ACM/ESE 118

Choice of regularization parameter

A regularized parameter estimate $\hat{\boldsymbol{\beta}}$ is affected by error due to the regularization (bias) and by perturbation error (variance) due to the error in the data. The regularization error usually increases with increasing regularization (smaller truncation parameter $\hat{r}$ in truncated singular value decomposition or larger regularization parameter h in ridge regression); the perturbation error usually decreases with increasing regularization. Methods for the choice of a regularization parameter attempt to find a good trade-off between the two kinds of error.

Discrepancy principle (Hansen 1998, chapter 7.2). If the error variance s^2 in the regression model

$$\mathbf{y} = \mathsf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} = \mathcal{N}(0, s^2\,\mathsf{I}), \tag{1}$$

is known, one may choose the regularization parameter such that the residual variance of the estimated regression model is consistent with the known error variance s^2. That is, one may choose the regularization parameter such that the discrepancy principle

$$\frac{\mathrm{RSS}}{n} = \frac{\|\mathsf{X}\hat{\boldsymbol{\beta}} - \mathbf{y}\|^2}{n} \approx s^2 \tag{2}$$

is satisfied. Here, $\hat{\boldsymbol{\beta}}$ is the regularized parameter estimate (regularized by TSVD, ridge regression, or some other regularization method). Note that the RSS is usually divided by the sample size n, not by an effective number of degrees of freedom (see below).

For the gravity surveying example of homework 5, the error variance is known to be $s^2 = 0.01$. Using ridge regression, the regularization parameter h that minimizes the difference between the left-hand side and the right-hand side of the discrepancy principle (2) is found to be $h = 0.47$. The regularized estimate of the density variations corresponding to this regularization parameter is shown in Figure 1.

In most practical situations, the error variance s^2 is not known. If a rough estimate of the error variance s^2 is available, the discrepancy principle (2) can be used to limit the range of regularization parameters over which to search for the "optimal" parameter. Below are two methods for finding good regularization parameters when the error variance s^2 is not known.

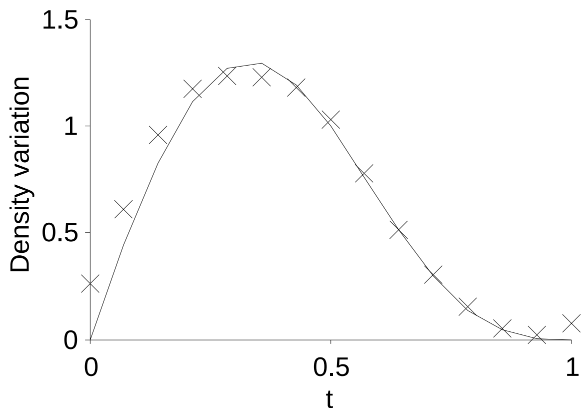

FIGURE 1: Ridge regression estimate of density variations (crosses) and exact solution (solid line). The ridge regression estimate is for the regularization parameter $h = 0.47$ determined with the discrepancy principle.

L-curve (Hansen 1998, chapters 4.6 and 7.5). The L-curve is a plot of the squared estimate norm $\|\hat{\boldsymbol{\beta}}\|^2$ against the residual sum of squares $\|\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{y}\|^2$. An optimal trade-off between estimate norm and RSS is found at the corner of the L-curve.

The L-curve for regularizing the gravity surveying example of homework 5 by ridge regression is shown in Figure 2. The corner of the L-curve is located approximately at the regularization parameter $h \approx 0.14$. Figure 3 shows the corresponding estimate of the density variations. In this example, choosing the regularization parameter h as the corner of the L-curve is seen to lead to undersmoothing of the estimate.

Generalized cross-validation (Hansen 1998, chapter 7.4; Wahba 1990, chapter 4). Generalized cross-validation is a form of leave-one-out cross-validation. It is based on successively leaving out elements y_i of the response vector $\mathbf{y}$, computing a regularized estimate of the parameter vector $\boldsymbol{\beta}$ from the reduced dataset, and predicting the left-out response variable y_i with the estimated model. The regularization parameter is chosen as the minimizer of the cross-validated prediction error. Generalized cross-validation differs from ordinary cross-validation in that the cross-validation is carried out after the data are transformed to a basis in which the individual rows of the design matrix $\mathbf{X}$ are strongly coupled; this transformation ensures that the regularization parameter determined by generalized cross-validation is independent of orthogonal transformations (rotations) of the response vector $\mathbf{y}$ (see Wahba 1990, chapter 4, for details).

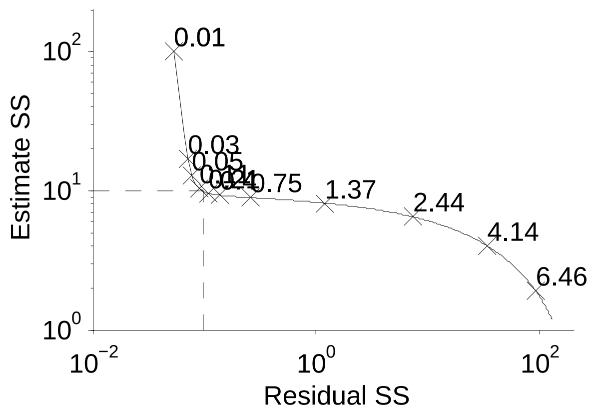

FIGURE 2: L-curve for gravity surveying example. The crosses are at points where the regularization parameter h is equal to one of the singular values σ_i of the design matrix $\mathbf{X}$. The corner of the L-curve is at $h \approx 0.14$.

