Penalized Likelihood Regression: Fast Computation and Direct Cross-Validation

Young-Ju Kim* Chong Gu†

Abstract

We consider penalized likelihood regression with exponential family responses. Parallel to recent development in Gaussian regression, the fast computation through asymptotically efficient low-dimensional approximations is explored, yielding algorithm that scales much better than the $O(n^3)$ algorithm for the exact solution. Also customizations of the direct cross-validation strategy for smoothing parameter selection in various distribution families are explored and evaluated.

KEY WORDS: Cross-validation; Kullback-Leibler; Penalized likelihood; Smoothing parameter.

1 Introduction

Suppose we have independent observations (x, Y) with a minus log likelihood $l(\eta; Y)$, where η is dependent on a covariate x. A classical parametric model restricts $\eta(x)$ to a low-dimensional function space through a certain function form $\eta(x, \beta)$, known up to a finite-dimensional parameter β, which are to be estimated from the data. To avoid possible model bias in a parametric model, one may allow η to vary in a high-dimensional function space which leads to nonparametric estimation techniques. We consider the penalized likelihood estimation of $\eta(x)$ through the minimization of

$$\frac{1}{n}\sum_{i=1}^{n} l_i(\eta(x_i); Y_i) + \frac{\lambda}{2} J(\eta), \tag{1.1}$$

where $J(\eta)$ is a roughness functional. The first term in (1.1) discourages the lack of fit of η to the data, the second term penalizes the roughness of η, and the smoothing parameter λ controls the trade-off between the two conflicting goals. The minimization of (1.1) is typically in an infinite-dimensional space $\mathcal{H} \subseteq \{\eta : J(\eta) < \infty\}$, with the minimizer denoted by η_λ.

Penalized likelihood regression with responses from exponential family distributions have been formulated and studied by O'Sullivan, Yandell, and Raynor (1986); see also Silverman (1978) and Green and Yandell (1985). The asymptotic convergence rates of η_λ have been established by Cox and O'Sullivan (1990), Gu and Qiu (1994), and Gu and Kim (2002), among others. The selection of λ and the computation of the estimates have been studied by Gu (1990, 1992), Xiang and Wahba (1996), and Gu and Xiang (2001).

The minimizer η_λ resides in an $n-$ dimensional space and the computation of η_λ is generally of the order $O(n^3)$ when the covariate x is multidimensional. One purpose of this article is to develop and illustrate tools for more scalable computation of penalized likelihood regression. Gu and Kim (2002) showed that the

*Department of Statistics, Kangwon National University, Chunchun Korea
†Department of Statistics, Purdue University, IN 47906 USA

minimizers of (1.1) in certain q-dimensional spaces share the same asymptotic convergence rates as η_λ, with q of the order as low as $n^{2/9}$ in commonly used settings; the computation of the q-dimensional estimates is of the order $O(nq^2)$, which amounts to $O(n^{13/9})$ for $q \asymp n^{2/9}$. This complements parallel development in Gaussian regression reported in Kim and Gu (2004).

The minimization of (1.1) for fixed λ typically requires the use of iterative methods such as the Newton iteration, and the selection of λ adds another dimension. Two methods have been developed in the literature for the selection of λ, with similar effectiveness but different numerical characteristics: the indirect cross-validation of Gu (1992) nests λ selection under the Newton iteration, whereas the direct cross-validation of Xiang and Wahba (1996) nests the Newton iteration under λ selection. The selection of λ through cross-validation is more involved than the Newton iteration, especially with the q-dimensional estimates for which analytical derivatives of cross-validation scores are not available. We hence choose to use the direct cross-validation. The direct cross-validation was shown to be highly effective for Bernoulli data in the simulations of Xiang and Wahba (1996) and Gu and Xiang (2001), but we find it necessary to customize the general method in various distribution families to achieve better performance/consistency, or to adapt the method to handle the complication of extra parameters. A major part of the article concerns the customizations of direct cross-validation for commonly used distribution families and the assessment of their empirical performances. Considered here are responses from exponential families.

