DEPARTMENT OF STATISTICS
University of Wisconsin
1210 West Dayton St.
Madison, WI 53706

TECHNICAL REPORT NO. 1083

October 10, 2003

Automatic Smoothing for Poisson Regression[1]

Ming Yuan [2]
Department of Statistics, University of Wisconsin, Madison WI

Key words and phrases: Penalized likelihood estimate, generalized approximate cross validation, unbiased risk estimate, Poisson regression.

[1]Research supported in part by NSF Grant DMS-0772292
[2]Email: yuanm@stat.wisc.edu

AUTOMATIC SMOOTHING FOR POISSON REGRESSION

Ming Yuan [1]

Department of Statistics

University of Wisconsin – Madison

1210 West Dayton Street

Madison, WI 53706

Keywords: Penalized likelihood estimate, generalized approximate cross validation, unbiased risk estimate, Poisson regression.

ABSTRACT

Adaptive choice of smoothing parameters for nonparametric Poisson regression (O'Sullivan *et. al.*, 1986) is considered in this paper. A computable approximation of the unbiased risk estimate (AUBR) for Poisson regression is introduced. This approximation can be used to automatically tune the smoothing parameter for the penalized likelihood estimator. An alternative choice is the generalized approximate cross validation (GACV) proposed by Xiang and Wahba (1996). Although GACV enjoys a great success in practice when applying for nonparametric logisitic regression, its performance for Poisson regression is not clear. Numerical simulations have been conducted to evaluate the GACV and AUBR based tuning methods.We found that GACV has a tendency to oversmooth the data when the intensity function is small. As a consequence, we suggest tuning the smoothing parameter using AUBR in practice.

[1]Email: yuanm@stat.wisc.edu

1 INTRODUCTION

Poisson regression is widely used to model the event count data (see Vermunt, 1996). A nonparametric estimate of the canonical parameter of a Poisson process based on penalized likelihood smoothing spline models was proposed by O'Sullivan, Yandell and Raynor (1986). In this paper, we are concerned with the adaptive choice of smoothing parameters in their smoothing spline models for nonparametric Poisson regression.

Suppose that $y_1, ..., y_n$ are n independent observations from a Poisson process with density of the form

$$f\left(y_i, \eta_0\left(x_i\right)\right) = \exp\left[y_i \eta_0\left(x_i\right) - e^{\eta_0(x_i)} - \ln\left(y_i!\right)\right], \tag{1.1}$$

where η_0 is the so-called canonical parameter and $\{x_i, i = 1, 2, ..., n\}$ are either univariate or multivariate covariates. We are interested in estimating $\eta_0\left(\cdot\right)$. When η_0 is of a linear form, it is the usual generalized linear model with a Poisson likelihood (McCullagh and Nelder, 1983). In this case, since a parametric form of η_0 is assumed, maximum likelihood methods may be used to estimate and asses the fitted models. Although the generalized linear model has been proved to be a very useful modeling approach in many applications, its linear assumption on η_0 is still very strict in many cases. Various parametric approaches have been proposed to allow more flexibility than this simple linear model (Wei, 1998). We will not review the general literature on parametric modeling of the count data. In contrast, we will focus on the nonparametric estimate of η_0 which allows η_0 to take on a more flexible form by only assuming that it is an element of some reproducing kernel Hilbert space $\mathcal{H}$ of smooth functions.

The smoothing spline model for Poisson regression was first introduced by O'Sullivan, Yandell and Raynor (1986). They used the penalized likelihood method to estimate η_0. Their smoothing spline estimator $\widehat{\eta}_\lambda\left(\cdot\right)$ of $\eta_0\left(\cdot\right)$ is defined as the minimizer in $\mathcal{H}$ of

$$-\sum_{i=1}^{n} l\left(y_i, \eta\left(x_i\right)\right) + \frac{n\lambda}{2} J\left(\eta\right). \tag{1.2}$$

where the smoothing parameter $\lambda \geq 0$ balances the tradeoff between minimizing the negative log likelihood function

$$L(y, \eta) \equiv -\sum_{i=1}^{n} l\left(y_i, \eta\left(x_i\right)\right) \equiv \sum_{i=1}^{n}\left[-y_i \eta\left(x_i\right) + e^{\eta(x_i)} - \ln\left(y_i!\right)\right], \tag{1.3}$$

and the "smoothness " $J\left(\eta\right)$. Here $J\left(\eta\right)$ is a quadratic penalty functional defined on $\mathcal{H}$.

Like the other smoothing spline models, the choice of the smoothing parameter λ considerably affects the performance of the estimate of η_0. If λ is very small, the estimator $\widehat{\eta}_\lambda$

would be very close to "interpolating" the observations. On the other hand if $J(\eta) = \int(\eta'')^2$ and $\lambda = \infty$, the estimator will boil down to an estimate of a generalized linear model. Thus, smoothing parameter selection is one of the most important practical issue for the nonparametric Poisson regression.

For regression with Gaussian type response, various criteria for selecting an optimal smoothing parameter have been proposed and extensively studied (see Wahba, 1990). Among which, cross-validation (CV, for short; Stone, 1974), generalized cross validation (GCV, for short; Craven and Wahba, 1979), Mallows C_p (Mallows, 1973) and their randomized versions (Girard, 1989,1991) are most commonly used. But for Poisson response, the situation is much more complicated because not only the mean but also the variance of the response are involved with the unknown function η_0. Another difficulty each smoothing parameter selection method should face is the computational problem. Even for a fixed smoothing parameter, we need iterations to get $\widehat{\eta}_\lambda$. Consequently, the tuning method must be computationally efficient.

Although many criteria for selecting smoothing parameters for Gaussian regression models have been extensively studied, the corresponding literature for generalized nonparametric regression is rather sparse. Two of the most often used techniques are a GCV based iterative procedure (Gu, 1990) and GACV (Xiang and Wahba, 1996). The first proposal for tuning smoothing parameters in non-Gaussian regression models is due to Gu (1990). He suggested adapting the GCV in an iterative procedure of solving (1.2). A major criticism of this approach is that it is not guaranteed to converge although in most applications it does. For that reason, the study here does not incorporate Gu's idea.

Since its introduction by Xiang and Wahba (1996), GACV has become the most popular tool to optimally choose the smoothing parameter for generalized spline-based regression. Although GACV was derived under a general setting where the conditional distribution of response variables given the covariates falls in the exponential family, most of its later applications focus on the Bernoulli data. Despite its great success for Bernoulli data, we are unclear about whether these good properties can be extended to Poisson regression.

An alternative approach to tune the smoothing parameter for the nonparametric Poisson regression is based on the unbiased risk (UBR, for short) estimate of a penalized likelihood estimator. Unlike the logisitic regression where a UBR does not exist according to Ye and Wong (1997), Poisson regression has an exact UBR. However, the problem is that naively using the UBR for tuning smoothing parameters is very computationally expensive.

The current paper has two purposes. First we want to derive a computable approximation

for the UBR so that it can be used to optimally choose smoothing parameters. We also want to answer the question how GACV can be applied to Poisson data, and how it compares to the AUBR for the Poisson regression.

In the next section, GACV for Poisson regression is briefly reviewed. An approximation of the unbiased risk estimate (AUBR) for Poisson regression is derived in Section 3. The computational problem for applying GACV and AUBR to the penalized likelihood smoothing spline model for Poisson regression is then formulized in Section 4. Several sets of simulations are conducted in Section 5 to evaluate both smoothing parameter tuning techniques. Finally, a real life example is given to illustrate the methods.

