AUTOMATIC SMOOTHING AND ESTIMATION IN SINGLE INDEX POISSON REGRESSION

Daniela Climov [1], Jeffrey Hart [2], Léopold Simar[1]

ABSTRACT. We address the problem of smoothing parameter (h) selection when estimating the direction vector (β_0) and the link function in the context of semiparametric, single index Poisson regression. The single index Poisson model (PSIM) differs from the classical nonparametric setting in two ways: first, the errors are heteroscedastic, and second, the direction parameter is unknown and has to be estimated. We propose two simple, automatic rules for simultaneously estimating β_0 and h in a PSIM. The first criterion, called weighted least squares (WLS_2), estimates the Kullback-Leibler risk function and has a penalty term to prevent undersmoothing in small samples. The second method, termed double smoothing (DS), is based on the estimation of an L_2 approximation of the Kullback-Leibler risk and makes use of a double smoothing idea as in Wand and Gutierrez (1997). Simulations are used to investigate the behavior of various criteria in the PSIM context. Our weighted least squares and double smoothing methods out-perform both a Kullback-Leibler version of cross-validation and the weighted least squares cross-validation criterion proposed by Härdle, Hall and Ichimura (1993).

KEYWORDS. Poisson data, single index model, risk estimation, kernel methods, nonparametric regression.

SHORT TITLE. Single Index Poisson Regression.

---

[1]Institut de Statistique, Université Catholique de Louvain, B 1348 Louvain-la-Neuve, Belgium. Constructive comments on previous versions from Michel Delecroix and research support from "Projet d'Actions de Recherche Concertées" (No. 98/03–217) from the Belgian Government are acknowledged.

[2]Department of Statistics, Texas A&M University, College Station, TX 77843, USA. Research supported by NSF Grant DMS 99-71755.

Introduction

In the context of high-dimensional regression models, single index models are more general than certain linear models (McCullagh and Nelder, 1989), but still provide a way of reducing the dimension of the predictor variable, thus avoiding the so-called "curse of dimensionality." A single index regression model for describing the dependence of a scalar variable Y_i upon a p-variate vector X_i has the form:

$$Y_i = g(\beta_0 X_i) + \epsilon_i, \qquad i = 1, \ldots, n, \tag{0.1}$$

where ϵ_i is a random variable with zero mean, conditional on X_i, β_0 is the "true" direction p-vector of unknown parameters, and g is the unknown link function. The scalar $\beta_0 X_i$ is called the *index*.

It has been proven that under the single index model, the vector β_0 can be estimated at an optimal parametric rate of $n^{-1/2}$, i.e., as if g were known (see Klein and Spady, 1993; Newey and Stoker, 1993; Sherman, 1994; Horowitz and Härdle, 1996). The nonparametric estimator of the link function g is constructed from a p-dimensional predictor variable, but it achieves the optimal one-dimensional nonparametric rate. For example, if g is assumed r times differentiable, then, under regularity conditions, the rate of $n^{-r/(2r+1)}$ is attained (Härdle and Stoker, 1989). The fact that these rates are independent of p, the dimension of the vector X of explanatory variables, illustrates well how the single index model avoids the curse of dimensionality.

In order to guarantee the good asymptotic properties just mentioned, the bandwidth parameter h used in the nonparametric smoother should be optimal. But the optimal bandwidth depends on g and β_0, and therefore data-driven methods of estimating this bandwidth are necessary. Such methods are the topic of this paper.

Section 1 reviews some existing methods for selecting a bandwidth parameter in the general setting of nonparametric regression. In the context of the Single Index Poisson regression Model (PSIM), Section 2 presents the derivation of cross-validation and double smoothing-based methods for selecting a bandwidth and estimating the direction vector of parameters. Finally, in Section 3 the behavior of the proposed estimators is studied by means of simulation.

1 Data-driven bandwidth selection methods

For notational simplicity, suppose for the moment that β_0 is known and denote by Z the index $\beta_0 X$. The regression function g is to be estimated from data (Z_i, Y_i), $i = 1, \ldots, n$, generated by the model

$$Y_i = g(Z_i) + \epsilon_i, \qquad i = 1, \ldots, n, \tag{1.2}$$

where ϵ_i are independent zero-mean random variables with variance $\sigma^2(Z_i)$, conditional on Z_i. We denote a nonparametric estimator of g by $\hat{g}_h$, where h is the smoothing parameter. In this paper we consider Nadaraya-Watson[1] (Nadaraya, 1964 and Watson, 1964) nonparametric estimates of g having the form

$$\hat{g}_h(z) = \frac{\sum_{j=1}^n Y_j K_h(z - Z_j)}{\sum_{j=1}^n K_h(z - Z_j)}, \tag{1.3}$$

where $K_h(x) = h^{-1} K(x/h)$ and K is a fixed kernel function (e.g., a probability density that is symmetric about 0).

The Nadaraya-Watson estimator belongs to the class of linear smoothers (such as local polynomials, splines, wavelet estimators), in the sense that

$$\hat{\mathbf{g}}_h = \mathbf{H} Y,$$

where

$$\hat{\mathbf{g}}_h = \begin{pmatrix} \hat{g}_h(Z_1) \\ \vdots \\ \hat{g}_h(Z_n) \end{pmatrix}$$

and the matrix $\mathbf{H}$, which depends on $Z = (Z_1, \ldots, Z_n)^T$ but not on $Y = (Y_1, \ldots, Y_n)^T$, is commonly called the hat matrix or smoother matrix. In the context of multiple linear regression, the trace of the hat matrix is equal to the number of regressors in the model, i.e., p, the dimension of the explanatory variable. By analogy, in the context of nonparametric regression, trace($\mathbf{H}$) may be interpreted as the number of effective parameters used in the smoothing fit (see Hastie and Tibshirani, 1990, Section 3.5). Thus, when h is very small,

[1] Any linear smoother may be used, e.g. the local linear estimator.

trace($\mathbf{H}$) tends to n, which corresponds to the situation where the estimate interpolates the data. When h tends to infinity, trace($\mathbf{H}$) tends to 1, corresponding to $\hat{Y} = \bar{Y}$.

Crucial to the performance of $\hat{g}_h$ is the choice of bandwidth h. One means of choosing h is the so-called plug-in method, which is typically based on the mean integrated squared error:

$$\text{MISE}(\hat{g}_h|Z_1,\ldots,Z_n) = E\left\{\int [\hat{g}_h(z) - g(z)]^2 f(z)\,dz \Big| Z_1,\ldots,Z_n\right\}, \tag{1.4}$$

where f is the density of Z_i and the expectation is with respect to the true conditional distribution of $(Y_1,\ldots,Y_n)$ given $(Z_1,\ldots,Z_n)$. When $\hat{g}_h$ is a Nadaraya-Watson estimator and g has two continuous derivatives, the smoothing parameter that minimizes an asymptotic approximation of MISE is

$$h_0^a = C_{model}\, C_K\, n^{-1/5}, \tag{1.5}$$

where

$$C_{model} = \left[\frac{\int \sigma^2(z)dz}{\int \left[g''(z) + \frac{2g'(z)f'(z)}{f(z)}\right]^2 f(z)dz}\right]^{1/5}$$

and

$$C_K = \left[\frac{\int K^2(u)du}{\left(\int u^2 K(u)du\right)^2}\right]^{1/5}.$$

A plug-in bandwidth $\hat{h}_0^a$ is given by the right-hand side of (1.5), with estimates of the unknown parameters "plugged into" C_{model} to obtain $\hat{C}_{model}$. Such a scheme has been proposed for Gasser-Müller type estimators by Gasser, Kneip and Köhler (1991) and for local polynomial estimators by Fan and Gijbels (1995).

