DEPARTMENT OF STATISTICS
University of Wisconsin
1300 University Ave.
Madison, WI 53706

TECHNICAL REPORT NO. 1156
September 28, 2009

Penalized likelihood regression in reproducing kernel Hilbert spaces with randomized covariate non-Gaussian data

Xiwen Ma[1]
Department of Statistics
University of Wisconsin, Madison

Grace Wahba[1]
Department of Statistics, Department of Computer Sciences
and Department of Biostatistics and Medical Informatics
University of Wisconsin, Madison

Bin Dai[1]
Department of Statistics
University of Wisconsin, Madison

[1] Research supported in part by NIH Grant EY09946, NSF Grant DMS-0604572, NSF Grant DMS-0906818 and ONR Grant N0014-06-00655. This work was a finalist in the ASA Nonparametric Statistics Section student paper competition and will appear in the Proceedings of JSM 2009. An Appendix with proofs and tuning details has been added here.

Penalized likelihood regression in reproducing kernel Hilbert spaces with randomized covariate non-Gaussian data

Xiwen Ma[*], Grace Wahba[*] and Bin Dai[*]

Department of Statistics, University of Wisconsin, Madison, WI 53706, USA

Abstract
Classical penalized likelihood regression problems deal with the case that the independent variables data are known exactly. In practice, however, it is common to observe data with incomplete covariate information. We are concerned with a fundamentally important case where some of the observations do not represent the exact covariate information, but only a probability distribution. In this case, penalized likelihood method is still applicable to estimate the regression function. We show that penalized likelihood estimate exists under a mild condition. In the computation, we propose a dimension reduction technique to minimize the penalized likelihood and a posterior version of GACV (Generalized Approximate Cross Validation) to choose the smoothing parameter. Our methodology can be extended to handle more complicated cases of incomplete covariate information. For example, covariate measurement error and partially missing covariates can be treated as special cases.

Key Words: penalized likelihood regression, reproducing kernel Hilbert space, randomized covariate data, generalized approximate cross validation, covariate measurement error, partially missing covariate data

1. Introduction

We are concerned with non or semi parametric regression for data from an exponential family without nuisance parameter. Suppose that (y_i, x_i) have been observed for n independent subjects, with y_i the response and x_i indexing the covariate information. The goal is to fit a probability mechanism, assuming that $[y_i|x_i]$ has a density with the canonical form

$$p(y_i|f(x_i)) = \exp\{y_i \cdot f(x_i) - b(f(x_i)) + c(y_i)\} \tag{1.1}$$

where $b(\cdot)$ and $c(\cdot)$ are given functions with $b(\cdot)$ strictly convex and $f(x_i)$ is the unknown natural model parameter. The regression function f will be

[*] Research supported in part by NIH Grant EY09946, NSF Grant DMS-0604572, NSF Grant DMS-0906818 and ONR Grant N0014-06-00655. This work was a finalist in the ASA Nonparametric Statistics Section student paper competition and will appear in the Proceedings of JSM 2009. An Appendix with proofs and tuning details has been added here.

estimated non or semi parametrically as an element of some reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ by solving a penalized likelihood problem

$$I_\lambda(f) = -\frac{1}{n}\sum_{i=1}^{n}\log p(y_i|f(x_i)) + \frac{\lambda}{2}J(f) \tag{1.2}$$

where the penalty $J(\cdot)$ is a norm or semi-norm in $\mathcal{H}$ with finite dimensional null space $\mathcal{H}_0 = \{f \in \mathcal{H} \mid J(f) = 0\}$ and λ is the smoothing parameter which balances the tradeoff between the model fitting and the smoothness. In this case, if the null space $\mathcal{H}_0$ satisfies some condition, saying that $I_\lambda(f)$ has a unique minimizer in $\mathcal{H}_0$, then the minimizer of $I_\lambda(f)$ in $\mathcal{H}$ exists in a known n-dimensional subspace spanned by $\mathcal{H}_0$ and functions of reproducing kernel. See Kimeldorf and Wahba (1971)[6], O'Sullivan (1983)[12], Wahba (1990)[15] and Xiang and Wahba (1996)[16].

This model building technique, known as penalized likelihood regression with RKHS penalty, allows for more flexibility than parametric regression models. Originated in the work of smoothing splines in Kimeldorf and Wahba (1970a[5], 1970b[7], 1971[6]), penalized likelihood regression is a durable statistical research topic and the applications are very broad. We will not review the general literature, other than to note two books and references therein. Wahba (1990)[15] offers a general introduction of spline models. Gu (2002)[2] comprehensively reviews the smoothing spline analysis of variance (SS-ANOVA), an important implementation of penalized likelihood regression in multivariate function estimation.

In this paper, the issue we are concerned about is the situation where components of x_i are not observable but only known to have come from a particular probability distribution. This concept of randomized covariates includes the common sense of covariate measurement error, but more flexible than that. In this case, we suggest to estimate the regression function by minimizing a randomized version of penalized likelihood which integrates out the random information in the covariates. This method, however, typically leads to a non-convex and infinite dimensional variational problem in the Hilbert space. Therefore we first prove that randomized penalized likelihood is minimizable under a mild condition. Then we propose a dimension reduction technique to minimize the penalized likelihood and a posterior version of GACV (Generalized Approximate Cross Validation) to tune the smoothing parameter.

Randomized covariate data is fundamentally important. The methodology can be extended to handle other cases of incomplete covariate information. For example, in the survey or medical research, it is common to obtain a data where the covariates are measured with errors. More specifically, x_i is not directly observed but instead $t_i = x_i + \epsilon_i, i = 1, ..., n$ is observed. Here $\epsilon_i, i = 1, ..., n$ are independent random perturbations, typically mean-zero Gaussian random variables (vectors). The distribution of ϵ_i is assumed to be known. Measurement error can be viewed as a special case of randomized covariates because x_i can be treated as a random variable identically distributed to $t_i - \epsilon_i$ with known distribution. Hence the methodology for randomized covariate data can be employed directly.