2 GACV

Denote $\widehat{\eta}$ as an estimator of η_0. The discrepancy between $\widehat{\eta}$ and the true parameter η_0 can be measured by the sample Kullback-Leibler (KL) distance

$$\begin{aligned} KL\left(\eta_0, \widehat{\eta}\right) &= \frac{1}{n}\sum_{i=1}^{n} E_z\left[z_i\left(\eta_0(x_i) - \widehat{\eta}(x_i)\right) - \left(e^{\eta_0(x_i)} - e^{\widehat{\eta}(x_i)}\right)\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[e^{\eta_0(x_i)}\left(\eta_0(x_i) - \widehat{\eta}(x_i)\right) - \left(e^{\eta_0(x_i)} - e^{\widehat{\eta}(x_i)}\right)\right], \end{aligned} \tag{2.1}$$

where z is an independent copy of $y = (y_1, \ldots, y_n)'$. Our goal is to adaptively choose an estimator such that the KL distance from $\widehat{\eta}$ to η_0 is minimized. Noting that term

$$\frac{1}{n}\sum_{i=1}^{n}\left[\eta_0(x_i)e^{\eta_0(x_i)} - e^{\eta_0(x_i)}\right]$$

is independent of the estimator $\widehat{\eta}$, we can simply consider the remaining terms of (2.1), which are the so-called comparative Kullback-Leibler (CKL) distance

$$CKL\left(\eta_0, \widehat{\eta}\right) = \frac{1}{n}\sum_{i=1}^{n}\left[-e^{\eta_0(x_i)}\widehat{\eta}(x_i) + e^{\widehat{\eta}(x_i)}\right]. \tag{2.2}$$

The involvement of the unknown parameter η_0 in CKL makes directly minimizing (2.2) infeasible. A commonly used technique to get around this problem is cross validation.

Define the leaving-out-one cross validation function $CV(\lambda)$ as

$$CV(\eta_0, \widehat{\eta}_\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left[-y_i\widehat{\eta}_\lambda^{(-i)}(x_i) + e^{\widehat{\eta}_\lambda(x_i)}\right], \tag{2.3}$$

where $\widehat{\eta}_\lambda^{(-i)}$ is the minimizer of (1.2) with the ith data point omitted. If η_0 is "smooth" and the data are dense, the leaving-out-one cross validation can be expected to be at least roughly unbiased for the CKL loss. Let

$$y^{(-i)} = \left(y_1, \ldots, y_{i-1}, e^{\widehat{\eta}_\lambda^{(-i)}(x_i)}, y_{i+1}, \ldots, y_n\right). \tag{2.4}$$

The leaving-out-one lemma (Xiang and Wahba, 1996) tells us that the minimizer of (1.2) with response vector $y^{(-i)}$ is also $\widehat{\eta}_\lambda^{(-i)}$. A linearization argument yields

$$\widehat{\eta}_\lambda(x_i) - \widehat{\eta}_\lambda^{(-i)}(x_i) \approx \frac{\partial \widehat{\eta}_\lambda(x_i)}{\partial y_i}(y_i - e^{\widehat{\eta}_\lambda^{(-i)}(x_i)}). \tag{2.5}$$

Thus,

$$\begin{aligned}
CV(\eta_0, \widehat{\eta}_\lambda) &= \frac{1}{n}\sum_{i=1}^{n}\left[-y_i\widehat{\eta}_\lambda^{(-i)}(x_i) + e^{\widehat{\eta}_\lambda(x_i)}\right] \\
&= L(y, \widehat{\eta}_\lambda) + \frac{1}{n}\sum_{i=1}^{n} y_i\left(\widehat{\eta}_\lambda(x_i) - \widehat{\eta}_\lambda^{(-i)}(x_i)\right) \\
&\approx L(y, \widehat{\eta}_\lambda) + \frac{1}{n}\sum_{i=1}^{n} y_i\frac{\partial \widehat{\eta}_\lambda(x_i)}{\partial y_i}(y_i - e^{\widehat{\eta}_\lambda^{(-i)}(x_i)}) \\
&= L(y, \widehat{\eta}_\lambda) + \frac{1}{n}\sum_{i=1}^{n} y_i\frac{\partial \widehat{\eta}_\lambda(x_i)}{\partial y_i}\frac{(y_i - e^{\widehat{\eta}_\lambda(x_i)})}{1 - \frac{e^{\widehat{\eta}_\lambda(x_i)} - e^{\widehat{\eta}_\lambda^{(-i)}(x_i)}}{y_i - e^{\widehat{\eta}_\lambda^{(-i)}(x_i)}}} \\
&\approx L(y, \widehat{\eta}_\lambda) + \frac{1}{n}\sum_{i=1}^{n} y_i\frac{\partial \widehat{\eta}_\lambda(x_i)}{\partial y_i}(y_i - e^{\widehat{\eta}_\lambda(x_i)})\frac{1}{1 - \frac{\partial \widehat{\eta}_\lambda(x_i)}{\partial y_i}e^{\widehat{\eta}_\lambda(x_i)}}
\end{aligned} \tag{2.6}$$

The last quantity of the above equation is called the approximate cross validation (ACV). Replacing $\partial\widehat{\eta}_\lambda(x_i)/\partial y_i$ by $tr(A(\widehat{\eta}_\lambda))$ and $e^{\widehat{\eta}_\lambda(x_i)}\partial\widehat{\eta}_\lambda(x_i)/\partial y_i$ by $tr(V(\widehat{\eta}_\lambda)^{1/2}A(\widehat{\eta}_\lambda)V(\widehat{\eta}_\lambda)^{1/2})$, we get the GACV for the Poisson regression.

$$GACV(\eta_0, \widehat{\eta}_\lambda) = L(y, \widehat{\eta}_\lambda) + \frac{tr(A(\widehat{\eta}_\lambda))}{n}\frac{\sum_{i=1}^{n} y_i(y_i - \exp(\widehat{\eta}_\lambda(x_i)))}{n - tr(V(\widehat{\eta}_\lambda)^{1/2}A(\widehat{\eta}_\lambda)V(\widehat{\eta}_\lambda)^{1/2})}, \tag{2.7}$$

where $A(\widehat{\eta}_\lambda) = \partial\widehat{\eta}_\lambda/\partial y$ and $V(\widehat{\eta}_\lambda)$ is a diagonal matrix with the $(i,i)th$ entry $e^{\widehat{\eta}_\lambda(x_i)}$. We usually call A the influence matrix or hat matrix for (1.2).

3 AUBR

Another way to avoid minimizing (2.2) is to minimize an unbiased estimator of CKL instead. Note that for any function f, the following identity holds (Hudson, 1978),

$$E_{y_i}\left[y_i f(y_i - 1)\right] = \sum_{k=0}^{\infty} e^{-\mu_i}\frac{\mu_i^k}{k!}k f(k-1)$$

$$
\begin{aligned}
&= \sum_{k=1}^{\infty} e^{-\mu_i} \frac{\mu_i^{k-1}}{(k-1)!} \mu_i f(k-1) \\
&= \mu_i \sum_{k=0}^{\infty} e^{-\mu_i} \frac{\mu_i^{k}}{k!} f(k) \\
&= \mu_i E_{y_i} \left[f(y_i) \right], \qquad (3.1)
\end{aligned}
$$

where $\mu_i = E(y_i) = \exp(\eta_0(x_i))$. Consequently, replacing f in (3.1) by $\widehat{\eta}$, an unbiased estimator for (2.2) is

$$
UBR\left(\eta_0, \widehat{\eta}\right) = \frac{1}{n} \sum_{i=1}^{n} \left[-y_i \widehat{\eta}^i(x_i) + e^{\widehat{\eta}(x_i)} \right] = L(y, \widehat{\eta}) + \sum_{i=0}^{n} y_i \left(\widehat{\eta}(x_i) - \widehat{\eta}^i(x_i) \right), \qquad (3.2)
$$

where $\widehat{\eta}^i$ is estimated in the same way as $\widehat{\eta}$ with response vector

$$
(y_1, \ldots, y_{i-1}, y_i - 1, y_{i+1}, \ldots, y_n)'.
$$

This unbiased risk estimator was first derived by Ye and Wong (1997) in a technical report.

