Nonlinear GCV and Quasi-GCV for Shrinkage Models

July 18, 2001

Abstract

The generalized cross validation (GCV) method proposed by Craven and Wahba has been a popular technique to select tuning parameters for smoothing and penalty in regression models. Its computational easiness and robustness compared with the cross-validation method makes it competitive for model selection as well. In recent works, it has been used as a standard tool to select tuning parameters in shrinkage models. It is well known that the GCV method performs well for linear estimators, such as the ridge estimator, which are a linear function of the response variable. However, it may not perform well for nonlinear estimators since the GCV emphasizes the linear characteristics by taking the trace of the projection matrix and ignores the nonlinear characteristics of the estimators. This paper aims to explore the GCV for nonlinear estimators and to further extend the results to correlated data in longitudinal studies. With the development of the nonlinear GCV and the quasi-GCV, we expect that the shrinkage models with nonlinear estimators and shrinkage models for correlated data have a similar technique to the GCV for the selection of the tuning parameters.

KEY WORDS: Lasso, Longitudinal studies, Ridge, Standard shrinkage rate, Weighted deviance.
1. INTRODUCTION

The generalized cross validation (GCV) method by Craven and Wahba (1979) has been widely used among others to select tuning parameters for statistical smoothing techniques and regression models with penalty. It not only provides a practical procedure to examine overfitting by applying penalties to a fitted model, but also provides a robust procedure for model selection. In regression analysis, the GCV method has been shown to achieve satisfactory results in selecting tuning parameters for penalty model, for example, the ridge regressions (Golub et al 1979), the kernal smoothing and spline smoothing (Hastie and Tibshirani 1990) and model selection (Shao 1997).

In a number of recent works (Tibshirani 1996 and Fu 1998), the GCV was used to select the tuning parameter for shrinkage models. However, some simulation studies showed that the GCV method did not achieve good result in selecting tuning parameters. For example, Fu (1998) considered the bridge penalty model with penalty function $\lambda \sum_{j=1}^{p} |\beta_j|^\gamma$ with $\gamma \geq 1$ and $\lambda \geq 0$. The GCV was employed to select the tuning parameter $\lambda$ and penalty parameter $\gamma$ to achieve the best performance in estimation and prediction of the shrinkage estimators. As it can be easily seen that the bridge penalty with $\gamma \geq 1$ and $\lambda \geq 0$ potentially outperforms its special cases, the ridge estimator ($\gamma = 2$) and the Lasso estimator ($\gamma = 1$). However, the simulation results in Fu (1998) showed that the ridge estimator performs better in estimation and prediction than the bridge estimator. This raises the question why the GCV method did not select better parameters even though it had the potential to outperform others.

In this paper, we explore the GCV function for nonlinear operators, which covers a wide range of shrinkage models in regressions. We modify the current GCV function to accommodate nonlinearity of certain estimators. We then further extend the modified GCV from independent observations to correlated observations in longitudinal studies. Specifically, we develop a quasi-GCV for the GEE model to select the tuning parameters of the penalized GEE developed in Fu (2001). The
paper is organized as follows. Section 2 explores the GCV for linear and nonlinear estimators. Section 3 compares the GCV with the nonlinear GCV. Section 4 extends the nonlinear GCV to quasi-GCV for correlated observations for the penalized GEE. Section 5 discusses the selection of $\lambda$ versus $\gamma$. Section 6 demonstrates the performance of the non-linear GCV and quasi-GCV through data analyses. Section 7 discusses the strength and weakness of the extension of the GCV to the nonlinear GCV and the quasi-GCV.

2. NONLINEAR GCV

For a linear regression problem of $n$ observations and $p$ covariates

$$Y = X\beta + \varepsilon$$

the ridge estimator is defined as

$$\hat{\beta}_{ridge} = (X^T X + \lambda I)^{-1} X^T y,$$  \hspace{1cm} (1)

while the Lasso estimator or the bridge estimator satisfies

$$\hat{\beta} = (X^T X + \lambda W^-)^{-1} X^T y.$$  \hspace{1cm} (2)

where $W^-$ is the generalized inverse of diagonal matrix

$$W = \text{diag}(2|\hat{\beta}_j|)$$
for the Lasso estimator ($\gamma = 1$) or

$$W = \text{diag} \left( \frac{2}{\gamma} |\hat{\beta}_j|^{2-\gamma} \right)$$

for the bridge estimator with $\gamma > 1$. Let $H$ be the symmetric projection matrix for the shrinkage estimator such that $X\hat{\beta} = Hy$. The GCV (Craven and Wahba 1979) is defined as