2 Penalized Likelihood Regression

The minimization of (1.1) is in a space $\mathcal{H} \subseteq \{\eta : J(\eta) < \infty\}$ in which $J(\eta)$ is a square (semi) norm, or a subspace therein. The evaluation $[x]f = f(x)$ appears in the first term, which is assumed to be continuous in $\mathcal{H}$. A space $\mathcal{H}$ in which the evaluation is continuous is called a reproducing kernel Hilbert space (RKHS) possessing a reproducing kernel (RK) $R(\cdot,\cdot)$, a non-negative definite function satisfying $\langle R(x,\cdot), f(\cdot)\rangle = f(x)$, $\forall \eta \in \mathcal{H}$, where $\langle\cdot,\cdot\rangle$ is the inner product in $\mathcal{H}$. The norm and the RK determine each other uniquely.

Let $\mathcal{N}_J = \{\eta : J(\eta) = 0\}$ be the null space of $J(\eta)$ and consider the tensor sum decomposition $\mathcal{H} = \mathcal{N}_J \oplus \mathcal{H}_J$. The space $\mathcal{H}_J$ is an RKHS with $J(\eta)$ as the square norm. For example, for $x \in [0,1]$, setting $J(\eta) = \int_0^1 \ddot{\eta}^2 dx$ yields the popular cubic splines.

For $\mathcal{X}$ a product domain, certain ANOVA decompositions can be built in using the tensor product spline technique. The decomposition can be characterized by $\mathcal{H} = \oplus_{\beta=0}^{q}\mathcal{H}_\beta$ and $J(\eta) = \sum_{\beta=1}^{q}\theta_\beta^{-1}J_\beta(\eta_\beta)$, where $\eta_\beta \in \mathcal{H}_\beta$, $0 < \theta_\beta < \infty$, and J_β is the square norm in $\mathcal{H}_\beta$, $\beta > 0$. One has $\mathcal{N}_J = \mathcal{H}_0$, $\mathcal{H}_J = \oplus_{\beta=1}^{q}\mathcal{H}_\beta$, and $R_J = \sum_{\beta=1}^{q}\theta_\beta R_\beta$, where R_β is the RK in $\mathcal{H}_\beta$. The θ_β's are an extra set of smoothing parameters to be selected, but they may not appear explicitly in the notation.

It is well known that the minimizer of (1.1) in $\mathcal{H}$ resides in the space $\mathcal{N}_J \oplus \text{span}\{R_J(x_i,\cdot), i = 1,\ldots,n\}$.

3 Asymptotic Convergence Rate

Let $f(x)$ be the limiting density of x_i on the domain $\mathcal{X}$, assumed to be bounded from above and below, and write $v_\eta(x) = E[w(\eta(x);Y)]$. Define the bilinear form $\tilde{V}(g,h) = \int_{\mathcal{X}} g(x)h(x)v_{\eta_0}(x)f(x)dx$. The asymptotic convergence rate of η_λ is characterized by an eigenvalue analysis of $J(\eta)$ with respect to $\tilde{V}(\eta) = \tilde{V}(\eta,\eta)$.

Let ψ_ν be the eigenfunctions satisfying $\tilde{V}(\psi_\nu,\psi_\mu) = \delta_{\nu,\mu}$ and $J(\psi_\nu,\psi_\mu) = \rho_\nu\delta_{\nu,\mu}$, where $J(g,h)$ is the bilinear form associated with $J(g)$ and $\delta_{\nu,\mu}$ is Kronecker's delta. Assume $\rho_\nu > C\nu^r$ for some $C > 0$, $r > 1$, and ν sufficiently large. Assuming $\sum_\nu \rho_\nu^p\eta_\nu^2 < \infty$, where $\eta_\nu = \tilde{V}(\psi_\nu,\eta)$ and $p \in [1,2]$, it can be shown that

as $\lambda \to 0$ and $n\lambda^{2/r} \to \infty$,

$$\tilde{V}(\eta_\lambda - \eta) = O_p(\lambda^p + n^{-1}\lambda^{-1/r}); \tag{3.1}$$

note that $J(\eta) = \sum_\nu \rho_\nu \eta_\nu^2$. If $\mathcal{H}^* \subset \mathcal{H}$ satisfies $\tilde{V}(\eta) = O_p(\lambda J(\eta))$, $\forall h \in \mathcal{H} \ominus \mathcal{H}^*$, then the rate of (3.1) also holds for the minimizer of (1.1) in $\mathcal{H}^*$. The optimal convergence rate is $O_p(n^{-pr/(pr+1)})$, achieved with $\lambda \asymp n^{-r/(pr+1)}$. For $\mathcal{H}_q = \mathcal{N}_J \oplus \text{span}\{R_J(z_j, \cdot), j = 1, \ldots, q\}$, where z_j have the limiting density $f(x)$, it can be shown that

