L^γ Penalty Models

Computation And Applications

Part II

Wenjiang Fu

wfu@stat.tamu.edu

http://stat.tamu.edu/$\sim$wfu.

Department of Statistics, Texas A&M University

SELECTION OF TUNING PARAMETER

Purpose: to achieve best estimation and prediction

Methods: leave–one–out cross–validation (CV), generalized cross-validation (GCV), etc.

Idea: to fit the model well while penalizing on the model size to prevent overfitting.

SELECTION OF TUNING PARAMETER

CV:

Given sample $S = [(x_1, y_1), \ldots, (x_n, y_n)]$. Leave one obs. (x_i, y_i) out, and fit model f based on remaining sample $S^{(-i)}$. Predict y_i with $y_i^* = f_{S^{(-i)}}(x_i)$.
Define CV $= n^{-1} \sum_{i=1}^{n} (y_i^* - y_i)^2$.

- ◇ CV is computationally expensive.
- ◇ CV is numerically unstable due to outliers, etc.

Selection of Tuning Parameter

GCV (Craven and Wahba 1979)

For shrinkage model (Fu 1998, Tibshirani 1996)

$$\mathrm{GCV} = \frac{(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})}{n\left(1 - \frac{\mathrm{tr}(\boldsymbol{H}) - n_0}{n}\right)^2}, \qquad (5)$$

where $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{W}^-)^{-1}\boldsymbol{X}^T$ is a projection matrix, $\boldsymbol{W}^-$ is generalized inverse of $\boldsymbol{W} = \mathrm{diag}\left(2|\widehat{\beta}_j|^{2-\gamma}/\gamma\right)$ for $\gamma \geq 1$. Let $n_0 = \#\{\widehat{\beta}_j = 0\}$ for lasso only.

Effective number of parameters $p(\lambda, \gamma) = \mathrm{tr}(\boldsymbol{H}) - n_0$.

$p(0, \gamma) = \mathrm{tr}(\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T) = p$, the number of parameters.

$p(\infty, \gamma) = 0$ as $\lambda \to \infty$.

Selection of Tuning Parameter

Select $\boldsymbol{\lambda} \geq 0$ for fixed $\gamma \geq 1$

For each fixed $\gamma \geq 1$, compute GCV for each of a sequence of $\boldsymbol{\lambda} \geq \mathbf{0}$ between 0 and a moderate number. Select the value of $\boldsymbol{\lambda}$ that minimizes GCV.

Select $\boldsymbol{\lambda} \geq 0$ and $\gamma \geq 1$

Compute GCV for each point $(\boldsymbol{\lambda}, \gamma)$ on a lattice of $[\mathbf{0}, \boldsymbol{\lambda}_\mathbf{0}] \times [\mathbf{1}, \mathbf{3}]$ with a moderate number $\boldsymbol{\lambda}_\mathbf{0}$. Select the values of $(\boldsymbol{\lambda}, \gamma)$ that minimize GCV surface.

Problem with GCV

GCV (5) favors lasso even if ridge performs better (Fu 1998).

Reason:

GCV (5) emphasizes linear part by taking $\mathbf{tr}(\boldsymbol{H})$, performs well for linear estimators, such as ridge.

$\widehat{\beta}_{\mathbf{brdg}}$ is nonlinear except for $\gamma = \mathbf{2}$.

For orthonormal $\boldsymbol{X}$ case, lasso is piece–wise linear.

GCV performs poorly in selecting $\boldsymbol{\lambda}$ for $\gamma \neq \mathbf{2}$.

By Taylor expansion,

$$\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{H}(\boldsymbol{y})\boldsymbol{y} =$$
$$\boldsymbol{H}(\boldsymbol{y}_0)\boldsymbol{y}_0 + \{\boldsymbol{H}(\boldsymbol{y}_0) + \boldsymbol{H}'(\boldsymbol{y}_0)\boldsymbol{y}_0\}(\boldsymbol{y} - \boldsymbol{y}_0) + \boldsymbol{o}(\boldsymbol{y} - \boldsymbol{y}_0).$$

Thus $\mathbf{tr}(\boldsymbol{H})$ is a linearization.

