

Cross-Validation in Function Estimation

Chong Gu

October 1, 2006

Cross-validation is an intuitive and effective technique for model selection in data analysis. In this discussion, I try to present a few incarnations of the general technique in a few nonparametric function estimation settings. Justifications of the technique in Gaussian regression settings will be discussed, along with possible reasons for the lack of similar justification in other settings. There will be discussions of some subtle conceptual issues which put certain widely adopted concepts/practice under scrutiny.

1 Cross-Validation and Related Techniques

1.1 PRESS and C_p

Consider a linear regression model with $P - 1$ predictors X_1, \dots, X_{P-1} ,

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_{P-1} X_{P-1} + \epsilon,$$

where $\epsilon \sim N(0, \sigma^2)$. It is assumed that $\mu_Y = \beta_0 + \beta_1 X_1 + \dots + \beta_{P-1} X_{P-1}$, but some of the $\beta_j X_j$'s may contribute very little or not at all. For model selection in this setting (a.k.a. variable selection), two effective techniques are PRESS and C_p .

Observing $(Y_i, X_{i,1}, \dots, X_{i,P-1})$, $n = 1, \dots, n$, one may calculate PRESS (Predicted RESidual Sum of Squares) for every of the $(2^{P-1} - 1)$ possible models,

$$\text{PRESS} = \sum_{i=1}^n (Y_i - \hat{Y}_{i(i)})^2,$$

where $\hat{Y}_{i(i)} = \hat{\beta}_0^{[i]} + \hat{\beta}_1^{[i]} X_{i,1} + \dots + \hat{\beta}_{P-1}^{[i]} X_{i,P-1}$ and $\hat{\beta}_j^{[i]}$'s are LS estimates using the $(n - 1)$ observations excluding $(Y_i, X_{i,1}, \dots, X_{i,P-1})$. One then can choose the model with the minimum PRESS score. PRESS is probably the first incarnation of cross-validation, and the objective of model selection is to achieve more precise prediction.

For every of the $(2^{P-1} - 1)$ possible models, one may also calculate the C_p statistic,

$$C_p = \frac{\text{SSE}}{\text{MSE}(X_1, \dots, X_{P-1})} - (n - 2p),$$

where SSE is for the model under consideration and p is its number of coefficients. One selects a model with a small C_p (so that the MSE $E(\hat{\mu} - \mu)^2$ is small) that is close to p (so that $E\hat{\mu} \approx \mu$). The objective of C_p -based selection is for the estimation precision of μ_Y but *not* for the identification of the “correct” model, as the full model is assumed to be correct to start with. The estimation of “negligible” β_j 's diverts resources and thus results in higher variance of the important $\hat{\beta}_j$'s, and the elimination of the minor terms, while introducing bias in $\hat{\mu}$, helps to reduce the variance.

1.2 Cross-validation, GCV, and C_L

Consider a regression model $Y = \eta(x) + \epsilon$ where $\epsilon \sim N(0, \sigma^2)$ and $f(x)$ is “smooth.” Two of the popular approaches to the nonparametric estimation of $f(x)$ are the kernel method and the penalized least squares method. To keep things simple, consider only univariate x .

Observing (Y_i, X_i) , $n = 1, \dots, n$, one may estimate $f(x)$ via

$$\hat{\eta}_h(x) = \frac{1}{nh} \sum_{i=1}^n Y_i K\left(\frac{X_i - x}{h}\right),$$

where $K(\cdot)$ is a given kernel function. Typically, $K(x)$ is unimodal and symmetric with respect to 0, $\int K(x)dx = 1$, and $\int x^j K(x)dx = 0$ for $j = 1, \dots, m$. This provides a family of estimates indexed by the *bandwidth* h , with smaller h yielding smaller bias but larger variance and larger h yielding smaller variance but larger bias.

Assuming $\int (\eta^{(m)}(x))^2 dx < \infty$, one may estimate $\eta(x)$ using the minimizer $\eta_\lambda(x)$ of

$$\frac{1}{n} \sum_{i=1}^n (Y_i - \eta(X_i))^2 + \lambda \int (\eta^{(m)}(x))^2 dx, \quad (1)$$

where the *smoothing parameter* λ plays similar role as the h for kernel estimates. With $\lambda = \infty$ one forces a parametric model of the form $\eta(x) = \beta_0 + \beta_1 x + \dots + \beta_{m-1} x^{m-1}$. With $\lambda = 0_+$ one obtains an interpolant with minimum $\int (\eta^{(m)}(x))^2 dx$. The solution is called a smoothing (natural) spline as η_λ is a piece-wise polynomial of order $2m - 1$.

Evaluating the estimated $\eta(x)$ at the data points X_i , one gets the predicted values \hat{Y}_i . For the kernel estimates, smoothing spline estimates, and other nonparametric estimates of $\eta(x)$ known as *linear smoothers*, one has

$$\hat{\mathbf{Y}} = A\mathbf{Y},$$

where A is the *smoothing matrix* indexed by h or λ or the like. In practice, one needs to select h or λ , for which the methods in the section title are designed; the smoothing matrix A plays an important role here.

