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.
ELECTION OF TUNING PARAMETER

GCV (Craven and Wahba 1979)
For shrinkage model (Fu 1998, Tibshirani 1996)

\[
GCV = \frac{(y - X\beta)^T(y - X\beta)}{n \left(1 - \frac{\text{tr}(H) - n_0}{n}\right)^2}, \tag{5}
\]

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

**Effective number of parameters** \( p(\lambda, \gamma) = \text{tr}(H) - n_0 \).
\( p(0, \gamma) = \text{tr}(X(X^TX)^{-1}X^T) = p \), the number of parameters.
\( p(\infty, \gamma) = 0 \) as \( \lambda \to \infty \).
Select $\lambda \geq 0$ for fixed $\gamma \geq 1$

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

Select $\lambda \geq 0$ and $\gamma \geq 1$

Compute GCV for each point $(\lambda, \gamma)$ on a lattice of $[0, \lambda_0] \times [1, 3]$ with a moderate number $\lambda_0$. Select the values of $(\lambda, \gamma)$ 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 $\text{tr}(H)$, performs well for linear estimators, such as ridge. $\hat{\beta}_{\text{brdg}}$ is nonlinear except for $\gamma = 2$.

For orthonormal $X$ case, lasso is piece–wise linear. GCV performs poorly in selecting $\lambda$ for $\gamma \neq 2$.

By Taylor expansion,

$$X\hat{\beta} = H(y)y =$$

$$H(y_0)y_0 + \{H(y_0) + H'(y_0)y_0\}(y - y_0) + o(y - y_0).$$

Thus $\text{tr}(H)$ is a linearization.
Account for nonlinearity
To account nonlinearity, modify GCV (5) through $p(\lambda, \gamma)$. RSS accounts the nonlinearity through the estimator $\hat{\beta}_{\text{brdg}}$. Instead of separating linear part from nonlinear part, we pool them together and consider the overall shrinkage effect through a standard shrinkage rate $s$.

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

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

\[ \text{NLGCV} = \frac{\text{RSS}}{n(1 - ps/n)^2}. \]  \hspace{1cm} (6)

Refer GCV (5) as linear GCV (LGCV).
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 $X$. $n = 10, p = 5.$

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

* $\text{MSE} = (\hat{\beta} - \beta)^T(X^TX)(\hat{\beta} - \beta)$.

** $\beta_1 = (0.5, 1, -0.2, 0, 0)$, $\beta_2 = (1, 0.2, -0.01, -0.5, 0.02)$ and $\beta_3 = (1, 0, 0, 0, 0)$. 

---

* Statistics

---

Penalty Models Lecture 2 by W. Fu
Table 3. Comparison of minimum NLGCV by $\gamma$
for prostate cancer data

<table>
<thead>
<tr>
<th>$\gamma$</th>
<th>NLGCV*</th>
<th>$\lambda^{**}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.5285</td>
<td>4.33</td>
</tr>
<tr>
<td>1.1</td>
<td>0.5300</td>
<td>4.51</td>
</tr>
<tr>
<td>2</td>
<td>0.5348</td>
<td>6.36</td>
</tr>
<tr>
<td>3</td>
<td>0.5346</td>
<td>9.15</td>
</tr>
</tbody>
</table>

* Value of the minimum NLGCV for fixed $\gamma$;
** Value of $\lambda$ that minimizes NLGCV for fixed $\gamma$.

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

- Bayesian interpretation of $L^\gamma$ penalty.
  - $\gamma = 2$: Gaussian prior.
  - $\gamma = 1$: Laplacian prior.
  - $\gamma > 1$: complex prior.

- Selecting $\lambda$ is to select window size for fixed $\gamma$.
- Selecting $\gamma$ is to select prior distribution.
  - For given data, $\beta$ may be generated from one prior, say $\gamma = 1.5$.
  - Prior distributions overlap largely.
  - Same $\beta$ may be generated from different priors.

- Conclusion: no selection between priors unless using Bayesian hierarchical model.
Penalty function as Bayesian prior

\[
(\beta | y) \sim C \exp \left\{ -\frac{1}{2} \left( \text{RSS} + \sum \left| \frac{\beta_j}{\lambda^{-1/\gamma}} \right|^{\gamma} \right) \right\}
\]
Selection of Tuning Parameter

Computation of NLGCV

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

$X$ not of full rank

- $\widehat{\beta}_{\text{ols}}$ is not unique.
- Compute the limit $\lim_{\lambda \to 0^+} \widehat{\beta}_{\text{rdg}}(\lambda) = \widehat{\beta}_{\text{rdg}}(0^+)$. Existence of the limit is guaranteed (Fu 2000).
- Define standard shrinkage rate $s$ similarly.
Ridge estimator with orthonormal $X$

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

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

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

Hence

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

Therefore, $LGCV = NLGCV$. 
Large sample behavior of $\hat{\beta}_{\text{brdg}}$

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

Large samples, is $\hat{\beta}_{\text{brdg}}$ 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 $\lambda_n$ and $\gamma > 0$ fixed.
Regularity conditions

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

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

nonnegative definite constant matrix.

$$\frac{1}{n} \max_{1 \leq i \leq n} x_i^T x_i \rightarrow 0.$$

$$\lambda_n / n \rightarrow \lambda_0 \geq 0 \quad (S1)$$

$$\lambda_n / \sqrt{n} \rightarrow \lambda_0 \geq 0 \quad (S2)$$

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

$(S2)$: $\lambda_n$ 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 \( \gamma > 1 \):

\[ V(u) = -2u^T W + u^T Cu + \lambda_0 \sum_{j=1}^{p} u_j \text{sgn}(\beta_j) |\beta_j|^{\gamma-1}. \]

For \( \gamma = 1 \):

\[ V(u) = -2u^T W + u^T Cu + \lambda_0 \sum_{j=1}^{p} \left[ u_j \text{sgn}(\beta_j) I(\beta_j \neq 0) + |u_j| I(\beta_j = 0) \right]. \]

\[ W \sim N(0, C\sigma^2). \]
Consistency

Theorem 2. (Knight and Fu 2000)
If $C$ is nonsingular and (S1) is satisfied, then
$$\hat{\beta}_n \rightarrow_p \text{argmin}(Z).$$

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

Theorem 3. (Knight and Fu 2000)
If $C$ is nonsingular and (S2) is satisfied, then
$$\sqrt{n}(\hat{\beta}_n - \beta) \rightarrow_d \text{argmin}(V).$$
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

$$\sqrt{n} (\hat{\beta}_n - \beta) \to_d \text{argmin}(V),$$

where

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

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

$L^{-\gamma}$ Penalty Models Lecture 2 by W. Fu
Asymptotic bias

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

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

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