We will as well be able to make a modest extension to treat the important situation where some components of some x_i's are completely missing. We adopt the idea of Ibrahim's method of weights (Ibrahim, 1990[3] and Ibrahim *et al.*, 1999[4]) which suggests to assume a parametric model for x and maximize the joint distribution of (y, x) by expectation-maximization (EM) algorithm. In the framework of method of weights, missing covariate data can be treated as a special case of randomized covariate data, allowing covariate distributions to be flexible. Therefore our methodology can be extended.

The rest of the paper is organized as follows. We introduce randomized covariate data in Section 2. Computation techniques are presented in Section 3. Section 4 discusses missing covariate data. Section 5 presents some simulations. We conclude our paper in Section 6.

2. Definition.

Consider the general smoothing spline set-up, where x_i is allowed to be from some arbitrary index set $\mathscr{T}$ on which a RHKS can be defined. Randomized covariate data is defined in the way that we "observe" for each subject a *probability space* $(\mathcal{X}_i, \mathcal{F}_i, v_{x_i})$, rather than a realization of x_i, where $\mathcal{X}_i \subseteq \mathscr{T}$, $\mathcal{F}_i$ is a $\sigma-$algebra and v_{x_i} is a probability measure over $(\mathcal{X}_i, \mathcal{F}_i)$.

In this case, we consider the following variational problem which integrated out the "randomness" of the covariates in the likelihood

$$I_\lambda^R(f) = -\frac{1}{n}\sum_{i=1}^{n} \log \int_{\mathcal{X}_i} p(y_i|f(u))dv_{x_i}(u) + \frac{\lambda}{2}J(f) \qquad (2.1)$$

Here f is restricted on the Borel measurable subset

$$\mathcal{H}_B = \{f \in \mathcal{H} \ : \ f \text{ is Borel measurable on } (\mathcal{X}_i, \mathcal{F}_i), i = 1, ..., n\} \qquad (2.2)$$

where the Lebesgue integrals in (2.1) can be defined. It can be shown that $\mathcal{H}_B$ is a subspace of $\mathcal{H}$.

PROPOSITION 2.1. *$\mathcal{H}_B$ is a subspace of $\mathcal{H}$.*

Remark 1. It is well known (see Wahba, 1978[13], 1983[14] and Nychka 1988[11]) that the penalty functional $J(\cdot)$ is equivalent to a mean zero partially improper Gaussian process prior for $f \in \mathcal{H}$, where the Gaussian process is diffuse in $\mathcal{H}_0$ and proper in $\mathcal{H}_1$ ($\mathcal{H}_0 \oplus \mathcal{H}_1 = \mathcal{H}$) with covariance function $K_1(\cdot, \cdot)$, the reproducing kernel for $\mathcal{H}_1$ associated with the norm $J(\cdot)$. Therefore minimizing $I_\lambda^R(f)$ is equivalent to maximizing a Bayesian posteriori.

$I_\lambda^R(f)$ is referred to as randomized penalized likelihood, where R denotes "randomness" of the covariates. In this case, however, computing a penalized likelihood estimate is extremely difficult. Firstly, since $p(y_i|f(x_i))$ is log-concave as a function of f, $I_\lambda^R(f)$ is in general not convex. Secondly, if at least one $(\mathcal{X}_i, \mathcal{F}_i, v_{x_i})$ has infinite supports, then the minimizer of $I_\lambda^R(f)$ will involve an infinite dimensional optimization problem in the Hilbert space, as can be concluded from the arguments in Kimeldorf and Wahba (1971)[6]. Therefore, we shall first prove that $I_\lambda^R(f)$ is minimizable and hence the phrase "penalized likelihood regression" is meaningful. Computation techniques will be described in Section 3.

Recall that for complete data penalized likelihood regression (1.2), the unique solution in the null space is sufficient to ensure the existence of penalized likelihood estimate. In the case of randomized covariate data, we extend this condition as follow

ASSUMPTION A.1 (Null space condition). There exist completely observed cases $(y_{k_1}, x_{k_1}), (y_{k_2}, x_{k_2}), ..., (y_{k_s}, x_{k_s})$ such that $L(f) = \sum_{i=1}^{s} \log p(y_{k_i} | f(x_{k_i}))$ has a unique maximizer in $\mathcal{H}_0$.

This null space condition can be satisfied easily when there are adequate complete observations. Now we state our main theorem which guarantees the minimizability of $I_\lambda^R(f)$.

THEOREM 2.2. *Under A.1,* $\exists f_\lambda \in \mathcal{H}_B$ *such that* $I_\lambda^R(f_\lambda) = \inf_{f\in\mathcal{H}_B} I_\lambda^R(f)$.

Theorem 2.2 shows the existence of penalized likelihood estimate, which justifies the title of the paper. In particular, if the penalty functional $J(\cdot)$ is the squared norm in a RKHS with null space containing only constants, then the estimator exists even if every data point is really a probability distribution.

Theorem 2.2 can be proved by combining the following results. Note that Proposition 2.3 is obtained from Theorem 7.3.7 in Kurdila and Zabarankin (2005)[8], Page 217.

PROPOSITION 2.3. *Let* $\mathcal{H}$ *be a Hilbert space. Suppose that* $\gamma : \mathcal{M} \subseteq \mathcal{H} \to \mathbb{R}$ *is positively coercive and weakly sequentially lower semicontinuous over the closed and convex set* $\mathcal{M}$*, then* $\exists f_0 \in \mathcal{M}$ *such that* $\gamma(f_0) = \inf_{f\in\mathcal{M}} \gamma(f)$.

LEMMA 2.4. *Under A.1, the penalized likelihood* $I_\lambda^R(f)$ *is positively coercive over* $\mathcal{H}_B$.

