Statistica Sinica 4(1994), 297-304

PENALIZED LIKELIHOOD REGRESSION: A SIMPLE ASYMPTOTIC ANALYSIS

Chong Gu and Chunfu Qiu

Purdue University and University of Illinois

Abstract: We conduct a simple asymptotic analysis of penalized likelihood regression for the analysis of data from exponential families. The convergence rates in terms of the integrated symmetrized Kullback-Leibler distance and a related mean square error are obtained.

Key words and phrases: Penalized likelihood, rate of convergence, smoothness, symmetrized Kullback-Leibler distance.

1. Introduction

Consider independent observations (x_i, y_i), $i = 1, \ldots, n$, where $y|x$ follows an exponential family distribution with density $\exp\{(y\eta(x) - b(\eta(x)))/a(\phi) + c(y, \phi)\}$ and x has a density $f(x) > 0$ on a generic domain $\mathcal{X}$. The $a(\phi)$, possibly known or otherwise considered as a nuisance parameter, is assumed common to all the observations, finite and positive. Of interest is the estimation of the function $\eta(x)$. The penalized likelihood method estimates $\eta(x)$ by the minimizer of the functional

$$-\frac{1}{n}\sum_{i=1}^{n}\Big\{y_i\eta(x_i) - b(\eta(x_i))\Big\} + (\lambda/2)J(\eta), \tag{1}$$

in a Hilbert space $\mathcal{H}$ in which $J(\eta)$ is a square (semi) norm. Evaluation $[x](\cdot) = (\cdot)(x)$ is assumed to be continuous so that the first term in (1) is continuous in η, and the null space of $J(\eta)$ should have dimension less than n so that the unconstrained model space dimension does not exceed the number of data. The first term in (1) is proportional to the minus log likelihood, which seeks a good fit of η to the data, the second term penalizes the roughness of the fit measured by $J(\eta)$, and the smoothing parameter λ controls the tradeoff. The minimizer of (1) can be shown to exist whenever the maximum likelihood estimate of η uniquely exists in the null space of J (cf. Gu and Qiu (1993, Theorem 4.1)).

A recent review of penalty smoothing, or smoothing splines, can be found in Wahba (1990). The specific formulation (1) for the analysis of data from exponential families is proposed and studied by O'Sullivan, Yandell and Raynor

(1986). See also Silverman (1978) and Green and Yandell (1985). A generic algorithm with automatic smoothing parameter selection is proposed by Gu (1990). Approximate Bayesian confidence intervals are illustrated in Gu (1992a). More discussion concerning the empirical choices of λ can be found in Gu (1992b). An asymptotic analysis of penalized likelihood estimation, of which (1) is a special case, is carried out by Cox and O'Sullivan (1990).

By standard exponential family theory (cf. McCullagh and Nelder (1989, §2.2.2)), $E(y|x) = \dot{b}(\eta(x)) = \mu(x)$ and $\mathrm{var}(y|x) = \ddot{b}(\eta(x))a(\phi) = v(x)a(\phi)$. We shall denote the true functions η, μ and v by a subscript 0, and the estimates by a hat ($\hat{}$) on top. The symmetrized Kullback-Leibler distance between two probability densities f and g is defined by $E_f \log(f/g) + E_g \log(g/f)$, which is always positive for $f \neq g$. When $a(\phi)$ is known, it is easy to verify that the symmetrized Kullback-Leibler distance between the true conditional distribution and the estimate at x, parameterized by $\eta_0(x)$ and $\hat{\eta}(x)$, is $\{(\hat{\eta} - \eta_0)(x)(\hat{\mu} - \mu_0)(x)\}/a(\phi)$. The weighted average

$$\int_{\mathcal{X}} (\hat{\eta} - \eta_0)(\hat{\mu} - \mu_0) f/a(\phi)\, dx \tag{2}$$