$$\tilde{V}(h) = (\tilde{V} + \lambda J)(h) O_p(q^{-1/2}\lambda^{-1/r});$$

random subsets $\{z_j\} \subset \{x_i\}$ have the limiting density $f(x)$. Setting $q \asymp \lambda^{-2/r-\tilde{\epsilon}} \asymp n^{2/(pr+1)+\epsilon}$, where $\lambda \asymp n^{-r/(pr+1)}$ is optimal and $\epsilon > 0$ is arbitrary, one has $\tilde{V}(\eta) = o_p(\lambda J(\eta))$, $\forall h \in \mathcal{H} \ominus \mathcal{H}_q$. See Gu and Kim (2002). The above results hold for the cubic splines with $r = 4$, and for the tensor product cubic splines with $r < 4$.

4 Computation

A exponential family distribution has a density of the form $\exp\{(y\theta - b(\theta))/a(\phi) + c(y, \phi)\}$, where θ is the canonical parameter and $a(\phi)$ is the dispersion parameter. Let η be a monotone transform of θ that takes values on $(-\infty, \infty)$; the unrestricted range of η avoids the complication of constrained minimization. The relation between η and $E[Y]$ is known as the link. One may take $l(\eta; Y) = -\{Y\theta(\eta) - b(\theta(\eta))\}$; the term $c(Y, \phi)$ is dropped as it does not depend on η, and the dispersion $a(\phi)$ can be absorbed into λ. Write $u(\eta; y) = dl/d\eta$ and $w(\eta; y) = d^2l/d\eta^2$. It can be shown that $E[u(\eta_0; Y)] = 0$ and $E[u^2(\eta_0; Y)] = \sigma^2 E[w(\eta_0; Y)]$, where η_0 is the true parameter and σ^2 is a constant; see Gu and Kim (2002).

Fixing the smoothing parameter λ (and the θ_β's hidden in $J(\eta)$, if present), (1.1) may be minimized via the Newton iteration. Write $\tilde{u}_i = u(\tilde{\eta}(x_i); Y_i)$ and $\tilde{w}_i = w(\tilde{\eta}(x_i); Y_i)$. The quadratic approximation of $l(\eta(x_i); Y_i)$ at $\tilde{\eta}(x_i)$ is seen to be

$$l(\tilde{\eta}(x_i); Y_i) + \tilde{u}_i\{\eta(x_i) - \tilde{\eta}(x_i)\} + \frac{1}{2}\tilde{w}_i\{\eta(x_i) - \tilde{\eta}(x_i)\}^2 = \frac{1}{2}\tilde{w}_i\{\eta(x_i) - \tilde{\eta}(x_i) + \frac{\tilde{u}_i}{\tilde{w}_i}\}^2 + C_i,$$

where C_i is independent of $\eta(x_i)$. The Newton iteration thus updates $\tilde{\eta}$ by the minimizer $\eta_{\lambda,\tilde{\eta}}$ of the penalized weighted least squares

$$\frac{1}{n}\sum_{i=1}^{n} \tilde{w}_i(\tilde{Y}_i - \eta(x_i))^2 + \lambda J(\eta) \tag{4.1}$$

where $\tilde{Y}_i = \tilde{\eta}(x_i) - \tilde{u}_i/\tilde{w}_i$. When $w(\tilde{\eta}(x_i); Y_i)$ is not assured to be positive, one may set $\tilde{w}_i = E[w(\tilde{\eta}(x_i); Y_i)]$. On convergence, one has $\eta_\lambda = \eta_{\lambda,\eta_\lambda}$.

The minimizer η_λ of (1.1) in $\mathcal{H}$ or $\mathcal{H}_q$ can be written as

$$\eta(x) = \sum_{\nu=1}^{m} d_\nu \phi_\nu(x) + \sum_{j=1}^{q} c_j R(z_j, x) = \boldsymbol{\phi}^T \mathbf{d} + \boldsymbol{\xi}^T \mathbf{c},$$

where $\{\phi_\nu\}$ is a basis of $\mathcal{N}_J$ and R_J is the RK in $\mathcal{H}_J$. $\boldsymbol{\phi}$ and $\boldsymbol{\xi}$ are vectors of functions and $\mathbf{d}$ and $\mathbf{c}$ are vectors of coefficients; $q = n$ for the exact solution in $\mathcal{H}$. Plugging this into (4.1), one has