Although the unbiased risk estimator (3.2) is very elegant, it is not directly applicable because it is usually very expensive to compute. For example, consider the penalized estimator. For each smoothing parameter λ, we must calculate $n+1$ penalized likelihood estimators $\widehat{\eta}_\lambda, \widehat{\eta}_\lambda^1, \ldots, \widehat{\eta}_\lambda^n$ so that $UBR\left(\eta_0, \widehat{\eta}_\lambda\right)$ could be obtained. This is impracticable even for medium sample sizes. Here, we shall try to find a good approximation of this UBR which will considerably reduce the computation.

Regard $\widehat{\eta}(x_i)$ as a continuously differentiable function of y_i. Using the mean value theorem, we have

$$
\widehat{\eta}(x_i) - \widehat{\eta}^i(x_i) = \frac{\partial \xi(x_i)}{\partial y_i}
$$

where ξ is the same estimator as $\widehat{\eta}$ with response vector $y - p1_i$ for some $0 \le p \le 1$ and 1_i is a vector with the ith entry being 1 and all the other elements being 0. Approximating $\partial \xi(x_i)/\partial y_i$ by $\partial \widehat{\eta}(x_i)/\partial y_i$, we get the following approximation of the UBR

$$
AUBR(\eta_0, \widehat{\eta}) \equiv L\left(y, \widehat{\eta}\right) + \frac{1}{n} \sum_{i=1}^{n} y_i \frac{\partial \widehat{\eta}(x_i)}{\partial y_i}. \qquad (3.3)
$$

Unlike GACV, the derivation of both UBR and AUBR is not limited to penalized likelihood spline estimator. In the next section, we shall restrict our attention to the penalized likelihood smoothing spline model. More details of the computational procedure will be formulized for this model and for different smoothing parameter tuning methods.

4 SMOOTHING PARAMETER SELECTION

For convenience, we shall use vector notation hereafter. For example, $y = (y_1, \ldots, y_n)'$, and $\eta_0 = (\eta_0(x_1), \ldots, \eta_0(x_n))'$. η, $\widehat{\eta}$, $\widehat{\eta}^i$ and other vectors are defined in the same manner.

The problem of adaptive choice of smoothing parameter λ can be described as finding the best approximation of η_0 within the class of penalized likelihood estimators $\mathcal{M} \equiv \{\widehat{\eta}_\lambda : \lambda \in \Lambda\}$. The difficulty is that we do not know η_0. Instead, we only observe random variables related to η_0. Either GACV or AUBR can serve as a proxy of the CKL distance from $\mathcal{M}$ to η_0.

Given λ, the computational problem is then to find η to minimize

$$I_\lambda(y, \eta) = -\frac{1}{n}\sum_{i=1}^{n} l(y_i, \eta_i) + \frac{\lambda}{2}\eta'\Sigma\eta, \tag{4.1}$$

Let $\widehat{\eta}_\lambda^{i,\delta}$ minimize $I_\lambda(y - \delta 1_i, \eta)$. Note that $\widehat{\eta}_\lambda$ minimizes $I_\lambda(y, \eta)$. Therefore,

$$\frac{\partial I_\lambda(y, \widehat{\eta}_\lambda)}{\partial \eta} = \frac{1}{n}(-y + \exp(\widehat{\eta}_\lambda)) + \lambda\Sigma\widehat{\eta}_\lambda = 0 \tag{4.2}$$

$$\frac{\partial I_\lambda(y - \delta 1_i, \widehat{\eta}_\lambda^{i,\delta})}{\partial \eta} = \frac{1}{n}\left(-y + \delta 1_i + \exp(\widehat{\eta}_\lambda^{i,\delta})\right) + \lambda\Sigma\widehat{\eta}_\lambda^{i,\delta} = 0 \tag{4.3}$$

Using a first order Taylor expansion to $\partial I_\lambda(y - \delta 1_i, \widehat{\eta}_\lambda^i)/\partial\eta$ at $(y, \widehat{\eta}_\lambda^i)$, we have the following equation:

$$\begin{aligned} 0 = \frac{\partial I_\lambda(y - \delta 1_i, \widehat{\eta}_\lambda^{i,\delta})}{\partial \eta} = \frac{\partial I_\lambda(y, \widehat{\eta}_\lambda)}{\partial \eta} &+ \frac{\partial^2 I_\lambda(y, \widehat{\eta}_\lambda)}{\partial \eta' \partial \eta}\left(\widehat{\eta}_\lambda^{i,\delta} - \widehat{\eta}_\lambda\right) - \delta\frac{\partial^2 I_\lambda(y, \widehat{\eta}_\lambda)}{\partial y' \partial \eta}1_i \\ &+ O\left(\left\|\widehat{\eta}_\lambda^{i,\delta} - \widehat{\eta}_\lambda\right\|^2 + \delta^2\right) \end{aligned} \tag{4.4}$$

If $\delta \to 0$, we get

$$\begin{aligned} \frac{\partial \widehat{\eta}_\lambda}{\partial y_i} &= -\left(\frac{\partial^2 I_\lambda(y, \widehat{\eta}_\lambda)}{\partial \eta' \partial \eta}\right)^{-1}\frac{\partial^2 I_\lambda(y, \widehat{\eta}_\lambda)}{\partial y' \partial \eta}1_i \\ &= (V(\widehat{\eta}_\lambda) + n\lambda\Sigma)^{-1}1_i. \end{aligned} \tag{4.5}$$

As defined before, $V(\widehat{\eta}_\lambda)$ is a diagonal matrix with the $(i,i)th$ element $\exp(\widehat{\eta}_\lambda(x_i))$. Thus, to compute GACV or AUBR, it suffices to evaluate Σ.

The representer theorem (Kimeldorf and Wahba, 1971) tells us that the exact minimizer of (1.2) has a finite representation when $J(\eta)$ is a semi norm in a reproducing kernel Hilbert space $\mathcal{H}$. A popular example is $J(\eta) = \int(\eta'')^2$. If $\mathcal{H}$ is decomposed into $\mathcal{H}_0 \oplus \mathcal{H}_1$, where $\mathcal{H}_0$

is the null space of J, then the minimizer of (1.2) in $\mathcal{H}$ has the following form

$$\widehat{\eta}_\lambda(x) = \sum_{v=1}^{m} d_v \phi_v(x) + \sum_{i=1}^{n} c_i K(x, x_i), \tag{4.6}$$

where $\{\phi_v\}$ is the basis of $\mathcal{H}_0$, and it is being assumed that an $n \times m$ matrix S with (i, v) entry $\phi_v(x_i)$ is of full column rank. $c = (c_1, \ldots, c_n)'$ satisfies $S'c = 0$, and K is the reproducing kernel for $\mathcal{H}_1$. For example, if we take $\mathcal{H}$ as the second order Sobolev space and $J(\eta) = \int (\eta'')^2$, then

$$K(u, v) = k_2(u)k_2(v) - k_4([u - v]),$$

where $n!k_n(u)$ is the nth Bernoulli polynomial and $[\tau]$ is the fractional part of τ. Furthermore, $J(\widehat{\eta}_\lambda) = c'Qc$ where Q is an $n \times n$ matrix with the $(i, j)th$ entry $K(x_i, x_j)$. Thus to minimize (1.2), it suffices to find c and $d = (d_1, \ldots, d_m)'$ that minimizes

$$-L(y, \sum_{v=1}^{m} d_v \phi_v(x_i) + \sum_{j=1}^{n} c_i K(x_i, x_j)) + \frac{n\lambda}{2} c'Qc. \tag{4.7}$$

In order to apply (4.4) to compute GACV and AUBR, we need to find Σ such that $\widehat{\eta}_\lambda' \Sigma \widehat{\eta}_\lambda = c'Qc$. It has been shown by Xiang and Wahba (1996) that such Σ has the form

$$\Sigma = \Delta \left(\Delta Q \Delta'\right)^{+} \Delta', \tag{4.8}$$

where Δ is any $n \times (n - m)$ matrix of orthonormal vectors whose columns are perpendicular to the columns of S, and $+$ represents the Moore-Penrose generalized inverse. In the case where Q is of full rank, we can also write

$$\Sigma = Q^{-1} - Q^{-1} S (S' Q^{-1} S)^{-1} S' Q^{-1}. \tag{4.9}$$

(4.5), (4.8) and (4.9) offer us a direct way to compute GACV and AUBR. However, when the sample size gets larger, the computation of Σ by (4.7) or (4.8) may not be stable. As an alternative, we suggest using the following procedure to approximate $\partial \widehat{\eta}_\lambda / \partial y$.