Account for nonlinearity

To account nonlinearity, modify GCV (5) through $\boldsymbol{p(\lambda,\gamma)}$. RSS accounts the nonlinearity through the estimator $\widehat{\boldsymbol{\beta}}_{\mathbf{brdg}}$. Instead of separating linear part from nonlinear part, we pool them together and consider the overall shrinkage effect through a standard shrinkage rate $\boldsymbol{s}$.

$$\boldsymbol{s} = \frac{||\widehat{\boldsymbol{\beta}}(\boldsymbol{\lambda},\boldsymbol{\gamma})||_{\gamma}}{||\widehat{\boldsymbol{\beta}^{0}}||_{\gamma}},$$

where $||\cdot||_{\gamma}$ is the $\boldsymbol{L}^{\gamma}$–norm of the shrinkage estimator $\widehat{\boldsymbol{\beta}}(\boldsymbol{\lambda},\boldsymbol{\gamma})$ or the no-shrinkage estimator $\widehat{\boldsymbol{\beta}^{0}}$ with $\boldsymbol{\gamma} \geq \mathbf{1}$. Apparently $\mathbf{0} \leq \boldsymbol{s} \leq \mathbf{1}$.

Nonlinear GCV

Modify the effective number of parameters

$$p(\lambda, \gamma) = ps$$

where p is the number parameters in the model, s is the standard shrinkage rate.

Define the nonlinear GCV as

$$\mathrm{NLGCV} = \frac{\mathrm{RSS}}{n(1 - ps/n)^2}. \qquad (6)$$

Refer GCV (5) as linear GCV (LGCV).

SELECTION OF TUNING PARAMETER

Comparison between LGCV and NLGCV for $\gamma = 1, 1.5, 2, 3$.
Solid – NLGCV (6); Dotted – LGCV (5).

