# ASYMPTOTIC ANALYSIS OF PENALIZED LIKELIHOOD REGRESSION

by

Chong Gu and Chunfu Qiu
Purdue University

Technical Report #91-50

Department of Statistics
Purdue University

September 1991

♡

# Asymptotic analysis of penalized likelihood regression

CHONG GU AND CHUNFU QIU

*Department of Statistics, Purdue University, West Lafayette, Indiana 47907, U.S.A.*

SUMMARY

We conduct an asymptotic analysis of the penalized likelihood regression for the analysis of data from exponential families. The asymptotic convergence rate in terms of the integrated symmetrized Kullback-Leibler is obtained.

*Some key words*: Symmetrized Kullback-Leibler; Penalized likelihood; Rate of convergence; Smoothness.

## 1. INTRODUCTION

Let $y$ be a random variable from an exponential family probability distribution with a log-likelihood $l(\eta, \phi|y)$, where $\eta$ is the parameter of interest and $\phi$ is a nuisance parameter. It is assumed that $\eta$ depends on a covariate $x$. A regression analysis seeks to estimate the dependence $\eta(x)$ based on the observed data $(x_i, y_i)$, $i = 1, \cdots, n$. The standard parametric approach models $\eta(x)$ in a low dimensional function space via a certain function form $\eta(x, \beta)$, where the function is known up to a parameter $\beta$. A generalized linear model results when there exists a monotone transformation of $\eta$, say $\theta(\eta)$, such that $\theta(\eta(x, \beta))$ is linear in $\beta$. Note that a parametric model implies rigid constraints on the form of the function $\eta(x)$. When knowledge is not sufficient to justify a parametric model, the constraints may introduce bias, and consequently lead to inaccurate conclusions. As an alternative, a nonparametric approach assumes less about the form of $\eta(x)$, and hence is more bias-robust, though at the expense of inferential efficiency were an adequate parametric model available.

The least assumption that most nonparametric methods make about $\eta(x)$, be it quantitative or qualitative, explicit or implicit, is that it is smooth. The penalized likelihood method quantifies the smoothness of functions and makes explicit use of smoothness in estimation. Specifically, the

method estimates $\eta$ by the minimizer of

$$-\sum_{i=1}^{n} l(\eta(x_i)|y_i) + \lambda J(\eta), \tag{1}$$

where the first term is the minus log-likelihood of the data with the nuisance parameter omitted, and the second term is a functional index of smoothness/roughness. The first term dictates a good fit to the data, the $J$ dictates a smooth estimate, and the smoothing parameter $\lambda$ controls the tradeoff.

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 & Raynor (1986). See also Green & Yandell (1985). A generic algorithm with automatic smoothing parameter selection is proposed by Gu (1990a). More discussion concerning the empirical choices of $\lambda$ can be found in Cox & Chang (1990) and Gu (1990b). An asymptotic analysis of penalized likelihood estimation, of which (1) is a special case, is conducted by Cox & O'Sullivan (1990).

In this article, we conduct a somewhat different asymptotic analysis of (1). The main difference between the Cox-O'Sullivan analysis and ours is in the form of the results: Cox & O'Sullivan (1990) work on certain functional space norms, which is a necessary choice for conceiving theorems applicable to both regression and density estimation, but is not necessarily the most natural choice for each of the individual problems; we work on the symmetrized Kullback-Leibler averaged over the covariate space, which is among the most natural scores for assessing the precision of the estimation of probability distributions. Besides, our approach is simpler. We do, however, draw heavily on the techniques used by Gu & Qiu (1991) in an asymptotic analysis of the penalized likelihood density estimation, where the target criterion is the symmetrized Kullback-Leibler for estimating a single probability distribution as appropriate in the context.

The remaining of the article is organized as follows. In Section 2, we formally formulate the model and discuss the smoothness assumptions which play a central role in the analysis. In Section 3, we outline the approach, present the convergence rate in $n$, $\lambda$, and the smoothness of the model space, and sketch the proofs. The analysis is conducted in a generic setup. An example of cubic spline logistic regression is presented as an illustration.