Despite having some good asymptotic properties, plug-in selectors target h_0^a (the minimizer of an asymptotic approximation of MISE) rather than h_0 (the minimizer of MISE itself), and so it is not always clear how well they behave in small or moderate samples. Also, from the practical point of view, plug-in selectors have the following disadvantage. Estimating C_{model} involves nonparametric estimation of a functional of g'', g', f and f'. This means, in the case of a Nadaraya-Watson estimator, that to obtain the plug-in bandwidth $\hat{h}_0^a$, four other, preliminary bandwidths have to be chosen. Furthermore, in the case of the single index model, g'' and g' depend on the unknown direction parameter β_0, so a preliminary direction has to be chosen too. After estimating h_0^a, another criterion has to be used for estimating β_0.

A rather straightforward method of bandwidth selection is cross-validation (Stone, 1974), the idea of which is to use a part of the data to construct an estimate of the regression model and then to predict the rest of the data with this estimate. The most often used form of cross-validation is the "least squares leave-one-out" cross-validation criterion:

$$\mathrm{LSCV}(h) = \frac{1}{n}\sum_{i=1}^{n}[Y_i - \hat{g}_h^i(Z_i)]^2, \tag{1.6}$$

where $\hat{g}_h^i(z)$ denotes an estimate of g computed without the ith data point. It can be shown that LSCV(h) is essentially an unbiased estimator of an empirical version of MISE, called mean average squared error:

$$\mathrm{MASE}(\hat{g}_h|Z_1,\ldots,Z_n) = E\left\{\frac{1}{n}\sum_{i=1}^{n}[\hat{g}_h(Z_i) - g(Z_i)]^2\Big|Z_1,\ldots,Z_n\right\}. \tag{1.7}$$

Alternative bandwidth selection methods may be based on minimization of an approximately unbiased estimator of a risk. Two popular risk functions are the MISE (as defined in (1.4)) and the expected Kullback-Leibler discrepancy between the true and estimated models, as defined below. Suppose that the conditional distribution of Y given Z is known up to g. The conditional cdf and pdf of Y given $Z = z$ will be denoted $F_0(y; g(z))$ and $f_0(y; g(z))$, respectively and the pdf of Z will be denoted $f(z)$. Under model (1.2), the Kullback-Leibler discrepancy between the true and estimated models is

$$\mathrm{KL}(\hat{g}_h) = \int\int \log\frac{f_0(y; g(z))}{f_0(y; \hat{g}_h(z))} dF_0(y; g(z)) f(z) dz. \tag{1.8}$$

The corresponding risk is the expected Kullback-Leibler discrepancy

$$\mathrm{MKL}(\hat{g}_h|Z_1,\ldots,Z_n) = E\left\{2KL(\hat{g}_h)|Z_1,\ldots,Z_n\right\},$$

where the factor 2 is introduced for mathematical convenience and the expectation is with respect to the true conditional pdf of $(Y_1,\ldots,Y_n)$ given $(Z_1,\ldots,Z_n)$.

The risk function MISE may be estimated by a criterion such as generalized cross-validation (GCV) (Craven and Wahba, 1979) and the risk MKL by using Akaike's Information Criterion (AIC) (Akaike, 1974). Having estimated a risk function, one may choose the bandwidth that minimizes this estimate. (See, e.g., Hart, 1997 for a comprehensive description of different selectors.) Asymptotic properties of risk estimation-based selectors have been studied and their asymptotic optimality has been proven in a nonparametric regression setting by Rice (1984) and by Härdle, Hall and Marron (1988). They argue that most bandwidth selectors based on risk estimation minimize a penalized version of the residual mean square, i.e.,

$$\mathrm{RMS}(h) = \frac{1}{n}\sum_{i=1}^{n}[Y_i - \hat{g}_h(Z_i)]^2.$$

As an estimator of the mean average squared error given in (1.7), the residual mean square is greatly biased. Minimizing it leads to a bandwidth that greatly undersmooths, or even interpolates, the data. To guard against undersmoothing, the residual mean square may be multiplied by a penalty term $\Psi(h)$. This penalty is designed to penalize undersmoothing in small or moderate samples, and thus increases with decreasing smoothness of $\hat{g}_h$. For example, multiplying $\mathrm{RMS}(h)$ by $\Psi_{GCV}(h) = [1 - \mathrm{trace}(\mathbf{H})/n)]^{-2}$ leads to the GCV criterion, by $\Psi_{AIC}(h) = \exp(2\mathrm{trace}(\mathbf{H})/n)$ leads to the AIC criterion and by $\Psi_T(h) = [1 - 2\mathrm{trace}(\mathbf{H})/n)]^{-1}$ leads to the "T" criterion (Rice, 1984).

Asymptotically, i.e., when $n \to \infty$, $h \to 0$ and $nh \to \infty$ (implying that $\mathrm{trace}(\mathbf{H})/n \to 0$), the risk-based selectors of the previous paragraph are equivalent. Therefore, in large samples one can expect these criteria to give similar choices for the bandwidth parameter, but in small samples (i.e., when $\mathrm{trace}(\mathbf{H})/n$ is large) they behave differently. Hurvich, Simonoff and Tsai (1998) provide a graph of the penalties for some of the risk based selectors as a function of $\mathrm{trace}(\mathbf{H})/n$, thus illustrating how the penalties vary with sample size.

Despite their favorable asymptotic properties, when applied to moderate or small samples, GCV and AIC present two major drawbacks: they have a tendency to undersmooth and lead to smoothing parameters with high sampling variability. These drawbacks inspired the formulation of an improved version of AIC, as proposed by Hurvich, Simonoff and Tsai (1998), for linear smoothers in the nonparametric regression setting of (1.2) with independent $N(0, \sigma^2)$ errors. The improved version of AIC, obtained by modifying the penalty term of the AIC criterion, is still asymptotically optimal, but leads to regression estimates that are undersmoothed less often than those obtained with GCV or AIC. The corrected AIC criterion could also be considered in the Poisson case, but, as discussed in Section 2.3.2, it turns out that in the absence of an unknown scale parameter no correction is forthcoming.

2 Methodology

Section 2.1 defines the Poisson single-index model that generates the data. Section 2.2 describes the derivation of the Kullback-Leibler (KL) risk for selecting a bandwidth and estimating the direction. Two popular methods of estimating the KL risk function are also presented: pseudo maximum likelihood (PML) and cross-validation (CV). Section 2.3 introduces an L_2-type approximation of the KL loss function, called Weighted Average Squared Error (WASE). For estimating the risk associated with WASE, we propose two methods: two stage weighted least squares (WLS_2) and double smoothing (DS).

