

Ridge, Bayesian Ridge and Shrinkage

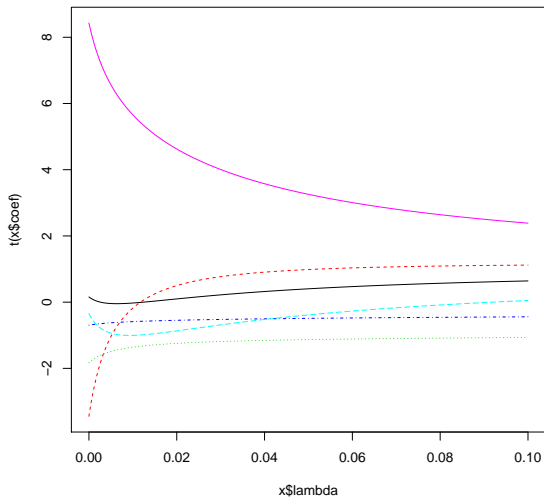
Readings Chapter 15 Christensen

STA721 Linear Models Duke University

Merlise Clyde

October 1, 2015

Ridge Trace



Generalized Cross-validation

```
> 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

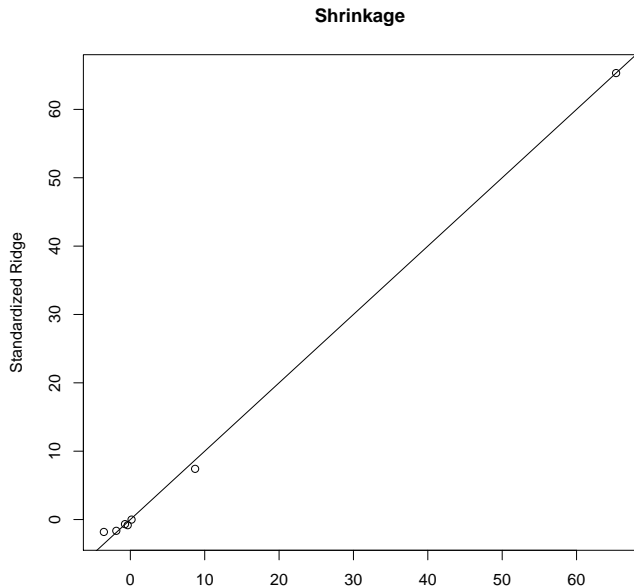
```
> longley.RReg = lm.ridge(Employed ~ ., data=longley,  
                          lambda=0.0028)
```

```
> 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

Shrinkage



Bayesian Ridge: Prior on k

Reparameterization:

$$\begin{aligned}\mathbf{Y} &= \mathbf{1}\alpha + (\mathbf{I} - \mathbf{P}_1)\mathbf{X}\mathbf{S}^{-1/2}\mathbf{S}^{1/2}\boldsymbol{\beta} + \boldsymbol{\epsilon} \\ &= \mathbf{1}\alpha + \mathbf{X}^s\boldsymbol{\beta}^s + \boldsymbol{\epsilon}\end{aligned}$$

$$\mathbf{Y}^c = \mathbf{X}^s\boldsymbol{\beta}^s + \boldsymbol{\epsilon}^s \quad \boldsymbol{\epsilon}^s \sim \mathbf{N}(\mathbf{0}, (\mathbf{I} - \mathbf{P}_1)/\phi)$$

$$\bar{\mathbf{Y}} \mid \alpha, \phi \sim N(\alpha, 1/(n\phi))$$

$$\mathbf{U}_p^T \mathbf{Y} = L\gamma + \epsilon_p \quad \epsilon_p \sim \mathbf{N}(\mathbf{0}, \mathbf{I}_p/\phi)$$

$$\text{SSE} \equiv \mathbf{Y}^T \mathbf{U}_{n-p-1} \mathbf{U}_{n-p-1}^T \mathbf{Y} \sim \mathbf{G}((n-p-1)/2, \phi/2)$$

Hierarchical prior

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

Posterior Distributions

Joint Distribution

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

Obtain marginal for γ via

- Numerical integration
- MCMC: Full conditionals

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

Set $t = 1$

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

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

Change of variables to get back to β

Rao-Blackwellization Model

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

- Loss $(\beta - \mathbf{a})^T(\beta - \mathbf{a})$ under action \mathbf{a}
- Decision Theory: Take action \mathbf{a} that minimizes posterior expected loss which is posterior mean of β .
- 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

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

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 $\gamma_j < \text{Var}(\hat{\gamma}_i) = \sigma^2 / l_i^2$ 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 $\gamma_j \stackrel{\text{iid}}{\sim} N(0, \sigma^2/k)$ take

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

Then Condition of Goldstein & Smith becomes

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

- If l_i is small almost any κ_i will improve over OLS
- if l_i^2 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 ?

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

which is not diagonal.

- Or start with

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

- 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 l_i 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 $l_i < \epsilon$

- 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

Full Conditionals

Full Conditionals

Full Conditionals