## 2. Penalized likelihood regression

### 2.1. *Formulation and preliminaries*

Consider independent observations $(x_i, y_i)$, $i = 1, \cdots, n$, where $y|x$ follows an exponential family distribution with density $\exp\{(y\eta(x) - b(\eta(x)))/a(\phi)\}$ 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, is assumed common to all the observations. Of interest is the estimation of the function $\eta(x)$. The penalized likelihood method estimates $\eta(x)$ by the minimizer of the functional

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

in a function space $\mathcal{H}$ in which $J(\eta)$ is defined and finite. Compared with (1), the $a(\phi)$ is absorbed into $\lambda$ in (2) and a divisor 2 of $\lambda$ is introduced for notational simplicity in later analysis.

By the standard exponential family theory (McCullagh & Nelder, 1989, §2.2.2),

$$\begin{aligned} E(y|x) &= \dot{b}(\eta(x)) = \mu(x) \\ var(y|x) &= \ddot{b}(\eta(x))a(\phi) = v(x)a(\phi). \end{aligned}$$

We shall denote the true functions $\eta$, $\mu$ and $v$ by a subscript 0, and the estimates by a hat on the top. The symmetrized Kullback-Leibler 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 know, it is easy to verify that the symmetrized Kullback-Leibler 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

$$\int_{\mathcal{X}} (\hat{\eta} - \eta_0)(\hat{\mu} - \mu_0) f/a(\phi) \tag{3}$$

defines an appropriate measure for the precision of the estimation of $\eta_0$ 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, (3) is the average symmetrized Kullback-Leibler between the distributions parameterized by $\{\hat{\eta}(x), a(\phi)\}$ and $\{\eta_0(x), a(\phi)\}$. Since $a(\phi)$ is a nuisance parameter, (3) remains an appropriate measure for the discrepancy between $\hat{\eta}$ and $\eta_0$. 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 $y_i$ adjusted by its variance, and that this approximation is exact for a Gaussian likelihood. For notational simplicity, we shall set $a(\phi) = 1$ in (3) and elsewhere in later analysis, and this will not impair the generality of the development.

We now specify further details about $J$ and $\mathcal{H}$ in (2). It is assumed that $\mathcal{H}$ is a Hilbert space in which $J$ is a square norm or a square seminorm with a finite dimensional null space, where a finite dimensional null space prevents interpolation. It is also assumed that evaluation $[x](\cdot) = (\cdot)(x)$ is continuous in $\mathcal{H}$, which ensures the continuity of (2) in its argument $\eta$. Under these assumptions, noting that $\ddot{b}(\eta) > 0$ so (2) is strictly convex in $\eta$, the minimizer of (2) exists whenever the maximum likelihood estimate exists in the null space of $J$; see Gu & Qiu (1991, Theorem 3.1).

As an example, let us consider the cubic spline logistic regression on a domain $\mathcal{X} = [0, 1]$. Binary responses $y_i$ are observed with covariates $x_i$, 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 $a(\phi) = 1$. $J(\eta) = \int_0^1 \ddot{\eta}^2$ and $\mathcal{H} = \{\eta : \ J(\eta) < \infty\}$. The null space of $J$ is the space of linear polynomials. It can be shown that evaluation is continuous in $\mathcal{H}$. The penalized likelihood estimate exists whenever the maximum likelihood estimate of the linear logistic model exists.

### 2.2. *Smoothness assumptions*

We now analyze the notion of smoothness defined by $J$. First define a distance to measure the deviation of $\hat{\eta}$ from $\eta_0$. Note that (3) is not a distance. Nevertheless, the quadratic form $V(\eta) = \int_{\mathcal{X}} \eta^2 v_0 f$ defines a distance $V(\hat{\eta} - \eta_0)$ which approximates (3), noting that $\dot{\mu}(\eta) = v$. $V(\eta)$ is an ordinary quadratic norm, and the smoothness defined by $J$ shall be characterized by an eigenvalue analysis of $J$ with respect to $V$.