With a non quadratic penalized negative log likelihood (1.2), iterations are needed to calculate the penalized likelihood fit for a fixed smoothing parameter. $\theta = (\theta_1, ..., \theta_N)'$ to minimize

$$I_\lambda = -\sum_{i=1}^{n} l\left(y_i, \eta_i\left(\theta\right)\right) + \frac{n\lambda}{2} \theta' \Sigma_\theta \theta, \tag{4.10}$$

where

$$\eta_i\left(\theta\right) = \sum_{j=1}^{N} \theta_j B_j\left(x_i\right)$$

and Σ_θ is defined by

$$\theta'\Sigma_\theta\theta = J\left(\sum_{j=1}^{n}\theta_j B_j\right).$$

Since I_λ is a convex function of θ, we may compute θ via a Newton-Ralphson iteration. Define $w_i = \exp(\eta(x_i)), u_i = y_i - w_i$. $\theta^{[k]}$ (the kth estimate of θ) can be updated by:

$$\theta^{[k+1]} = \min_\theta \left[\frac{1}{n}\sum_{i=1}^{n} w_i^{[k]}\left(y_i^{[k]} - \eta_i(\theta)\right)^2 + \frac{\lambda}{2}\theta'\Sigma_\theta\theta\right], \tag{4.11}$$

where $y_i^{[k]} = \eta_i(\theta^{[k]}) - u_i^{[k]}/w_i^{[k]}$ and $u_i^{[k]}, w_i^{[k]}$ are the values of u, w evaluated at $\theta^{[k]}$. The influence matrix for (4.10) is defined as an $n\times n$ matrix $A^{[k+1]}$ such that $\eta\left(\theta^{[k+1]}\right) = A^{[k+1]}y^{[k]}$. Denote $A^{[\infty]}$ the influence matrix after the convergence of this algorithm. $A^{[\infty]}$ is a natural approximation of $\partial\theta^{[\infty]}/\partial y^{[\infty]}$. The chain rule gives us

$$A(\widehat{\eta}_\lambda) = \frac{\partial\widehat{\eta}_\lambda}{\partial y} = \frac{\partial\widehat{\eta}_\lambda}{\partial y^{[\infty]}}\frac{\partial y^{[\infty]}}{\partial y} \approx A^{[\infty]}V(\widehat{\eta}_\lambda)^{-1}, \tag{4.12}$$

Then, A can be evaluated via an efficient and stable algorithm for computing $A^{[k]}$ (Wahba, 1990).

5 SIMULATIONS

Several sets of simulations will be presented in this section to evaluate the performance of GACV and AUBR as smoothing parameter tuning techniques for nonparametric Poisson regression.

5.1 GACV and AUBR as Approximations to CKL

As argued before, AUBR approximates an unbiased estimator of CKL. The main motivation for GACV is also to approximate the CKL. Asymptotically, if η_0 is "smooth", one may expect the GACV curve and AUBR curve ($GACV(\eta_0, \widehat{\eta}_\lambda)$ or $AUBR(\eta_0, \widehat{\eta}_\lambda)$ vs λ) to be close to the CKL curve. The first set of simulations is conducted to evaluate how close the GACV and AUBR computed via either (4.9) or (4.12) are to the CKL curve. For this purpose, we consider the following test function

$$\eta_0(x_i) = 2\sin(2\pi x_i), \qquad i = 1, \ldots, 100,$$

where $x_i = (i - 0.5)/100$. $y_i, i = 1, \ldots, 100$ are independently sampled from a Poisson distribution with intensity $\exp(\eta_0(x_i))$. The above experiment has been repeated for 50

times. For each simulated dataset, the CKL curve, the GACV and AUBR curves computed via both (4.9) and (4.12) have been recorded.

The top left panel of figure 1 gives the average curves from 50 replications of the above experiment. AUBR, calculated via either (4.9) or (4.12), provides an unbiased estimator of the CKL from this figure. Although the average GACV curves are not so close to CKL as the AUBR curves, they do capture the main shape of CKL curves in average. To get a better understanding of the individual behaviour of each criteria, the remaining five panels in Figure 1 give the CKL curve, the GACV and AUBR curve computed by either (4.9) or (4.12) for each of the 50 replications. An interesting feature is that GACV and AUBR from (4.9) are much more variable than those from (4.12). For this reason, we shall use (4.12) to evaluate GACV and AUBR in the rest of the paper. Another interesting observation from Figure 1 is that GACV penalizes small smoothing parameters too much.

To investigate the possible dependence on the magnitude of η_0 of the performance of AUBR and GACV, we also conduct the following simulation. Consider the following 5 test functions

$$\begin{aligned}
\eta_0(x_i) &= 2\sin(2\pi x_i) + 3; \\
\eta_0(x_i) &= 2\sin(2\pi x_i) + 2; \\
\eta_0(x_i) &= 2\sin(2\pi x_i) + 1; \\
\eta_0(x_i) &= 2\sin(2\pi x_i); \\
\eta_0(x_i) &= 2\sin(2\pi x_i) - 1.
\end{aligned}$$

Again, the sample size is 100 and covariate x is defined as before. For each test function, the previous experiment is repeated 50 times. The average CKL curve, AUBR curve and GACV curve are depicted in Figure 2. One can see that both AUBR and GACV are very close to the average CKL for the first three test functions. For the last two test functions, however, biases appear for both criteria. AUBR tends to give too small penalty for small smoothing parameters. GACV, on the other hand, gives too much penalty for small smoothing parameters. This observation might suggest a tendency of GACV to oversmooth when the intensity function is small.

5.2 Tuning Smoothing Parameters with GACV and AUBR

The performance of GACV and AUBR based smoothing parameter tuning methods depends on how well they approximate the CKL around the optimal smoothing parameter. Fortunately, although GACV and AUBR are biased estimators of CKL when the intensity function

is very small, they are still reasonably unbiased around the minima of the CKL curve in general. To compare the smoothing parameter tuning method based on GACV and AUBR, we conducted two sets of simulations.

The first set consists of 4 test functions of different magnitude. To get a better impression of the magnitude of the intensity functions, we present them in terms of the intensity function $\mu = exp(\eta_0)$.

$$\begin{aligned}
\mu_1(x) &= \exp\left(\frac{sin(2\pi x)}{2 - sin(2\pi x)}\right) \\
\mu_2(x) &= 7 + 7x^5 + 7(x-1)^5 \\
\mu_3(x) &= 10000\left(x^8(1-x)^2 + x^2(1-x)^8\right) \\
\mu_4(x) &= 5\exp\left(-100(x-0.75)^2\right) + 480(x-0.75)^2
\end{aligned}$$

The last three test functions were used by Climov, Hart and Simar (2002) and they have different signal to noise ratios, which are defined as

$$SNR = \frac{\max_{0\le x\le 1}\mu(x) - \min_{0\le x\le 1}\mu(x)}{\sqrt{E\mu(X)}},$$

where the expectation is taken over covariate X, in our example, it is uniformly distributed over $[0, 1]$. The values of SNR are 9,30,5 for μ_2, μ_3, μ_4, respectively. All 4 test functions are displayed in Figure 3.

For each test function, the following procedure was repeated for one hundred times:

(1) One hundred $x's$ are drawn from $U[0, 1]$.