Write $\hat{Y}_i = \eta_\lambda(X_i)$ with $\eta_\lambda(x)$ the minimizer of (1), and $\hat{Y}_{i(i)} = \eta_\lambda^{[i]}(X_i)$, where $\eta_\lambda^{[i]}(x)$ minimizes the delete-one version of (1),

$$\frac{1}{n} \sum_{j \neq i} (Y_j - \eta(X_j))^2 + \lambda \int (\eta^{(m)}(x))^2 dx. \quad (2)$$

It can be shown that $\eta_\lambda^{[i]}(x)$ is the minimizer of (1) with $\eta_\lambda^{[i]}(X_i)$ replacing Y_i , thus $a_{i,i}(Y_i - \hat{Y}_{i(i)}) = \hat{Y}_i - \hat{Y}_{i(i)}$, or

$$Y_i - \hat{Y}_{i(i)} = (Y_i - \hat{Y}_i)/(1 - a_{i,i}).$$

This leads to the ordinary cross-validation score

$$V_0(\lambda) = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_{i(i)})^2 = \frac{1}{n} \sum_{i=1}^n \frac{(Y_i - \hat{Y}_i)^2}{(1 - a_{i,i})^2}. \quad (3)$$

An invariance argument suggests the replacement of $a_{i,i}$ by their average value $\text{trace}A/n$, yielding the generalized cross-validation (GCV) of Craven and Wahba (1979),

$$V(\lambda) = \frac{n^{-1} \mathbf{Y}^T (I - A(\lambda))^2 \mathbf{Y}}{\{n^{-1} \text{trace}(I - A(\lambda))\}^2}. \quad (4)$$

Note that although the derivation of GCV is through (1), it can be used and do get used for all linear smoothers.

Closely related to GCV is the C_L score,

$$U(\lambda) = \frac{1}{n} \mathbf{Y}^T (I - A(\lambda))^2 \mathbf{Y} + 2 \frac{\sigma^2}{n} \text{trace} A(\lambda), \quad (5)$$

which assumes a known σ^2 .

1.3 Optimality of $U(\lambda)$ and $V(\lambda)$

Before presenting the theoretical justification of C_L and GCV, let me try to clarify a few conceptual issues.

In parametric statistics, one has a discrete collection of tentative models and assumes that the correct model is among them, and a model selection method is consistent if it zeroes in to this correct model as $n \rightarrow \infty$. The consistency property is not prediction-oriented as in the design of PRESS, nor is it finite-sample MSE-oriented as in the practical use of C_p ; with $n \rightarrow \infty$ for fixed number of parameters, variance is of no concern and all that matters is the bias. [I am not familiar with this line of literature, so I do not know if there are results when none of the tentative models is correct, or whether one could formulate the problem in such a way that the number of parameters increase with n so bias-variance trade-off remains relevant.]

In nonparametric function estimation, the family of tentative estimates form a “trajectory” in the function space and one could not and does not assume that the true function is on the trajectory. Instead, one looks for the estimate that is “closest” to the true function in some sense. The objective of such “model selection” is *not* to identify the “correct” model (there is none), but rather to locate the “best” estimate given the observed data. Bias-variance trade-off is the central issue here (though not always explicitly), and as $n \rightarrow \infty$, the optimal choice would have smaller and smaller bias. For Gaussian regression, a natural measure for estimation precision is the MSE on the data points,

$$L(\lambda) = \frac{1}{n} \sum_{i=1}^n (\hat{\eta}_\lambda(X_i) - \eta(X_i))^2,$$

and the estimate on the trajectory that minimizes this MSE would be the optimal choice. Note that the trajectory is dependent on the observed data, so is the MSE loss and its minimizer. The C_L score of (5) is actually an unbiased “estimate” of relative loss, with “relative” meaning the dropping of terms that do not depend on λ .

The optimality of $U(\lambda)$ and $V(\lambda)$ were established by Ker-Chau Li (1986): Let λ_o , λ_u , and λ_v be the minimizers of $L(\lambda)$, $U(\lambda)$, and $V(\lambda)$, respectively, then

$$\frac{L(\lambda_o)}{L(\lambda_u)} \xrightarrow{p} 1, \quad \frac{L(\lambda_o)}{L(\lambda_v)} \xrightarrow{p} 1. \quad (6)$$

The key results leading to these are that as $\lambda \rightarrow 0$ at certain rates that contain the optimal one,

$$U(\lambda) - L(\lambda) - n^{-1} \boldsymbol{\epsilon}^T \boldsymbol{\epsilon} = o_p(L(\lambda)), \quad V(\lambda) - L(\lambda) - n^{-1} \boldsymbol{\epsilon}^T \boldsymbol{\epsilon} = o_p(L(\lambda));$$

The main condition for these is that $nR(\lambda) \rightarrow \infty$, where $R(\lambda) = E[L(\lambda)]$, which states that the parametric \sqrt{n} -consistency is not achievable in the setting. Note that $U(\lambda) \sim V(\lambda) \stackrel{p}{\sim} 1$ and $L(\lambda) = o_p(1)$, so these are delicate results.