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 $l_1, \cdots, l_k$ such that $l_j(\eta) = 0$, $j = 1, \cdots, k$, implies that $B(\eta) \leq \epsilon A(\eta)$; see Weinberger (1974, §3.3). To avoid interpolation, $\lambda J$ in (2) has to restrict the estimate to an effectively finite dimensional space, and to obtain the sought-after flexibility, the effective model space dimension has to be increased via reducing $\lambda$ as the sample size $n$ increases. These considerations make the following assumption necessary for a sensible (2).

**Assumption A.1.** $V$ is completely continuous with respect to $(V + J)$.

Under A.1, using Theorem 3.1 of Weinberger (1974, p.52), it can be shown that there exist $\phi_\nu \in \mathcal{H}$ and $0 \leq \rho_\nu \uparrow \infty$, $\nu = 1, 2, \cdots$, 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 Gu & Qiu (1991, §4). The notion of smoothness is characterized by the rate of growth of $\rho_\nu$.

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

For the cubic spline logistic regression, A.1 and A.2 are satisfied when $\log(v_0 f)$ is bounded from both above and below on $[0,1]$, and $r = 4$ in A.2; see, e.g., Silverman (1982, p.802).

## 3. Asymptotic analysis

We first introduce a quadratic approximation to (2), whose minimizer is an approximation of $\hat{\eta}$ linear in $y_i$. The convergence rate of such a linear approximation is obtained via a Fourier analysis with $\phi_\nu$ in §2.2 as basis. We then calculate a bound for the distance between $\hat{\eta}$ and the linear approximation under two extra assumptions. The main convergence results in $V(\hat{\eta} - \eta_0)$ and in (3) follow simply by combining the results obtained in the two steps.

### 3.1. *Linear approximation*

Assume $\eta_0 \in \mathcal{H}$. Let $V(g,h)$ be the inner product associated with the quadratic norm $V$. Let $\eta_1$ be the minimizer of the quadratic functional

$$-\frac{1}{n}\sum_{i=1}^{n}\{y_i\eta(x_i) - \mu_0(x_i)\eta(x_i)\} + (1/2)V(\eta - \eta_0) + (\lambda/2)J(\eta). \tag{4}$$

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 $\eta$ with basis $\phi_\nu$. Substituting these into (4) and solving for $\eta_1$, 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 $E\beta_\nu^2 = n^{-1}$. It then follows that

$$\begin{aligned} EV(\eta_1 - \eta_0) &= E\sum_{i=1}^{n}(\eta_{\nu,1} - \eta_{\nu,0})^2 = O(n^{-1}\lambda^{-1/r} + \lambda) \\ E\lambda J(\eta_1 - \eta_0) &= E\lambda\sum_{i=1}^{n}\rho_\nu(\eta_{\nu,1} - \eta_{\nu,0})^2 = O(n^{-1}\lambda^{-1/r} + \lambda), \end{aligned}$$

as $n \to \infty$ and $\lambda \to 0$; see Gu & Qiu (1991, Theorem 4.1). 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 $\lambda J(\eta_1 - \eta_0) = O_p(n^{-1}\lambda^{-1/r} + \lambda)$.*

### 3.2. *Main result*

We need two more assumptions in further analysis.

**Assumption A.3.** For $\eta$ in a convex set $B_0$ around $\eta_0$ containing $\hat{\eta}$ and $\eta_1$,
$\exists c_1, c_2 \in (0, \infty)$ such that $c_1 v_0(x) \leq v(x) \leq c_2 v_0(x)$ 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 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 $\int_{\mathcal{X}} \phi_\nu^2 \phi_\mu^2 v_0^2 f \leq c_3$, $\forall \nu, \mu$.