$$\text{GCV} = \frac{\text{RSS}}{n(1 - \text{tr}(H)/n)^2} - n_0$$

(3)

where RSS is the residual sum of squares and $n_0$ is the number of parameters $\hat{\beta}_j = 0$. $\text{tr}(H)$, the trace of matrix $H$, is called the effective number of parameters and is denoted by $p(\lambda, \gamma)$ with

$$p(\lambda, \gamma) = \text{tr}(H) \leq p.$$

The trace of the projection matrix $H$ in the definition of the GCV clearly indicates that the linearization of the projection operator is emphasized and the nonlinear part is ignored in the effective number of parameters of the shrinkage estimator. As pointed out by Fu (1998), many shrinkage estimators are non-linear, i.e. $\hat{\beta}$ is not a linear function of $y$. For example, in the bridge shrinkage family, the ridge estimator is the only linear estimator while the Lasso is piece-wise linear for orthonormal regression matrix. Thus, the GCV may not select the best estimator among the bridge family if the nonlinear characteristics of the shrinkage estimator is ignored. Consequently the bridge estimator selected by the linearized GCV (3) may not out-perform the ridge estimator.

To accommodate nonlinearity of shrinkage estimators, we propose the following modification to the GCV. Since the RSS is calculated with the shrinkage estimator itself, it reflects both linear and nonlinear characteristics. Only the effective number of parameters $p(\lambda, \gamma)$ needs to be modified.
Instead of separating the linear from the nonlinear characteristics of the shrinkage estimator, we pool them together by emphasizing the shrinkage effect. We define a standard shrinkage rate by

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

where $||\cdot||_\gamma$ is the $\gamma$-norm of the shrinkage estimator $\hat{\beta}(\lambda, \gamma)$ and the non-shrinkage estimator $\hat{\beta}(0)$ with $\gamma \geq 1$. Apparently $0 \leq s \leq 1$. Tibshirani (1996) defined this standard shrinkage rate for the Lasso and used it as a tuning parameter. We then define the effective number of parameters for the shrinkage model by

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

where $p$ is the number of covariates in the model. Therefore the GCV for a generalized linear model is calculated as

$$GCV = \frac{\text{Dev}}{n(1 - sp/n)^2}$$

(4)

where Dev is the model deviance replacing RSS for non-Gaussian distribution in the exponential family.

3. COMPARISON BETWEEN LINEARIZED GCV AND NONLINEAR GCV

It is well known that the GCV method works well for linear estimators, such as the ridge estimator (1). With the above modification, we expect the nonlinear GCV still works well for linear estimators. We compare the linearized GCV (3) with the nonlinear GCV (4) using the prostate cancer data analyzed in Tibshirani (1996), Fu (1998) and Osborne et. al (2000). Figure 1 shows that the effective number of parameters and the GCV curves calculated with the linearized GCV (dotted curves) and the nonlinear GCV (solid curves) for different $\gamma$ values, $\gamma = 1, 1.1, 2,$ and
3. It can be seen that the two GCV curves behave differently for nonlinear estimators $\gamma = 1$ (the Lasso), 1.1 and 3, while they behave very similarly for linear estimator $\gamma = 2$ (the ridge estimator).

In fact, the two different versions of GCV are equal to each other for ridge estimator with orthonormal regression matrix as shown in the following. With orthonormal matrix $X$, $X^T X = I$.

Thus

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

Similarly

$$||\hat{\beta}_{ridge}||_2 = \frac{1}{1 + \lambda} \sqrt{y^T y}$$

$$||\hat{\beta}(0)||_2 = \sqrt{y^T y}$$

Hence

$$ps = p \frac{||\hat{\beta}_{ridge}||_2^2}{||\hat{\beta}(0)||_2^2} = \frac{p}{1 + \lambda} = tr(H).$$

The formulation of the effective number of parameters through the standard shrinkage rate simplifies the calculation. Particularly it simplifies the complex calculation in (3) for the Lasso estimator when several parameters are set to zero (Fu, 1998). It does not require the calculation of the trace of the weighted regression matrix in the IRLS procedure for the generalized linear models.

The standard shrinkage rate method also smooths the GCV curve for some nonlinear shrinkage estimators as shown in the top and upper-mid panels in Figure 1, where the linearized GCV presents rather “rugged” shape. This smoothing effect makes the selection of the tuning parameters stable for nonlinear estimators.