LEMMA 2.5. *Functional* $\log \int_{\mathcal{X}_i} p(y_i|f(u))dv_{x_i}(u) : \mathcal{H}_B \to \mathbb{R}$ *is weakly sequentially continuous.*

LEMMA 2.6. *The penalty functional* $J(\cdot)$ *is weakly sequentially lower semicontinuous.*

3. Computation.

In the variational problem (2.1), f was restricted on the Borel measurable subspace $\mathcal{H}_B$. In practical applications, however, we often face the case that all functions in the RKHS are Borel measurable. In this case, we no longer need the restriction mentioned in (2.2). In order to derive more applicable results, we would like to proceed our discussion under the following condition

ASSUMPTION A.2. Mapping $\psi_K(x) = K(\cdot, x)$ is Borel measurable for all $(\mathcal{X}_i, \mathcal{F}_i)$, $i = 1, ...n$. Here $K(\cdot, \cdot)$ is the reproducing kernel of $\mathcal{H}$.

Under A.2, by Theorem 90 of Berlinet and Thomas-Agnan (2004)[1], Page 195, every function in $\mathcal{H}$ is Borel measurable. It can be verified that if $\mathscr{T} \subseteq \mathcal{R}^d$ and every $\mathcal{F}_i$ is a Borel σ-field, then A.2 is satisfied with all continuous kernels and radial basis kernels $R(s,t) = \phi(||s-t||_d)$ with $\phi(\cdot)$ continuous at 0. Here $||\cdot||_d$ is the usual Euclidian norm.

3.1 Quadrature EM algorithm.

In general, the exact minimizer f_λ is infinite dimensional and hence not computable. In this case, we shall find a finite dimensional space and compute an estimator in this space. We consider the following class of penalized likelihood

$$I_\lambda(\mathbf{Z}, \Pi; f) = -\frac{1}{n}\sum_{i=1}^{n} \log \sum_{j=1}^{m_i} \pi_{ij} p(y_i | f(z_{ij})) + \frac{\lambda}{2} J(f) \tag{3.1}$$

where $\mathbf{Z} = \{z_{11}, z_{12}, ..., z_{1m_1}, z_{21}, ..., z_{nm_n}\} \subseteq \mathscr{T}$ and $\Pi = \{\pi_{11}, \pi_{12}, ..., \pi_{1m_1}, \pi_{21}, ..., \pi_{nm_n}\}$ is a collection of positive values. In words, when we evaluate the integrals on the right hand side of (2.1), each $\{\mathcal{X}_i, \mathcal{F}_i, v_{x_i}\}$ is replaced with a discrete probability distribution defined over $\{z_{i1}, z_{i2}, ..., z_{im_i}\}$ with probability mass function $P(x_i = z_{ij}) = \pi_{ij}, \ j = 1, ..., m_i$. Hence, $\mathbf{Z}$ and Π are referred to as nodes and weights of the quadrature for probability distributions.

In (3.1), f is only evaluated on the finite quadrature nodes $\mathbf{Z}$. Under A.1, it can be seen from the arguments in Kimeldorf and Wahba (1971)[6] that the minimizer of $I_\lambda(\mathbf{Z}, \Pi; f)$ in $\mathcal{H}$ is in a finite dimensional subspace spanned by $\mathcal{H}_0$ and $\{K(\cdot, z_{ij}) : z_{ij} \in \mathbf{Z}\}$. Therefore $I_\lambda(\mathbf{Z}, \Pi; f)$ can be viewed as a projection of $I_\lambda^R(f)$ onto this subspace.

$I_\lambda(\mathbf{Z}, \Pi; f)$ can be minimized via EM algorithm. It can be verified that the E-step at iteration $t+1$ has the form of

$$Q(f|f^t) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i} w_{ij}^t \cdot \log p(y_i | f(z_{ij})) - \frac{\lambda}{2} J(f) \tag{3.2}$$

where f^t is estimated at iteration t and $w_{ij}^t = \pi_{ij} p(y_i | f^t(z_{ij})) / \sum_j \pi_{ij} p(y_i | f^t(z_{ij}))$ indicates the posterior distribution of z_{ij} given y_i and f^t. The M-step updates f by maximizing $Q(f|f^t)$ in $\mathcal{H}$ which is computationally straightforward because $-Q(f|f^t)$ is seen to be a weighted complete data penalized likelihood. The computed minimizer $\hat{f}_\lambda$ will be treated as an approximation of the true estimator.

This computation method, which combines dimension reduction and EM algorithm, is referred to as *quadrature EM algorithm.*

3.2 Tuning method.

So far the smoothing parameter λ, is assumed to be fixed. In this section, we outline our posterior tuning strategy.

Let f^* be the true regression function and $\hat{f}_\lambda$ denote the computed estimator. For non-Gaussian exponential family data, Kullback-Leibler (KL) distance is perhaps the most popular criterion for smoothing parameter selection

$$KL(f^*(x_i), \hat{f}_\lambda(x_i)) = E_{y_i^0|f^*(x_i)} \left\{ \log \frac{p(y_i^0|f^*(x_i))}{p(y_i^0|\hat{f}_\lambda(x_i))} \right\} \tag{3.3}$$

where the expectation is taken over $y_i^0 \sim p(y|f^*(x_i))$ independent of y_i. However, KL–loss is not directly applicable in the presence of randomized covariate information. In this case, we suggest to compute the posterior mean KL distance

$$E\left\{KL(\lambda)\right\} = \frac{1}{n}\sum_{i=1}^{n} E_{x_i|y_i,f^*}\left\{E_{y_i^0|f^*(x_i)}\left\{\log \frac{p(y_i^0|f^*(x_i))}{p(y_i^0|\hat{f}_\lambda(x_i))}\right\}\right\} \tag{3.4}$$

where the expectation $E_{x_i|y_i,f^*}$ is taken w.r.t. the posterior distribution of x_i given y_i and f^*. In practice, $[x_i|y_i, f^*]$ can be estimated by $[x_i|y_i, \hat{f}_\lambda]$.

