf{0}_p & L \\ \mathbf{0}_{n-p-1} & \mathbf{0}_{n-p-1\times p} \end{bmatrix} \begin{pmatrix} \alpha \\ \gamma \end{pmatrix} + \boldsymbol{\epsilon}^*
\end{aligned}
$$

- $y_0^* \equiv \hat{\alpha} = \bar{y}$
- $\hat{\gamma} = (L^T L)^{-1} L^T \mathbf{U}_p^T \mathbf{Y}$ or $\hat{\gamma}_i = y_i^*/l_i$ for $i = 1, \ldots, p$
- $\text{Var}(\hat{\gamma}_i) = \sigma^2/l_i^2$

Directions in $\mathbf{X}$ space $\mathbf{U}_j$ with small eigenvectors l_i have the largest variances. Unstable directions.

Ridge Regression & Independent Prior

(Another) Normal Conjugate Prior Distribution on γ:

$$\boldsymbol{\gamma} \mid \phi \sim \mathsf{N}(\mathbf{0}_p, \frac{1}{\phi k}\mathbf{I}_p)$$

Ridge Regression & Independent Prior

(Another) Normal Conjugate Prior Distribution on γ:

$$\boldsymbol{\gamma} \mid \phi \sim \mathsf{N}(\mathbf{0}_p, \frac{1}{\phi k}\mathbf{I}_p)$$

Posterior mean

$$\tilde{\gamma} = (L^T L + k\mathbf{I})^{-1} L^T \mathbf{U}_p^T \mathbf{Y} = (L^T L + k\mathbf{I})^{-1} L^T L \hat{\gamma}$$

Ridge Regression & Independent Prior

(Another) Normal Conjugate Prior Distribution on γ:

$$\boldsymbol{\gamma} \mid \phi \sim \mathrm{N}(\mathbf{0}_p, \frac{1}{\phi k}\mathbf{I}_p)$$

Posterior mean

$$\tilde{\gamma} = (L^T L + k\mathbf{I})^{-1} L^T \mathbf{U}_p^T \mathbf{Y} = (L^T L + k\mathbf{I})^{-1} L^T L \hat{\boldsymbol{\gamma}}$$

$$\tilde{\gamma}_i = \frac{l_i^2}{l_i^2 + k}\hat{\gamma}_i = \frac{\lambda_i}{\lambda_i + k}\hat{\gamma}_i$$

Ridge Regression & Independent Prior

(Another) Normal Conjugate Prior Distribution on γ:

$$\boldsymbol{\gamma} \mid \phi \sim \mathsf{N}(\mathbf{0}_p, \frac{1}{\phi k}\mathbf{I}_p)$$

Posterior mean

$$\tilde{\gamma} = (L^T L + k\mathbf{I})^{-1} L^T \mathbf{U}_p^T \mathbf{Y} = (L^T L + k\mathbf{I})^{-1} L^T L \hat{\boldsymbol{\gamma}}$$

$$\tilde{\gamma}_i = \frac{l_i^2}{l_i^2 + k}\hat{\gamma}_i = \frac{\lambda_i}{\lambda_i + k}\hat{\gamma}_i$$

- When $\lambda_i \to 0$ then $\tilde{\gamma}_i \to 0$
- When $k \to 0$ we get OLS back but if k gets too big posterior mean goes to zero.

Transform

- Transform back $\tilde{\boldsymbol{\beta}} = \mathbf{V}\tilde{\gamma}$

Transform

- Transform back $\tilde{\boldsymbol{\beta}} = \mathbf{V}\tilde{\gamma}$

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}$$

Transform

- Transform back $\tilde{\boldsymbol{\beta}} = \mathbf{V}\tilde{\gamma}$

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}$$

- importance of standardizing

Transform

- Transform back $\tilde{\boldsymbol{\beta}} = \mathbf{V}\tilde{\gamma}$

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}$$

- importance of standardizing
- Is there a value of k for which ridge is better in terms of Expected MSE than OLS?

Transform

- Transform back $\tilde{\boldsymbol{\beta}} = \mathbf{V}\tilde{\boldsymbol{\gamma}}$

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}$$

- importance of standardizing
- Is there a value of k for which ridge is better in terms of Expected MSE than OLS?
- Choice of k?

MSE

Can show that

$$\mathrm{E}[(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})^T(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})] = \mathrm{E}[(\boldsymbol{\gamma} - \tilde{\boldsymbol{\gamma}})^T(\boldsymbol{\gamma} - \tilde{\boldsymbol{\gamma}})]$$

MSE

Can show that

$$\mathrm{E}[(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})^T(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})] = \mathrm{E}[(\boldsymbol{\gamma} - \tilde{\boldsymbol{\gamma}})^T(\boldsymbol{\gamma} - \tilde{\boldsymbol{\gamma}})]$$

- $\mathrm{Var}(\tilde{\gamma}_i) = \sigma^2 l_i^2/(l_i^2 + k)^2$

MSE

Can show that

$$E[(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})^T(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})] = E[(\boldsymbol{\gamma} - \tilde{\boldsymbol{\gamma}})^T(\boldsymbol{\gamma} - \tilde{\boldsymbol{\gamma}})]$$

- $\mathrm{Var}(\tilde{\gamma}_i) = \sigma^2 l_i^2/(l_i^2 + k)^2$
- Bias of $\tilde{\boldsymbol{\gamma}}$ is $-k\gamma_i/(l_i^2 + k)$

MSE

Can show that

$$E[(\beta - \tilde{\beta})^T(\beta - \tilde{\beta})] = E[(\gamma - \tilde{\gamma})^T(\gamma - \tilde{\gamma})]$$

- $\mathrm{Var}(\tilde{\gamma}_i) = \sigma^2 l_i^2/(l_i^2 + k)^2$
- Bias of $\tilde{\gamma}$ is $-k\gamma_i/(l_i^2 + k)$
- MSE
$$\sigma^2 \sum_i \frac{l_i^2}{(l_i^2 + k)^2} + k^2 \sum_i \frac{\gamma_i^2}{(l_i^2 + k)^2}$$

The derivative with respect to k is negative at $k = 0$, hence the function is decreasing.