The definition of the nonlinear GCV involves the shrinkage estimator and the non-shrinkage estimator, and thus implicitly involves the response variable $y$, while the linearized GCV seems to have only the projection matrix involved, but not $y$. In fact, the latter is true only for linear
Figure 1: Comparison between the linearized GCV and the nonlinear GCV (right panels) and their corresponding effective number of parameters (left panels) for different values of $\gamma$. Top panels: $\gamma = 1$ (the Lasso); upper-mid panels: $\gamma = 1.1$; lower-mid panels: $\gamma = 2$ (the ridge); bottom panels: $\gamma = 3$. Solid curves: nonlinear GCV; dotted curves: linearized GCV.
shrinkage estimators, such as the ridge estimator, but not true for nonlinear shrinkage estimators, such as the Lasso, in which the projection matrix $H$ also depends on the shrinkage estimator $\hat{\beta}(\lambda, \gamma)$. Thus, the response $y$ is also involved in the GCV calculation. Another computational difference between the two versions of the GCV is that the nonlinear GCV also requires the calculation of the non-shrinkage estimator while the linearized GCV does not. In fact, this is not a disadvantage for the nonlinear GCV at all and does not make the computation of the nonlinear GCV more complicated. Instead, it simplifies the computation because the trace of the projection matrix can be very complicated for some nonlinear shrinkage estimators, while the non-shrinkage estimator is fairly easy to compute in most cases except for singular designs.

Since a singular design matrix leads to an infinite number of estimators, the standard shrinkage rate needs to be modified. It has been proven by Fu (2000) and Knight and Fu (2000) that the bridge shrinkage estimator $\hat{\beta}(\lambda, \gamma)$ of singular designs converges to a unique limiting estimator $\hat{\beta}(0+)$ as $\lambda \rightarrow 0+$ for fixed $\gamma \geq 1$. The standard shrinkage rate for singular designs can thus be modified as

$$s = \frac{||\hat{\beta}(\lambda, \gamma)||_2}{||\hat{\beta}(0+)||_\gamma}$$

This leads to a well defined nonlinear GCV for nonlinear estimators in singular designs.

4. QUASI-GCV

In longitudinal studies, the observations within subjects are correlated. Hence the model deviance in the GCV (4) is not well defined. In order to employ a convenient method similarly to select tuning parameters for shrinkage models in longitudinal studies, such as the penalized GEE models (Fu 2001), we need to modify the model deviance and the number of observations.

To modify the model deviance for correlated observations, we incorporate the within subject
correlation into the deviance. Recall that the model deviance of a generalized linear model is the
sum of squares of the deviance residuals. Even if the model deviance does not exist due to the lack
of joint likelihood in the GEE models by Liang and Zeger (1986) and Zeger and Liang (1986), the
deviance residual can still be calculated at each observation in the same way with the marginal
distribution specified in the GEE.

\[ r = \text{sign}(y - \hat{\mu}) \sqrt{-2 \log L(y, \hat{\mu})} \]

where \( L \) is the likelihood function of the specified marginal distribution, and \( \hat{\mu} \) is the fitted value.

Then a weighted deviance is defined as

\[ W\text{Dev}(\lambda, \gamma) = \sum_{k=1}^{K} r_k^T [R_k]^{-1} r_k \]

(5)

where \( r_k \) is the deviance residual vector of length \( n_k \) for subject \( k = 1, \ldots, K \) and \( R_k \) is the working
correlation matrix with dimension \( n_k \times n_k \) in the GEE model.

We also need to modify the number of observations, or the degrees of freedom of correlated
observations. If the within subject observations are independent, the degrees of freedom is equal
to the number of observations. However, the within subject observations are usually correlated in
longitudinal studies, and positive correlation leads to less degrees of freedom. Thus the degrees of
freedom of the correlated observations needs to be modified to reflect the correlation. We define
the effective number of degrees of freedom of correlated observations as

\[ N = \sum_{k=1}^{K} \frac{n_k^2}{|R_k|} \]

(6)

where \( K \) is the number of subjects, \( n_k \) the number of observations of subject \( k \). \(|R_k| = \sum \rho_{ij} \), the
sum of all elements of the working correlation matrix $R_k = (\rho_{ij})$.