2.1 Model

The single index model we will consider is based on the Poisson conditional distribution family. The observed data (X_i, Y_i), $i = 1, \ldots, n$, are independent and identically distributed $(p+1)$-vectors generated by the model

$$\begin{aligned} Y \mid (X = x) &\sim Po(m(x)), \qquad (2.9) \\ m(x) &= E(Y \mid X = x) = g(\beta_0\, x), \end{aligned}$$

where g is a smooth function of a single variable and β_0 is the p-variate direction vector. As usual in regression problems, analysis will proceed conditionally on the observed X values.

When g is unspecified, the best one can do is to estimate the parameter β_0 up to a multiplicative scalar, because any scale change in $\beta_0 X$ can be absorbed into the link function. So, we will restrict the parameter space of dimension p to a space of dimension $p-1$ by fixing the first component of β_0 to 1. The magnitude of the component β_{0j}, $1 < j \leq p$, has, in this $p-1$ dimensional parameter space, a simple interpretation: it measures the change in X_j required to match the effect of a unit change in X_1.

Specifying the model as in (2.9) is equivalent to specifying a link function g and an error distribution in model (0.1). Based on this model, the objective is to estimate the direction β_0 and the link function g. In doing so, a crucial step is the choice of the smoothing parameter, which involves the usual compromise between a model's smoothness and how closely it fits the data.

2.2 Derivation of the Kullback-Leibler risk

Model (2.9) is completely specified once g and β_0 are known. Let f_X denote the density of the covariate X, let μ be an arbitrary positive function and let β be any p-vector with first component equal to 1. The KL discrepancy as defined in (1.8) between the true model (g, β_0) and the candidate (μ, β) is

$$\begin{aligned} \mathrm{KL}(\mu, \beta) = \int [\mu(\beta x) &- g(\beta_0 x) \\ &+ g(\beta_0 x) \log(g(\beta_0 x)/\mu(\beta x))] f_X(x) dx, \end{aligned} \tag{2.10}$$

which may be viewed as a loss function.

Ignoring terms that do not depend on either μ or β, an equivalent loss is

$$\Delta_{KL}(\mu, \beta) = \int [\mu(\beta x) - g(\beta_0 x) \log(\mu(\beta x))] \, f_X(x) dx.$$

An empirical version of this loss requiring no knowledge of f_X is

$$\tilde{\Delta}_{KL}(\mu, \beta) = \frac{1}{n} \sum_{i=1}^{n} [\mu(\beta X_i) - g(\beta_0 X_i) \log(\mu(\beta X_i))] \,. \tag{2.11}$$

We shall adopt $\tilde{\Delta}_{KL}$ as our loss function, which seems sensible inasmuch as we are conditioning on the observed X's.

In the single index model, we will replace the candidate function μ by a Nadaraya-Watson estimate (given in (1.3)) of the functions $g(\beta x) = E(Y \mid \beta X = \beta x)$, where z is replaced by βx and Z_j by βX_j. Choosing a candidate model (μ, β) thus amounts to selecting values for the smoothing parameter h and the direction vector β. Ideally we would select h and β to minimize our loss function $\tilde{\Delta}_{KL}(\hat{g}_h, \beta)$. Note that $\hat{g}_h$ depends upon β, but we supress this fact in our notation. The corresponding KL risk function is

$$E\left(2\tilde{\Delta}_{KL}(\hat{g}_h, \beta)\right) = \frac{2}{n} \sum_{i=1}^{n} \{E[\hat{g}_h(\beta X_i)] - g(\beta_0 X_i) E[\log(\hat{g}_h(\beta X_i))]\} \,,$$

where the expectation is with respect to the conditional distribution of $(Y_1, \ldots, Y_n)$ given $(X_1, \ldots, X_n)$. Unfortunately, this risk depends upon the unknown quantities g and β_0. A solution to this problem is to derive appropriate estimators of the KL risk function.

An approximately unbiased estimator of the risk $E\left(2\tilde{\Delta}_{KL}(\hat{g}_h, \beta)\right)$ is the following cross-validation criterion:

$$\mathrm{CV}(h,\beta) = \frac{2}{n}\sum_{i=1}^{n}\left\{\hat{g}_h(\beta X_i) - Y_i \log(\hat{g}_h^i(\beta X_i))\right\}, \tag{2.12}$$

where $\hat{g}_h$ is as in (1.3) and $\hat{g}_h^i$ has the same form as $\hat{g}_h$ except it is computed without the data point (X_i, Y_i), $i = 1, \ldots, n$. The leave-one-out Nadaraya-Watson estimator in the second term of (2.12) is used in order for Y_i and $\hat{g}_h^i(\beta X_i)$ to be independent, $i = 1, \ldots, n$, implying that the expectation of the product $Y_i \log(\hat{g}_h^i(\beta X_i))$ is the product of expectations. As a consequence, $\mathrm{CV}(h,\beta)$ is an approximately unbiased estimator of $E\left(2\tilde{\Delta}_{KL}(\hat{g}_h,\beta)\right)$.

Another possible risk estimator is based on pseudo maximum likelihood, as proposed in a semiparametric setting by Bonneu and Delecroix (1992):

$$\mathrm{PML}(h,\beta) = \frac{2}{n}\sum_{i=1}^{n}\left\{\hat{g}_h^i(\beta X_i) - Y_i \log(\hat{g}_h^i(\beta X_i))\right\}. \tag{2.13}$$

The PML criterion is obtained by replacing in the log-likelihood function of the PSIM model the unknown link function g by the leave-one-out Nadaraya-Watson estimator $\hat{g}_h^i$. For any given value of h, the pseudo-likelihood function is viewed as a function of the direction parameter β, and an estimator of β_0 is obtained by minimizing this function with respect to β. Delecroix, Hristache, and Patilea (1999) proved that, when PML is considered as a function of both β and h, the minimizer $(\hat{\beta}, \hat{h})$ yields a consistent estimator for β_0 and an optimal bandwidth choice, respectively.

2.3 L_2-type approximation of the KL loss function

To derive alternative methods, we reconsider the loss function $\tilde{\Delta}_{KL}(\hat{g}_h,\beta)$ from (2.11). For notational convenience let $m_i = g(\beta_0 X_i)$ and $\hat{m}_i = \hat{g}_h(\beta X_i)$. Expanding $\log \hat{m}_i$ in a Taylor series about m_i and regrouping terms yields

$$\begin{aligned} 2\tilde{\Delta}_{KL}(\hat{g}_h,\beta) = \frac{2}{n}\sum_{i=1}^{n}[m_i - m_i \log m_i \quad &+ \quad (2m_i)^{-1}(\hat{m}_i - m_i)^2 \\ &- \quad m_i(\hat{m}_i - m_i)^3/(3\tilde{m}_i^3)], \end{aligned}$$

where $\tilde{m}_i$ is between m_i and $\hat{m}_i$, $i = 1, \ldots, n$. The term $(m_i - m_i \log m_i)$ is free of h and β, and the last term within the brackets will generally be small in comparison to the second. This suggests that minimizing the loss

$$\mathrm{WASE}(\hat{g}_h,\beta) = \frac{1}{n}\sum_{i=1}^{n}\frac{(\hat{m}_i - m_i)^2}{m_i} \tag{2.14}$$

is asymptotically equivalent to minimizing the Kullback-Leibler loss in our Poisson model.