For the record, as $n \rightarrow \infty$ and $\lambda \rightarrow 0$, $L(\lambda) = O_p(\lambda + n^{-1} \lambda^{1/2m})$ for the minimizer of (1).

1.4 Cross-validation for density estimation

Consider a density estimation problem with independent samples $X_i \sim f(x)$, $i = 1, \dots, n$. To estimate $f(x)$, one may use the kernel estimate,

$$f_h(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{X_i - x}{h}\right),$$

where $K(\cdot)$ is as given earlier. To assess the performance of $f_h(x)$ as an estimate of $f(x)$, one may use the Kullback-Leibler loss

$$L(h) = E_f[\log\{f(X)/f_h(X)\}],$$

which, after dropping the term $E_f[\log f(X)]$ not involving h , reduces to the relative KL discrepancy

$$-E_f[\log f_h(X)].$$

Estimating the relative KL by cross-validated sample mean, one has the KL cross-validation score,

$$V(h) = -\frac{1}{n} \sum_{i=1}^n \log f_h^{[i]}(X_i), \quad (7)$$

where $f_h^{[i]}(x)$ is based on the $(n-1)$ samples excluding X_i . It was shown by Peter Hall (1987) that if the tails of the kernel $K(x)$ are no thinner than the tails of $f(x)$, then

$$\frac{L(h_o)}{L(h_v)} \xrightarrow{p} 1, \quad (8)$$

where h_o and h_v minimizes $L(h)$ and $V(h)$, respectively.

Parallel to (1), one may assume a finite support \mathcal{X} for $f(x)$ and employ the logistic density transform $f(x) = e^{\eta(x)} / \int_{\mathcal{X}} e^{\eta}$, and estimate $\eta(x)$ by the minimizer $\eta_\lambda(x)$ of the penalized likelihood functional,

$$-\frac{1}{n} \sum_{i=1}^n \left\{ \eta(X_i) - \log \int_{\mathcal{X}} e^{\eta(x)} \right\} + \frac{\lambda}{2} \int (\eta^{(m)}(x))^2 dx. \quad (9)$$

The Kullback-Leibler distance is now

$$L(\lambda) = E_f[\log\{f(X)/f_\lambda(X)\}] = E_f[\eta(X) - \eta_\lambda(X)] + \left\{ \log \int_{\mathcal{X}} e^{\eta_\lambda(x)} - \log \int_{\mathcal{X}} e^{\eta(x)} \right\}, \quad (10)$$

with a relative KL discrepancy

$$\log \int_{\mathcal{X}} e^{\eta_\lambda(x)} - E_f[\eta_\lambda(X)],$$

where the first term can be computed and the second term can be estimated through a cross-validated sample mean, $n^{-1} \sum_{i=1}^n \eta_\lambda^{[i]}(X_i)$, with $\eta_\lambda^{[i]}(x)$ minimizing the delete-one version of (9),

$$-\frac{1}{n-1} \sum_{j \neq i} \left\{ \eta(X_j) - \log \int_{\mathcal{X}} e^{\eta(x)} \right\} + \frac{\lambda}{2} \int (\eta^{(m)}(x))^2 dx.$$

This yields a cross-validation score

$$V(\lambda) = \log \int_{\mathcal{X}} e^{\eta_\lambda(x)} - \frac{1}{n} \sum_{i=1}^n \eta_\lambda^{[i]}(X_i). \quad (11)$$

While empirical results strongly suggest optimality similar to those established by Li (1806) and Hall (1987), attempts on the theoretical analysis have not been successful.

It can be shown that the symmetrized KL loss,

$$E_f[\log\{f(X)/f_\lambda(X)\}] - E_{f_\lambda}[\log\{f(X)/f_\lambda(X)\}],$$

which roughly doubles the KL loss $L(\lambda)$ of (10), is of the order $O_p(\lambda + n^{-1}\lambda^{1/2m})$, so the minimum KL loss is of order $O_p(n^{-2m/(1+2m)}) = o_p(n^{-1/2})$. On the other hand, the estimation of $E_f[g(X)]$ through the sample mean is at best of the order $O_p(n^{-1/2})$. One seems to need more delicate term grouping for any success at a theoretical analysis of $V(\lambda)$.

1.5 Cross-validation for non-Gaussian regression and hazard estimation

For non-Gaussian regression and hazard estimation, scores similar to (11) were derived following similar lines. The empirical performances of these scores suggest optimality properties similar to (6) and (8), but theoretical analysis is lacking.

1.6 Cross-validation for regression with correlated errors

For regression with correlated data, two scenarios have been treated in two dissertations by my former students, Ping Ma and Chun Han, respectively.