Finally, the quasi-GCV is defined as

$$QGCV = \frac{WDev(\lambda, \gamma)}{K(1 - p(\lambda, \gamma)/N)^2}$$

(7)

where $p(\lambda, \gamma)$ is the effective number of parameters in the penalty model defined as

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

$p$ is the number of covariates in the model, $||\hat{\beta}(\lambda, \gamma)||_{\gamma}$ and $||\hat{\beta}(0)||_{\gamma}$ are the $\gamma$-norms of the penalized GEE estimator and the non-shrinkage GEE estimator with $\gamma \geq 1$, respectively.

Notice that for independent observations, the weighted deviance (5) reduces to deviance, the effective number of degrees of freedom (6) reduces to the number of observations as the correlation matrix $R_k$ becomes identity matrix. Consequently, the quasi-GCV (7) reduces to the GCV (4).

Also notice that since the correlation is incorporated into the deviance residuals, the deviance residuals should be used to estimate the working correlation structure in the GEE model instead of the Pearson residuals when the quasi-GCV is used to select the tuning parameter.

**Motivation of quasi-GCV**

The quasi-GCV is motivated by the following derivations of the weighted deviance and the effective number of degrees of freedom.

The weighted deviance is derived from the correlated Gaussian responses as follows. Assume $Y$ are correlated responses from model $Y = X\beta + \varepsilon$ with $\varepsilon \sim N(0, \Sigma)$, where $\Sigma$ is a non-diagonal variance-covariance matrix.

In order to apply the GCV method for independent responses, we take a transformation $Z =$
PY, where \( P = \Lambda^{-1/2}Q \) satisfying \( Q\Sigma Q^T = \Lambda \), a diagonal matrix. Then \( Z \) follows a normal distribution of \( N(PX\beta, I) \). In calculating the GCV for independent observations \( Z \), one has

\[
RSS = (Z - PX\beta)^T (Z - PX\beta) = (Y - X\beta)^T P^T P (Y - X\beta) = (Y - X\beta)^T \Sigma^{-1} (Y - X\beta)
\]

Thus, the GCV can be applied to the correlated observations \( Y \) by incorporating the correlation structure in the residuals. Similarly, one incorporates the correlation structure into the deviance residuals to achieve the same effect.

The effective number of degrees of freedom is derived from the correlated Gaussian observations. Assume \( Y \) follows a distribution \( N(0, \sigma^2 R) \), where matrix \( R = (\rho_{ij}) \) has diagonal elements \( \rho_{ii} = 1 \). Consider the variance of the sample mean \( \bar{Y} \).

\[
\text{var}(\bar{Y}) = \frac{1}{n^2} \text{var}(\sum Y_i) = \frac{1}{n^2} \text{cov}(\sum Y_i, \sum Y_i) \\
= \frac{1}{n^2} \sum_i \sum_j \text{cov}(Y_i, Y_j) \\
= \frac{\sigma^2}{n^2} \sum_i \sum_j \rho_{ij} = \frac{\sigma^2}{n^2} |R| \\
= \frac{\sigma^2}{n^2/|R|}.
\]

Notice that for the special case where \( Y_i \)'s are independent, \( R \) is thus an identity matrix, \( \text{var}(\bar{Y}) = \sigma^2/n \). The denominator \( n \) is the number of degrees of freedom of the \( n \) independent observations \( Y_1, \ldots, Y_n \). By analogy, we define \( n^2/|R| \), the denominator of (8), to be the effective number of degrees of freedom of the correlated observations \( Y_1, \ldots, Y_n \). For non-negative correlation coefficient \( \rho_{ij} \geq 0 \), this effective number of degrees of freedom is between \( 1 \) and \( n \), as the former is for \( n \) duplicates of \( Y_1 \) and the latter for \( n \) independent observations \( Y_1, \ldots, Y_n \).

Certain problem may arise with negative correlations. However, it is rare in practice to have a
series of observations within subjects with negative correlation. Especially in longitudinal studies, one expects positively correlated responses within subjects. Therefore, the effective number of degrees of freedom works well for longitudinal studies in general.