Recall that $\hat{f}_\lambda$ is calculated from the model at the last iteration of the EM algorithm

$$-\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i} w_{ij}^\lambda \cdot \log p(y_i|f(z_{ij})) + \frac{\lambda}{2}J(f) \tag{3.5}$$

where w_{ij}^λ denotes the weights at the last EM iteration. Since $\{z_{i1}, ..., z_{im_i}\}$ and $\{w_{i1}^\lambda, ..., w_{im_i}^\lambda\}$ is a discrete approximation for $[x_i|y_i, \hat{f}_\lambda]$, we have that

$$E\{KL(\lambda)\} \approx \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i} w_{ij}^\lambda \cdot E_{y_i^0|f^*(z_{ij})}\left\{\log \frac{p(y_i^0|f^*(z_{ij}))}{p(y_i^0|\hat{f}_\lambda(z_{ij}))}\right\} \tag{3.6}$$

Note that the right hand side of (3.6) is exactly the KL–loss for the model (3.5). Therefore $E\{KL(\lambda)\}$ can be evaluated at the last EM iteration. In this paper, we suggest to compute for each final model (3.5) a generalized approximate cross validation (GACV) score proposed in Xiang and Wahba (1996)[16] and choose the λ with minimum GACV score. We refer to this tuning method as *posterior GACV*, since (3.5) can be viewed as posterior mean of penalized likelihood.

4. Missing covariate data.

In this section we describe penalized likelihood regression with missing covariate data. We assume the missing mechanism to be missing at random.

4.1 Notations and Models.

Let $x_i = (x_{i1}, ..., x_{id})$ denote the vector of covariates ranging over $\mathscr{T} \subseteq \mathbb{R}^d$. Write $x_i = (x_i^{mis}, x_i^{obs})$ where x_i^{mis} is $d_i \times 1$ vector of missing components and x_i^{obs} is vector of observed components. Let $\mathcal{X}_i = \mathscr{T} \cap \{(x_i^{mis}, x_i^{obs}) : x_i^{mis} \in \mathbb{R}^{d_i}\}$ denote the set of feasible x_i. When x_i is fully observed, $x_i^{mis} = \emptyset$ and $\mathcal{X}_i = \{x_i\}$.

By the idea of Ibrahim's method of weights, we assume a parametric model for the covariates, with a density $p(x|\theta) > 0$, where $\theta \in \Theta \subseteq \mathbb{R}^p$ is a real vector of indexing parameter.

In this case, the likelihood of (f, θ) can be obtained by integrating or summing out the missing components in the joint density (Little and Rubin, (2002)[10])

$$L(f, \theta) = \sum_{i=1}^{n} \log \int_{\mathcal{X}_i} p(y_i|f(x_i))p(x_i|\theta)dx_i \tag{4.1}$$

where $dx_i = dx_i^{mis}$ if x_i has missing components, otherwise

$$\int_{\mathcal{X}_i} p(y_i|f(x_i))p(x_i|\theta)dx_i = p(y_i|f(x_i))p(x_i|\theta) \tag{4.2}$$

Now it is straightforward to define missing data penalized likelihood of (f, θ) by

$$I_\lambda^M(f, \theta) = -\frac{1}{n}L(f, \theta) + \frac{\lambda}{2}J(f) \tag{4.3}$$

In the consideration of the relationship with randomized covariate data, let $v_{x_i}^\theta$ be the probability measure w.r.t. the density $p(x_i|\theta)/\int_{\mathcal{X}_i} p(x_i|\theta)dx_i,\ x_i \in \mathcal{X}_i$. Note that $\int_{\mathcal{X}_i} p(x_i|\theta)dx_i < \infty$ from the Fubini's Theorem. It is not hard to see that

$$I_\lambda^M(f, \theta) = -\frac{1}{n}\sum_{i=1}^{n} \log \int_{\mathcal{X}_i} p(y_i|f(u))dv_{x_i}^\theta(u) + \frac{\lambda}{2}J(f) - \frac{1}{n}\sum_{i=1}^{n} \log \int_{\mathcal{X}_i} p(x_i|\theta)dx_i \tag{4.4}$$

is composed of randomized penalized likelihood and likelihood for the covariate distribution. Hence missing covariate data can be treated as a special case of randomized covariate data, allowing covariate distributions to be flexible.

4.2 Results.

We first study the existence of penalized likelihood estimate for (f, θ). The following assumptions can be easily satisfied in the most settings.

ASSUMPTION M.1. $\mathcal{X}_i, i = 1, ..., n$ are compact.

ASSUMPTION M.2. $p(x|\theta)$ is continuous in θ and the parameter space Θ is compact.

COROLLARY 4.1. *Under A.1, M.1 and M.2, there exist $f_\lambda \in \mathcal{H}$ and $\theta_\lambda \in \Theta$ such that $I_\lambda^M(f_\lambda, \theta_\lambda) = \inf_{f\in\mathcal{H},\theta\in\Theta} I_\lambda^M(f, \theta)$.*

In the computation, the quadrature EM algorithm is slightly more flexible. The E-step can be written as

$$Q(f,\theta|f^t,\theta^t) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i} w_{ij}^t \cdot \log p(y_i|f(z_{ij}^t)) - \frac{\lambda}{2}J(f) + \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i} w_{ij}^t \cdot \log p(z_{ij}^t|\theta) \tag{4.5}$$

where (f^t, θ^t) is obtained in the preceding iteration and w_{ij}^t is computed by

$$w_{ij}^t = \frac{\pi_{ij}^t p(y_i|f^t(z_{ij}^t))}{\sum_j \pi_{ij}^t p(y_i|f^t(z_{ij}^t))} \tag{4.6}$$

Here $\{z_{ij}^t\}_{j=1,\ldots,m_i}$ and $\{\pi_{ij}^t\}_{j=1,\ldots,m_i}$ are the nodes and weights for the quadrature based on $\mathcal{X}_i$ and $p(x|\theta^t)$. Then the M-step updates f and θ separately by maximizing

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i} w_{ij}^t \cdot \log p(y_i|f(z_{ij}^t)) - \frac{\lambda}{2}J(f) \tag{4.7}$$

and

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i} w_{ij}^t \cdot \log p(z_{ij}^t|\theta) \tag{4.8}$$

which are computationally straightforward assuming the log-concavity of $p(x|\theta)$ (as a function of θ).

In order to select the smoothing parameter, we note that θ is a nuisance parameter and the choice of λ only depends on the goodness of fit of $\hat{f}_\lambda$. Therefore, similar to the discussions in Section 3.2, we may select λ by computing the GACV score of the model (4.7) at the last EM iteration.