(2) For each covariate $x_i, i = 1, \ldots, 100$, y_i is sampled from a Poisson distribution with intensity $\mu_k(x_i)$ for $k = 1, \ldots, 4$.

(3) Estimate η by $\widehat{\eta}_\lambda$ with λ chosen to minimize the GACV score, compute the true KL distance from the estimator of the true function.

(4) Estimate η by $\widehat{\eta}_\lambda$ with λ chosen to minimize the AUBR score, compute the true KL distance from the estimator of the true function.

Figure 4 presents the scatter plots of $KL\left(\eta_0, \widehat{\eta}_{\lambda_{GACV}}\right)$ versus $KL\left(\eta_0, \widehat{\eta}_{\lambda_{AUBR}}\right)$. For test functions μ_2 and μ_4, AUBR performs better than GACV. For test functions μ_1 and μ_3, AUBR and GACV perform similarly. But AUBR still is slightly better than GACV in terms of outperforming GACV more often.

In another set of simulations, we want to extend our comparison to two-dimensional test functions. Two test functions were used:

$$\begin{aligned}\mu_5(x_1,x_2) &= \exp(2\sin(2\pi x_1)-\sin(2\pi x_2))\\ \mu_6(x_1,x_2) &= \exp(8(\exp(-\frac{1}{3.38\left((x_1-2)^2+x_2^2\right)})+\exp(-\frac{1}{3.38\left((x_1+2)^2+x_2^2\right)}))-1)-46\end{aligned}$$

200 triples of (x_1, x_2, y) are sampled according to the following laws: (x_1, x_2) is independently sampled from $U[0,1]^2$, then y is sampled from a Poisson distribution with intensity $\mu(x_1, x_2)$. After getting these 200 observations, we estimate the log intensity function by a penalized likelihood estimator with thin plate splines where

$$J(\eta)=\int_0^1\int_0^1\left[\left(\frac{\partial^2\eta}{\partial x_1^2}\right)^2+\left(\frac{\partial^2\eta}{\partial x_1\partial x_2}\right)^2+\left(\frac{\partial^2\eta}{\partial x_2^2}\right)^2\right]dx_1dx_2$$

and smoothing parameter λ chosen to minimize either AUBR or GACV. We repeat this experiment one hundred times.

The range of μ_5 is from 0.05 to 20, while the range of μ_6 is from 0.6 to 8.7. To investigate the possible effects of the magnitude of the intensity functions, we also consider two other test functions with intensities

$$\mu_7(x_1,x_2)=7.4\times\mu_5(x_1,x_2),\qquad \mu_8(x_1,x_2)=7.4\times\mu_6(x_1,x_2).$$

We refer to μ_5 and μ_6 as test functions with small intensities and to μ_7 and μ_8 as test functions with large intensities. The top left panel of Figure 5 depicts the common shape of μ_5 and μ_7. The top right panel gives the common shape of μ_6 and μ_8. The KL distance based comparisons between GACV and AUBR are given in the remaining four panels. When the intensity functions are relatively small, AUBR outperforms GACV. When the intensity gets larger, this difference diminishes.

5.3 Conclusion

From the above simulations, we find that both GACV and AUBR are good approximations of the CKL loss and both can be applied to choose smoothing parameters for nonparametric Poisson regression if the intensity function is large. However, when the intensity functions are small, GACV has a tendency to overestimate the CKL loss for small smoothing parameters. As a consequence, we conjectured that oversmoothing might occur for small intensity functions if we choose smoothing parameters by GACV. This is confirmed by the observation that AUBR outperforms GACV when intensity functions are small.

6 AN APPLICATION

In this section, we apply the smoothing parameter tuning methods based on GACV and AUBR to a real dataset. The dataset represents a time series of medical count data. The surveyors recorded the number of (new) polio infections per month during the time from January 1st, 1970 till December 12th, 1987. The data were recorded by the US Ministry for Health. The dataset contains the number of infections per month, and thus a total of $18 \times 12 = 216$ count data. We are interested in how the incidence rate evolves over time.

For this purpose, we consider a Poisson regression where the canonical parameter η_0 is a function of time. Figure 6 gives the AUBR curve and GACV curve for the Poisson regression. Also provided are the observed counts and the fitted intensity functions using the smoothing parameter chosen by minimizing the AUBR or GACV. From the figure, we observe a suspicious oversmoothing of GACV based tuning technique. This further confirms what we discovered in the simulations.

ACKNOWLEDGMENTS

This work was supported in part by NSF Grant DMS-0072292. The author thanks his PhD advisor, Prof. Grace Wahba for suggesting the problem and her guidance during the preparation of this paper.

References

[1] Craven, P., Wahba, G. (1979), Smoothing noisy data with spline functions. Estimating the correct degree of smoothing by the method of generalized cross-validation, *Numer. Math.* **31(4)**, 377–403.

[2] Girard, D. A. (1989), A fast "Monte Carlo cross-validation" procedure for large least squares problems with noisy data, *Numer. Math.* **56(1)**, 1–23.

[3] Girard, D. A. (1991), Asymptotic optimality of the fast randomized versions of GCV and C_L in ridge regression and regularization, *Ann. Statist.* **19(4)**, 1950–1963.

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

[5] Hudson, H. M. (1978), A Natural Identity for Exponential Families with Application in Multiparameter Estimation, *Ann. Statist.* **6**, 473-484.

[6] Kimeldorf, G. and Wahba, G. (1971), Some Results on Techebycheffian Spline Functions, *J. Math. Anal. Applic.*, **33**, 82-95.

[7] Mallows, C. L. (1973), Some Comments on C_p, *Technometrics* **15**, 661–675.

[8] McCullagh, P., Nelder, J. A. (1983), *Generalized linear models*, **Monographs on Statistics and Applied Probability**, London: Chapman & Hall.

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

[10] Stone, M. (1974), Cross-validation and multinomial prediction, *Biometrika* **61**, 509–515.

[11] Vermunt, J. K. (1996), *Log-linear event history analysis*, **Series on Work and Organization**, Tilburg: Tilburg University Press.

[12] Wahba, G. (1990), *Spline models for observational data*, **CBMS-NSF Regional Conference Series in Applied Mathematics, 59**, Philadelphia: Society for Industrial and Applied Mathematics.

[13] Wei, B. C. (1998), *Exponential family nonlinear models*, **Lecture Notes in Statistics, 130**, Singapore: Springer-Verlag Singapore.

[14] Xiang, D. and Wahba, G. (1996), A generalized approximate cross validation for smoothing splines with non-Gaussian data, *Statist. Sinica* **6(3)**, 675–692.

[15] Ye, J. and Wong, W. H. (1997), Evaluation of Highly Complex Modeling Procedures with Binomial and Poisson Data, *Technical Report*, Department of Statistics, University of Chicago.