5. SELECTION OF TUNING PARAMETER $\lambda$ VERSUS PENALTY PARAMETER $\gamma$

Fu (1998) selected both the tuning parameter $\lambda$ and the penalty parameter $\gamma$ through the linearized GCV method, but claimed that the GCV favored the Lasso by achieving much smaller value of

$$\min_{\lambda} GCV$$

for $\gamma = 1$ than other values of $\gamma$. This kind of phenomenon is corrected by the nonlinear GCV as shown in Figure 1. It shows that the linearized GCV (dotted curves) has a smaller minimum than the nonlinear GCV (solid curve) for small values of $\gamma = 1$ and 1.1 (top and upper-middle panels), but a larger minimum for large values of $\gamma = 2$ and 3 (bottom and lower-middle panels). Table 1 shows that the nonlinear GCV does not favor the Lasso any more. The Lasso achieves the minimum GCV at $\lambda = 4.33$ with GCV = 0.5285, the bridge of $\gamma = 1.1$ at $\lambda = 4.51$ with GCV = 0.5300, the ridge at $\lambda = 6.36$ with GCV = 0.5348 and the bridge of $\gamma = 3$ at $\lambda = 9.15$ with GCV = 0.5346. With the range of $\gamma$ from 1 to 3, the value of minimum nonlinear GCV achieved by the selection of $\lambda$ varies only slightly. Therefore, no value of $\gamma$ is dominating the nonlinear GCV, i.e.,

$$\min_{\lambda} GCV(\lambda, \gamma)$$

as a function of $\gamma$ is flat or does not vary largely in a moderate range of $\gamma$. It thus raises a question: whether the penalty parameter $\gamma$ can be simultaneously selected with the tuning parameter $\lambda$
Table 1: Comparison of minimum GCV by $\gamma$ for the prostate cancer data

<table>
<thead>
<tr>
<th>$\gamma$</th>
<th>GCV*</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 minimum nonlinear GCV for fixed $\gamma$;  
** Value of $\lambda$ to achieve the minimum GCV for fixed $\gamma$.

To answer this question, we recall that the bridge penalty function can be regarded as a prior distribution of the model parameters from a Bayesian point of view, as discussed in Fu (1998). Different value of $\gamma$ corresponds to different prior distribution, $\gamma = 2$ yields the Gaussian prior while $\gamma = 1$ yields the Laplacian prior. With the continuum of the penalty parameter $\gamma$, the prior distributions of close values of $\gamma$ overlap largely. Therefore, one set of parameters can be generated from several different prior distributions (in fact, an infinite number of priors with all values of $\gamma$ between any two values close to each other). From this point of view, it is difficult to select the penalty parameter $\gamma$ using the nonlinear GCV or the quasi-GCV if no value of $\gamma$ is dominating. Therefore, the GCV or the quasi-GCV method is not recommended to select the penalty parameter $\gamma$ for the bridge penalty models if no $\gamma$ value is dominating. It will be demonstrated in next section through data analyses of the prostate cancer data with the nonlinear GCV and the pollution data with the quasi-GCV.

6. DATA ANALYSES

6.1 An example of linear model with nonlinear GCV
We demonstrate the nonlinear GCV by analyzing the prostate cancer data. The description of the data set can be found in several papers (Stamey et al 1989, Tibshirani 1996 and Fu 1998).

We fit a series of linear models with the bridge penalty of different $\gamma$ values, the no-penalty model (OLS), the Lasso model ($\gamma = 1$), the bridge with $\gamma = 1.5$, the ridge ($\gamma = 2$) and the bridge with $\gamma = 3$. A $\lambda$ value is selected for each model by minimizing the nonlinear GCV.

The nonlinear GCV selects $\lambda = 4.33$ for $\gamma = 1$ with GCV = 0.5285, $\lambda = 5.32$ for $\gamma = 1.5$ with GCV = 0.5334, $\lambda = 6.36$ for $\gamma = 2$ with GCV = 0.5348 and $\lambda = 9.15$ for $\gamma = 3$ with GCV = 0.5346 at the minimum. Apparently, no $\gamma$ value is dominating since the minimum GCV as a function of $\gamma$ does not vary largely. Therefore, the GCV does not help select the penalty parameter $\gamma$. Figure 2 demonstrates the nonlinear GCV curves for fixed $\gamma = 1, 1.5, 2$ and 3. Figure 3 shows the GCV for a range of parameters $\lambda$ and $\gamma$ in $[0, 10] \times [1, 3]$. It is demonstrated that the nonlinear GCV value does not vary largely in this entire range of the parameters $\lambda$ and $\gamma$.