5. Simulations

In this section, our method is examined by two simulated examples: (1) cubic spline Poisson regression with covariate measurement error; and (2) thin plate spline logistic regression with partially missing covariates. We generate 100 datasets in each example and for each simulated dataset we estimate the regression function by: (a) our proposed method; (b) full data analysis before measurement error or missing covariates; and (c) naive method by ignoring the noisy data or leaving out the observations with missing covariates. We compare these three methods graphically and by Kullback-Leibler distance between the estimated and true regression function. All the simulations are conducted using R-2.9.1 installed in Red Hat Enterprise Linux 5.

5.1 Example 1.

Consider the Poisson distribution $p(y|\lambda(x)) = \lambda(x)^y e^{-\lambda(x)}/y!,\ y = 0, 1...$ where

$$\lambda(x) = 10^6(x^{11}(1-x)^6) + 10^4(x^3(1-x)^{10}) + 1 \tag{5.1}$$

is chosen from Example 5.5 of Gu (2002)[2]. We take $x \sim U(0,1)$ and generate a sample of $n = 100$ pairs of (x, y). Then we randomly select five (x_i, y_i) as completely observed cases and for the rest data points, we generate random noise by $x_i + \epsilon_i$, where ϵ_i are $iid \sim N(0, 0.1^2)$. The noise-signal-ratio here is $Var(\epsilon)/Var(x) = 0.12$.

In this example, we apply cubic spline regression for estimating $\lambda(x)$. To implement the quadrature EM algorithm, we use *statmod* package in R-2.9.1 to compute a Gaussian quadrature for the covariate distribution. $m = 9$ quadrature nodes are created for each noisy covariate. In Figure 1, we show a graph with the true curve of $\lambda(x)$ and three estimate curves. We also present the boxplot of KL–loss from 100 simulations.

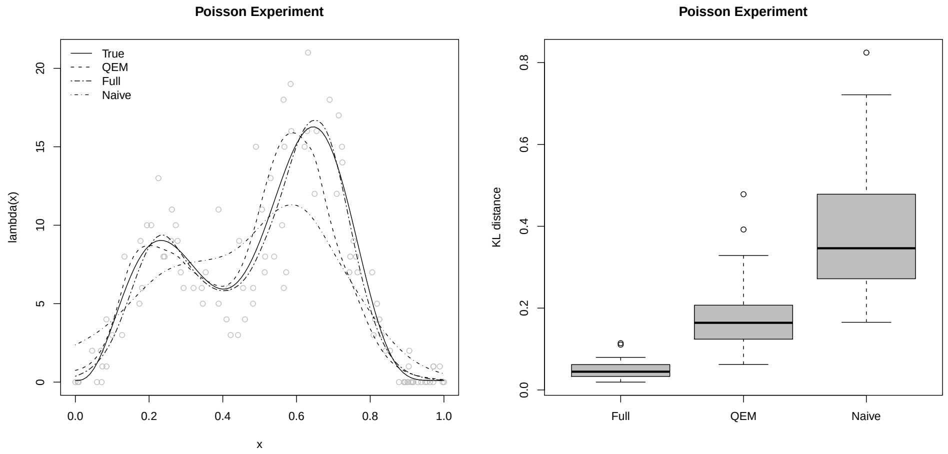

Figure 1. Estimates of cubic spline regression and boxplot of KL–loss. True: the true curve of $\lambda(x)$. QEM: estimate from quadrature EM algorithm. Full: data analysis before measurement error. Naive: data analysis by ignoring the measurement error. On the left panel, circles represent the original sample before measurement error.

5.2 Example 2.

Consider the binomial distribution $\begin{pmatrix} 5 \\ y \end{pmatrix} p(x)^y (1 - p(x))^{5-y}$, where $p(x) = p(x_1, x_2)$ is a modification of Franke's test function

$$\begin{aligned} p(x) = \frac{1}{1.24}(&\frac{3}{4}e^{-((9x_1-2)^2+(9x_2-2)^2)/4} + \frac{3}{4}e^{-((9x_1+1)^2/49+(9x_2+1)^2/10)} \\ &+ \frac{1}{2}e^{-((9x_1-7)^2+(9x_2-3)^2)/4} - \frac{1}{5}e^{-((9x_1-4)^2+(9x_2-7)^2)} + 0.2) \end{aligned} \quad (5.2)$$

Take $x \sim N(\mu, \Sigma)$ where $\mu = (0.3, 0.3)^T$ and $\Sigma = ((0.4, 0.24)^T, (0.24, 0.4)^T)$. In each simulation, we generate $n = 160$ covariates within the unit square $[0, 1] \times [0, 1]$ and then generate y from the binomial distribution. Afterwards, we create missing data in the way that if $y > 3$ then we randomly delete either x_1 or x_2.

In this example, we use thin plate spline regression to fit the data. To implement our method, we first assume a bivariate normal distribution $N((a_1,$

$a_2)^T, ((a_3, a_4)^T, (a_4, a_5)^T))$ for x, where $a_i, i = 1, ...5$ are nuisance parameters. Then, similar to Example 1, we use Gaussian quadrature for fixed $a_1, ..., a_5$. $m = 11$ quadrature nodes are created for each missing covariate. Simulation results are presented by Figure 2.

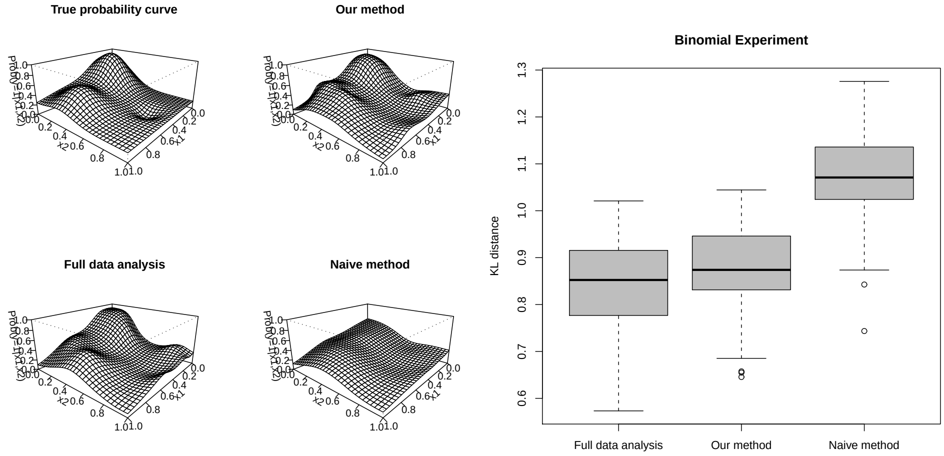

Figure 2. Estimates of thin plate spline regression and boxplot of KL–loss from 100 simulations.