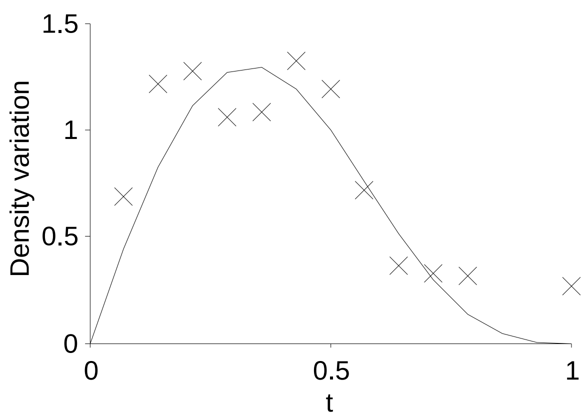

FIGURE 3: Ridge regression estimate of density variations (crosses) and exact solution (solid line). The ridge regression estimate is for the regularization parameter $h = 0.14$ estimated from the L-curve (Figure 2).

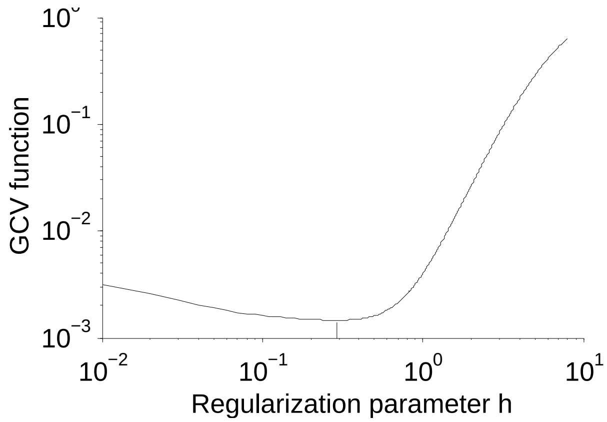

FIGURE 4: GCV function for ridge regression. The minimum of the GCV function is at $h = 0.29$.

In generalized cross-validation as in ordinary leave-one-out cross-validation, it is not necessary to compute n individual estimates of the parameter vector $\boldsymbol{\beta}$ to compute the cross-validated mean square prediction error. It can be shown that choosing a regularization parameter by generalized cross-validation amounts to choosing the regularization parameter that minimizes the GCV function

$$\mathcal{G} = \frac{\mathrm{RSS}}{\mathcal{T}^2}, \tag{3}$$

where

$$\mathcal{T} = n - \sum_{i=1}^{p} f_i, \tag{4}$$

is an effective number of degrees of freedom and f_i are the filter factors of the regularization method employed. For TSVD with truncation parameter $\hat{r}$ ($f_i = 1$ for $i \leq \hat{r}$ and $f_i = 0$ for $i > \hat{r}$), the effective number of degrees of freedom is $\mathcal{T} = n - \hat{r}$, corresponding to $\hat{r}$ effective predictors. For ridge regression, the effective number of degrees of freedom need not be an integer. Both the effective number of degrees of freedom and the residual sum of squares generally increase with increasing regularization. The GCV function (3) attempts to find a good trade-off between the two.

The GCV function (3) for regularizing the gravity surveying example of homework 5 by ridge regression is shown in Figure 4. The minimum of the GCV function is at $h = 0.29$. Figure 5 shows the corresponding estimate of the density variations. Generalized cross-validation appears to give a relatively good estimate

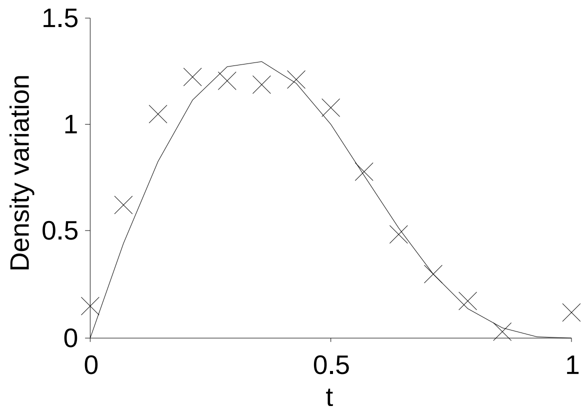

FIGURE 5: Ridge regression estimate of density variations (crosses) and exact solution (solid line). The ridge regression estimate is for the regularization parameter $h = 0.29$ chosen by generalized cross-validation.

of the regularization parameter in this example. Generalized cross-validation is one of the most successful methods for choosing a regularization parameter.

REFERENCES

Hansen, P. C., 1998: *Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion*. SIAM Monogr. on Mathematical Modeling and Computation, Society for Industrial and Applied Mathematics.

Wahba, G., 1990: *Spline Models for Observational Data*. CBMS-NSF Regional Conference Series in Applied Mathematics, Vol. 59, Society for Industrial and Applied Mathematics.