The first thesis was by Ping Ma, which concerns mixed-effect/variance-component models of the form

$$Y = \eta(x) + \mathbf{z}^T \mathbf{b} + \epsilon,$$

where $\mathbf{b} \sim N(0, B)$ and $\epsilon \stackrel{i.i.d.}{\sim} N(0, \sigma^2)$, with the dimension p of \mathbf{b} much smaller than n and its variance-covariance matrix partly or entirely unknown; the error variance-covariance matrices are low-rank modifications of $\sigma^2 I$. To estimate $\eta(x)$, one may minimize

$$\frac{1}{n} \sum_{i=1}^n (Y_i - \eta(x_i) - \mathbf{z}_i^T \mathbf{b})^2 + \lambda \int (\eta^{(m)}(x))^2 dx + \frac{1}{n} \mathbf{b}^T \Sigma \mathbf{b},$$

where (λ, Σ) are tuning parameters; Σ should reflect the structure of B^{-1} . The joint estimation of $(\eta(x), \mathbf{b})$ yields $\hat{Y} = \hat{\eta}(x) + \mathbf{z}^T \hat{\mathbf{b}}$, and one still has an expression $\hat{\mathbf{Y}} = A(\lambda, \Sigma) \mathbf{Y}$. The selection of (λ, Σ) can be done using C_L and GCV, and we were able to establish the optimality of such practice similar to (6), with the losses

$$L_1(\lambda, \Sigma) = \frac{1}{n} \sum_{i=1}^n (\hat{Y}_i - \eta(x_i) - \mathbf{z}_i^T \mathbf{b})^2$$

and

$$L_2(\lambda, \Sigma) = \frac{1}{n} (\hat{\boldsymbol{\eta}} - \boldsymbol{\eta})^T P_Z^\perp (\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}),$$

where $\boldsymbol{\eta}^T = (\eta(x_1), \dots, \eta(x_n))$, $P_Z^\perp = I - Z(Z^T Z)^{-1} Z^T$, and $Z^T = (\mathbf{z}_1, \dots, \mathbf{z}_n)$.

Note that we are not concerned with the estimation of B in this exercise, but only use Σ (or B^{-1}) as tuning parameters for the estimation of $\eta(x)$ or $\eta(x) + \mathbf{z}^T \mathbf{b}$.

For the optimality with respect to L_1 , one needs $p = O(\sqrt{n})$. For L_2 , one needs p fixed.

Similar practice was applied to non-Gaussian regression with satisfactory empirical performance, but theory is lacking.

The second thesis was by Chun Han, which concerns stationary time series models or mixed-effect models with the dimension of \mathbf{b} growing with n ; the error variance-covariance matrices are no longer low-rank modifications of $\sigma^2 I$. Formally, one has

$$Y_i = \eta(x_i) + \epsilon_i, \quad \epsilon \sim N(0, \sigma^2 W^{-1}),$$

where W is known up to a set of parameter γ . One may estimate $\eta(x)$ via the minimization of

$$(\mathbf{Y} - \boldsymbol{\eta})^T W (\mathbf{Y} - \boldsymbol{\eta}) + n\lambda \int (\eta^{(m)}(x))^2 dx,$$

with (λ, γ) as tuning parameters. For the joint selection of (λ, γ) , we derived the C_L like score

$$U(\lambda, \gamma) = \frac{1}{n\sigma^2} \mathbf{Y}^T W^{1/2} (I - A)^2 W^{1/2} \mathbf{Y} - \frac{1}{n} \log |W| + \frac{2}{n} \text{trace} A,$$

where $W^{1/2} \hat{\mathbf{Y}} = AW^{1/2} \mathbf{Y}$, and the GCV like score

$$V(\lambda, \gamma) = \log \{ n^{-1} \mathbf{Y}^T W^{1/2} (I - A)^2 W^{1/2} \mathbf{Y} \} - \frac{1}{n} \log |W| + \frac{2 \text{trace} A}{n - \text{trace} A}.$$

Remember that W depends on γ and A depends on (λ, γ) . Using the Kullback-Leibler loss for the joint estimation of $(\eta_0(x), \gamma_0)$ by $(\eta(x), \gamma)$,

$$\begin{aligned} L(\lambda, \gamma) &= E_0 \left[\frac{1}{2\sigma^2} (\mathbf{Y} - \boldsymbol{\eta})^T W (\mathbf{Y} - \boldsymbol{\eta}) - \frac{1}{2} \log |W| - \frac{1}{2\sigma^2} (\mathbf{Y} - \boldsymbol{\eta}_0)^T W_0 (\mathbf{Y} - \boldsymbol{\eta}_0) + \frac{1}{2} \log |W_0| \right] \\ &= \frac{1}{2\sigma^2} (\boldsymbol{\eta}_0 - \boldsymbol{\eta})^T W (\boldsymbol{\eta}_0 - \boldsymbol{\eta}) + \frac{1}{2} \text{tr} (W W_0^{-1} - I) - \frac{1}{2} \log |W W_0^{-1}|, \end{aligned}$$

the optimality similar to (6) was established for U and V , and the resulting $\hat{\gamma}$ is \sqrt{n} -consistent. A key condition for the theory is that $W^{-1} > cI$ for some $c > 0$ uniformly over the tentative γ . The theory applies to the standard stationary and invertible ARMA models with γ in a compact set, and to a mixed-effect model with $p \asymp n$.