In any event, the loss $\mathrm{WASE}(\hat{g}_h, \beta)$ is a reasonable discrepancy measure independent of the above argument, since it may be interpreted as a weighted average squared error between $\hat{g}_h(\beta x)$ and $g(\beta_0 x)$. The corresponding WASE risk function is defined by

$$R(h, \beta; m) = E\left\{\mathrm{WASE}(\hat{g}_h, \beta)\right\} = E\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{(\hat{m}_i - m_i)^2}{m_i} \mid X_1, \ldots, X_n\right\}. \tag{2.15}$$

Now, $\hat{m}_i$ is the Nadaraya-Watson estimator of $g(\beta X_i)$ and may be written

$$\hat{m}_i = \sum_{k=1}^{n} H_{ik} Y_k, \tag{2.16}$$

where H_{ik} is the ik component of the hat matrix $\mathbf{H}$, which for the Nadaraya-Watson smoother is

$$H_{ij} = \frac{K_h(\beta X_i - \beta X_j)}{\sum_{r=1}^{n} K_h(\beta X_i - \beta X_r)}, \quad i, j = 1, \ldots, n.$$

Then, using the linearity (2.16), the fact that $\mathrm{Var}(Y_i|X_i) = m_i$, $i = 1, \ldots, n$, and the independence between Y_i and Y_k for $i, k = 1, \ldots, n$, $i \neq k$, we have

$$R(h, \beta; m) = \frac{1}{n}\sum_{i=1}^{n}\frac{E\left\{(Y_i - \hat{m}_i)^2 | X_1, \ldots, X_n\right\}}{m_i} + \frac{2}{n}\mathrm{tr}(\mathbf{H}) - 1. \tag{2.17}$$

Again, in practice we cannot evaluate the risk $R(h, \beta; m)$, since it depends on the unknown link function g and on β_0. Therefore, we derive an approximately unbiased estimator of $R(h, \beta; m)$.

2.3.1 A classical style risk estimator

An obvious candidate estimator of $R(h, \beta; m)$ is the following Weighted Least Squares criterion:

$$\begin{aligned} \mathrm{WLS}(h, \beta) &= \frac{1}{n}\sum_{i=1}^{n}\frac{(Y_i - \hat{m}_i)^2}{\hat{m}_i} + \frac{2}{n}\mathrm{tr}(\mathbf{H}) - 1 \qquad (2.18) \\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{[Y_i - \hat{g}_h(\beta X_i)]^2}{\hat{g}_h(\beta X_i)} + \\ &\quad \frac{2}{n}\left[\sum_{i=1}^{n}\frac{K_h(0)}{\sum_{j=1}^{n} K_h(\beta X_i - \beta X_j)}\right] - 1, \end{aligned}$$

which is similar to classical risk estimators such as Mallows' C_p (Mallows, 1973). Unfortunately, this estimator has a substantial bias due to the use of $1/\hat{m}_i$ as a weight in place of $1/m_i$. The inadequacy of using a weight depending on the smoothing parameter being selected has been pointed out already by Härdle, Hall and Ichimura (1993). To remedy this problem, they proposed a two-stage cross-validation procedure. Analogously, we propose the following two-stage *risk estimation* procedure.

Stage 1

Minimize an unweighted least squares cross-validation criterion (as in (1.6)) to obtain pilot estimates of $(h,\, \beta)$:

$$(\hat{h}_1, \hat{\beta}_1) = \mathrm{argmin}_{h,\beta} \frac{1}{n} \sum_i \left[Y_i - \hat{g}_h^i(\beta X_i) \right]^2 ,$$

and then estimate the link function g by the Nadaraya-Watson estimator $\hat{g}_{\hat{h}_1}(\hat{\beta}_1 x)$.

Stage 2

The final estimator of $(h,\, \beta)$ will then be obtained by minimizing the following risk estimator:

$$\mathrm{WLS}_2(h, \beta) = \frac{1}{n} \sum_i \frac{[Y_i - \hat{g}_h(\beta X_i)]^2}{\hat{g}_{\hat{h}_1}(\hat{\beta}_1 X_i)} + \frac{2}{n} \mathrm{tr}(\mathbf{H}) - 1. \tag{2.19}$$

Note that in the definition of the above criterion, the weight function $\hat{g}_{\hat{h}_1}(\hat{\beta}_1 X_i)^{-1}$ is free of (h, β). Intuitively speaking, $\mathrm{WLS}_2(h, \beta)$ should therefore more closely mimic the risk $R(h, \beta; m)$ than should the one-stage criterion $\mathrm{WLS}(h, \beta)$.

2.3.2 Double smoothing risk estimator

The stability of risk estimation criteria such as Mallows' C_p and AIC have been called into question by a number of authors, including Hall and Johnstone (1992), Hurvich and Tsai (1995) and Hurvich, Simonoff and Tsai (1998). The latter two references propose "corrected" versions of AIC that involve a modified penalty term and turn out to be considerably more stable than the classical AIC. It can be shown that the motivation behind the modified penalty term rests entirely on the model having an unknown scale parameter. In the classical, homoscedastic regression setting, if one uses the same derivation as in Hurvich, Simonoff and Tsai (1998) but assumes a *known* scale parameter, the resulting risk estimation criterion is just Mallows' C_p in which the scale estimate is the known scale parameter. In particular, no modification of the C_p penalty term arises in this derivation.

The relevance of the preceding comments to our setting is that, effectively, there is no unknown scale parameter in the Poisson model. Since the mean and variance are one and the same, there is no scale parameter above and beyond the mean function. As a result, the methodology of Hurvich, Simonoff and Tsai (1998) offers no insight on improving the stability of our risk estimator $\mathrm{WLS}_2(h, \beta)$, at least when the probability model really is Poisson. We thus consider a completely different approach for stabilizing a data-driven smoothing parameter.

Evaluating the risk (2.15) in the case of a Poisson model leads to the following expression for $R(h, \beta; m)$, which is known up to the function m:

$$R(h, \beta; m) = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{\tilde{m}_i^2}{m_i} + \frac{\rho_i}{m_i} - 2\tilde{m}_i + m_i \right],$$

where

$$\tilde{m}_i = \sum_{j=1}^{n} H_{ij} m_j \qquad \text{and} \qquad \rho_i = \sum_{j=1}^{n} H_{ij}^2 m_j.$$

As proposed by Wand and Gutierrez (1997), $R(h, \beta; m)$ may be estimated by $R(h, \beta; m_{pilot})$, where m_{pilot} is some pilot estimate of m. The resulting estimates of bandwidth and direction parameter are

$$(\hat{h}_{DS}, \hat{\beta}_{DS}) = \mathrm{argmin}_{h,\beta} R(h, \beta; m_{pilot}).$$

Wand and Gutierrez (1997) suggest that m_{pilot} be chosen to optimize the resulting smoothing parameter. Here we will not be quite so ambitious, settling instead for a simple solution where $m_{pilot}(x) \equiv \hat{g}_{\hat{h}_1}(\hat{\beta}_1 x)$, the stage 1 estimate in the scheme proposed in Section 2.3.1.