defines a natural measure for the precision of the estimation of η_0 by $\hat{\eta}$, where the weight function $f(x)$ is the proportion of data allocated to the neighborhood of x. When $a(\phi)$ is unknown, (2) is the average symmetrized Kullback-Leibler distance between the distributions parameterized by $\{\hat{\eta}(x), a(\phi)\}$ and $\{\eta_0(x), a(\phi)\}$. Since $a(\phi)$ is a nuisance parameter, (2) remains a reasonable measure for the discrepancy between $\hat{\eta}$ and η_0. Note that $(\hat{\eta} - \eta_0)(\hat{\mu} - \mu_0)$ is approximately equal to $(\hat{\mu} - \mu_0)^2 v_0^{-1}$, the mean square error in the mean space of y adjusted by its variance, and that this approximation is exact for a Gaussian likelihood. For notational simplicity, we shall set $a(\phi) = 1$ in (2) and elsewhere, and this will not impair the generality of the convergence rate calculation.

The purpose of this article is to conduct a simpler asymptotic analysis of (1). In contrast to the super generic theory of Cox and O'Sullivan (1990) which provides unified convergence rates in a class of functional space norms under several different stochastic structures, we concentrate on the specific distance (2) derived naturally from the specific stochastic structure in regression. To compensate for the loss of generality, we are able to trim the lengthy intermediate analyses and less comprehensible regularity conditions, hopefully making the structure of the problem transparent and the theory accessible to a broader audience. The approach we follow parallels that in Gu and Qiu (1993) in an analysis of penalized likelihood density estimation, and has its root in Silverman's (1982) earlier work.

As a running example in our analysis we consider cubic spline logistic regression on a domain $\mathcal{X} = [0, 1]$. Binary responses y_i are observed with co-

variates x_i, where $y|x$ is Bernoulli with $P(y = 1|x) = \mu(x) = e^\eta/(e^\eta + 1)$, $\eta = \log\{\mu/(1-\mu)\}$, $v = \mu(1-\mu) = e^\eta/(e^\eta+1)^2$, and $a(\phi) = 1$. $J(\eta) = \int_0^1 \ddot{\eta}^2\, dx$ and $\mathcal{H} = \{\eta : J(\eta) < \infty\}$. The null space of J is the space of linear polynomials, of dimension 2. It can be shown that evaluation is continuous in $\mathcal{H}$ (cf. Wahba (1990)). A penalized likelihood estimate exists whenever the maximum likelihood estimate of the linear logistic model exists.

2. Asymptotic Analysis

2.1. Smoothness assumptions

The quadratic form $V(\eta) = \int_{\mathcal{X}} \eta^2 v_0 f\, dx$ defines a normed distance $V(\hat{\eta} - \eta_0)$ which approximates (2), noting that $\dot{\mu}(\eta) = v$. $V(\eta)$ is an ordinary quadratic norm, and the smoothness defined by J will be characterized by an eigenvalue analysis of J with respect to V. In what follows we shall use $V(\cdot,\cdot)$ and $J(\cdot,\cdot)$ to indicate the (semi) inner products associated with the square (semi) norms V and J.

A bilinear form B is said to be completely continuous with respect to another bilinear form A, if for any $\epsilon > 0$, there exist finite number of linear functionals $l_1, \ldots, l_k$ such that $l_j(\eta) = 0$, $j = 1, \ldots, k$, implies that $B(\eta) \leq \epsilon A(\eta)$; see Weinberger (1974, §3.3).

Assumption A.1. V is completely continuous with respect to J.

Under A.1, using Theorem 3.1 of Weinberger (1974, p.52), it can be shown that there exist $\phi_\nu = \mathcal{H}$ and $0 \leq \rho_\nu \uparrow \infty$, $\nu = 1, 2, \ldots$, such that $V(\phi_\nu, \phi_\mu) = \delta_{\nu,\mu}$ and $J(\phi_\nu, \phi_\mu) = \rho_\nu \delta_{\nu,\mu}$, where $\delta_{\nu,\mu}$ is the Kronecker delta; see, e.g., Gu and Qiu (1993, §5). The notion of smoothness is characterized by the rate of growth of ρ_ν.