6. Concluding remarks.

By working with randomized covariate data, we have provided a general framework for handling incomplete covariate data in the context of penalized likelihood regression. Penalized likelihood estimate exists if the null space condition holds for the completely observed cases. Numerically, we propose a dimension reduction technique to minimize the penalized likelihood and a posterior tuning strategy to choose the tuning parameter. We note that quadrature EM algorithm is computationally friendly as it does not require a large number of quadrature nodes to get a good estimate. Our methodology has been applied to both simulated and real datasets with excellent results. More details of the data analysis will appear in our final manuscript.

Acknowledgments

This work is supported in part by NIH Grant EY09946, NSF Grant DMS-0604572, NSF Grant DMS-0906818 and ONR Grant N0014-06-00655.

7. Appendix

7.1 Technical proofs.

Proof of proposition 2.1. Any linear combination of measurable functions is still measurable. Therefore it suffices to prove that $\mathcal{H}_B$ is complete. Let $f_1, f_2, ...$ be a Cauchy sequence in $\mathcal{H}_B$ and f^* be its limit in $\mathcal{H}$. Then $f_1, f_2, ...$ converge pointwisely to f^*. Note that the pointwise limit of measurable functions is still a measurable function. Therefore $f^* \in \mathcal{H}_B$. □

Now to simply the notation in the proofs of Lemma 2.4-2.6, let's define

$$l_i(t) = \log p(y_i|t) = y_i \cdot t - b(t) + c(y_i) \tag{7.1}$$

the log-density as a function of the natural parameter. Then $l_i(t)$ is strictly concave and bounded from above. Therefore there are three possible cases of the limit of $l_i(t)$:

$$\text{Type 1:} \quad \lim_{t\to-\infty} l_i(t) = \bar{l}_i \ ; \ \lim_{t\to+\infty} l_i(t) = -\infty \tag{7.2}$$

$$\text{Type 2:} \quad \lim_{t\to-\infty} l_i(t) = -\infty \ ; \ \lim_{t\to+\infty} l_i(t) = \bar{l}_i \tag{7.3}$$

$$\text{Type 3:} \quad \lim_{t\to-\infty} l_i(t) = -\infty; \lim_{t\to+\infty} l_i(t) = -\infty \tag{7.4}$$

where $\bar{l}_i = \sup_t l_i(t) = \sup_t \log p(y_i|t) < \infty$.

Proof of Lemma 2.4. Without loss of generality, we suppose that A.1 is satisfied with the first m cases (hence they are completely observed). In order to show Lemma 2.4, we first prove that under A.1, $L(f) = \sum_{i=1}^m \log p(y_i|f(x_i))$ is positively coercive over $\mathcal{H}_0$. Suppose to the contrary that this is not true. Then there exists a constant $U > 0$ and a sequence $\{g_k\}_{k\in\mathbb{N}} \subseteq \mathcal{H}_0$ with $||g_k||_{\mathcal{H}} = 1$ such that

$$-\sum_{i=1}^{m} l_i(k \cdot g_k(x_i)) \leq U, \quad k \in \mathbb{N} \tag{7.5}$$

Since the unit sphere $\{g \in \mathcal{H}_0 \ : \ ||g||_{\mathcal{H}} = 1\}$ is sequence compact, there exists a subsequence $\{g_{k_j}\}_{j\in\mathbb{N}}$ converging to some g^* with $||g^*||_{\mathcal{H}} = 1$. We claim that

$$g^*(x_i) \begin{cases} \leq 0, & \text{if } i \text{ belongs to Type 1 as (7.2)} \\ \geq 0, & \text{if } i \text{ belongs to Type 2 as (7.3)} \\ = 0, & \text{if } i \text{ belongs to Type 3 as (7.4)} \end{cases} \tag{7.6}$$

Suppose to the contrary that (7.6) is not true. If i belong to Type 1, then $g^*(x_i) = a > 0$. Since $\{g_{k_j}\}_{j\in\mathbb{N}}$ converges to g^*, there exists $N > 0$ such that

$$g_{k_j}(x_i) \geq a/2, \quad \text{for all } j > N \tag{7.7}$$

From (7.5), we have

$$l_i(k_j \cdot g_{k_j}(x_i)) \geq U - \sum_{s\neq i} \bar{l}_s < \infty, \quad j \in \mathbb{N} \tag{7.8}$$

This is a contradiction of (7.2) since when $j > N$

$$k_j \cdot g_{k_j}(x_i) \geq k_j \cdot a/2 \to +\infty \tag{7.9}$$

Similar contradiction can be observed when i belongs to Type 2 or Type 3. Therefore the claim in Equation (7.6) follows.

Now let g_0 be the unique minimizer of $-\sum_{i=1}^m l_i(g(x_i))$ in $\mathcal{H}_0$. Consider $g_0 + rg^*$ with $r > 0$. Combining (7.2)–(7.4) and (7.6), we can see that

$$-\sum_{i=1}^m l_i(g_0(x_i) + rg^*(x_i)) \leq -\sum_{i=1}^m l_i(g_0(x_i)), \quad \forall r > 0 \tag{7.10}$$

But this is a contradiction. Hence $L(f)$ is positively coercive over $\mathcal{H}_0$, which means that

$$||g||_{\mathcal{H}} \to \infty \Rightarrow -\sum_{i=1}^m l_i(g(x_i)) \to +\infty, \quad g \in \mathcal{H}_0 \tag{7.11}$$

Consider the orthogonal decomposition $f = g + h$ where $g \in \mathcal{H}_0 \bigcap \mathcal{H}_B$ and $h \in \mathcal{H}_1 \bigcap \mathcal{H}_B$. The Lemma can be proved in steps.

