Semiparametric Estimation of a Class of Generalized Linear Models Without Smoothing

Alessio Sancetta[*]

April 18, 2014

Abstract

In a generalized linear model, the mean of the response variable is a possibly nonlinear function of a linear combination of explanatory variables. When the nonlinear function is unknown and is estimated nonparametrically from the data, these models are known as single index models. Using the relation of generalized linear models with the exponential family model, this paper shows how to use a modified version of the empirical cumulant generating function to estimate the linear function of the explanatory variables with no need of smoothing techniques. The resulting estimator is consistent and normally distributed. Extensive simulations, partially reported here, show that the method works in practice. The method can also be seen as complementary to existing fully nonparametric methods. In fact, it can provide an initial value that can be used to fine tune a nonparametric estimator of the link function in the first step of the estimation.

Key words: Empirical cumulant generating function, Exponential dispersion model, Generalized linear model, Single index model.

1 Introduction

Generalized linear models (McCullagh and Nelder, 1989) allow the expectation of the response Y given the explanatory variables X to be non linear, through what is called

[*] Address: Department of Economics, Royal Holloway, Egham TW20 0EX. E-mail: <asancetta@gmail.com>. URL: <http://sites.google.com/site/wwwsancetta/>. I am very grateful to the referee for comments that substantially improved the paper and for correcting technical errors.

the the link function, e.g. $\mathbb{E}[Y|X] = G(X'\beta)$, for some univariate function G, whose inverse is called link function, i.e. $X'\beta = G^{-1}(\mathbb{E}[Y|X])$, as it links a linear function of the predictors to the the conditional expectation. In some situations, it is not obvious what G should be. If G is not specified it can then be estimated from the data. Then, one calls this semiparametric model the single index model.

The generalized linear model make direct reference to the exponential family model and the exponential dispersion family model (Jørgensen , 1986, 1987). On the other hand, the single index model makes reference to neither the specific functional form of G nor to the distribution of the errors, hence it is more general.

The literature on estimation of single index models abound. One approach is the average derivatives method, where one exploits the fact that

$$d\mathbb{E}[Y|X=x]/dx = dG\left(x'\beta\right)/dx \propto \beta,$$

and the prime $'$ stands for transposition. This requires a high dimensional kernel smoother and consequently is subject to the so called curse of dimensionality (Powell et al., 1989, Härdle and Stoker, 1989 , see Hristache et al., 2001, for an improved method and references therein). Another approach is to estimate G nonparametrically based on some initial estimate of β and then estimate β using the estimator for G, cycling through the procedure until convergence (e.g. Härdle et al., 1993, Horowitz and Härdle, 1996, Xia, 2006, Cui et al. 2011, Fan et al., 2013, and references therein). One of such nonparametric models is the Estimating Function Method (EFM) approach of Cui et al. (2011). This method achieves the same if not smaller variance than the estimator in Carroll et al. (1997). For the EFM and other approaches, the first guess of β can be crucial for convergence to a global maximum. The problem is made even harder by the fact that the initial amount of smoothing used to estimate G strongly depends on the starting value of β. This initial problem could be avoided if one had a reasonably good estimate of the index parameter that does not require previous estimation of G based on some fully nonparametric approach. Recently, Fan et al. (2013) have considered estimation of the quantile regression for the single index model in the presence of large number of regressors via penalization, essentially incorporating variable selection into the kernel smoothing estimation.

The goal of this paper is to impose the semiparametric restriction that the density of Y conditional on X belongs to the exponential dispersion family model with canonical link, and use this to estimate β. The estimation takes advantage of the fact that - under the aforementioned restrictions- the only infinite dimensional parameter is related to the conditional cumulant generating function of the response variables. Direct estimation of this would require nonparametric methods. However, this paper shows that it is possible to find a particular relation between the conditional mean and the unconditional expectation of some known function of the data. To the author knowledge this relation is new. Estimating unconditional expectations of known functions does not require any smoothing. Hence, in this context, the estimation of β can be turned into a nonlinear least square problem and estimated by Generalized Method of Moments (GMM). The resulting estimator is shown to be normally distributed. The method is applicable to continuous and binary dependent variables.