Assumption A.2. $\rho_\nu = c_\nu \nu^r$, where $r > 1$, $c_\nu \in (\beta_1, \beta_2)$, and $0 < \beta_1 < \beta_2 < \infty$.

Intuitively, A.1 implies that λJ in (1) restricts the estimate to an effectively finite dimensional space in terms of the V norm, and A.2 specifies the rate at which the effective model space dimension is expanded as $\lambda \to 0$.

For cubic spline logistic regression, A.1 and A.2 are satisfied when V is equivalent to the L_2 norm $\int_0^1 \eta^2 dx$, and $r = 4$ in A.2 (cf. Utreras (1981), Silverman (1982)).

2.2. Linear approximation

Assume $\eta_0 \in \mathcal{H}$. Let η_1 be the minimizer of the quadratic functional

$$-\frac{1}{n}\sum_{i=1}^{n}\Big\{y_i\eta(x_i) - \mu_0(x_i)\eta(x_i)\Big\} + (1/2)V(\eta - \eta_0) + (\lambda/2)J(\eta). \qquad (3)$$

Write $\eta = \sum_\nu \eta_\nu \phi_\nu$ and $\eta_0 = \sum_\nu \eta_{\nu,0}\phi_\nu$, where $\eta_\nu = V(\eta, \phi_\nu)$ are the Fourier coefficients of η with basis ϕ_ν. Substituting these into (3) and solving for η_1, one obtains $\eta_{\nu,1} = (\beta_\nu + \eta_{\nu,0})/(1 + \lambda\rho_\nu)$, where $\beta_\nu = (1/n)\sum_{i=1}^n (y_i - \mu_0(x_i))\phi_\nu(x_i)$. It is easy to verify that $E\beta_\nu = 0$ and $E\beta_\nu^2 = n^{-1}$. The following theorem can then be proved parallel to Theorem 5.1 of Gu and Qiu (1993); see also Silverman (1982, §6).

Theorem 1. *Under* A.1 *and* A.2, *as* $n \to \infty$ *and* $\lambda \to 0$, $V(\eta_1 - \eta_0) = O_p(n^{-1}\lambda^{-1/r} + \lambda)$ *and* $\lambda J(\eta_1 - \eta_0) = O_p(n^{-1}\lambda^{-1/r} + \lambda)$.

2.3. Approximation error and main result

We need two more assumptions in further analysis.

Assumption A.3. For η in a convex set B_0 around η_0 containing $\hat{\eta}$ and η_1, $\exists\, c_1, c_2 \in (0, \infty)$ such that $c_1 v_0(x) \le v(x) \le c_2 v_0(x)$ uniformly on $\mathcal{X}$.

Since $(\hat{\eta} - \eta_1)(\hat{\mu} - \mu_1) = (\hat{\eta} - \eta_1)^2 v_{\alpha\hat{\eta}+(1-\alpha)\eta_1}$ where $\alpha \in [0, 1]$, A.3 leads to the equivalence of the V distance and the symmetrized Kullback-Leibler distance in B_0. It is also worth noting that A.3 is trivial in penalized least squares regression for Gaussian data where $v \equiv 1$.

Assumption A.4. $\exists\, c_3 < \infty$ such that $\int_{\mathcal{X}} \phi_\nu^2 \phi_\mu^2 v_0^2 f\, dx \le c_3, \quad \forall \nu, \mu$.

Note that $\int_{\mathcal{X}} \phi_\nu^2 v_0 f\, dx = 1$. A.4 will follow when $(\phi_\nu v_0^{1/2})(x)$ have bounded kurtosis under the density f, especially when $\phi_\nu v_0^{1/2}$ are uniformly bounded on $\mathcal{X}$.

Theorem 2. *Under* A.1–A.4, *as* $\lambda \to 0$ *and* $n\lambda^{2/r} \to \infty$, $V(\hat{\eta} - \eta_1) = O_p(n^{-1}\lambda^{-1/r} + \lambda)$.