(i) $||h||_{\mathcal{H}} \to +\infty$. In this case

$$I_\lambda^R(f) \geq -\frac{1}{n}\sum_{i=1}^n \bar{l}_i + \frac{1}{2}\lambda||h||_{\mathcal{H}} \to +\infty \tag{7.12}$$

(ii) $||h||_{\mathcal{H}} \leq U$ for some $U > 0$ but $||g||_{\mathcal{H}} \to +\infty$. In this case

$$|h(x_i)| = |\langle h, K(\cdot, x_i)\rangle| \leq ||h||_{\mathcal{H}} K^{1/2}(x_i, x_i) \leq U \cdot K^{1/2}(x_i, x_i), \quad i = 1, 2, ...m \tag{7.13}$$

This means that

$$f(x_i) = g(x_i) + h(x_i) = g(x_i) + O(1), \quad i = 1, ..., m, \ ||h||_{\mathcal{H}} \leq U \tag{7.14}$$

Let $||g||_{\mathcal{H}} \to \infty$, we have

$$\begin{aligned} I_\lambda^R(f) &\geq -\frac{1}{n}\sum_{i=1}^n \log \int_{\mathcal{X}_i} p(y_i|f(u))dv_{x_i}(u) \\ &\geq -\frac{1}{n}\sum_{i=1}^m l_i(g(x_i) + h(x_i)) - \frac{1}{n}\sum_{j=m+1}^n \bar{l}_j \\ &= -\frac{1}{n}\sum_{i=1}^m l_i(g(x_i) + O(1)) - \frac{1}{n}\sum_{j=m+1}^n \bar{l}_j \\ &\to +\infty \end{aligned} \tag{7.15}$$

where (7.15) follows from the claim in Equation (7.11).

The Lemma is now proved by combining (i) and (ii). □

Proof of Lemma 2.5. Let $\{f_k\}_{k\in\mathbb{N}}$ be a sequence in $\mathcal{H}_B$ which converges weakly to f^*. Since pointwise limit of measurable functions is still a measurable function, $f^* \in \mathcal{H}_B$. From the continuity of $l_i(t)$, $\{e^{l_i(f_k(x_i))}\}_{k\in\mathbb{N}}$ pointwise converges to $e^{l_i(f^*(x_i))}$ over $\mathcal{X}_i$. Note that $e^{l_i(f_k(x_i))} \leq e^{\bar{l}_i}$ and every constant is integrable with respect to $(\mathcal{X}_i, \mathcal{F}_i, v_{x_i})$. By Dominated Convergence Theorem, we have that

$$\lim_{k\to\infty} \int_{\mathcal{X}_i} e^{l_i(f_k(u))} dv_{x_i}(u) = \int_{\mathcal{X}_i} e^{l_i(f^*(u))} dv_{x_i}(u) \tag{7.16}$$

The Lemma now follows since $\log(\cdot)$ is continuous. □

Proof of Lemma 2.6. Let $\{f_k\}_{k\in\mathbb{N}}$ be a sequence in $\mathcal{H}_B$ which weakly converges to f^*. Consider the orthogonal decomposition of each f_k by $f_k = g_k + h_k$ with $g_k \in \mathcal{H}_0 \bigcap \mathcal{H}_B$ and $h_k \in \mathcal{H}_1 \bigcap \mathcal{H}_B$. It is straightforward to see that $\{h_k\}_{k\in\mathbb{N}}$ weakly converges to h^*, the smooth part of f^*. Therefore we can write

$$0 \leq ||h_k - h^*||^2_{\mathcal{H}} = ||h_k||^2_{\mathcal{H}} + ||h^*||^2_{\mathcal{H}} - 2\langle h_k, h^*\rangle \tag{7.17}$$

Let $k \to \infty$, we observe that

$$0 \leq \liminf_k ||h_k||^2_{\mathcal{H}} - ||h^*||^2_{\mathcal{H}} \tag{7.18}$$

and the Lemma is proved by definition. □

Proof of Theorem 2.2. Consider the functional $I^R_\lambda : \mathcal{H}_B \subseteq \mathcal{H} \to \mathbb{R}$. Then Theorem 2.2 is proved by combining Proposition 2.2, Lemma 2.4-2.6 and Proposition 2.3 □.

Proof of Corollary 4.1. Fixed $\theta \in \Theta$, by (4.4) and Theorem 2.2, $I^M_\lambda(f,\theta)$ is minimizable in $\mathcal{H}$. Let

$$T(\theta) \triangleq \min_{f\in\mathcal{H}} I^M_\lambda(f,\theta) \tag{7.19}$$

denote the minimum penalized likelihood given θ. We claim that $T(\theta)$ is continuous.

By Assumption M.1 and M.2, there exists $U > 0$ such that $p(x|\theta) < U$ for all $x \in \mathcal{X}_i$ and $\theta \in \Theta$. Now for any sequence $\{\theta_k\}_{k\in\mathbb{N}} \in \Theta$ converges to θ^*. $p(y_i|f(x_i))p(x_i|\theta_k)$ pointwise converges to $p(y_i|f(x_i))p(x_i|\theta^*)$. Note that $p(y_i|f(x_i))p(x_i|\theta_k) \leq e^{\bar{l}_i} \cdot U$ and every constant is integrable on the compact domain $\mathcal{X}_i$. By the continuity of $\log(\cdot)$ and Dominate Convergence Theorem, we conclude that

$$\lim_{k\to\infty} -\log \int_{\mathcal{X}_i} p(y_i|f(x_i))p(x_i|\theta_k)dx_i = -\log \int_{\mathcal{X}_i} p(y_i|f(x_i))p(x_i|\theta^*)dx_i \tag{7.20}$$

which means that $I^M_\lambda(f,\theta)$ is continuous in θ for any fixed f. This is sufficient to prove the continuity of $T(\theta)$. The corollary now follows from the compactness of Θ. □