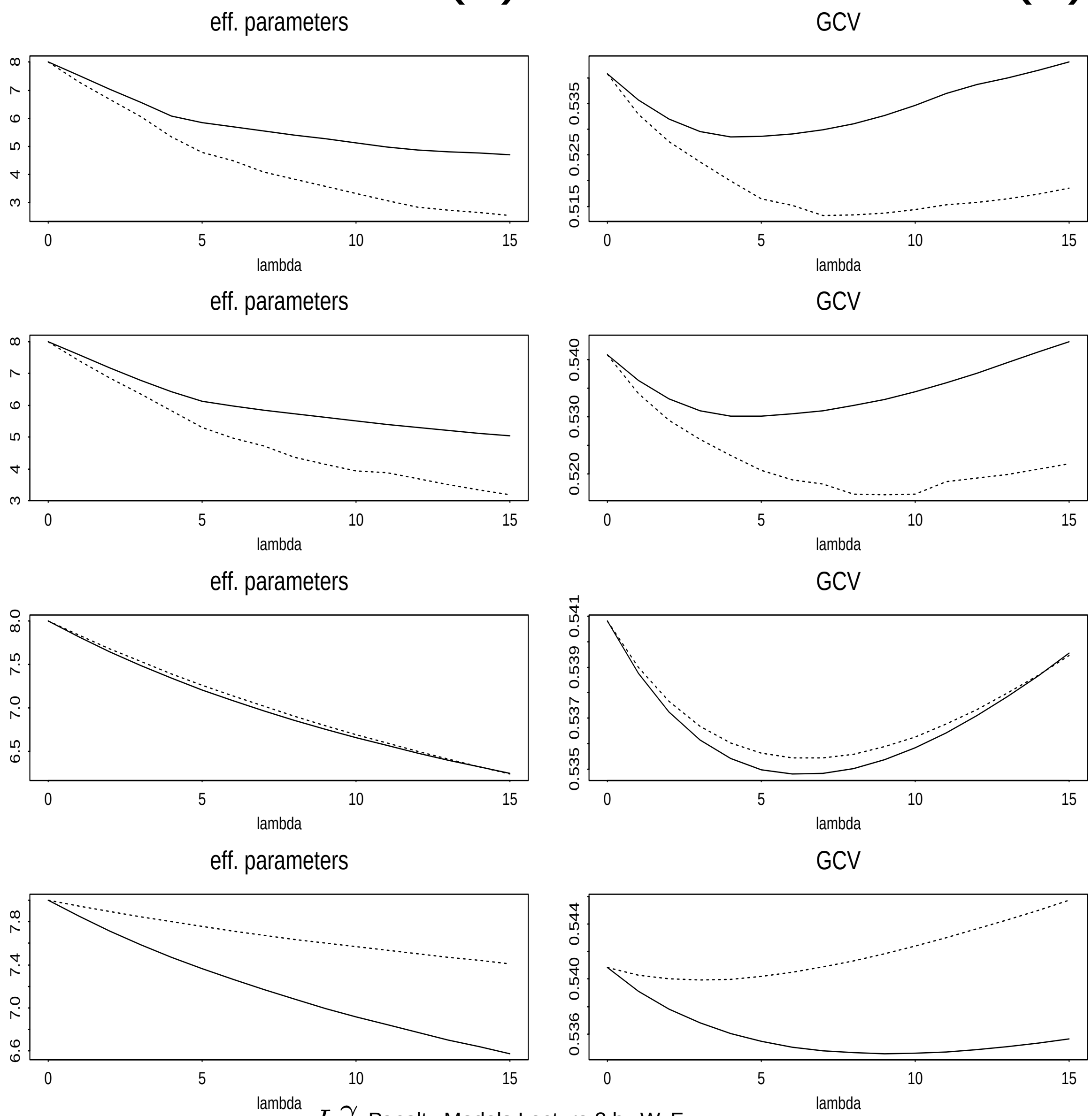


Table 2. MSE* in simulation studies with highly collinear X. $n = 10, p = 5$.

model	$\boldsymbol{\beta}$**	OLS	LGCV	NLGCV
Lasso	$\boldsymbol{\beta_1}$	.1759(.0115)	.1468(.0099)	.0977(.0118)
	$\boldsymbol{\beta_2}$	.0159(.0001)	.0149(.0001)	.0146(.0001)
	$\boldsymbol{\beta_3}$	.0618(.0015)	.0534(.0014)	.0389(.0014)
ridge	$\boldsymbol{\beta_1}$	.1679(.0111)	.0898(.0120)	.0821(.0114)
	$\boldsymbol{\beta_2}$	.0162(.0001)	.0132(.0001)	.0125(.0001)
	$\boldsymbol{\beta_3}$	.0613(.0015)	.0346(.0015)	.0314(.0013)

* $\mathbf{MSE} = (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T(\boldsymbol{X}^T\boldsymbol{X})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$.
** $\boldsymbol{\beta_1} = (0.5, 1, -0.2, 0, 0)$, $\boldsymbol{\beta_2} = (1, 0.2, -0.01, -0.5, 0.02)$ and $\boldsymbol{\beta_3} = (1, 0, 0, 0, 0)$.

Table 3. Comparison of minimum NLGCV by γ

for prostate cancer data

γ	NLGCV*	λ**
1	0.5285	4.33
1.1	0.5300	4.51
2	0.5348	6.36
3	0.5346	9.15

* Value of the minimum NLGCV for fixed γ;
** Value of λ that minimizes NLGCV for fixed γ.

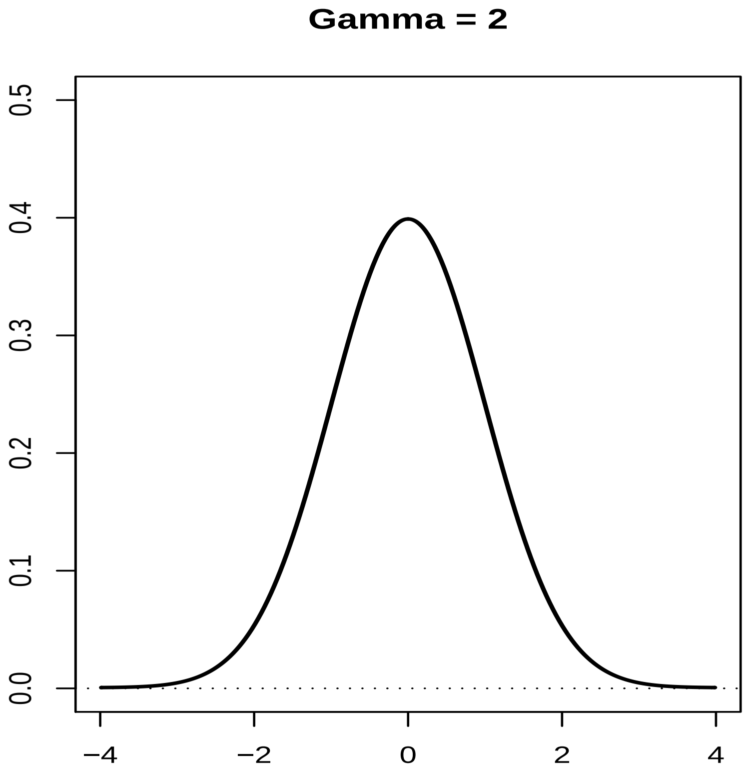
Conclusion: no γ value dominates the NLGCV. No selection for $\gamma \geq 1$.