The next section presents the relation between the conditional mean and variance of the response and the unconditional expectation of some function of the data. This relation is the motivation for the estimator. Having defined the estimator, the asymptotic properties are derived under regularity conditions. Section 3 contains a discussion of the results and the conditions. The proofs are deferred to Section 4.

2 Statement of the Problem

For some $\lambda > 0$, let P_λ be a probability measure with cumulant generating function $\lambda\psi\left(t\right) = \ln\left(\int e^{yt} dP_\lambda\left(y\right)\right)$ supposed to be finite for $t \in \mathcal{T}$ and $\mathcal{T}$ is some set containing the origin (called the effective domain of ψ, e.g. Jørgensen, 1987). Then,

$$\frac{dP_\lambda\left(y|\eta\right)}{dP_\lambda\left(y\right)} = \exp\left\{\lambda\left(\eta y - \psi\left(\eta\right)\right)\right\} \tag{1}$$

is a density in the exponential dispersion family with respect to (w.r.t.) the dominating measure P_λ. The family is very large as it is essentially defined through any probability measure P_λ having a finite moment generating function around the origin. Hence, the parameter space can be restricted to be the set of values $\eta \in \mathbb{R}$ and $\lambda > 0$ for which $\psi\left(\eta\right)$ is finite, and $\lambda\psi\left(\bullet\right)$ is the cumulant generating function of some P_λ. Throughout it is assumed that λ and η are inside the parameter space, assumed to be nonempty, so that $\lambda\psi(t)$ is always finite.

Here, interest is restricted to the canonical parameter $\eta := x'\beta$, for some explanatory variable $x \in \mathcal{X} \subseteq \mathbb{R}^K$ and a conformable vector β. This shall be a maintained condition throughout the paper. In its full generality, the exponential dispersion model assumes the canonical parameter to be a possibly non linear function of $x'\beta$. As discussed in Nelder and Wedderburn (1972), McCullagh and Nelder (1989), for $\eta = x'\beta$, (1) is a subset of the generalized linear model such that, given a sample $\{Y_i, X_i : i = 1, 2, ..., n\}$, a sufficient statistic for β is given by $\sum_{i=1}^n X_i Y_i$. Here, interest is restricted to this case only, where however $\lambda (> 0)$ is unrestricted. The effective domain of ψ implicitly define restrictions on x and β via η.

Example 1 *Consider $\psi(\eta) = \eta^2/2$ and set $\lambda = \sigma^{-2}$ for some $\sigma^2 \in (0, \infty)$, so that the exponential dispersion model is the linear Gaussian model*

$$\exp\left\{\frac{1}{\sigma^2}\left(yx'\beta - \frac{(x'\beta)^2}{2}\right)\right\} P_\lambda(y)$$

where $P_\lambda(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{y^2}{2\sigma^2}\right\}$. Then, $\{\eta \in \mathbb{R} : \psi(\eta) < \infty\} = \mathbb{R}$ so that the only restriction on x and β is that $x'\beta \in \mathbb{R}$.

In the above example, the parameters are essentially unrestricted. This is often not the case.

Example 2 *Let $\psi(\eta) = -\ln(-\eta)$ and $\lambda > 0$ so that the exponential dispersion model is the gamma model*

$$\exp\left\{\lambda\left(yx'b + \ln\left(-x'b\right)\right)\right\} P_\lambda(y)$$

where $P_\lambda(y) = \exp\{(\lambda - 1)\ln(\lambda y) + \ln\lambda - \ln\Gamma(\lambda)\}$, and $\Gamma(\lambda)$ is the gamma function. Hence, the model is defined for $\eta < 0$ only in order to make sure that $\psi(\eta) < \infty$. In this case, it is convenient to reparametrise in terms of $\tilde{b} = -b$ so that $\eta < 0$ is for example satisfied restricting x and $\tilde{b}$ to have only positive entries.

Another implication is that the restriction on η does restrict the distribution of the regressors when they are stochastic, or their range of values when deterministic. In the

Gaussian example, X can take values in $\mathbb{R}^K$, but its distribution needs to be tight to avoid infinities.

