

Bayesian Ridge and Shrinkage

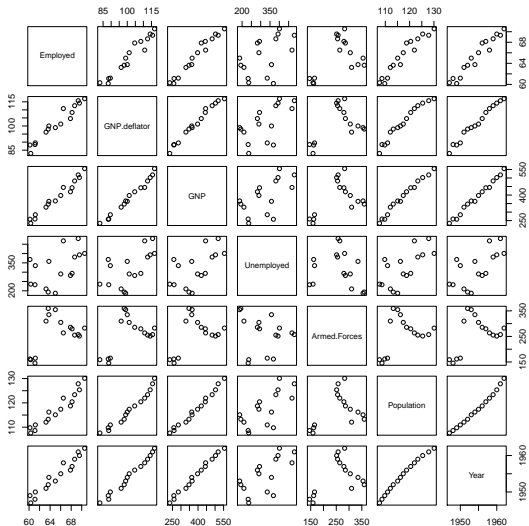
Readings Chapter 15 Christensen

STA721 Linear Models Duke University

Merlise Clyde

October 23, 2018

Longley Data



OLS

```
> longley.lm = lm(Employed ~ ., data=longley)
> summary(longley.lm)
```

Coefficients:

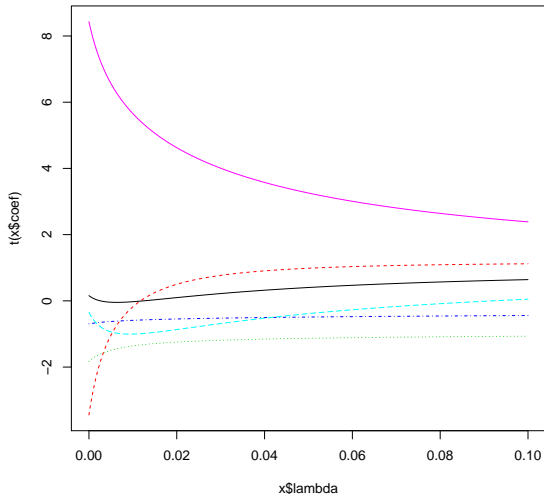
	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-3.482e+03	8.904e+02	-3.911	0.003560	**
GNP.deflator	1.506e-02	8.492e-02	0.177	0.863141	
GNP	-3.582e-02	3.349e-02	-1.070	0.312681	
Unemployed	-2.020e-02	4.884e-03	-4.136	0.002535	**
Armed.Forces	-1.033e-02	2.143e-03	-4.822	0.000944	***
Population	-5.110e-02	2.261e-01	-0.226	0.826212	
Year	1.829e+00	4.555e-01	4.016	0.003037	**

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3049 on 9 degrees of freedom
Multiple R-squared: 0.9955, Adjusted R-squared: 0.9925
F-statistic: 330.3 on 6 and 9 DF, p-value: 4.984e-10

Ridge Trace

```
plot(MASS::lm.ridge(Employed ~ ., data=longley))
```



Generalized Cross-validation: Golub et al (1979)

```
> select(lm.ridge(Employed ~ ., data=longley,  
  lambda=seq(0, 0.1, 0.0001)))
```

modified HKB estimator is 0.004275357

modified L-W estimator is 0.03229531

smallest value of GCV at 0.0028

```
> longley.RReg = lm.ridge(Employed ~ ., data=longley,  
  lambda=0.0028)
```

```
> coef(longley.RReg)
```

	GNP.deflator	GNP	Unemployed	Armed.Forces
	-2.950e+03	-5.381e-04	-1.822e-02	-1.76e-02
				-9.607e-03