We anticipate that this method will be more stable than the one proposed in the previous section, inasmuch as $R(h, \beta; m_{pilot})$ is in error only through m_{pilot}. By contrast, $\mathrm{WLS}_2(h, \beta)$ differs from $R(h, \beta; m)$ due to errors in the pilot estimate *and* errors in the estimate of $E[(Y_i - \hat{m}_i)^2 | X_1, \ldots, X_n]$.

2.3.3 A two-stage cross-validation criterion

For the heteroscedastic single index models, Härdle, Hall, Ichimura (1993) propose the following two stage cross-validation estimation scheme. They suppose that the conditional variance of Y given X is only a function of the index $\beta_0 X$:

$$\mathrm{Var}(Y|x) = \sigma^2 G\{g(\beta_0 x)\},$$

where G is a known, smooth function and σ is a (possibly unknown) overdispersion parameter. The first stage provides the pilot estimates $(\hat{h}_1, \hat{\beta}_1)$ minimizing the unweighted LSCV criterion defined in Section 2.3.1. In our PSIM model, $\mathrm{Var}(Y|x) = g(\beta_0 x)$, thus at the second stage the final estimator of (h, β) is obtained by minimizing the following weighted cross-validation criterion:

$$\mathrm{WLSCV}_2(h, \beta) = \frac{1}{n} \sum_i \frac{\left[Y_i - \hat{g}_h^i(\beta X_i)\right]^2}{\hat{g}_{\hat{h}_1}(\hat{\beta}_1 X_i)}, \tag{2.20}$$

which is the same as the first term of the WLS_2 criterion defined in (2.19), except for the leave-one-out estimator in the numerator.

WLSCV_2 may be considered as an alternative estimator of the risk associated with WASE, defined in (2.15). Unless n is very small, $R(h, \beta; m)$ will be approximately:

$$R^*(h, \beta; g(\beta_0 X)) = E\left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{\left[\hat{g}_h^i(\beta X_i) - g(\beta_0 X_i)\right]^2}{g(\beta_0 X_i)} \mid X_1, \ldots, X_n \right\}.$$

Using the same arguments as for deriving the expression (2.17), it may be shown that

$$R^*(h, \beta; g(\beta_0 X)) = \frac{1}{n} \sum_{i=1}^{n} \frac{E\left\{(Y_i - \hat{g}_h^i(\beta X_i))^2 | X_1, \ldots, X_n\right\}}{g(\beta_0 X_i)} - 1,$$

for which a candidate estimator is the WLSCV_2 criterion.

3 Simulation study

In this section we present results of a simulation study that investigates various methods of estimating bandwidth and direction parameter. The methods include the cross-validation and pseudo-maximum-likelihood estimators of the Kullback -Leibler risk presented in Section 2.2 and the WASE risk estimation schemes of Section 2.3.

3.1 Monte Carlo setup

The simulated data (X_i, Y_i), $i = 1, \ldots, n$, are independent and identically distributed observations of (X, Y), where X is a bivariate vector having components $X_j \sim U[0, 1]$, $j = 1, 2$, and Y is Poisson distributed with conditional mean depending on the index $Z = \beta_0 X/10$. The direction vector used to generate the data is $\beta_0 = (1, 9)$. Four link functions were used in our study:

$$
\begin{aligned}
g_1(z) &= 5 + 900z^2(1-z)^2 \\
g_2(z) &= 10000(z^8(1-z)^2 + z^2(1-z)^8) \\
g_3(z) &= 5\exp(-100(z-0.75)^2) + 480(z-0.75)^2 \\
g_4(z) &= 7 + 7z^5 + 7(z-1)^5.
\end{aligned}
$$

The four links correspond to different signal to noise ratios (SNR), which we define as

$$\text{SNR} = \frac{\max_{0\le z\le 1} g(z) - \min_{0\le z\le 1} g(z)}{\sqrt{E(g(Z_1)}}.$$

Note that in our Poisson setting $g(z) = \text{Var}(Y|Z = z)$, which motivates our use of $\sqrt{E(g(Z_1)}$ in the denominator of SNR. The values of SNR are $9, 9, 30, 5$ for g_1, g_2, g_3, g_4, respectively.

Two sample sizes were used ($n = 50$, 150) and 500 replications were conducted for each combination of sample size and link function. Estimation of h and β was carried out exactly as described in Section 2. The Nadaraya-Watson estimate of g, whose smoothing parameter is to be chosen, is computed using a normal kernel.

3.2 The results

The data-driven criteria computed for each simulated sample are as follows:

CV, the likelihood cross-validation criterion, defined in (2.12).

PML, the pseudo maximum likelihood criterion, defined in (2.13).

WLS_2, the two stage weighted least squares criterion, defined in (2.19).

DS, the double smoothing estimation scheme of Section 2.3.2.

WLSCV_2, the two stage cross-validation scheme of Section 2.3.3.

For a given data set, the optimal choice of (h, β) is that which minimizes the empirical version of the Kullback-Leibler discrepancy given in (2.11). Minimizing this loss function is asymptotically equivalent to minimizing the WASE loss defined in (2.14). All the data-driven criteria studied and the two optimal losses KL and WASE are computed for each data set and minimized with respect to both h and β. The results obtained for the 500 replications are presented in Tables 1-4; each table corresponds to a particular link function.

n	DS	PML	CV	WLS2	WLSCV2	WASE	KL
				$Mean(\hat{h})$			
50	0.034	0.027	0.023	0.034	0.030	0.032	0.034
150	0.027	0.025	0.024	0.026	0.024	0.025	0.026
				$Std.dev.(\hat{h}) \times 10$			
50	0.086	0.110	0.112	0.094	0.106	0.077	0.079
150	0.050	0.056	0.069	0.053	0.055	0.039	0.040
				$Mean(\hat{h} - \hat{h}_{KL})^2 \times 10^3$			
50	0.143	0.229	0.296	0.157	0.195	0.012	
150	0.049	0.054	0.070	0.049	0.053	0.002	
				$Mean(\hat{\beta})$			
50	9.233	10.389	10.560	9.343	9.491	8.934	8.958
150	9.289	9.313	9.412	9.225	9.231	8.911	8.921
				$Std.dev.(\hat{\beta})$			
50	2.374	2.748	3.085	2.384	2.527	1.003	1.036
150	1.490	1.426	1.762	1.270	1.308	0.349	0.341
				$Mean(\hat{\beta} - \beta_0)^2$			
50	5.680	9.464	11.932	5.789	6.614	1.009	1.072
150	2.298	2.128	3.267	1.661	1.761	0.130	0.122
				$Mean(WASE(\hat{h}, \hat{\beta}))$			
50	0.274	0.326	0.359	0.276	0.295	0.213	
150	0.115	0.119	0.128	0.115	0.117	0.096	
				$Mean(KL(\hat{h}, \hat{\beta}))$			
50	0.129	0.160	0.177	0.131	0.142		0.101
150	0.055	0.057	0.062	0.055	0.056		0.046