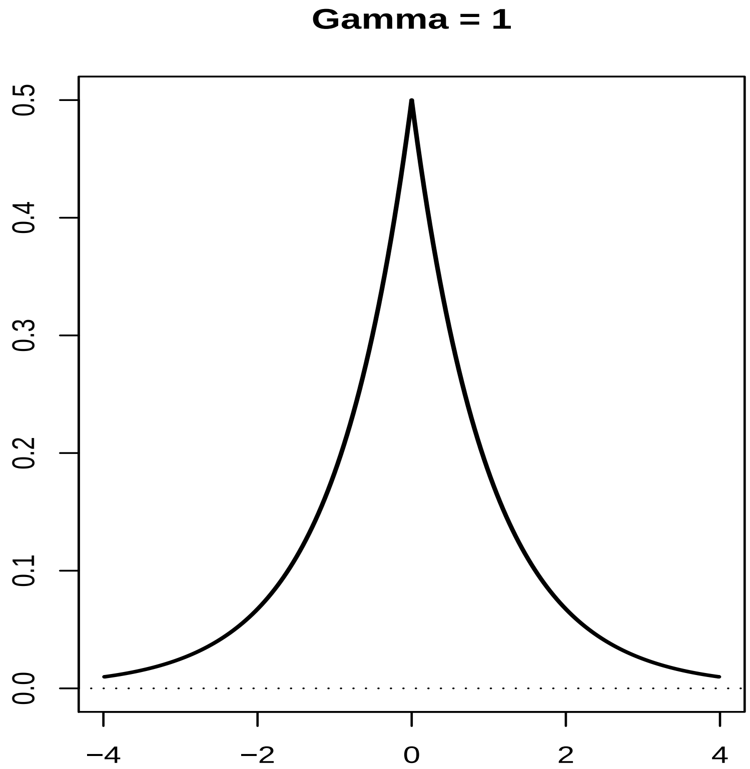
Why no selection for γ.

- Bayesian interpretation of L^γ penalty.
  - ◇ $\gamma = 2$: Gaussian prior.
  - ◇ $\gamma = 1$: Laplacian prior.
  - ◇ $\gamma > 1$: complex prior.
- Selecting λ is to select window size for fixed γ.
- Selecting γ is to select prior distribution.
  - ◇ For given data, β may be generated from one prior, say $\gamma = 1.5$.
  - ◇ Prior distributions overlap largely.
  - ◇ Same β may be generated from different priors.
- Conclusion: no selection between priors unless using Bayesian hierarchical model.

Penalty function as Bayesian prior

$$(\beta|y) \sim C \exp\left\{-\frac{1}{2}\left(\mathrm{RSS} + \sum \left|\frac{\beta_j}{\lambda^{-1/\gamma}}\right|^{\gamma}\right)\right\}$$





Computation of NLGCV

- ◇ Compute $\widehat{\boldsymbol{\beta}}_{\mathbf{ols}}$ with no penalty.
- ◇ Compute $\widehat{\boldsymbol{\beta}}_{\mathbf{brdg}}(\boldsymbol{\lambda}, \gamma)$.
- ◇ Compute the ratio of their L^{γ} norms for $\boldsymbol{s}$.
- ◇ Compute NLGCV (6).

$\boldsymbol{X}$ not of full rank

- ◇ $\widehat{\boldsymbol{\beta}}_{\mathbf{ols}}$ is not unique.
- ◇ Compute the limit $\lim_{\boldsymbol{\lambda}\to \mathbf{0}+} \widehat{\boldsymbol{\beta}}_{\mathbf{rdg}}(\boldsymbol{\lambda}) = \widehat{\boldsymbol{\beta}}_{\mathbf{rdg}}(\mathbf{0}+)$.
  Existence of the limit is guaranteed (Fu 2000).
- ◇ Define standard shrinkage rate $\boldsymbol{s}$ similarly.

Ridge estimator with orthonormal X

For ridge estimator with orthonormal matrix, $X^T X = I$.

$tr(H) = tr\{X^T(X^T X + \lambda I)^{-1} X\} = p/(1+\lambda)$.

$||\widehat{\beta}_{\mathbf{rdg}}||_2 = (1+\lambda)^{-1}\sqrt{y^T y}$, $||\widehat{\beta}^0||_2 = \sqrt{y^T y}$.
Hence

$$ps = p\frac{||\widehat{\beta}_{rdg}||_2}{||\widehat{\beta}^0||_2} = \frac{p}{1+\lambda} = tr(H).$$

Therefore, LGCV = NLGCV.