1.7 References

PRESS is in nearly every textbook on regression analysis and linear models, though I don't know exactly who first proposed it. To read about C_p and C_L , check out the classical reference of Mallows.

Mallows, C. L. (1973). "Some comments on C_p ," *Technometrics* 15, 661–675.

For the motivation and derivation of GCV along with some attempt on the theoretical justification of $U(\lambda)$ and $V(\lambda)$, check out the seminal work of Craven and Wahba (1979); the risk-based theory does not really justify its practical use, however. The really relevant loss-based justification of $U(\lambda)$ and $V(\lambda)$ was by Li (1986).

Craven, P. and G. Wahba (1979). "Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation," *Numer. Math.* 31, 377–403.

Li, K.-C. (1986). "Asymptotic optimality of C_L and generalized cross-validation in the ridge regression with application to spline smoothing," *Ann. Statist.* 14, 1101–1112.

Kernel density estimation and the associated bandwidth selection was once a business by itself. The work of Hall (1987) for the justification of KL-based cross-validation score (7) was an important contribution, but it was taken by many as negative on the KL distance as performance measure and negative on CV as bandwidth selector, and partly inspired the developments of the so-called plug-in methods for bandwidth selection; I consider this an unfortunate turn of event.

Hall, P. (1987). “On Kullback-Leibler loss and density estimation,” *Ann. Statist.* 15, 1491–1519.

The implementation and empirical performance of (11) can be found in my joint work with Jingyuan Wang and in Chapter 6 of my book. Discussions concerning non-Gaussian regression and hazard estimation are in Chapters 5 and 7 of my book.

Gu, C. and J. Wang (2003). “Penalized likelihood density estimation: Direct cross-validation and scalable approximation,” *Statist. Sin.* 13, 811–826.

Gu, C. (2002). *Smoothing Spline ANOVA Models*. New York: Springer-Verlag.

Details concerning cross-validation for regression with correlated errors are to be found in the following articles.

Gu, C. and P. Ma (2005). “Optimal smoothing in nonparametric mixed-effect models,” *Ann. Statist.* 33, 1357–1379.

Gu, C. and P. Ma (2005). “Generalized nonparametric mixed-effect models: Computation and smoothing parameter selection,” *J. Comput. Graph. Statist.* 14, 485–504.

Han, C. and C. Gu (2006). “Optimal smoothing with correlated data,” manuscript.

2 Problems (Real and Perceived), Concepts, and Controversies

2.1 A simple simulation setting

We will use a simple simulation setting to illustrate some of the issues to be discussed. The issues are not specific to this particular setting, as similar simulations in other settings demonstrate the same qualitative characteristics.

On $x_i = (i - .5)/n, i = 1, \dots, n$, we generate 100 replicates of data from

$$Y_i = \eta(x_i) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2),$$

with $\eta(x) = 1 + 3 \sin(2\pi x - \pi)$ and $\sigma^2 = 1$. For λ on a fine grid of $\log_{10} n\lambda = (-5)(.05)(-1)$, we calculate the minimizers of (1) with $m = 2$ for each of the replicates and evaluate various quantities associated with them. The grid was broad enough to bracket the λ of interest for all the 100 replicates.

2.2 Undersmoothing and modifications of cross-validation

One problem suffered by cross-validation methods is undersmoothing: in up to 10% of the cases, the methods lead to very small λ or h or the like, resulting in undersmoothing or even interpolation. The problem doesn’t seem to go away with larger n , at least not for n up to 500.

An alternative to $V(\lambda)$ of (4) is the so-called generalized maximum likelihood (GML) method derived under the empirical Bayes interpretation of smoothing splines, which is simply the restricted