The proof of this theorem is given in §2.4.

Theorem 3. *Under* A.1–A.4, *as* $\lambda \to 0$ *and* $n\lambda^{2/r} \to \infty$, $V(\hat{\eta} - \eta_0) = O_p(n^{-1}\lambda^{-1/r} + \lambda)$ *and* $\int_{\mathcal{X}} (\hat{\mu} - \mu_0)(\hat{\eta} - \eta_0) f\, dx = O_p(n^{-1}\lambda^{-1/r} + \lambda)$.

Proof. Use Theorems 1, 2, and Assumption A.3.

For cubic spline logistic regression, A.3 is satisfied when $\mu(x)$ is uniformly bounded away from 0 and 1 on $x \in [0, 1]$ for members of B_0. Direct verification of A.4 is rather inconvenient if not impossible, since explicit formulas for ϕ_ν are in general not available. A suggestive special case does exist, however, when $v_0 f \propto 1$ and when $\mathcal{H}$ is reduced to the periodic restriction of $\{\eta : J(\eta) < \infty\}$, which has $\sin(2\pi\mu x)$ and $\cos(2\pi\mu x)$ as the basis and hence satisfies A.4 when v_0 is bounded.

2.4. Proof of Theorem 2

Write (1) as $L(\eta)+(\lambda/2)J(\eta)$ and define $A_{\eta,h}(\alpha) = L(\eta+\alpha h)+(\lambda/2)J(\eta+\alpha h)$. It can be shown that

$$\dot{A}_{\eta,h}(0) = -\frac{1}{n}\sum_{i=1}^{n}\Big\{y_i h(x_i) - \mu(x_i)h(x_i)\Big\} + \lambda J(\eta, h).$$

Setting $\eta = \hat{\eta}$ and $h = \hat{\eta} - \eta_1$, one obtains

$$\begin{aligned} 0 &= \dot{A}_{\hat{\eta},\hat{\eta}-\eta_1}(0) \\ &= -\frac{1}{n}\sum_{i=1}^{n}\Big\{y_i(\hat{\eta}-\eta_1)(x_i) - \hat{\mu}(x_i)(\hat{\eta}-\eta_1)(x_i)\Big\} + \lambda J(\hat{\eta}, \hat{\eta}-\eta_1). \end{aligned} \tag{4}$$

Similarly, denote (3) by $L_1(\eta) + (\lambda/2)J(\eta)$ and define $B_{\eta,h}(\alpha) = L_1(\eta+\alpha h) + (\lambda/2)J(\eta+\alpha h)$. It follows that

$$\dot{B}_{\eta,h}(0) = -\frac{1}{n}\sum_{i=1}^{n}\Big\{y_i h(x_i) - \mu_0(x_i)h(x_i)\Big\} + V(\eta-\eta_0, h) + \lambda J(\eta, h).$$

Hence,

$$\begin{aligned} 0 &= \dot{B}_{\eta_1,\hat{\eta}-\eta_1}(0) \\ &= -\frac{1}{n}\sum_{i=1}^{n}\Big\{y_i(\hat{\eta}-\eta_1)(x_i) - \mu_0(x_i)(\hat{\eta}-\eta_1)(x_i)\Big\} \\ &\quad + V(\eta_1-\eta_0, \hat{\eta}-\eta_1) + \lambda J(\eta_1, \hat{\eta}-\eta_1). \end{aligned} \tag{5}$$

On equating (4) and (5), and after some algebra, one obtains

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^{n}(\hat{\mu}-\mu_1)(x_i)(\hat{\eta}-\eta_1)(x_i) + \lambda J(\hat{\eta}-\eta_1) \\ &= V(\eta_1-\eta_0, \hat{\eta}-\eta_1) - \frac{1}{n}\sum_{i=1}^{n}(\mu_1-\mu_0)(x_i)(\hat{\eta}-\eta_1)(x_i). \end{aligned} \tag{6}$$