Notice that the selected $\lambda = 4.33$ for $\gamma = 1$ by the nonlinear GCV is smaller than the value $\lambda = 7.2$ by the linearized GCV reported in Fu (1998). Table 2 compares the model parameter estimates between the OLS, the Lasso model by the nonlinear GCV and the Lasso model by the linearized GCV. The standard errors of the parameter estimates were obtained by 10000 bootstrap samples (Efron and Tibshirani 1993) with fixed $\lambda = 4.33$ and $\gamma = 1$ for the nonlinear GCV method. The standard errors for the Lasso estimator by the linearized GCV were obtained from Fu (1998). Clearly, the Lasso estimator by the nonlinear GCV has larger absolute values with its smaller $\lambda$ than the one by the linearized GCV. The nonlinear GCV shrinks the parameter of $lcp$ to 0, but not the gleason score as by the linearized GCV. The standard errors show that there is no major difference in terms of statistical significance between the two Lasso estimators, which are consistent to the subset selection by the leaps and bounds (L-P) method (Furnival and Wilson 1974) as discussed in Fu (1998).
Figure 2: Selection of tuning parameter $\lambda$ for fixed $\gamma$ by the nonlinear GCV for the prostate cancer data. Upper left panel: $\gamma = 1$; Upper right panel: $\gamma = 1.5$; Lower left panel: $\gamma = 2$; Lower right panel: $\gamma = 3$. 
Select lambda and gamma via GCV

Figure 3: Nonlinear GCV with parameters $\lambda$ and $\gamma$ for the prostate cancer data.
Table 2: Comparison of the estimates for the prostate cancer data *

<table>
<thead>
<tr>
<th></th>
<th>OLS</th>
<th>Lasso (linear GCV)</th>
<th>Lasso (nonlin. GCV)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>λ = 0</td>
<td>λ = 7.2</td>
<td>λ = 4.33</td>
</tr>
<tr>
<td>intercept</td>
<td>2.478(.072)</td>
<td>2.478(.072)</td>
<td>2.478(.116)</td>
</tr>
<tr>
<td>leavol</td>
<td>0.688(.103)</td>
<td>0.618(.090)</td>
<td>0.625(.092)</td>
</tr>
<tr>
<td>lweight</td>
<td>0.225(.084)</td>
<td>0.190(.076)</td>
<td>0.203(.082)</td>
</tr>
<tr>
<td>age</td>
<td>-0.145(.082)</td>
<td>-0.048(.046)</td>
<td>-0.081(.069)</td>
</tr>
<tr>
<td>lbph</td>
<td>0.155(.084)</td>
<td>0.103(.066)</td>
<td>0.122(.080)</td>
</tr>
<tr>
<td>svi</td>
<td>0.316(.100)</td>
<td>0.245(.087)</td>
<td>0.253(.096)</td>
</tr>
<tr>
<td>lcp</td>
<td>-0.147(.126)</td>
<td>0(.068)</td>
<td>0(.072)</td>
</tr>
<tr>
<td>gleason</td>
<td>0.032(.112)</td>
<td>0(.047)</td>
<td>0.009(.060)</td>
</tr>
<tr>
<td>pgg45</td>
<td>0.127(.123)</td>
<td>0.063(.056)</td>
<td>0.071(.083)</td>
</tr>
</tbody>
</table>

* The standard errors of the Lasso estimators in the parentheses were obtained through 10000 bootstrap samples with fixed λ value selected by the GCV.

6.2 An example of GEE penalty model with quasi-GCV

We demonstrate the quasi-GCV method by analyzing the pollution data in a longitudinal study of the impact of air pollution on asthma patients. The data were described in details in Fu (2001). It has 39 asthma patients, each of them were observed for the presence (1) or absence (0) of asthmatic symptoms together with levels of air pollutants: nitrogen oxide (NO), nitrogen dioxide (NO2), a combined index of nitrogen oxide and nitrogen dioxide (NOX), total reduced sulphur (TRS), ozone (OZ), carbon-monoxide (CO), sulphur dioxide (SO2), coefficient of haze (COH), the mean temperature and the mean humidity of the day for 21 consecutive days. The square root scale is employed for NO and NOX in the model. Some of the covariates are highly correlated, for example, the correlation is 0.91 between NO and NOX and 0.79 between NOX and NO2. With the collinearity among the covariates, the GEE model with a common working correlation to the immediate before and after observation specified reveals a significant positive dependence on the combined index NOX, the
Table 3: Comparison of minimum quasi-GCV by $\gamma$ for the pollution data