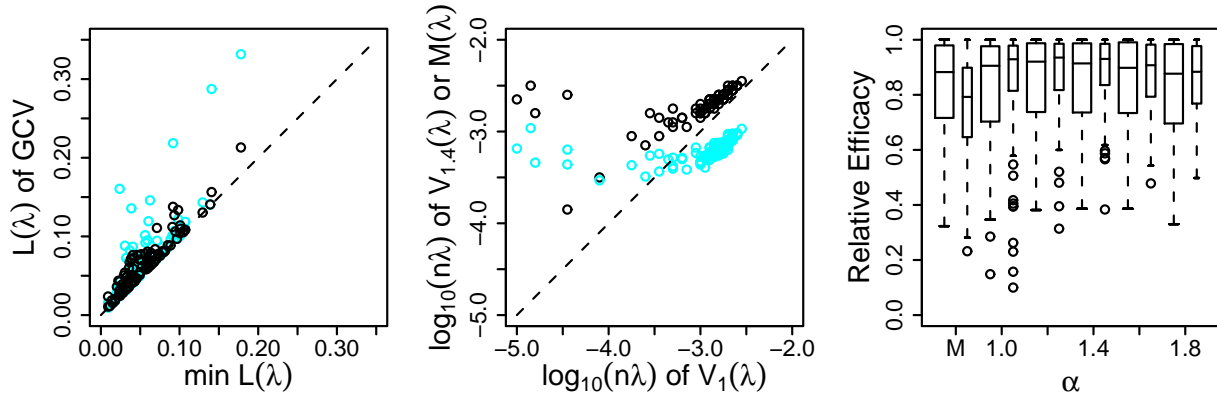


Figure 1: Left: Performances of $V_\alpha(\lambda)$ with $\alpha = 1$ (faded) and $\alpha = 1.4$ for $n = 100$. Center: the λ minimizing $V_1(\lambda)$ versus that minimizing $M(\lambda)$ (faded) or $V_{1.4}(\lambda)$, for $n = 100$. Right: $\min L(\lambda)/L(\hat{\lambda})$ with $\hat{\lambda}$ minimizing $M(\lambda)$ or $V_\alpha(\lambda)$ at $\alpha = 1, 1.2, 1.4, 1.6, 1.8$, for $n = 100$ (fatter boxes) and $n = 500$ (thinner boxes).

maximum likelihood (REML) method for mixed-effect/variance component models. The GMLscore is given by

$$M(\lambda) = \frac{n^{-1}\mathbf{Y}^T(I - A(\lambda))\mathbf{Y}}{|I - A(\lambda)|_+^{1/(n-m)}},$$

where $|I - A|_+$ is the product of the $n - m$ positive eigenvalues of $(I - A)$. The GML method never interpolates, but consistently undersmooths for $\eta(x)$ “super smooth.”

A simple modification seems to cure the undersmoothing problem for cross-validation. For $V(\lambda)$ of (4), one may use

$$V_\alpha(\lambda) = \frac{n^{-1}\mathbf{Y}^T(I - A(\lambda))^2\mathbf{Y}}{\{n^{-1}\text{trace}(I - \alpha A(\lambda))\}^2}, \quad (12)$$

where $\alpha > 1$, and for $V(\lambda)$ of (11), one may use

$$V_\alpha(\lambda) = \log \int_{\mathcal{X}} e^{\eta_\lambda(x)} - \frac{1}{n} \sum_{i=1}^n \eta_\lambda(X_i) + \alpha \frac{1}{n} \sum_{i=1}^n \{\eta_\lambda(X_i) - \eta_\lambda^{[i]}(X_i)\}. \quad (13)$$

Simulation studies suggest that an α in the range $1.2 \sim 1.4$ would be the most effective.

In the setting of §2.1, one may calculate the loss $L(\lambda)$ as well as the selection scores $V_\alpha(\lambda)$ and $M(\lambda)$ on the λ grid for all the replicates, and identify λ_o , λ_v along with the associated $L(\lambda)$. Figure 1 summarizes some of the empirical results.

A so-called extended exponential (EE) method was proposed by Kou and Efron (2002) to “combine the strengths of C_p and GML,” but I am not able to follow the arguments.

2.3 Negative correlation and model indexing

One major criticism of cross-validation in the literature was the famous negative correlation between the optimal and cross-validated bandwidths, as demonstrated in the middle frame of Figure 2. The negative correlation is bothersome only when the index λ is meaningful across-replicate, however, which will be analyzed below.

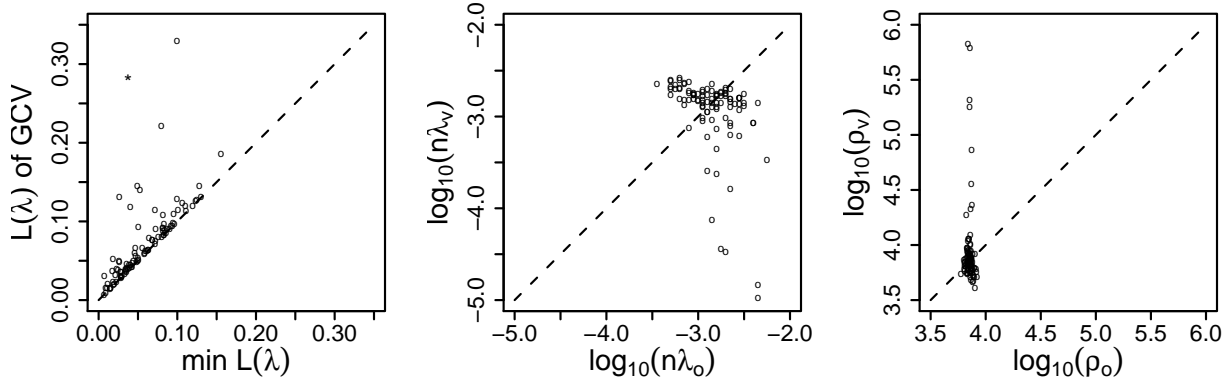


Figure 2: Left: Performances of $V_1(\lambda)$ for $n = 100$. Center: $\log_{10}(n\lambda_o)$ versus $\log_{10}(n\lambda_v)$. Right: $\log_{10}(\rho_o)$ versus $\log_{10}(\rho_v)$.