By A.3,

$$c_1\frac{1}{n}\sum_{i=1}^{n}(\hat{\eta}-\eta_1)^2(x_i)v_0(x_i) \le \frac{1}{n}\sum_{i=1}^{n}(\hat{\mu}-\mu_1)(x_i)(\hat{\eta}-\eta_1)(x_i). \tag{7}$$

From the Fourier expansion of $\hat{\eta}-\eta_1$,

$$\left|\frac{1}{n}\sum_{i=1}^{n}(\hat{\eta}-\eta_1)^2(x_i)v_0(x_i) - V(\hat{\eta}-\eta_1)\right|$$

$$
\begin{aligned}
&= \left| \sum_\nu \sum_\mu (\hat{\eta}_\nu - \eta_{\nu,1})(\hat{\eta}_\mu - \eta_{\mu,1}) \left\{ \frac{1}{n} \sum_{i=1}^n \phi_\nu(x_i)\phi_\mu(x_i)v_0(x_i) - \int \phi_\nu \phi_\mu v_0 f dx \right\} \right| \\
&\leq \left[\sum_\nu \sum_\mu (1+\lambda\rho_\nu)^{-1}(1+\lambda\rho_\mu)^{-1} \left\{ \frac{1}{n} \sum_{i=1}^n \phi_\nu(x_i)\phi_\mu(x_i)v_0(x_i) \right.\right. \\
&\quad \left.\left. - \int \phi_\nu \phi_\mu v_0 f dx \right\}^2 \right]^{1/2} \left[\sum_\nu \sum_\mu (1+\lambda\rho_\nu)(1+\lambda\rho_\mu)(\hat{\eta}_\nu - \eta_{\nu,1})^2(\hat{\eta}_\mu - \eta_{\mu,1})^2 \right]^{1/2} \\
&= O_p(n^{-1/2}\lambda^{-1/r})(V+\lambda J)(\hat{\eta}-\eta_1) \\
&= o_p(1)(V+\lambda J)(\hat{\eta}-\eta_1), \qquad (8)
\end{aligned}
$$

where the Cauchy-Schwarz inequality, A.4, and the fact that $\sum_\nu (1+\lambda\nu^r)^{-1} = O(\lambda^{-1/r})$ (cf. Gu and Qiu (1993, Lemma 5.2)) are used. Combining (7) and (8), a lower bound for the left-hand-side of (6) is given by

$$(c_1 V + \lambda J)(\hat{\eta} - \eta_1)(1 + o_p(1)). \qquad (9)$$

On the right hand side of (6), A.3 leads to

$$\frac{1}{n}\sum_{i=1}^n (\mu_1 - \mu_0)(x_i)(\hat{\eta}-\eta_1)(x_i) = c\frac{1}{n}\sum_{i=1}^n (\eta_1-\eta_0)(x_i)(\hat{\eta}-\eta_1)(x_i)v_0(x_i), \qquad (10)$$

where $c \in [c_1, c_2]$. Similar to (8), it can be shown that

$$
\begin{aligned}
&\left| \frac{1}{n}\sum_{i=1}^n (\eta_1-\eta_0)(x_i)(\hat{\eta}-\eta_1)(x_i)v_0(x_i) - V(\eta_1-\eta_0, \hat{\eta}-\eta_1) \right| \\
&= o_p(1)(V+\lambda J)^{1/2}(\eta_1-\eta_0)(V+\lambda J)^{1/2}(\hat{\eta}-\eta_1). \qquad (11)
\end{aligned}
$$

Combining (10) and (11), an upper bound is given by

$$|1-c|V^{1/2}(\eta_1-\eta_0)V^{1/2}(\hat{\eta}-\eta_1) + o_p(1)(V+\lambda J)^{1/2}(\eta_1-\eta_0)(V+\lambda J)^{1/2}(\hat{\eta}-\eta_1). \qquad (12)$$

Joining (9) and (12) and applying Theorem 1 yield the result of the theorem.