Table 1: Simulation results for g_1 link function

n	DS	PML	CV	WLS2	WLSCV2	WASE	KL
				$Mean(\hat{h})$			
50	0.022	0.016	0.015	0.021	0.019	0.019	0.020
150	0.016	0.015	0.015	0.015	0.014	0.014	0.015
				$Std.dev.(\hat{h}) \times 10$			
50	0.066	0.069	0.071	0.069	0.077	0.052	0.053
150	0.021	0.027	0.034	0.022	0.024	0.020	0.023
				$Mean(\hat{h} - \hat{h}_{KL})^2 \times 10^3$			
50	0.068	0.091	0.108	0.065	0.081	0.004	
150	0.009	0.011	0.015	0.009	0.011	0.003	
				$Mean(\hat{\beta})$			
50	9.071	10.890	10.962	9.026	9.336	8.709	8.713
150	8.852	9.229	9.432	8.934	9.132	8.877	8.839
				$Std.dev.(\hat{\beta})$			
50	2.128	2.787	3.138	2.091	2.203	1.007	0.972
150	0.929	1.346	1.846	0.925	1.176	0.472	0.368
				$Mean(\hat{\beta} - \beta_0)^2$			
50	4.524	11.326	13.679	4.363	4.955	1.096	1.025
150	0.883	1.860	3.586	0.859	1.397	0.237	0.161
				$Mean(WASE(\hat{h}, \hat{\beta}))$			
50	0.435	0.509	0.516	0.425	0.452	0.334	
150	0.225	0.219	0.237	0.210	0.214	0.178	
				$Mean(KL(\hat{h}, \hat{\beta}))$			
50	0.198	0.242	0.255	0.198	0.212		0.158
150	0.091	0.096	0.104	0.091	0.095		0.081

Table 2: Simulation results for g_2 link function

n	DS	PML	CV	WLS2	WLSCV2	WASE	KL
				$Mean(\hat{h})$			
50	0.029	0.026	0.024	0.029	0.030	0.029	0.029
150	0.023	0.023	0.022	0.023	0.023	0.023	0.023
				$Std.dev.(\hat{h}) \times 10$			
50	0.079	0.102	0.120	0.084	0.097	0.073	0.072
150	0.040	0.047	0.075	0.044	0.047	0.033	0.033
				$Mean(\hat{h} - \hat{h}_{KL})^2 \times 10^3$			
50	0.116	0.167	0.236	0.126	0.146	0.004	
150	0.029	0.043	0.083	0.039	0.042	0.000	
				$Mean(\hat{\beta})$			
50	9.413	10.136	11.262	9.277	9.244	8.992	8.982
150	9.120	9.063	9.564	9.030	9.047	8.992	8.984
				$Std.dev.(\hat{\beta})$			
50	2.042	2.403	2.917	1.813	1.762	0.821	0.837
150	1.086	0.895	1.483	0.841	0.841	0.238	0.242
				$Mean(\hat{\beta} - \beta_0)^2$			
50	4.333	7.055	13.611	3.358	3.160	0.672	0.699
150	1.192	0.804	2.513	0.707	0.708	0.057	0.059
				$Mean(WASE(\hat{h}, \hat{\beta}))$			
50	0.323	0.362	0.428	0.321	0.326	0.260	
150	0.133	0.134	0.153	0.132	0.133	0.114	
				$Mean(KL(\hat{h}, \hat{\beta}))$			
50	0.161	0.183	0.218	0.160	0.163		0.129
150	0.067	0.067	0.077	0.066	0.067		0.057

Table 3: Simulation results for g_3 link function

n	DS	PML	CV	WLS2	WLSCV2	WASE	KL
				$Mean(\hat{h})$			
50	0.062	0.054	0.047	0.061	0.058	0.061	0.068
150	0.053	0.048	0.047	0.050	0.048	0.040	0.046
				$Std.dev.(\hat{h}) \times 10$			
50	0.300	0.336	0.334	0.318	0.331	0.259	0.248
150	0.274	0.278	0.286	0.273	0.277	0.147	0.154
				$Mean(\hat{h} - \hat{h}_{KL})^2 \times 10^3$			
50	1.643	1.902	2.285	1.731	1.841	0.216	
150	1.061	1.025	1.037	1.002	1.014	0.099	
				$Mean(\hat{\beta})$			
50	7.179	8.225	8.325	7.723	8.055	8.123	8.155
150	8.686	9.132	9.230	9.031	9.118	8.013	8.192
				$Std.dev.(\hat{\beta})$			
50	3.605	4.264	4.311	4.126	4.217	3.221	3.049
150	3.858	4.111	4.130	4.139	4.104	2.013	1.912
				$Mean(\hat{\beta} - \beta_0)^2$			
50	16.285	18.746	19.005	18.621	18.637	11.124	9.992
150	14.957	16.885	17.076	17.100	16.825	5.017	4.301
				$Mean(WASE(\hat{h}, \hat{\beta}))$			
50	0.187	0.208	0.225	0.184	0.194	0.109	
150	0.088	0.090	0.092	0.087	0.090	0.058	
				$Mean(KL(\hat{h}, \hat{\beta}))$			
50	0.079	0.102	0.113	0.081	0.091		0.049
150	0.035	0.039	0.040	0.037	0.039		0.025

Table 4: Simulation results for g_4 link function

The values of $Mean(\hat{h})$ make it clear that in small samples ($n = 50$) CV and PML have a tendency to undersmooth. The mean values of $\hat{h}_{PML}$ and $\hat{h}_{CV}$ are, for all link functions considered, smaller than the average of the WASE and KL optimal $\hat{h}$ values. The DS and WLS_2 criteria yield values of $\hat{h}$ very close to the optimal WASE and KL values for both sample sizes. This seems to indicate that in the PSIM setting considered, DS and WLS_2 are less subject to undersmoothing than CV and PML. We also calculate the mean trace of the hat matrix $\mathbf{H}$ for each criterion, which can be interpreted as the average number of effective parameters used in smoothing. The values we obtain (not reported in the tables) using DS and WLS_2 are about the same as the optimal values gotten by minimizing the KL distance. This result agrees with the first two lines of the tables, where the average $\hat{h}$ values for DS and WLS_2 are comparable to the optimal KL values.