Population	Year
-1.185e-01	1.557e+00

https://www.jstor.org/stable/1268518?seq=1#metadata_info_tab_contents

Bayesian Ridge: Prior on k

Reparameterization:

$$\begin{aligned}\mathbf{Y} &= \mathbf{1}\alpha + (\mathbf{I} - \mathbf{P}_1)\mathbf{X}S^{-1/2}S^{1/2}\boldsymbol{\beta} + \boldsymbol{\epsilon} \\ &= \mathbf{1}\alpha + \mathbf{X}^s\boldsymbol{\beta}^s + \boldsymbol{\epsilon} \\ S &= \text{diag}[(n-1)\text{Var}(X_j)] \\ (\mathbf{X}^s)^T\mathbf{X}^s &= \text{Corr}(\mathbf{X})\end{aligned}$$

Hierarchical prior

- ▶ $p(\alpha \mid \phi, \boldsymbol{\beta}^s, \kappa) \propto 1$
- ▶ $\boldsymbol{\beta}_s \mid \phi, \kappa \sim \mathbf{N}(\mathbf{0}, \mathbf{I}(\phi\kappa)^{-1})$
- ▶ $p(\phi \mid \kappa) \propto 1/\phi$
- ▶ prior on κ ? Take $\kappa \mid \phi \sim \mathbf{G}(1/2, 1/2)$

Posterior Distributions

Joint Distribution

- ▶ $\alpha, \beta_s, \phi \mid \kappa, \mathbf{Y}$ Normal-Gamma family given \mathbf{Y} and κ
- ▶ $\kappa \mid \mathbf{Y}$ not tractable

Obtain marginal for β_s via

- ▶ Numerical integration
- ▶ MCMC: Full conditionals

Pick initial values $\alpha^{(0)}, \beta_s^{(0)}, \phi^{(0)}$,

Set $t = 1$

1. Sample $\kappa^{(t)} \sim p(\kappa \mid \alpha^{(t-1)}, \beta_s^{(t-1)}, \phi^{(t-1)}, \mathbf{Y})$
2. Sample $\alpha^{(t)}, \beta_s^{(t)}, \phi^{(t)} \mid \kappa^{(t)}, \mathbf{Y}$
3. Set $t = t + 1$ and repeat until $t > T$

Use Samples $\alpha^{(t)}, \beta_s^{(t)}, \phi^{(t)}, \kappa^{(t)}$ for $t = B, \dots, T$ for inference

Change of variables to get back to β

Full Conditional for κ

Rao-Blackwellization

What is “best” estimate of β_s from Bayesian perspective?

- ▶ Loss $(\beta_s - \mathbf{a})^T(\beta_s - \mathbf{a})$ under action \mathbf{a}
- ▶ Decision Theory: Take action \mathbf{a} that minimizes posterior expected loss which is posterior mean of β_s .
- ▶ Estimate of posterior mean is Ergodic Average of MCMC:
 $\sum_i \beta_s^{(t)} / T \rightarrow$
- ▶ Posterior mean given κ

$$\tilde{\beta}_s(\kappa) = (\mathbf{X}^{sT} \mathbf{X}^s + \kappa \mathbf{I})^{-1} \mathbf{X}^{sT} \mathbf{X}^s \hat{\beta}_s$$

- ▶ Rao-Blackwell Estimate

$$\frac{1}{T} \sum_t (\mathbf{X}^{sT} \mathbf{X}^s + \kappa^{(t)} \mathbf{I})^{-1} \mathbf{X}^{sT} \mathbf{X}^s \hat{\beta}_s$$

Testimators & Canonical Model

$$\mathbf{U}_p \mathbf{Y} = L V^T \boldsymbol{\beta}_s + \epsilon_p \Leftrightarrow \mathbf{U}_p \mathbf{Y} = L \boldsymbol{\gamma} + \epsilon_p$$

Goldstein & Smith (1974) have shown that if

1. $0 \leq h_i \leq 1$ and $\tilde{\gamma}_i = h_i \hat{\gamma}_i$
2. $\frac{\gamma_i^2}{\text{Var}(\hat{\gamma}_i)} < \frac{1+h_i}{1-h_i}$

then $\tilde{\gamma}_i$ has smaller MSE than $\hat{\gamma}_i$

Case: If $\gamma_j^2 < \text{Var}(\hat{\gamma}_i) = \sigma^2 / l_i^2$ then $h_i = 0$ and $\tilde{\gamma}_i$ is better.