7.2 GACV estimate of λ for weighted data

Consider the general weighted data penalized likelihood regression problem

$$I_\lambda(f) = -\frac{1}{n}\sum_{i=1}^{N} w_i \log p(y_i|f(x_i)) + \frac{\lambda}{2}J(f) \tag{7.21}$$

or equivalently

$$I_\lambda(f) = \frac{1}{n}\sum_{i=1}^{N} w_i[-y_i f(x_i) + b(f(x_i))] + \frac{\lambda}{2}J(f) \tag{7.22}$$

Here w_i stands for the weight of each data point, satisfying $\sum_{i=1}^{N} w_i = n$. In this case, the CKL criterion for tuning λ is

$$CKL(\lambda) = \frac{1}{n}\sum_{i=1}^{N} w_i[-\mu_i^* f_\lambda(x_i) + b(f_\lambda(x_i))] \tag{7.23}$$

where f_λ denotes the minimizer of (7.22) and μ_i^* denotes the expectation of $y_i^0 \sim p(y|f^*(x_i))$ independent of y_i. In order to derive the GACV estimate, denote $\tilde{y}_i = w_i y_i$ and $\tilde{b}(\cdot) = w_i b(\cdot)$. Then (7.22) can be written as

$$I_\lambda(f) = \frac{1}{n}\sum_{i=1}^{N} -\tilde{y}_i f(x_i) + \tilde{b}(f(x_i)) + \frac{\lambda}{2}J(f) \tag{7.24}$$

By using this trick, the GACV estimate of (7.23) can be obtained directly from Xiang and Wahba (1996)[16]:

$$GACV(\lambda) = OBS(\lambda) + \frac{trH}{n}\frac{\sum_{i=1}^{N} w_i y_i (y_i - \mu_\lambda(x_i))}{N - tr(W^{1/2}HW^{1/2})} \tag{7.25}$$

where

$$OBS(\lambda) = \frac{1}{n}\sum_{i=1}^{N} w_i[-y_i f_\lambda(x_i) + b(f_\lambda(x_i))] \tag{7.26}$$

is the observed likelihood, $\mu_\lambda(x_i) = b'(f_\lambda(x_i))$ is the expectation of $y_i^0 \sim p(y|f_\lambda(x_i))$, $W = W(f_\lambda)$ is the $N \times N$ diagonal matrix with $b''(f(x_i))$ in the iith position and $H = T[TW + n\Sigma_\lambda]^{-1}$ is so-called the influence matrix of the variational problem (7.22). Here T is the $N \times N$ diagonal matrix with w_i in the iith position and Σ_λ is the matrix when $J_\lambda(f)$ is represented as a quadratic form in $(f(x_1), f(x_2), ..., f(x_N))'$ (see Xiang and Wahba (1996)[16]). Following Lin *et al.* (2000)[9], we may derive the randomized version of (7.25):

$$ranGACV(\lambda) = OBS(\lambda) + \frac{1}{R}\sum_{r=1}^{R}\frac{\epsilon_r'(f^{Y+\epsilon_r} - f^Y)}{n}\frac{\sum_{i=1}^{N} w_i y_i (y_i - \mu_\lambda(x_i))}{\epsilon_r'\epsilon_r - \epsilon_r' W(f_\lambda^Y)(f^{Y+\epsilon_r} - f^Y)} \tag{7.27}$$

where $\epsilon_1, ..., \epsilon_R$ are R replicate vectors independently drawn from $N(0, \sigma_\epsilon^2 I_{N\times N})$. This randomized GACV excludes the evaluation of the influence matrix H and therefore is more computationally friendly.

References

[1] A. Berlinet and C. ThomasAgnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publishers, 2004.

[2] C. Gu. *Smoothing Spline ANOVA Models*. Springer, 2002.

[3] J. G. Ibrahim. Incomplete data in generalized linear models. *Journal of the American Statistical Association*, 85:765–769, 1990.

[4] J. G. Ibrahim, S. R. Lipsitz, and M.-H. Chen. Missing covariates in generalized linear models when the missing data mechanism is nonignorable. *J. Roy. Statist. Soc. Ser. B*, 61:173–190, 1999.

[5] G. Kimeldorf and G. Wahba. A correspondence between bayesian estimation of stochastic processes and smoothing by splines. *Ann. Math. Statist.*, 41:495–502, 1970.

[6] G. Kimeldorf and G. Wahba. Some results on tchebycheffian spline functions. *J. Math. Anal. Appl.*, 33:82–95, 1971.

[7] G. Kimeldorf and G. Wahba. Spline functions and stochastic processes. *Sankhya Ser*, A 32:173–180, 1971.

[8] A. Kurdila and M. Zabarankin. *Convex Functional Analysis (Systems and Control: Foundations and Applications)*. Birkhauser, 2005.

[9] X. Lin, G. Wahba, D. Xiang, F. Gao, R. Klein, and B. E. K. Klein. Smoothing spline anova models for large data sets with bernoulli observations and the randomized gacv. *Ann. Statist.*, 28:1570–1600, 2000.

[10] R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data (second edition)*. Wiley, New York, 2002.

[11] D. Nychka. Bayesian confidence intervals for a smoothing spline. *Journal of the American Statistical Association*, 83:1134–1143, 1988.

[12] F. O'Sullivan. The analysis of some penalized likelihood estimation schemes. Technical Report 726, Department of statistics, University of Wisconsin, Madison, WI, 1983.

[13] G. Wahba. Improper priors, spline smoothing and the problem of guarding against model errors in regression. *J. Roy. Statist. Soc. Ser. B*, 40:364–372, 1978.

[14] G. Wahba. Bayesian "confidence intervals" for the cross-validated smoothing spline. *J. Roy. Statist. Soc. Ser. B*, 45:133–150, 1983.

[15] G. Wahba. *Spline Models for Observational Data*. SIAM, Philadelphia, 1990.

[16] D. Xiang and G. Wahba. A generalized approximate cross validation for smoothing splines with non-gaussian data. *Statistica Sinica*, 6:675–692, 1996.