For all four link functions considered, the standard deviation of $\hat{h}_{DS}$ is the smallest among all data-driven criteria for both sample sizes, with $\hat{h}_{WLS2}$ and $\hat{h}_{WLSCV2}$ having the second, respectively third smallest standard deviation. This shows that the second stage in the double smoothing procedure acts like a shrinkage operator on the bandwidth: it gives the same average values of $\hat{h}$ as WLS_2 but with reduced variance. For $n = 150$, the standard deviation of $\hat{h}$ for CV is the largest among all the criteria studied for all four link functions, with PML having the second largest standard deviation. For $n = 50$, the standard deviation of $\hat{h}$ for CV is the largest among all data-driven criteria for the g_1 and g_3 link functions and $\hat{h}_{WLSCV2}$, $\hat{h}_{PML}$ have the largest standard deviation for g_2, g_4, respectively.

Our simulation results clearly indicate that the double smoothing criterion gives the best results for bandwidth selection, followed by the two-stage weighted least-squares criterion, which leads to smoothing parameters with slightly higher variance. For small samples, CV, PML and $WLSCV_2$ are biased towards undersmoothing and give more variable bandwidth choice (especially CV and PML). This is also illustrated in the two upper plots of Figures 1 and 2, representing kernel density estimates of the data-driven DS, PML, WLS_2 and $WLSCV_2$ bandwidths, for g_1 and respectively g_3 link functions. The fifth density estimate in the upper plots represents the optimal KL bandwidth. For $n = 50$ and even for $n = 150$, the bandwidths minimizing DS and WLS_2 have a more nearly symmetric sampling distribution than bandwidths minimizing PML and $WLSCV_2$. The estimated density curves for $\hat{h}_{DS}$ and $\hat{h}_{WLS2}$ have almost the same form as the density of the optimal value $\hat{h}_{KL}$, but with larger variability, whereas the density curves for $\hat{h}_{WLSCV2}$ and, especially, $\hat{h}_{PML}$ have modes shifted to the left, indicating undersmoothing.

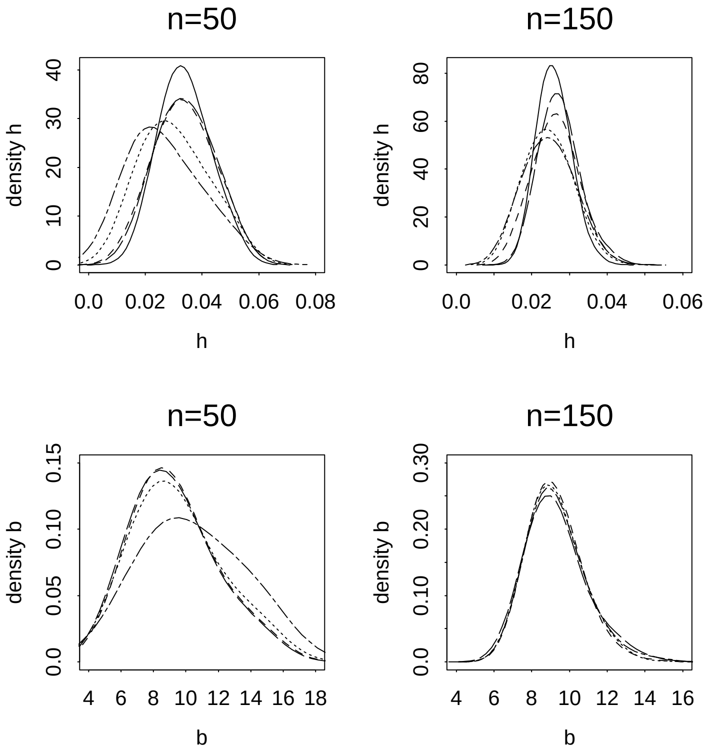

Figure 1: Density estimates for estimated h and b for function g_1: KL(solid line), DS(long dashed line), WLS2(dashed line), WLSCV2(dotted line), PML(long-short dashed line).

For estimating the direction parameter from samples with $n = 50$, for g_1 and g_2 links, WLS$_2$ and DS had the smallest bias and smallest variance, followed by WLSCV$_2$, PML and CV. For link g_3, the best direction results are obtained by WLSCV, followed by WLS$_2$, DS, PML and CV, whereas for link g_4, DS had the best direction results, followed by WLS$_2$, WLSCV$_2$, PML and CV. For samples with $n = 150$, the differences between the different criteria are less marked. We observe though that WLS$_2$ gives the best results for all the link functions considered, except for g_4, for which DS has the best results for direction estimation. CV exhibits the largest bias and variance among all criteria for both sample sizes. For the function g_4, the distance between the "true" direction and the direction estimate given by the data-driven criteria and even by the two optimal losses WASE and KL is larger than for the other functions. This may be explained by the higher level of noise for this link function

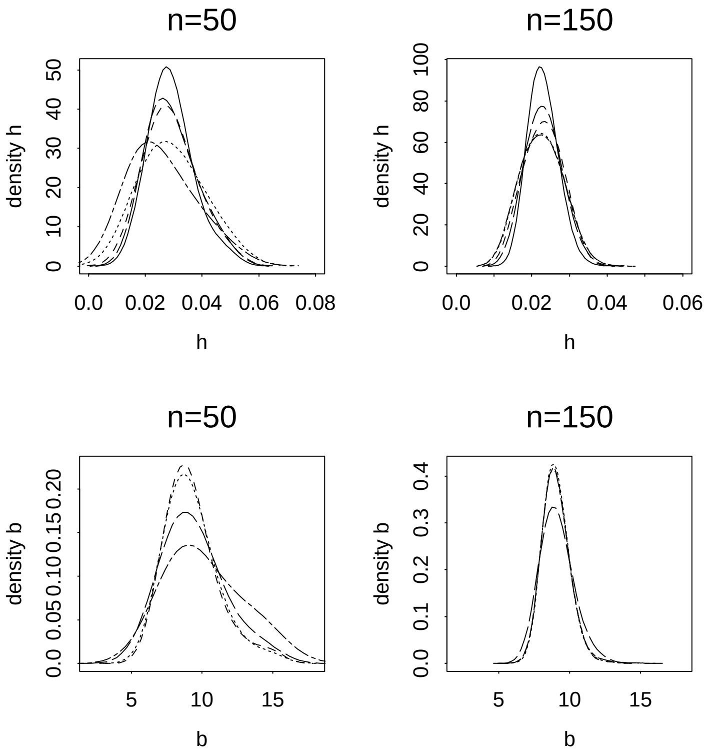

Figure 2: Density estimates for estimated h and b for function g_3: KL(solid line), DS(long dashed line), WLS2(dashed line), WLSCV2(dotted line), PML(long-short dashed line).

and also by its particular sigmoid shape (nearly linear in the middle and curved at the ends). A local-linear estimate would be more suitable than the Nadaraya-Watson in this case, as it automatically adjusts for edge effects.

Kernel density estimates of $\hat{\beta}$ for data-driven DS, PML, WLS$_2$ and WLSCV$_2$ criteria are given in the lower plots of Figures 1 and 2. For legibility reasons, the density curve for the optimal values $\hat{\beta}_{KL}$ is not represented, as it is much more peaked around the true value $\beta_0 = 9$ than all data-driven criteria. For $n = 50$ the values $\hat{\beta}_{PML}$ present more variability around the value β_0, whereas the density curve for $\hat{\beta}_{WLS2}$ is the most concentrated around β_0. For $n = 150$, the density curves obtained with the four data-driven criteria are nearly superposed and peaked around β_0. The results obtained indicate that the best direction estimates are given by the WLS$_2$ criterion.