$$(\tilde{\mathbf{Y}} - S\mathbf{d} - R\mathbf{c})^T \tilde{W}(\tilde{\mathbf{Y}} - S\mathbf{d} - R\mathbf{c}) + n\lambda \mathbf{c}^T Q \mathbf{c}, \tag{4.2}$$

as $\lambda \to 0$ and $n\lambda^{2/r} \to \infty$,

$$\tilde{V}(\eta_\lambda - \eta) = O_p(\lambda^p + n^{-1}\lambda^{-1/r}); \tag{3.1}$$

note that $J(\eta) = \sum_\nu \rho_\nu \eta_\nu^2$. If $\mathcal{H}^* \subset \mathcal{H}$ satisfies $\tilde{V}(\eta) = O_p(\lambda J(\eta))$, $\forall h \in \mathcal{H} \ominus \mathcal{H}^*$, then the rate of (3.1) also holds for the minimizer of (1.1) in $\mathcal{H}^*$. The optimal convergence rate is $O_p(n^{-pr/(pr+1)})$, achieved with $\lambda \asymp n^{-r/(pr+1)}$. For $\mathcal{H}_q = \mathcal{N}_J \oplus \text{span}\{R_J(z_j, \cdot), j = 1, \ldots, q\}$, where z_j have the limiting density $f(x)$, it can be shown that

$$\tilde{V}(h) = (\tilde{V} + \lambda J)(h) O_p(q^{-1/2}\lambda^{-1/r});$$

random subsets $\{z_j\} \subset \{x_i\}$ have the limiting density $f(x)$. Setting $q \asymp \lambda^{-2/r-\tilde{\epsilon}} \asymp n^{2/(pr+1)+\epsilon}$, where $\lambda \asymp n^{-r/(pr+1)}$ is optimal and $\epsilon > 0$ is arbitrary, one has $\tilde{V}(\eta) = o_p(\lambda J(\eta))$, $\forall h \in \mathcal{H} \ominus \mathcal{H}_q$. See Gu and Kim (2002). The above results hold for the cubic splines with $r = 4$, and for the tensor product cubic splines with $r < 4$.

4 Computation

A exponential family distribution has a density of the form $\exp\{(y\theta - b(\theta))/a(\phi) + c(y, \phi)\}$, where θ is the canonical parameter and $a(\phi)$ is the dispersion parameter. Let η be a monotone transform of θ that takes values on $(-\infty, \infty)$; the unrestricted range of η avoids the complication of constrained minimization. The relation between η and $E[Y]$ is known as the link. One may take $l(\eta; Y) = -\{Y\theta(\eta) - b(\theta(\eta))\}$; the term $c(Y, \phi)$ is dropped as it does not depend on η, and the dispersion $a(\phi)$ can be absorbed into λ. Write $u(\eta; y) = dl/d\eta$ and $w(\eta; y) = d^2l/d\eta^2$. It can be shown that $E[u(\eta_0; Y)] = 0$ and $E[u^2(\eta_0; Y)] = \sigma^2 E[w(\eta_0; Y)]$, where η_0 is the true parameter and σ^2 is a constant; see Gu and Kim (2002).

Fixing the smoothing parameter λ (and the θ_β's hidden in $J(\eta)$, if present), (1.1) may be minimized via the Newton iteration. Write $\tilde{u}_i = u(\tilde{\eta}(x_i); Y_i)$ and $\tilde{w}_i = w(\tilde{\eta}(x_i); Y_i)$. The quadratic approximation of $l(\eta(x_i); Y_i)$ at $\tilde{\eta}(x_i)$ is seen to be

$$l(\tilde{\eta}(x_i); Y_i) + \tilde{u}_i\{\eta(x_i) - \tilde{\eta}(x_i)\} + \frac{1}{2}\tilde{w}_i\{\eta(x_i) - \tilde{\eta}(x_i)\}^2 = \frac{1}{2}\tilde{w}_i\{\eta(x_i) - \tilde{\eta}(x_i) + \frac{\tilde{u}_i}{\tilde{w}_i}\}^2 + C_i,$$

where C_i is independent of $\eta(x_i)$. The Newton iteration thus updates $\tilde{\eta}$ by the minimizer $\eta_{\lambda,\tilde{\eta}}$ of the penalized weighted least squares

$$\frac{1}{n}\sum_{i=1}^{n} \tilde{w}_i(\tilde{Y}_i - \eta(x_i))^2 + \lambda J(\eta) \tag{4.1}$$

where $\tilde{Y}_i = \tilde{\eta}(x_i) - \tilde{u}_i/\tilde{w}_i$. When $w(\tilde{\eta}(x_i); Y_i)$ is not assured to be positive, one may set $\tilde{w}_i = E[w(\tilde{\eta}(x_i); Y_i)]$. On convergence, one has $\eta_\lambda = \eta_{\lambda,\eta_\lambda}$.