<table>
<thead>
<tr>
<th>$\gamma$</th>
<th>QGCV*</th>
<th>$\lambda^{**}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10.7469</td>
<td>3.0</td>
</tr>
<tr>
<td>1.5</td>
<td>10.7436</td>
<td>2.7</td>
</tr>
<tr>
<td>2</td>
<td>10.7397</td>
<td>3.2</td>
</tr>
<tr>
<td>3</td>
<td>10.7336</td>
<td>5.0</td>
</tr>
</tbody>
</table>

* Value of minimum quasi-GCV for fixed $\gamma$;  
** Value of $\lambda$ to achieve the minimum quasi-GCV for fixed $\gamma$.

carbon-monoxide (CO) and the sulphur dioxide (SO2), but significant negative dependence on the nitrogen oxide (NO) and nitrogen dioxide (NO2) and marginally significant negative dependence on ozone (OZ) as shown in Table 4. All other factors are not statistically significant. The standard errors of the no-penalty GEE model were obtained through the “sandwich” estimator by Liang and Zeger (1986). Since nitrogen oxide, nitrogen dioxide and ozone are well known pollutants affecting human expiratory system, the negative significance of NO, NO2 and OZ is misleading. Therefore, the no-penalty GEE model does not yield good estimation of the air pollutants. To achieve better estimation, we fit the penalized GEE models (Fu 2001) to the data with different bridge penalty parameter $\gamma = 1, 1.5, 2$ and 3. The tuning parameter $\lambda$ is selected for each fixed $\gamma$ through the quasi-GCV method.

Figures 4 shows the minimum quasi-GCV is achieved at different $\lambda$ value for each fixed $\gamma = 1, 1.5, 2$ and 3. $\lambda = 3.0$ for $\gamma = 1$ with QGCV = 10.7469, $\lambda = 2.7$ for $\gamma = 1.5$ with QGCV = 10.7436, $\lambda = 3.2$ for $\gamma = 2$ at QGCV = 10.7397 and $\lambda = 5.0$ for $\gamma = 3$ at QGCV = 10.7336.

As shown in Table 3, no $\gamma$ value is dominating the quasi-GCV. Therefore, no penalty parameter $\gamma$ can be selected through the quasi-GCV method in this case.

Table 4 shows that the Lasso shrinks the parameter for NOX to zero and yields significant
Table 4: Comparison of the estimates for the pollution data *

<table>
<thead>
<tr>
<th></th>
<th>No penalty</th>
<th>Lasso penalty $\lambda = 0$</th>
<th>Lasso penalty $\gamma = 1, \lambda = 3.0$</th>
<th>Ridge penalty $\gamma = 2, \lambda = 3.2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-2.659(.464)</td>
<td>-2.634(.644)</td>
<td>-2.640(.658)</td>
<td></td>
</tr>
<tr>
<td>meanhumd</td>
<td>0.044(.062)</td>
<td>0.021(.077)</td>
<td>0.052(.081)</td>
<td></td>
</tr>
<tr>
<td>meantemp</td>
<td>-0.438(.374)</td>
<td>-0.430(.567)</td>
<td>-0.428(.499)</td>
<td></td>
</tr>
<tr>
<td>NO</td>
<td>-1.015(.405)</td>
<td>-0.265(.308)</td>
<td>-0.335(.213)</td>
<td></td>
</tr>
<tr>
<td>NO2</td>
<td>-0.787(.175)</td>
<td>-0.340(.393)</td>
<td>-0.373(.215)</td>
<td></td>
</tr>
<tr>
<td>NOX</td>
<td>0.975(.464)</td>
<td>0(.094)</td>
<td>0.159(.200)</td>
<td></td>
</tr>
<tr>
<td>TRS</td>
<td>-0.082(.116)</td>
<td>-0.075(.166)</td>
<td>-0.087(.159)</td>
<td></td>
</tr>
<tr>
<td>OZ</td>
<td>-0.269(.174)</td>
<td>-0.219(.236)</td>
<td>-0.201(.155)</td>
<td></td>
</tr>
<tr>
<td>CO</td>
<td>0.591(.155)</td>
<td>0.468(.190)</td>
<td>0.489(.160)</td>
<td></td>
</tr>
<tr>
<td>COH</td>
<td>-0.106(.426)</td>
<td>-0.056(.513)</td>
<td>-0.139(.419)</td>
<td></td>
</tr>
<tr>
<td>SO2</td>
<td>0.413(.106)</td>
<td>0.374(.127)</td>
<td>0.381(.113)</td>
<td></td>
</tr>
</tbody>
</table>