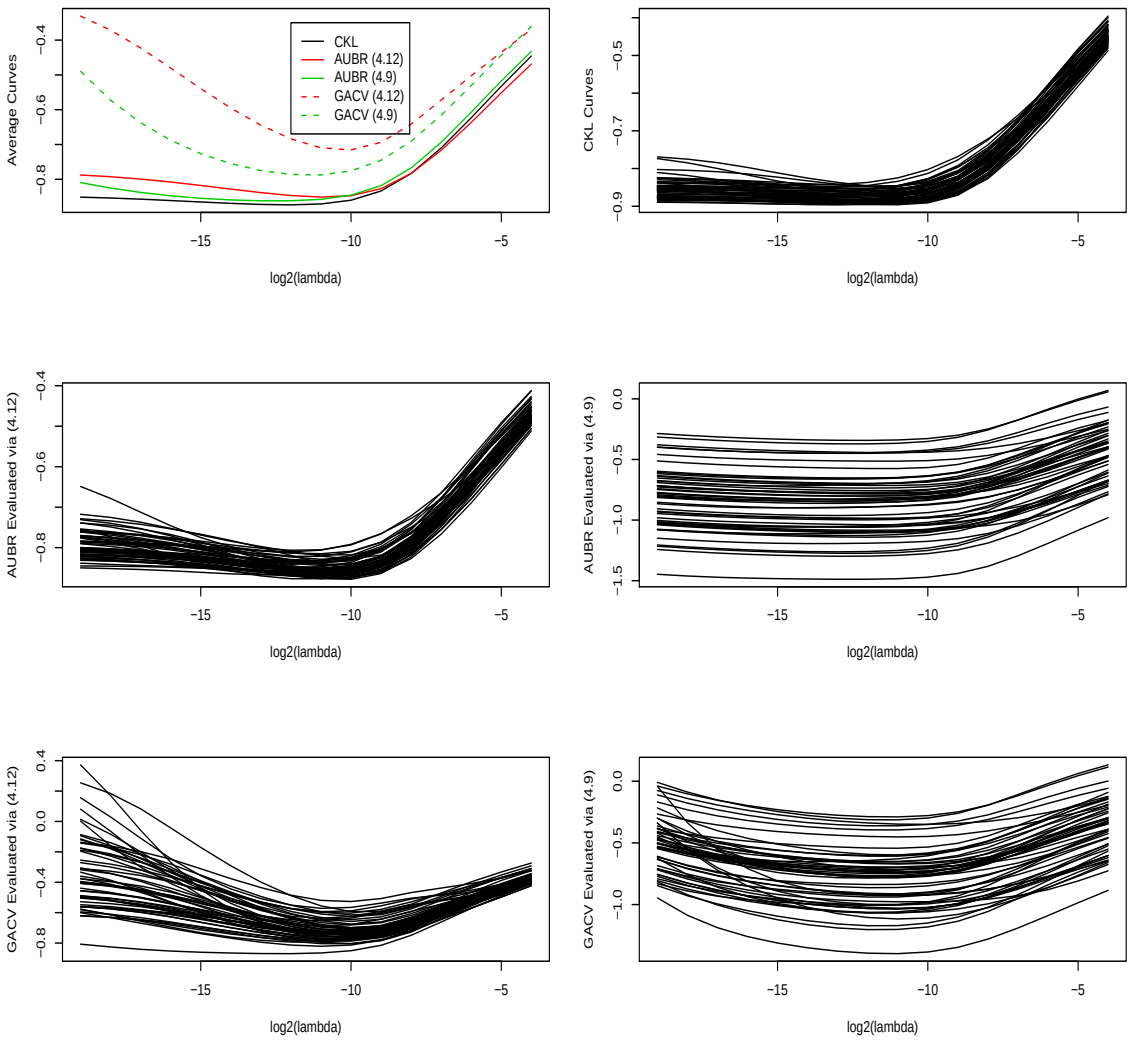

Figure 1: AUBR Curve, GACV Curve and CKL Curve: 50 simulated datasets are generated. The GACV and AUBR curves are computed via either (4.9) or (4.12). The CKL curves are also recorded for each dataset. In the top left panel, the average curves are given. In the rest 5 panels, CKL, GACV and AUBR via (4.9) and (4.12) curves from all 50 replications are given.

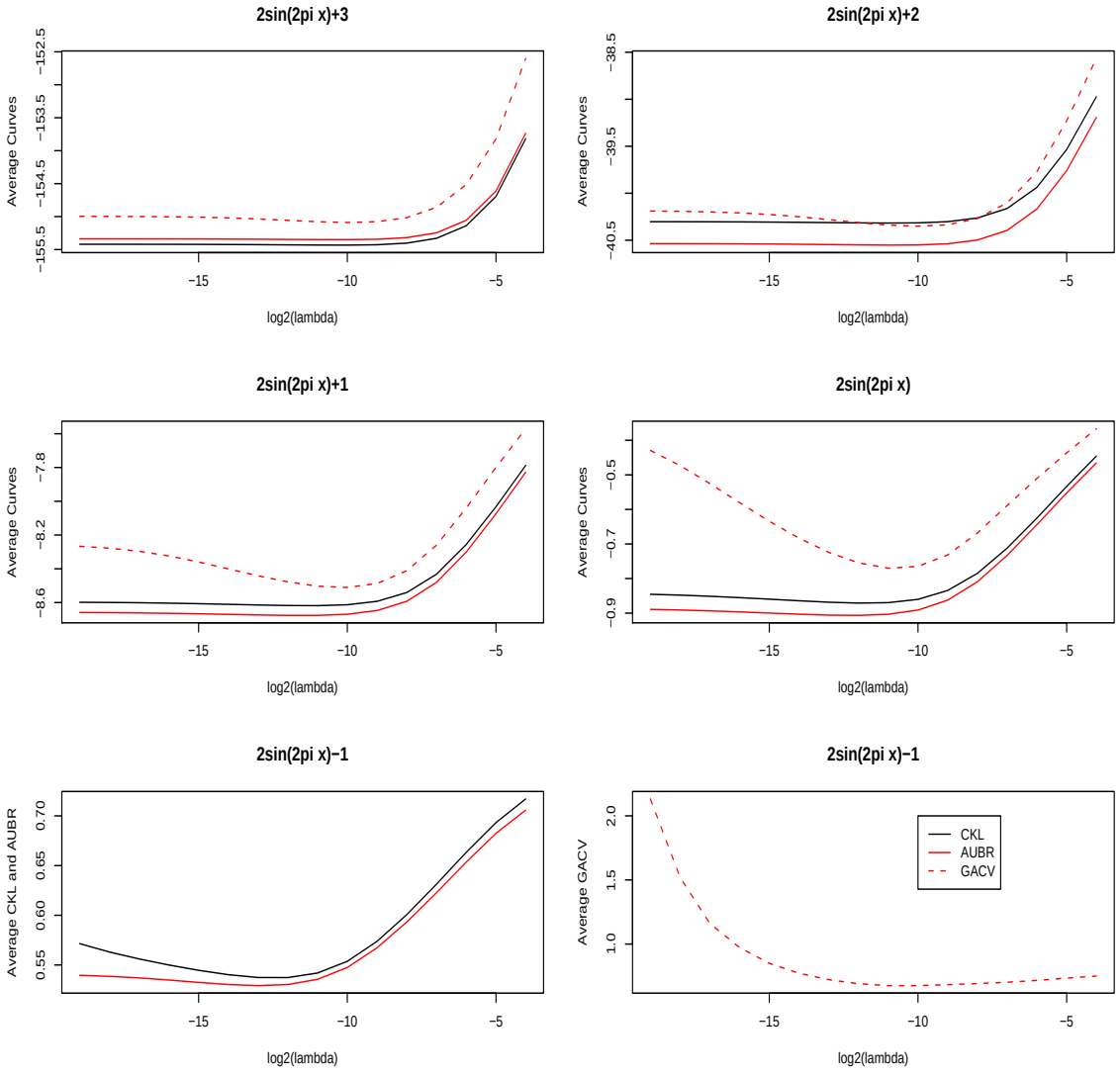

Figure 2: AUBR and GACV Approximate CKL: 50 simulated datasets are generated for 6 test functions with different magnitude. The GACV and AUBR curves are computed via (4.12). The CKL curves are also recorded for each dataset. The average curves for each test function are provided. In each panel, black solid line corresponds to average CKL curve, red solid line stands for average AUBR curve and red dashed line for average GACV curve.