Note that even under the current restriction on η, $\mathbb{E}[Y|X=x] = G(x'b)$, for possibly non-linear but monotonic G (McCullagh and Nelder, 1989, p. 20). In particular, $d\psi(\eta)/d\eta$ is the inverse canonical link function, i.e. for $\eta = x'\beta$, $d\psi(\eta)/d\eta = \mathbb{E}[Y|X=x]$ (McCullagh and Nelder, 1989, p. 24).

Restricting attention to canonical links does have non-trivial implications. For example, for binary response data, the canonical link is a logit, e.g. probit is ruled out. However, the model does allow for consistent estimation for binary response and heteroskedasticity of unknown form. It is well known that in this case the estimator for a standard logit is inconsistent (e.g. Davidson and MacKinnon, 1984). In the case of a continuous real valued response, if the conditional distribution of Y is Gaussian, a canonical link implies a linear regression. Different specifications for the dominating measure P do lead to nonlinear regression for response variables with values in $\mathbb{R}$. However, no closed form solutions for ψ are known in these cases.

To better understand the derivation of the estimator, it is convenient to start with a population version, which is then used to derive the feasible estimator.

2.1 Unfeasible Estimator

The following observation is the basis for the estimator proposed here.

Lemma 1 *Let the density of Y conditional on $X = x$ be in the dispersion exponential model as in (1) with $\eta = x'\beta$. Suppose that,*

$$\mathbb{E}\exp\left\{-\lambda\psi\left(X'\beta\right)\right\} + \mathbb{E}\exp\left\{Y\left(t - X'\beta\right)\right\} < \infty$$

Define

$$\mu(t,b) := \frac{d\ln\mathbb{E}\exp\{Y(t - X'b)\}}{dt},$$

and

$$\sigma^2(t,b) := \frac{d^2\ln\mathbb{E}\exp\{Y(t - X'b)\}}{dt^2}.$$

Then,

$$\begin{aligned} \mathbb{E}\left[Y|X=x\right] &= \psi^{(1)}\left(x'\beta\right) \\ &= \mu\left(\lambda x'\beta, \lambda\beta\right) \end{aligned}$$

and

$$\begin{aligned} Var\left(Y|X=x\right) &= \psi^{(2)}\left(x'\beta\right)/\lambda \\ &= \sigma^2\left(\lambda x'\beta, \lambda\beta\right), \end{aligned}$$

where $\psi^{(j)}(t) := d^j\psi(t)/dt^j$, with ψ as in (1).

Proof. Let P_X be the law of X. Then,

$$\begin{aligned} \mathbb{E}\exp\left\{Y\left(t-\lambda X'\beta\right)\right\} &= \mathbb{E}\mathbb{E}\left[\exp\left\{Y\left(t-\lambda X'\beta\right)\right\}|X\right] \\ &= \int\int\exp\left\{y\left(t-\lambda x'\beta\right)\right\}\exp\left\{\lambda\left(x'\beta y-\psi\left(x'\beta\right)\right)\right\}dP_\lambda(y)\,dP_X(x) \\ &\quad \text{[Using (1) to take expectation]} \\ &= \int\int\exp\left\{\lambda\left(\frac{ty}{\lambda}-\psi\left(x'\beta\right)\right)\right\}dP_\lambda(y)\,dP_X(x) \\ &= \exp\left\{\lambda\psi\left(\frac{t}{\lambda}\right)\right\}\int\exp\left\{-\lambda\psi\left(x'\beta\right)\right\}dP_X(x) \\ &\quad \text{[by the properties of (1), e.g. eq. (2.6) in Jorgensen (1987)]} \\ &=: \exp\left\{\lambda\psi\left(\frac{t}{\lambda}\right)\right\}C_\psi, \end{aligned}$$

by obvious definition of C_ψ. Taking logs, differentiating w.r.t. t, and evaluating at $t=\lambda x'\beta$,

$$\begin{aligned} \left.\frac{d\ln\mathbb{E}\exp\left\{Y\left(t-\lambda X'\beta\right)\right\}}{dt}\right]_{t=\lambda x'\beta} &= \left.\lambda\frac{d\psi\left(t/\lambda\right)}{dt}\right]_{t=\lambda x'\beta} \\ &= \psi^{(1)}\left(x'\beta\right), \end{aligned}$$