3. Discussion and Concluding Remarks

The essence penalty smoothing, bias-variance tradeoff through controlling the effective model space dimension, is reflected in the smoothness assumptions A.1 and A.2. These assumptions are intrinsic to practical penalty smoothing and can usually be verified when V is equivalent to $\int_{\mathcal{X}} \eta^2 dx$ since the latter is usually completely continous with respect to J. The Fourier analysis based on A.1 and A.2 is the key to our analysis. Messy technicalities are effectively obviated by the

regularity conditions A.3 and A.4, which, in general, may not be verifiable from more primitive conditions. These assumptions, however, appear highly plausible. When v_0 is uniformly bounded from below and above, A.3 will fail to hold only when η_1 and $\hat{\eta}$ systematically move away from η_0 as $n \to \infty$, so it appears mild. Although the explicit forms of ϕ_ν are, in general, not available, we do know that $V(\phi_\nu) = 1$ and that ϕ_ν represent more wiggliness or higher frequencies as $\nu \to \infty$; thus the magnitude of ϕ_ν is unlikely to grow indefinitely, and hence A.4 looks reasonable.

In the foregoing development, the smoothness assumptions are made on the canonical parameter η of the exponential family likelihood. Since smoothness assumptions are much less restrictive than parametric assumptions, the choice of modeling parameter, or link, as known in the generalized linear models literature, has much less impact on penalized likelihood regression than on parametric regression. The choice of the canonical parameter as modeling parameter has several advantages: First, there is, in general, no numerically awkward constraint on the possible values that η can take; second, (1) is guaranteed to be convex; third, a convenient and effective empirical choice of λ is available and theoretically justifiable (cf. Gu (1990, 1992b)); and fourth, a simple asymptotic analysis is possible as in this article. If circumstance demands a modeling parameter θ other than the canonical parameter η, however, the techniques used in this article may still be applicable in a similar analysis with a $V(\theta)$ defined by $\int_{\mathcal{X}} \theta^2 (d\eta/d\theta)_0^2 v_0 f dx$, but the conditions and proofs could become much messier.

Acknowledgements

Chong Gu's research was supported by NSF grants DMS-9101730 and DMS-9301511. Chunfu Qiu's research was supported by a David Ross grant at Purdue University. The authors thank a referee for comments which helped to improve the presentation.

References

Cox, D. D. and O'Sullivan, F. (1990). Asymptotic analysis of penalized likelihood and related estimators. *Ann. Statist.* **18**, 1676–1695.

Green, P. and Yandell, B. (1985). Semi-parametric generalized linear models. In *GLIM85: Proceedings of the International Conference on Generalized Linear Models. Lecture Notes in Statist.* **32** (Edited by R. Gilchrist), 44–55, Springer-Verlag, New York.

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

Gu, C. (1992a). Penalized likelihood regression: A Bayesian analysis. *Statistica Sinica* **2**, 255–264.

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

Gu, C. and Qiu, C. (1993). Smoothing spline density estimation: Theory. *Ann. Statist.* **21**, 217–234.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd edition. Chapman and Hall, London.

O'Sullivan, F., Yandell, B. S. and Raynor, W. J. (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.

Silverman, B. W. (1982). On the estimation of a probability density function by the maximum penalized likelihood method. *Ann. Statist.* **10**, 795–810.

Utreras, F. D. (1981). Optimal smoothing of noisy data using spline functions. *SIAM J. Sci. Statist. Comput.* **2**, 349–362.

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

Weinberger, H. F. (1974). *Variational Methods for Eigenvalue Approximation.* CBMS-NSF Regional Conference Series in Applied Mathematics, Vol. 15. SIAM, Philadelphia.

Department of Statistics, Purdue University, West Lafayette, IN 47907, U.S.A.
Department of Statistics, University of Illinois, Champaign, IL 61820, U.S.A.

(Received June 1992; accepted August 1993)