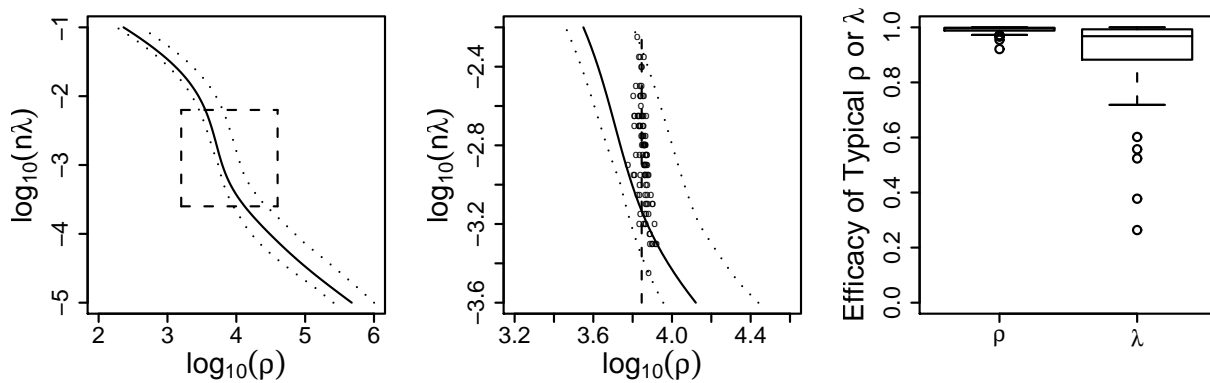


Figure 3: Left and Center: The ρ - λ mapping and the (ρ_o, λ_o) pair of 100 replicates. Right: Efficacy of “typical” ρ_o or λ_o .

Remember that unlike in the parametric settings where one selects from a discrete set of models, we are choosing from a continuum of tentative estimates $\eta_\lambda(x)$. An important issue is how to “align” estimates based on different data. Mathematically, the minimizer of (1) is the solution to a constrained LS problem,

$$\min \sum_{i=1}^n (Y_i - \eta(x_i))^2, \quad \text{s.t.} \quad \int_0^1 (\eta^{(m)}(x))^2 dx \leq \rho$$

for some $\rho > 0$. There is a one-to-one correspondence between λ and ρ given the data. Intuitively, ρ is meaningful across-replicate as it imposes the same constraint on $\eta(x)$ in the estimation process regardless the data one observes, whereas the same λ implies different constraints for different observations. A simple simulation will confirm that it is indeed the case, that $\eta_\lambda(x)$ based on different data should be aligned by the corresponding ρ , but we first observe in the right frame of Figure 2 that the negative correlation disappears on the ρ scale.

In the setting of §2.1, one may calculate $\rho = \int_0^1 (\eta^{(2)}(x))^2 dx$ for all estimates $\eta_\lambda(x)$ on the λ grid, for all replicates, and for the true function $\eta(x) = 1 + 3 \sin(2\pi x - \pi)$; the true function has $\rho = 10^{3.846}$. In the left and middle frames of Figure 3, we demonstrate the ρ - λ mapping for 100

replicates of samples of size $n = 100$ and identify the 100 optimal (ρ_o, λ_o) on the grid; it is reassuring to see that $10^{3.846}$ is in the middle of the ρ_o 's. To assess the across-replicate interpretability of ρ and λ , we pick a “typical” optimal $\tilde{\rho} = 10^{3.846}$, and a “typical” optimal $\tilde{\lambda} = \text{median}(\lambda_o)$, and calculates their efficacy over the replicates via $L(\rho_o)/L(\tilde{\rho})$ and $L(\lambda_o)/L(\tilde{\lambda})$ which is shown in the right frame of Figure 3. The tight spread of ρ_o 's confirms the intuition that ρ is the proper model index here.

Besides smoothing splines, there does not seem to exist a “ ρ -index” for other smoothing methods. Nevertheless, model indexing remains an important issue, which has subtle implications in the theory and practice of bandwidth selection.

Even for smoothing splines, the ρ -index is difficult to work with, and the identification of it as the proper model index does *not* offer any operational help. It however helps to identify some popular but questionable concepts and practices.

One of the misleading concept is the “degree-of-freedom” in regression, which is defined as $\text{trace}A$: given x_i 's, λ - $\text{trace}A$ is a one-to-one mapping, so the selection of λ through $\text{trace}A$ simply leaves things in the hands of random noise ϵ_i . In classical parametric statistics, the degree-of-freedom is defined in terms of model dimension, and it is relevant in settings other than regression. The “coincidence” that the trace of hat matrix in linear regression matches the model dimension does *not* automatically qualify the trace as a valid model index. In fact, $\text{trace}A$ is much worse than λ , as it allows one to compare $A(\lambda)$ with $A(h)$ while they are not comparable.

Resampling is widely used in many phases of statistical analysis. Working with a index such as λ that is not meaningful across-replicate, however, one should avoid using resampling for bandwidth selection.