Note that $\int_{\mathcal{X}} \phi_\nu^2 v_0 f = 1$. 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 the theorem is given in §3.3.

**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* $\int_{\mathcal{X}} (\hat{\mu} - \mu_0)(\hat{\eta} - \eta_0) f = O_p(n^{-1}\lambda^{-1/r} + \lambda)$.

*Proof*: Use Theorems 1, 2, and Assumption A.3. □

For the cubic spline logistic regression, A.3 is satisfied when $\mu(x)$ is uniformly bounded away from both 0 and 1 on $[0, 1]$ for members of $B_0$. A direct verification of A.4 is rather inconvenient if not impossible, since explicit formulas of $\phi_\nu$ are in general not available. A suggestive special case does exist, however, when $v_0 f = 1$ 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 $\phi_\nu$ and hence grants A.4 when $v_0$ is bounded.

### 3.3. *Proof of Theorem 2*

Write (2) 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

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

where $J(g,h)$ is the semi inner product associated with the quadratic seminorm $J$. Setting $\eta = \hat{\eta}$ and $h = \hat{\eta} - \eta_1$, one obtains

$$0 = \dot{A}_{\hat{\eta},\hat{\eta}-\eta_1}(0) = -\frac{1}{n}\sum_{i=1}^{n}\{y_i(\hat{\eta}-\eta_1)(x_i) - \hat{\mu}(x_i)(\hat{\eta}-\eta_1)(x_i)\} + \lambda J(\hat{\eta}, \hat{\eta}-\eta_1). \tag{5}$$

Similarly, denote (4) 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

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

Hence,

$$0 = \dot{B}_{\eta_1,\hat{\eta}-\eta_1}(0) = -\frac{1}{n}\sum_{i=1}^{n}\{y_i(\hat{\eta}-\eta_1)(x_i) - \mu_0(x_i)(\hat{\eta}-\eta_1)(x_i)\} + V(\eta_1-\eta_0, \hat{\eta}-\eta_1) + \lambda J(\eta_1, \hat{\eta}-\eta_1). \tag{6}$$

Equating (5) and (6), some algebra yields

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{\mu}-\mu_1)(x_i)(\hat{\eta}-\eta_1)(x_i) + \lambda J(\hat{\eta}-\eta_1) = V(\eta_1-\eta_0, \hat{\eta}-\eta_1) - \frac{1}{n}\sum_{i=1}^{n}(\mu_1-\mu_0)(x_i)(\hat{\eta}-\eta_1)(x_i). \tag{7}$$

By A.3,

$$c_1\frac{1}{n}\sum_{i=1}^{n}(\hat{\eta}-\eta_1)^2(x_i)v_0(x_i) \le \frac{1}{n}\sum_{i=1}^{n}(\hat{\mu}-\mu_1)(x_i)(\hat{\eta}-\eta_1)(x_i). \tag{8}$$

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

$$\begin{aligned}
&|\frac{1}{n}\sum_{i=1}^{n}(\hat{\eta}-\eta_1)^2(x_i)v_0(x_i) - V(\hat{\eta}-\eta_1)| \\
= \;& |\sum_{\nu}\sum_{\mu}(\hat{\eta}_\nu-\eta_{\nu,1})(\hat{\eta}_\mu-\eta_{\mu,1})\{\frac{1}{n}\sum_{i=1}^{n}\phi_\nu(x_i)\phi_\mu(x_i)v_0(x_i) - \int\phi_\nu\phi_\mu v_0 f\}| \\
\le \;& [\sum_{\nu}\sum_{\mu}(1+\lambda\rho_\nu)^{-1}(1+\lambda\rho_\mu)^{-1}\{\frac{1}{n}\sum_{i=1}^{n}\phi_\nu(x_i)\phi_\mu(x_i)v_0(x_i) - \int\phi_\nu\phi_\mu v_0 f\}^2]^{1/2} \\
& [\sum_{\nu}\sum_{\mu}(1+\lambda\rho_\nu)(1+\lambda\rho_\mu)(\hat{\eta}_\nu-\eta_{\nu,1})^2(\hat{\eta}_\mu-\eta_{\mu,1})^2]^{1/2} \\
= \;& O_p(n^{-1/2}\lambda^{-1/r})(V+\lambda J)(\hat{\eta}-\eta_1) \\
= \;& o_p(1)(V+\lambda J)(\hat{\eta}-\eta_1),
\end{aligned} \tag{9}$$