and the left most side term above is just $\mu(t,b)$, as defined in the statement of the lemma, with $t=x'b$ and $b=\lambda\beta$. From the properties of the exponential dispersion model (e.g. Jørgensen , 1987) or by direct calculation, it follows that the right most hand side of the above display is $\mathbb{E}\left[Y|X=x\right]$. Differentiating once again, gives the conditional variance

$$\begin{aligned}\left.\frac{d^2\ln\mathbb{E}\exp\left\{Y\left(t-\lambda X'\beta\right)\right\}}{dt^2}\right]_{t=\lambda x\beta} &= \left.\lambda\frac{d^2\psi\left(t/\lambda\right)}{dt^2}\right]_{t=\lambda x'\beta}\\ &= \frac{\psi^{(2)}\left(x'\beta\right)}{\lambda},\end{aligned}$$

where again the left hand side element is just $\sigma^2\left(t,b\right)$ where $t=x'b$ and $b=\lambda\beta$. ■

In Lemma 1, $\mathbb{E}\exp\left\{Y\left(t-\lambda X'\beta\right)\right\}$ is neither the unconditional or the conditional moment generating function of the response, as the expectation is w.r.t. both Y and X. The conditional mean is found as the first derivative w.r.t. t of the log of this expression and then evaluating at $t=\lambda x'\beta$. Given that the expression uses only unconditional expectation, it is amenable of estimation with no need of smoothing techniques by replacing expectations with empirical ones.

If in Lemma 1 we knew $\mu\left(t,b\right)$, we could derive an unfeasible estimator for $\lambda\beta$. Note that β is not identifiable from the function μ alone. However, the structure of (1) with $\eta=x'\beta$ does make $\lambda\beta$ uniquely identifiable. Furthermore, the inclusion of an intercept in the estimation becomes redundant. This does not mean that the model cannot have mean different from zero.

Example 3 *Suppose that $X=1$, i.e. the intercept only case. Then, $\beta=b/\lambda$ becomes the intercept in the model. Consequently,*

$$\begin{aligned}\mu\left(b,b\right): &= \left.\frac{d\ln\mathbb{E}\exp\left\{Y\left(t-b\right)\right\}}{dt}\right]_{t=b}\\ &= \left.\frac{\mathbb{E}Y\exp\left\{Y\left(t-b\right)\right\}}{\mathbb{E}\exp\left\{Y\left(t-b\right)\right\}}\right]_{t=b}\\ &= \mathbb{E}Y.\end{aligned}$$

The b parameter drops and is not recoverable. The more general case of regressors plus intercept preserves the entries in b that do not correspond to the intercept, but makes the intercept unidentifiable.

Maximization of the log-likelihood from (1) gives the following moment vector valued equation

$$m_n(b) := \frac{1}{n}\sum_{i=1}^{n}\left[Y_i - \mu\left(X_i'b, b\right)\right]X_i = 0_K, \tag{2}$$

satisfied by $b = \lambda\beta$, where 0_K is the K-dimensional vector of zeros. The above display is well known in the theory of quasi-maximum likelihood estimation, as it requires sample orthogonality between the error term and the regressors. The form of this moment condition relies on the canonical parameter being linear.

Then, from the theory of optimal estimating functions and GMM, an estimator of $\lambda\beta$ is given by minimizing the following - unfeasible- objective function

$$m_n(b)' W^{-1} m_n(b) \tag{3}$$

where

$$W = \lim_n \frac{1}{n}\sum_{i=1}^{n}\frac{\psi^{(2)}\left(X_i'\beta\right)}{\lambda}X_iX_i', \tag{4}$$

pretending that the limit exists; note that

$$\lim_n \mathbb{E}\left[m_n(\lambda\beta)\, m_n(\lambda\beta)' \,|X\right] = W, \tag{5}$$

e.g. Jørgensen (1987). The unfeasible estimator is the starting point for the construction of a feasible estimator, as described in the next section.