Large sample behavior of $\widehat{\beta}_{\mathbf{brdg}}$

Finite samples, $\widehat{\beta}_{\mathbf{brdg}}$ is biased and performs well in estimation and prediction.

Large samples, is $\widehat{\beta}_{\mathbf{brdg}}$ consistent?

Need to study the asymptotics under penalized least squares criterion: to minimize

$$\sum_{i=1}^{n}(Y_i - x_i^T\phi)^2 + \lambda_n \sum_{j=1}^{p} |\phi_j|^\gamma.$$

for given λ_n and $\gamma > 0$ fixed.

Regularity conditions

Design $X = (x_1, \ldots, x_n)$. x_i are row vectors.

$$C_n = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T \to C,$$

nonnegative definite constant matrix.

$$\frac{1}{n}\max_{1\le i\le n} x_i^T x_i \to 0.$$

$$\lambda_n/n \to \lambda_0 \ge 0 \qquad (S1)$$

$$\lambda_n/\sqrt{n} \to \lambda_0 \ge 0 \qquad (S2)$$

$(S1)$: λ_n grows fast but not faster than n.

$(S2)$: λ_n grows slowly and not faster than $\sqrt{n}$.

Limiting distributions

$$Z(\phi) = (\phi - \beta)^T C(\phi - \beta) + \lambda_0 \sum_{j=1}^{p} |\phi_j|^\gamma.$$

For $\gamma > 1$:

$$V(u) = -2u^T W + u^T C u + \lambda_0 \sum_{j=1}^{p} u_j \mathrm{sgn}(\beta_j)|\beta_j|^{\gamma-1}.$$

For $\gamma = 1$:

$$V(u) = -2u^T W + u^T C u + \lambda_0 \textstyle\sum_{j=1}^{p} \Big[u_j \mathrm{sgn}(\beta_j) I(\beta_j \neq 0) + |u_j| I(\beta_j = 0)\Big].$$

$W \sim N(0, C\sigma^2)$.

Consistency

Theorem 2. (Knight and Fu 2000)
If C is nonsingular and (S1) is satisfied, then

$$\widehat{\beta}_n \rightarrow_p \mathrm{argmin}(Z).$$

So if $\lambda_n = o(n)$, $\widehat{\beta}_n$ is consistent.

Theorem 3. (Knight and Fu 2000)
If C is nonsingular and (S2) is satisfied, then

$$\sqrt{n}(\widehat{\beta}_n - \beta) \rightarrow_d \mathrm{argmin}(V).$$

Consistency

Theorem 4. (Knight and Fu 2000)
If C is nonsingular and $\lambda_n/n^{\gamma/2} \to \lambda_0 \geq 0$ for $\gamma < 1$, then

$$\sqrt{n}(\widehat{\beta}_n - \beta) \to_d \operatorname{argmin}(V),$$

where

$$V(u) = -2u^T W + u^T C u + \lambda_0 \sum_{j=1}^{p} |u_j|^\gamma I(\beta_j = 0)$$

with $W \sim N(0, C\sigma^2)$.

Asymptotic bias

For $\boldsymbol{\lambda_0 > 0}$, asymptotic bias exists for $\boldsymbol{\gamma \geq 1}$.
For example, ridge $(\boldsymbol{\gamma = 2})$,

$$\sqrt{n}(\widehat{\beta}_n - \beta) \to_d C^{-1}(W - \lambda_0 \beta) \sim N(-\lambda_0 C^{-1}\beta, \sigma^2 C^{-1}).$$

But for $\boldsymbol{\gamma < 1}$, it is very different. Non-zero $\boldsymbol{\beta_j}$ can be estimated without asymptotic bias, meanwhile there is a positive mass to shrink $\boldsymbol{\beta_j = 0}$ to 0.

References

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions. *Numerische Mathematik* 31, 377-403.

Fu, W.J. (1998) Penalized regressions: the Bridge versus the Lasso, *J. Comp. Grap. Statist.* 7: 397-416.

Fu, W.J. (2000) Ridge estimator in singular design with application to age-period-cohort analysis of disease rates, *Comm. Statist. - Theor. Meth.*, 29:263-278.

Fu, W.J. (2005) Nonlinear GCV and quasi–GCV for shrinkage models, *Statist. Plann. Infer.*, in press.

Knight, K. and Fu, W.J. (2000). Asymptotics for Lasso-type estimators, *Ann. Statist.,* 28:1356-1378.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso, *J. Roy. Statist. Soc. B* 58:267-288.