MSE

Can show that

$$E[(\beta - \tilde{\beta})^T(\beta - \tilde{\beta})] = E[(\gamma - \tilde{\gamma})^T(\gamma - \tilde{\gamma}]$$

- $\mathrm{Var}(\tilde{\gamma}_i) = \sigma^2 l_i^2/(l_i^2 + k)^2$
- Bias of $\tilde{\gamma}$ is $-k\gamma_i/(l_i^2 + k)$
- MSE

$$\sigma^2 \sum_i \frac{l_i^2}{(l_i^2 + k)^2} + k^2 \sum_i \frac{\gamma_i^2}{(l_i^2 + k)^2}$$

The derivative with respect to k is negative at $k = 0$, hence the function is decreasing.

Since $k = 0$ is OLS, this means that is a value of k that will always be better than OLS

Alternative Motivation

- If $\hat{\beta}$ is unconstrained expect high variance with nearly singular $\mathbf{X}$

Alternative Motivation

- If $\hat{\beta}$ is unconstrained expect high variance with nearly singular $\mathbf{X}$
- Let $\mathbf{Y}^c = (\mathbf{I} - \mathbf{P}_1)\mathbf{Y}$ and $\mathbf{X}^c$ the centered and standardized $\mathbf{X}$ matrix

Alternative Motivation

- If $\hat{\beta}$ is unconstrained expect high variance with nearly singular $\mathbf{X}$
- Let $\mathbf{Y}^c = (\mathbf{I} - \mathbf{P}_1)\mathbf{Y}$ and $\mathbf{X}^c$ the centered and standardized $\mathbf{X}$ matrix
- Control how large coefficients may grow

Alternative Motivation

- If $\hat{\boldsymbol{\beta}}$ is unconstrained expect high variance with nearly singular $\mathbf{X}$
- Let $\mathbf{Y}^c = (\mathbf{I} - \mathbf{P}_1)\mathbf{Y}$ and $\mathbf{X}^c$ the centered and standardized $\mathbf{X}$ matrix
- Control how large coefficients may grow

$$\min_{\boldsymbol{\beta}} (\mathbf{Y}^c - \mathbf{X}^c\beta)^T(\mathbf{Y}^c - \mathbf{X}^c\beta)$$

subject to

$$\sum \beta_j^2 \leq t$$

Alternative Motivation

- If $\hat{\beta}$ is unconstrained expect high variance with nearly singular $\mathbf{X}$
- Let $\mathbf{Y}^c = (\mathbf{I} - \mathbf{P}_1)\mathbf{Y}$ and $\mathbf{X}^c$ the centered and standardized $\mathbf{X}$ matrix
- Control how large coefficients may grow

$$\min_{\boldsymbol{\beta}} (\mathbf{Y}^c - \mathbf{X}^c \beta)^T (\mathbf{Y}^c - \mathbf{X}^c \beta)$$

subject to

$$\sum \beta_j^2 \leq t$$

- Equivalent Quadratic Programming Problem

$$\min_{\boldsymbol{\beta}} \|\mathbf{Y}^c - \mathbf{X}^c \boldsymbol{\beta}\|^2 + k\|\boldsymbol{\beta}\|^2$$

Alternative Motivation

- If $\hat{\beta}$ is unconstrained expect high variance with nearly singular $\mathbf{X}$
- Let $\mathbf{Y}^c = (\mathbf{I} - \mathbf{P}_1)\mathbf{Y}$ and $\mathbf{X}^c$ the centered and standardized $\mathbf{X}$ matrix
- Control how large coefficients may grow

$$\min_{\boldsymbol{\beta}} (\mathbf{Y}^c - \mathbf{X}^c \beta)^T (\mathbf{Y}^c - \mathbf{X}^c \beta)$$

  subject to

$$\sum \beta_j^2 \leq t$$

- Equivalent Quadratic Programming Problem

$$\min_{\boldsymbol{\beta}} \|\mathbf{Y}^c - \mathbf{X}^c \boldsymbol{\beta}\|^2 + k \|\boldsymbol{\beta}\|^2$$

- "penalized" likelihood

Alternative Motivation

- If $\hat{\beta}$ is unconstrained expect high variance with nearly singular $\mathbf{X}$
- Let $\mathbf{Y}^c = (\mathbf{I} - \mathbf{P}_1)\mathbf{Y}$ and $\mathbf{X}^c$ the centered and standardized $\mathbf{X}$ matrix
- Control how large coefficients may grow

$$\min_{\boldsymbol{\beta}} (\mathbf{Y}^c - \mathbf{X}^c\beta)^T(\mathbf{Y}^c - \mathbf{X}^c\beta)$$

subject to

$$\sum \beta_j^2 \leq t$$

- Equivalent Quadratic Programming Problem

$$\min_{\boldsymbol{\beta}} \|\mathbf{Y}^c - \mathbf{X}^c\boldsymbol{\beta}\|^2 + k\|\boldsymbol{\beta}\|^2$$

- "penalized" likelihood

Picture