where Cauchy-Schwartz, A.4, and the fact that $\sum_\nu(1+\lambda\nu^r)^{-1} = O(\lambda^{-1/r})$ (Gu & Qiu, 1991, Lemma 4.2) are used. Combining (8) and (9), a lower bound for the left-hand-side of (7) is given by

$$(c_1 V + \lambda J)(\hat{\eta}-\eta_1)(1+o_p(1)). \tag{10}$$

On the right hand side of (7), A.3 leads to

$$\frac{1}{n}\sum_{i=1}^{n}(\mu_1-\mu_0)(x_i)(\hat{\eta}-\eta_1)(x_i)=c\frac{1}{n}\sum_{i=1}^{n}(\eta_1-\eta_0)(x_i)(\hat{\eta}-\eta_1)(x_i)v_0(x_i), \tag{11}$$

where $c\in[c_1,c_2]$. Similar to (9), it can be shown that

$$|\frac{1}{n}\sum_{i=1}^{n}(\eta_1-\eta_0)(x_i)(\hat{\eta}-\eta_1)(x_i)v_0(x_i)-V(\eta_1-\eta_0,\hat{\eta}-\eta_1)|=o_p(1)(V+\lambda J)^{1/2}(\eta_1-\eta_0)(V+\lambda J)^{1/2}(\hat{\eta}-\eta_1). \tag{12}$$

Combining (11) and (12), an upper bound is given by

$$|1-c|V^{1/2}(\eta_1-\eta_0)V^{1/2}(\hat{\eta}-\eta_1)+o_p(1)(V+\lambda J)^{1/2}(\eta_1-\eta_0)(V+\lambda J)^{1/2}(\hat{\eta}-\eta_1). \tag{13}$$

Joining (10) and (13) and applying Theorem 1 yield the theorem. □

## 4. Remarks

In the foregoing development, the smoothness assumptions are made of the canonical parameter $\eta$ 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 the penalized likelihood regression than on the 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 $\eta$ can take; second, (2) is guaranteed to be convex; third, a convenient and effective empirical choice of $\lambda$ is available and theoretically justifiable (Gu, 1990a, b); and fourth, a simple generic asymptotic analysis is possible as in this article. If circumstance demands a modeling parameter $\theta$ other than the canonical parameter $\eta$, however, the techniques used in this article may still be applicable in a similar analysis with a $V(\theta)$ defined by $\int_{\mathcal{X}}\theta^2(d\eta/d\theta)_0^2v_0f$, but the conditions and proofs could become much messier.

## Acknowledgements

This research was supported by a grant from the U.S. National Science Foundation and by a David Ross grant at Purdue University.

## References


Cox, D. D. & Chang, Y.-F. (1990). Iterated state space algorithms and cross validation for generalized smoothing splines. Technical Report 49, University of Illinois, Dept. of Statistics.

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

Green, P. & Yandell, B. (1985). Semi-parametric generalized linear models. In *GLIM85: Proceedings of the International Conference on Generalized Linear Models, Lecture Notes in Statistics (Vol. 32)*, pp. 44 – 55. Ed. R. Gilchrist, Springer-Verlag.

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

—— (1990b). A note on cross-validating non Gaussian data. Technical Report 96, University of British Columbia, Dept. of Statistics.

Gu, C. & Qiu, C. (1991). Smoothing spline density estimation: Theory. Technical Report 91-19, Purdue University, Dept. of Statistics.

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

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

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

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.