Besides the negative correlation, another criticism against cross-validation is that the cross-validated bandwidth “converges” very slowly asymptotically. Two aspects in this argument need close scrutiny: i) if the target of “convergence” is something like a fixed λ , then it is no worry because the optimal λ_o changes with data and λ_v may simply be “chasing” λ_o . ii) The bandwidth is *not* part of the stochastic setting but a *tuning* parameter in the estimation process, and its meaning is only through $L(\lambda)$, so it's okay for λ_v to be far from λ_o as long as $L(\lambda_o)/L(\lambda_v) \approx 1$.

2.4 Loss versus risk

Risk calculation is a basic exercise in the theoretical analysis of statistical procedures, but when done using a model index not interpretable across-replicate, the proper use of it can be tricky.

Imagine that you are studying the MSE performance of LS regression empirically, and you have 3 models $\mathcal{M}_1 = \{\mu = E[Y] = \beta_0 + \beta_1 x\}$, $\mathcal{M}_2 = \{\mu = \beta_0 + \beta_1 x + \beta_2 x^2\}$, and $\mathcal{M}_3 = \{\mu = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3\}$. After generating data, you calculate $\hat{\mu}^{(j)}$, $j = 1, 2, 3$, under \mathcal{M}_j and calculate the MSE $n^{-1} \sum_{i=1}^n (\hat{\mu}_i^{(j)} - \mu_i)^2$, and you average over replicates the MSEs corresponding to the same $\hat{\mu}^{(j)}$. You'd be crazy to take $\hat{\mu}^{(1)}$ from the first 20 replicates to average with $\hat{\mu}^{(2)}$ from the next 40 replicates.

While to a much minor extent, the calculation of $R(\lambda) = E[L(\lambda)]$ is like mixing $\hat{\mu}^{(1)}$'s from some replicates with $\hat{\mu}^{(2)}$'s from other replicates in the above example. You may notice that I never call $\eta_\lambda(x)$ an estimator but only an estimate given the data. Many would call the minimizer of $R(\lambda)$ “optimal,” but since I never observe an “average” sample, I consider a “risk-optimal” λ meaningless, and don't care about any convergence towards it.

While derivative concepts based on $R(\lambda)$ have no meaning in my books, $R(\lambda)$ serves as an important analytical device. For example, to prove $V(\lambda) - L(\lambda) - n^{-1} \epsilon^T \epsilon = o_p(L(\lambda))$, one proceeds by showing that $V(\lambda) - L(\lambda) - n^{-1} \epsilon^T \epsilon = o_p(R(\lambda))$ and that $L(\lambda) - R(\lambda) = o_p(R(\lambda))$; to establish $L(\lambda) = O_p(\lambda + n^{-1} \lambda^{-1/2m})$, one simply show that $R(\lambda) = O(\lambda + n^{-1} \lambda^{-1/2m})$.

Risk calculation using the ρ -index would be conceptually meaningful, only if it were possible.

Although no one questions the appeal of loss-optimal bandwidth and most do simulations in terms of it (negative correlation wouldn't be there otherwise), many authors claim that the data do not contain enough information for one to pursue the loss-optimal bandwidth. In view of the results by Li (1986) and Hall (1987), such claims are misleading.

2.5 References

For the derivation of GML and its asymptotic analysis, check Wahba (1985). The GML always pick $\lambda \asymp n^{-2m/(2m+1)}$ but for $\eta(x)$ "super smooth" the optimal one is $\lambda \asymp n^{-2m/(4m+1)}$.

Wahba, G. (1985). "A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem," *Ann. Statist.* 13, 1378–1402.

The modifications of cross-validation in (12) and (13) and the empirical performances can be found in the articles based on the theses of Jingyuan Wang and Young-Ju Kim.

Gu, C. and J. Wang (2003). "Penalized likelihood density estimation: Direct cross-validation and scalable approximation," *Statist. Sin.* 13, 811–826.

Kim, Y.-J. and C. Gu (2004). "Smoothing spline Gaussian regression: More scalable computation via efficient approximation," *J. Roy. Statist. Soc. Ser. B* 66, 337–356.

Details of the EE method can be found in the reference below.

Kou, S. and B. Efron (2002). "Smoothers and the C_p , GML and EE criteria: A geometric approach," *J. Amer. Statist. Assoc.* 97, 766–782.

The negative correlation between cross-validated bandwidth and the optimal bandwidth was observed by many, and was publicized by Scott and Terrell (1987) and Hall and Johnstone (1992), though all took it at the face value and tried to "fix" it, which is unnecessary.

Scott, D. W. and G. R. Terrell (1987). "Biased and unbiased cross-validation in density estimation," *J. Amer. Statist. Assoc.* 82, 1131–1146.

Hall, P. and I. Johnstone (1992). "Empirical functionals and efficient smoothing parameter selection" (with discussion) *J. Roy. Statist. Soc. Ser. B* 54, 475–530.

Detailed discussion of model indexing and the ramifications in bandwidth selection can be found in the following reference.

Gu, C. (1998). "Model indexing and smoothing parameter selection in nonparametric function estimation" (with discussion), *Statist. Sin.* 8, 607–646.