* The standard errors of the Lasso and ridge estimators in the parentheses were obtained through the delete-one-subject jackknife method with fixed $\lambda$ value selected by the quasi-GCV. The standard errors of the no penalty GEE model were obtained through the “sandwich” estimator by Liang and Zeger (1986).
Figure 4: Selection of tuning parameter $\lambda$ for fixed $\gamma$ by the quasi-GCV for the pollution data. Upper left panel: $\gamma = 1$; Upper right panel: $\gamma = 1.5$; Lower left panel: $\gamma = 2$; Lower right panel: $\gamma = 3$. 
positive dependence on the CO and SO2 while it leaves all other parameters nonsignificant. The ridge model yields significant positive dependence on CO and SO2 and marginally significant negative dependence on NO and NO2. The standard errors of the penalized GEE models were obtained via the delete-one-subject jackknife method (Shao and Tu 1995). It clearly demonstrates the effect of the penalized GEE model with the selection of tuning parameter through the quasi-GCV.

7. CONCLUSIONS AND DISCUSSIONS

The GCV method has been a popular technique to select tuning parameters for smoothing and shrinkage models. It performs well for linear estimators, such as the ridge estimator. However, since the GCV takes the linearization of the estimator into consideration, it may not select the best tuning parameter for nonlinear shrinkage estimators. As shown in Fu (1998), it always favors the Lasso by selecting $\gamma = 1$ even when the Lasso is out-performed by the ridge estimator in terms of the MSE and PSE in certain cases.

To modify the GCV for nonlinear estimators, we proposed the nonlinear GCV by taking the shrinkage effect into consideration through the standard shrinkage rate of the estimators. It combines both linear and nonlinear characteristics of the shrinkage estimators and potentially selects better tuning parameter for nonlinear estimators than the linearized GCV.

We have shown that the standard shrinkage rate smooths the rugged shape of the linearized GCV for some nonlinear estimators, including the Lasso estimator, and makes the selection of the tuning parameters stable. The standard shrinkage rate also simplifies the calculation of the GCV function. It only requires the calculation of the shrinkage estimator and the non-shrinkage estimator. It does not require the trace of the projection matrix $H$, which could be complex to calculate for the generalized linear models and the penalized GEE models.

The correlation of the within subject observations in the GEE models makes the GCV ill-defined
due to the lack of joint likelihood. To accommodate the correlation, we proposed the weighted deviance and the effective number of degrees of freedom of correlated observations. Both were motivated from correlated Gaussian observations. Thus, the quasi-GCV is defined to extend the GCV to correlated data by incorporating the correlation structure. With independent observations, the weighted deviance reduces to the deviance, the effective number of degrees of freedom reduces to the number of observations. Therefore, the quasi-GCV becomes the nonlinear GCV.

The nonlinear GCV also has some drawbacks. It involves the response variable \( y \) in the calculation. Even though the effect of \( y \) can be cancelled for the ridge estimators in the orthonormal cases, it is not true in general. Thus the nonlinear GCV depends on the response \( y \), and is not intrinsic to the shrinkage operator itself as the linearized GCV for the ridge estimators. The nonlinear GCV also requires calculation of the non-shrinkage estimator, which does not exist in singular designs. Thus the standard shrinkage rate has to be modified using the limiting shrinkage estimator \( \hat{\beta}(0+) \) for bridge regressions in singular designs. This makes the nonlinear GCV well defined for singular designs.

We have demonstrated that the selection of the penalty parameter \( \gamma \) of the bridge penalty is not recommended by the nonlinear GCV or the quasi-GCV simultaneously with the selection of the tuning parameter \( \lambda \) if no \( \gamma \) value is dominating the nonlinear GCV or the quasi-GCV. From a Bayesian point of view, the bridge penalty represents prior distribution of the model parameters. Different value of \( \gamma \) yields different prior, \( \gamma = 2 \) yields Gaussian prior, while \( \gamma = 1 \) yields Laplacing prior. With the continuum of \( \gamma \), these prior distributions overlap largely. Thus one set of true model parameters can be generated from many different priors. Thus, it is difficult and problematic to select the penalty parameter \( \gamma \) using the nonlinear GCV or the quasi-GCV method when no \( \gamma \) value is dominating.

It is hoped that the nonlinear GCV and the quasi-GCV perform well for the selection of tuning
parameters in shrinkage models.

ACKNOWLEDGEMENTS

I would like to thank Robert Tibshirani for his supervision and generous support while part of this work was conducted during my Ph.D. program.

REFERENCES