Apply: Estimate σ^2 with $\text{SSE} / (n - p - 1)$ and γ_i with $\hat{\gamma}_i$. Set $h_i = 0$ if t-statistic is less than 1.

“testimator” - see also Sclove (JASA 1968) and Copas (JRSSB 1983)

Generalized Ridge

Instead of $\gamma_j \stackrel{\text{iid}}{\sim} \text{N}(0, \sigma^2/\kappa)$ take

$$\gamma_j \stackrel{\text{ind}}{\sim} \text{N}(0, \sigma^2/\kappa_j)$$

Then Condition of Goldstein & Smith becomes

$$\gamma_i^2 < \sigma^2 \left[\frac{2}{\kappa_i} + \frac{1}{\hat{\rho}_i} \right]$$

- ▶ If l_i is small almost any κ_i will improve over OLS
- ▶ if $\hat{\rho}_i$ is large then only very small values of κ_i will give an improvement.
- ▶ Prior on κ_i ?
- ▶ Prior that can capture the feature above?

- Induced prior on β_s ?

$$\gamma_j \mid \sigma^2, \kappa_j \stackrel{\text{ind}}{\sim} \text{N}(0, \sigma^2 / \kappa_j) \Leftrightarrow \beta_s \sim \text{N}(\mathbf{0}, \sigma^2 \mathbf{V} \mathbf{K}^{-1} \mathbf{V}^T)$$

which is not diagonal.

- Or start with

$$\beta_s \mid \sigma^2, \mathbf{K} \sim \text{N}(0, \sigma^2 \mathbf{K}^{-1})$$

- loss of invariance with linear transformations of \mathbf{X}^s
- $\mathbf{X}^s \mathbf{A} \mathbf{A}^{-1} \beta = \mathbf{Z} \alpha$ where $\mathbf{A}^{-1} \beta = \alpha$

Related Regression on PCA

- ▶ Principal Components of \mathbf{X} may be obtained via the Singular Value Decomposition:

$$\mathbf{X} = \mathbf{U}_p \mathbf{L} \mathbf{V}^T$$

- ▶ the λ_i^2 are the eigenvalues of $\mathbf{X}^T \mathbf{X}$

$$\begin{aligned}\mathbf{Y} &= \mathbf{1}\alpha + \mathbf{U}\mathbf{L}\mathbf{V}^T\boldsymbol{\beta} + \boldsymbol{\epsilon} \\ &= \mathbf{1}\alpha + \mathbf{F}\boldsymbol{\gamma} + \boldsymbol{\epsilon}\end{aligned}$$

- ▶ Columns $\mathbf{F}_i \propto \mathbf{U}_i$ are the principal components of the data multivariate data $\mathbf{X}_1, \dots, \mathbf{X}_p$
- ▶ If the direction \mathbf{F}_i is ill-defined ($l_i = 0$ or $\lambda_i < \epsilon$ then we may decide to not use \mathbf{F}_i in the model.
- ▶ equivalent to setting
 - ▶ $\tilde{\gamma}_i = \hat{\gamma}_i$ if $l_i \geq \delta$
 - ▶ $\tilde{\gamma}_i = 0$ if $l_i < \epsilon$

How to choose δ ? Why should \mathbf{Y} be related to first k principal

Summary

- ▶ OLS can clearly be dominated by other estimators for estimating β
- ▶ Lead to Bayes like estimators
- ▶ choice of penalties or prior hyper-parameters
- ▶ hierarchical model with prior on κ_j
- ▶ Shrinkage, dimension reduction & variable selection ?
- ▶ what loss function? Estimation versus prediction? Copas 1983