The global performance results given by the optimal KL loss classify the criteria as follows: the best results (for $n = 50$) are obtained by minimizing DS and WLS_2 criteria, followed closely by WLSCV_2, then by PML and CV. While the differences in KL performances are less marked for $n = 150$, they still follow the same pattern. The global performance results obtained using the WASE error measure are consistent with the results obtained using the KL distance. In fact, WASE seems to be a good approximation for KL, even for $n = 50$. Values $(\hat{h}, \hat{\beta})$ obtained minimizing WASE are very similar to the values $(\hat{h}, \hat{\beta})$ minimizing KL.

4 Conclusion

In this paper we propose two simple and automatic methods for simultaneously estimating the direction parameter β_0 and the smoothing parameter h in a Poisson regression based on the single index model. The first criterion, called weighted least squares (WLS_2) estimates the Kullback-Leibler risk function and has a penalty term to prevent undersmoothing in small samples. The second method, termed double smoothing (DS), is based on the estimation of the WASE risk function (which is an L_2 approximation of the Kullback-Leibler risk) and makes use of a double smoothing idea as in Wand and Gutierrez (1997). We used simulations to compare these two methods to the cross-validation criterion (CV), the pseudo-maximum likelihood (PML) criterion and the weighted least-squares cross-validation criterion (WLSCV_2) proposed by Härdle, Hall and Ichimura (1993).

The proposed WLS_2 criterion gave the best results among all criteria for estimating the direction parameter of the single index model, whereas the DS rule was the best among all investigated methods for estimating the bandwidth parameter.

Even better results can probably be obtained with DS and WLS_2 methods if at the first stage presented in Section 2.3.1, the pilot estimates are obtained by minimizing, for example, the PML criterion instead of LSCV.

The two methods we propose in this paper can be used together for estimating the unknown link function g as follows: the bandwidth parameter may be estimated by using the double smoothing criterion DS and the direction by minimizing the weighted least-squares WLS_2. The link function may then be estimated using the Nadaraya-Watson estimator $\hat{g}_{\hat{h}_{DS}}(\hat{\beta}'_{WLS2}X)$.

The methodology proposed in this paper can be easily generalized to deal with the case of overdispersed Poisson data. Suppose that $\mathrm{Var}(Y_i|x_i) = \sigma^2 g(\beta_0 x_i) = \sigma^2 m_i$, where σ^2 is the unknown overdispersion parameter. In this case, using the same arguments as for obtaining the risk function in (2.17) (corresponding to the Poisson case where $\mathrm{Var}(Y_i|x_i) = m_i$), we obtain the following risk function:

$$R(h, \beta; m) = \frac{1}{n} \sum_{i=1}^{n} \frac{E\left\{(Y_i - \hat{m}_i)^2 | X\right\}}{m_i} + \frac{2}{n} \mathrm{tr}(\mathbf{H}) \sigma^2.$$

We may now use either an analog of WLS_2 or DS to estimate this quantity. A candidate for the estimate of σ^2 required in the second stage risk estimate is

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \frac{(Y_i - \hat{m}_i)^2}{\hat{m}_i},$$

where $\hat{m}_i$ is a first stage estimate of m_i obtained by one of the methods in this paper.

REFERENCES

Akaike, H. (1970). Statistical Predictor Information. *Annals of the Institute of Statistical Mathematics.* **22**, 203-217.

Akaike, H. (1974). A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control* **19**, 716-723.

Bonneu, M. and Delecroix, M. (1992). Estimation sèmiparametrique dans les modéles explicatifs conditionnels à indice simple" Cahiers GREMAQ, no 92.09.256, Toulouse I.

Craven, P. and Wahba, G. (1979) Smoothing Noisy Data With Spline Functions *Numerische Mathematik* **31**, 377-403.

Fan, J. and Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *J. Roy. Statist. Soc. Ser.B* **57**, 371–394.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications.* London: Chapman and Hall.

Gasser, T., Kneip, A. and Köhler, W. (1991). A flexible and fast method for automatic smoothing. *J. Amer. Statist. Assoc.* **86**, 643–652.

Hall, P. and Johnstone, I. (1992). Empirical functionals and effiecient smoothing parameter selection (with discussion). *J. Roy. Statist. Soc. Ser. Ser.B* **54**, 475-530.

Hart, J.D. (1997). *Nonparametric Smoothing and Lack-of-Fit Tests.* New York: Springer-Verlag.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models.* London: Chapman and Hall.

Härdle, W., Hall, P. and Ichimura, H. (1993). Optimal smoothing in single index models. *Ann. Statist.* **21**, 157-178.

Härdle, W., Hall, P. and Marron, J.S. (1988). How far are automatically chosen regression smoothing parameters from their optimum? *J. Amer. Statist. Assoc.* **83**, 86-101.

Härdle, W. and Stoker, T. (1989). Investigating smooth multiple regression by the method of average derivatives. *J. Amer. Statist. Assoc.* **84**, 986-995.

Horowitz, J. and Härdle, W. (1996)). Direct Semiparametric Estimation of Single-Index Models with Discrete Covariates. *J. Amer. Statist. Assoc.* **91**, 1632-1640.

Hurvich, C.M., Simonoff, J.S. and Tsai, C.L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike Information Criterion. *J. Roy. Statist. Soc. Ser.B* **60**, 271-293.

Hurvich, C.M. and Tsai, C-L. (1989). Regression and time series model selection in small samples. *Biometrika* **76**, 297-307.

Hurvich, C.M. and Tsai, C-L. (1995). Model Selection for Extended Quasi-Likelihood Models in Small Samples. *Biometrics* **51**, 1077-1084.

Klein, R.W. and Spady R.H. (1993). An Efficient Semiparametric Estimator for Binary Response Models. *Econometrica* **61**, 387-421.

Mallows, C. L. (1973). Some comments on C_p. *Technometrics* **15**, 661–675.

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

Nadaraya, E. A. (1964). On estimating regression. *Theory Probab. Appl.* **9**, 141–142.

Newey, W.K. and Stoker, T.M. (1993). Efficiency of weighted average derivative estimators and index models. *Econometrica.* **61**, 1199-1223.

Rice, J. (1984). Bandwidth choice for nonparametric regression. *Ann. Statist.* **12**, 1215-1230.

Sherman, R.P. (1994). U-processes in the analysis of a generalized semiparametric regression estimator. *Econometric theory*, **10**, 372-395.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions(with discussion). *J. Roy. Statist. Soc. Ser.*, *Ser. B* **36**, 111-147.

Wand, M.P. and Gutierrez, R.G. (1997). Exact risk approaches to smoothing parameter selection. *Journal of Nonparametric Statistics*, **8**, 337–354.

Watson, G. S. (1964). Smooth regression analysis. *Sankhya Ser. A* **26**, 359-372.