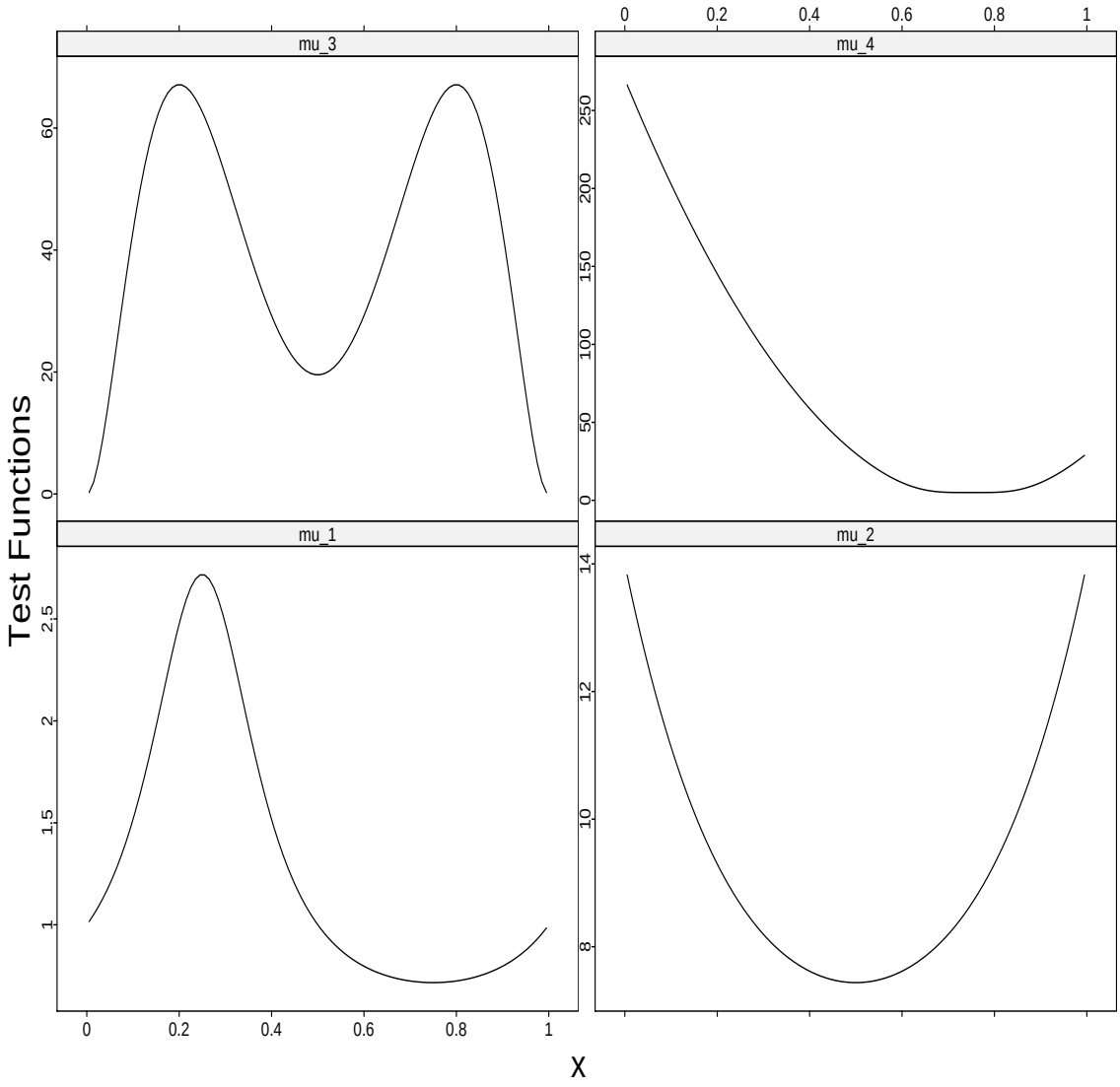

Figure 3: Test Functions: these are the four test functions we used to compare AUBR with GACV. They correspond to different magnitudes.

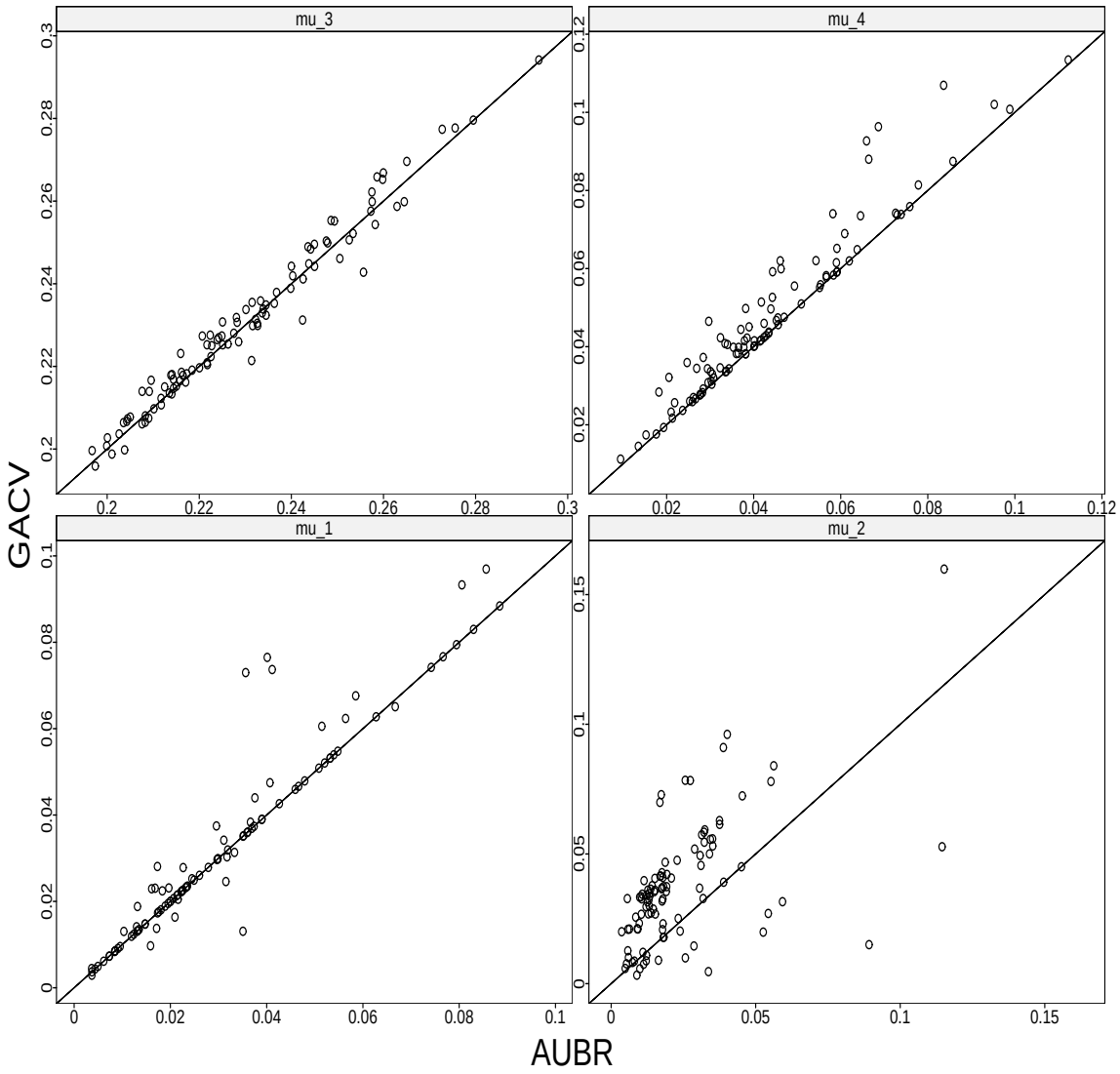

Figure 4: KL Comparison: For each of the four test functions presented in Figure 3, 100 sets of $(y_i, x_i), i = 1, \ldots, 100$ are sampled as described in Section 5. The true KL distances from the true intensity function are recorded for AUBR and GACV based fitting.