The minimizer η_λ of (1.1) in $\mathcal{H}$ or $\mathcal{H}_q$ can be written as

$$\eta(x) = \sum_{\nu=1}^{m} d_\nu \phi_\nu(x) + \sum_{j=1}^{q} c_j R(z_j, x) = \boldsymbol{\phi}^T \mathbf{d} + \boldsymbol{\xi}^T \mathbf{c},$$

where $\{\phi_\nu\}$ is a basis of $\mathcal{N}_J$ and R_J is the RK in $\mathcal{H}_J$. $\boldsymbol{\phi}$ and $\boldsymbol{\xi}$ are vectors of functions and $\mathbf{d}$ and $\mathbf{c}$ are vectors of coefficients; $q = n$ for the exact solution in $\mathcal{H}$. Plugging this into (4.1), one has

$$(\tilde{\mathbf{Y}} - S\mathbf{d} - R\mathbf{c})^T \tilde{W}(\tilde{\mathbf{Y}} - S\mathbf{d} - R\mathbf{c}) + n\lambda \mathbf{c}^T Q \mathbf{c}, \tag{4.2}$$

with $V_{\xi,\phi}$ a $q \times m$ matrix having the (j,ν)th entry

$$\frac{1}{N}\sum_{i=1}^{n} e^{\bar{\eta}(x_i)}\xi_j(x_i)\phi_\nu(x_i) - \frac{1}{N}\sum_{i=1}^{n} e^{\bar{\eta}(x_i)}\xi_j(x_i)\frac{1}{N}\sum_{i=1}^{n} e^{\bar{\eta}(x_i)}\phi_\nu(x_i)$$

and other V matrices similarly defined; remember that $\sum_{i=1}^{n} e^{\bar{\eta}(x_i)} = N$. When chosen properly, the fudge factor α proves to be quite successful in preventing severe undersmoothing while preserving the overall effectiveness of cross-validation; a good default value is $\alpha = 1.4$.

For the Gamma regression with $Y_i \sim \text{Gamma}(\nu, \beta(x_i))$, a slight modification of (5.1) with the fudge factor α is considered.

We evaluate the empirical performances of (5.1) and of variations thereof for various distribution families.

References

Cox, D. D. and F. O'Sullivan (1990). Asymptotic analysis of penalized likelihood and related estimators. *Ann. Statist. 18*, 124–145.

Green, P. J. and B. Yandell (1985). Semi-parametric generalized linear models. In R. Gilchrist, B. Francis, and J. Whittaker (Eds.), *Proceedings of the GLIM85 Conference*, pp. 44–55. Berlin: Springer-Verlag.

Gu, C. (1990). Adaptive spline smoothing in non Gaussian regression models. *J. Amer. Statist. Assoc. 85*, 801–807.

Gu, C. (1992). Cross-validating non-Gaussian data. *J. Comput. Graph. Statist. 1*, 169–179.

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

Gu, C. and Y.-J. Kim (2002). Penalized likelihood regression: General formulation and efficient approximation. *Can. J. Statist. 30*, 619–628.

Gu, C. and C. Qiu (1994). Penalized likelihood regression: A simple asymptotic analysis. *Statist. Sin. 4*, 297–304.

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

Gu, C. and D. Xiang (2001). Cross-validating non-Gaussian data: Generalized approximate cross-validation revisited. *J. Comput. Graph. Statist. 10*, 581–591.

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.

O'Sullivan, F., B. Yandell, and W. Raynor (1986). Automatic smoothing of regression functions in generalized linear models. *J. Amer. Statist. Assoc. 81*, 96–103.

Silverman, B. W. (1978). Density ratios, empirical likelihood and cot death. *Appl. Statist. 27*, 26–33.

Wahba, G. (1990). *Spline Models for Observational Data*, Volume 59 of CBMS-NSF Regional Conference Series in Applied Mathematics. Philadelphia: SIAM.

Xiang, D. and G. Wahba (1996). A generalized approximate cross validation for smoothing splines with non-Gaussian data. *Statist. Sin. 6*, 675–692.