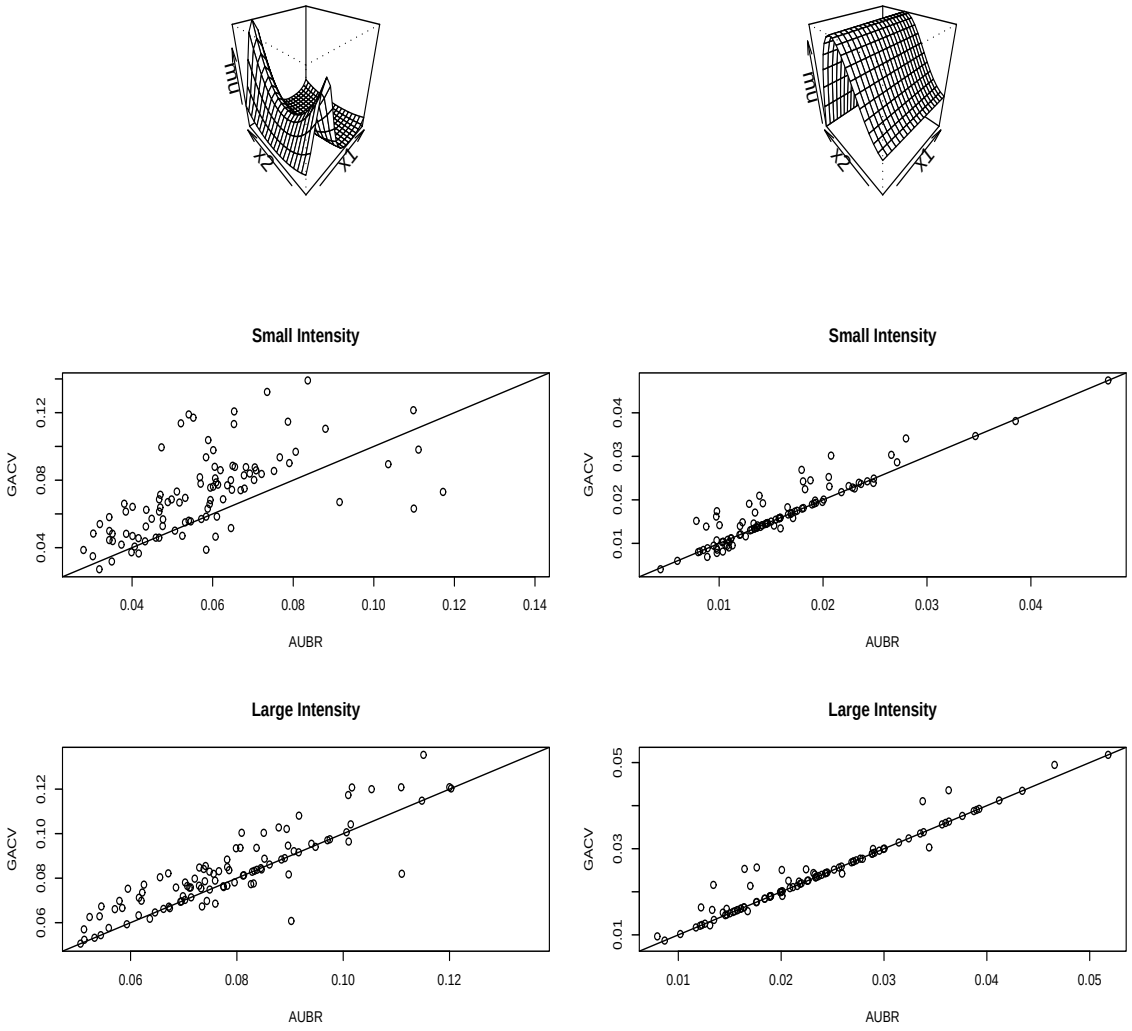

Figure 5: Two-Dimensional Comparison: Four 2-dimensional test functions are used to compare AUBR and GACV. Two of these test functions share the same shape depicted in the top left panel. They correspond to a relatively small intensity function and large intensity function. The rest two panel in the left column are the KL distance based test for these two test function respectively. The rest panel are designed in the fashion.

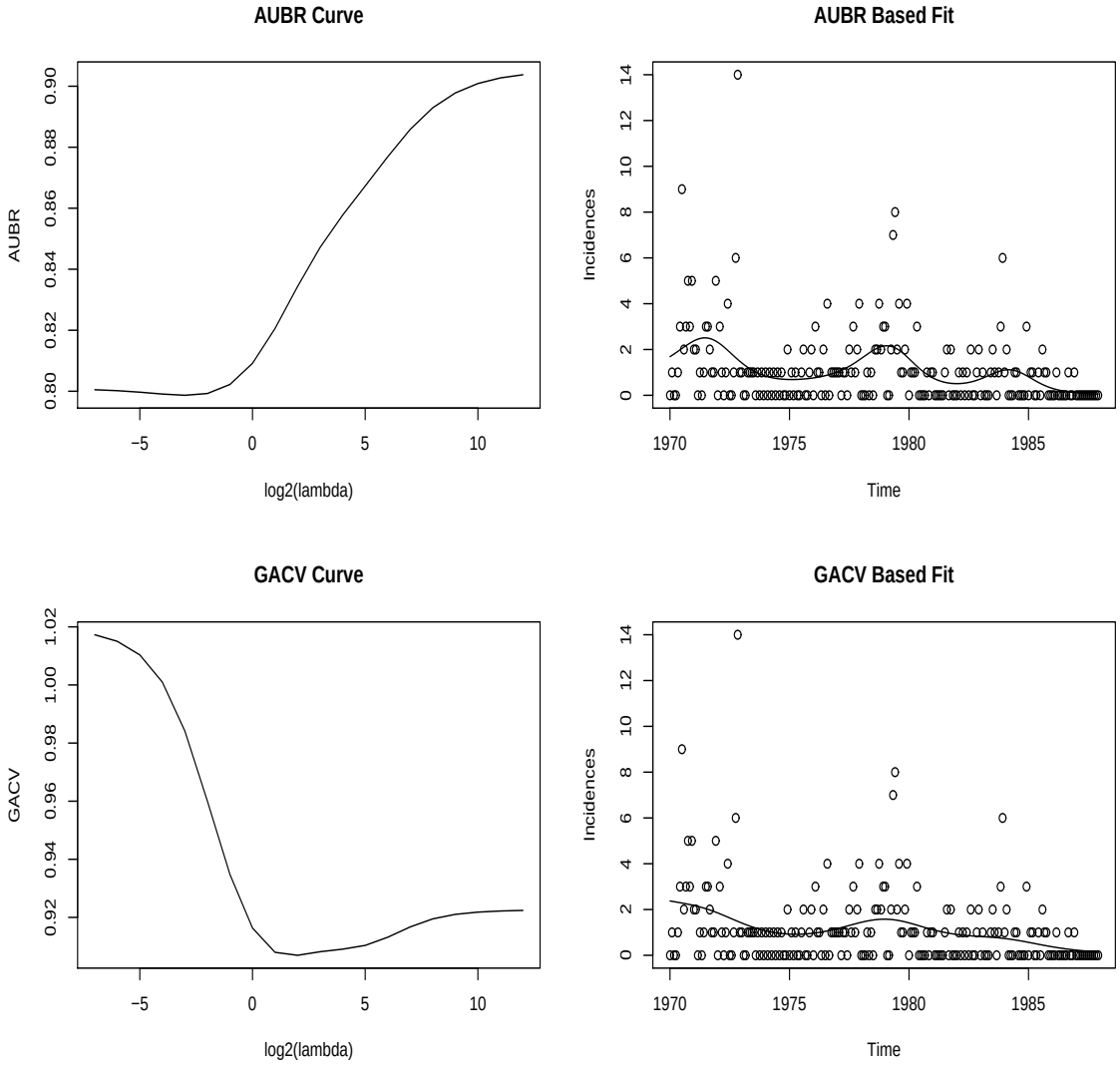

Figure 6: Polio Infection Data: AUBR and GACV values for spline estimator with different smoothing parameters for the polio infection data by modeling the log monthly mortality rate as a smooth function of time are given in the left column. Penalized likelihood estimator for the polio infection data with smoothing parameter chosen by minimizing the AUBR or GACV score are given in the right panel.