2.2 Feasible Estimator

The unfeasible estimator is based on $\mu(t,b)$ and $\sigma^2(t,b)$, evaluated at $t = \lambda x'\beta$ and $b = \lambda\beta$, as defined in Lemma 1. The following shows that these quantities are actually the mean and the variance of Y w.r.t. a suitable change of measure.

Lemma 2 *Suppose that there are compact sets $\mathcal{T} \subset \mathbb{R}$ and $\mathcal{B} \subset \mathbb{R}^K$ such that for some $\epsilon > 0$,*

$$\sup_{t\in\mathcal{T}, b\in\mathcal{B}} \mathbb{E}\exp\left\{(1+\epsilon)\, Y\left(t - X'b\right)\right\} < \infty.$$

Then, for $t \in \mathcal{T}, b \in \mathcal{B}$,

$$\mu(t,b) = \mathbb{E}^{P(t,b)}[Y],$$

$$\sigma^2(t,b) = Var^{P(t,b)}(Y),$$

where $\mathbb{E}^{P(t,b)}$ *and* $Var^{P(t,b)}$ *are mean and variance with respect to the law defined by*

$$dP_{YX}(y,x;t,b) = L(y,x;t,b)\, dP_{YX}(y,x),$$

where $P_{YX}(y,x)$ *is the joint law of* Y *and* X *and*

$$L(y,x;t,b) := \frac{\exp\{y(t-x'b)\}}{\mathbb{E}\exp\{Y(t-X'b)\}}.$$

Proof. By the condition in the lemma, it is possible to interchange between integral and derivatives, as the integrand and its partial derivative w.r.t. t are continuous. Then, by the definition of μ as in Lemma 1, interchanging the order of expectation and differentiation,

$$\mu(t,b) = \frac{\mathbb{E}Y\exp\{Y(t-X'b)\}}{\mathbb{E}\exp\{Y(t-X'b)\}}.$$

Similarly, for σ^2 as in Lemma 1,

$$\sigma^2\left(\lambda x'\beta, \lambda\beta\right) = \frac{\mathbb{E}Y^2\exp\{Y(t-X'b)\}}{\mathbb{E}\exp\{Y(t-X'b)\}} - \left[\frac{\mathbb{E}Y\exp\{Y(t-X'b)\}}{\mathbb{E}\exp\{Y(t-X'b)\}}\right]^2.$$

Let

$$L(t,b) := L(Y,X;t,b) := \frac{\exp\{Y(t-X'b)\}}{\mathbb{E}\exp\{Y(t-X'b)\}}.$$

These equations say that $\mu(t,b)$ and $\sigma^2(\lambda x'\beta, \lambda\beta)$ are the unconditional mean and variance of $L(t,b)Y$, respectively, where $L(t,b)$ has values in $[0,\infty)$ and has mean one. Hence, by the Radon-Nikodym Theorem,

$$dP_{YX}(y,x;t,b) = L(y,x;t,b)\, dP_{YX}(y,x),$$

where the quantities on the r.h.s. are as defined in the statement of the lemma. ■

By replacing expectation with sample averages, Lemma 2 allow us estimate the population quantities in terms of the following empirical counterparts:

$$p_{in}(X_j,b) := \frac{\exp\left\{Y_i(X_j-X_i)'b\right\}}{\sum_{i=1}^n \exp\left\{Y_i(X_j-X_i)'b\right\}}, \tag{6}$$

$$\mu_n\left(X_j'b, b\right) := \sum_{i=1}^{n} Y_i p_{in}\left(X_j, b\right),$$

$$\sigma_n^2\left(X_j'b, b\right) := \sum_{i=1}^{n} Y_i^2 p_{in}\left(X_j, b\right) - \left[\mu_n\left(X_j'b, b\right)\right]^2,$$

$$g_n(b) = \frac{1}{n}\sum_{j=1}^{n}\left[Y_j - \mu_n\left(X_j'b, b\right)\right] X_j,$$

$$W_{n0} := \frac{1}{n}\sum_{j=1}^{n} X_j X_j'.$$

$$W_n(b) := \frac{1}{n}\sum_{j=1}^{n} \sigma_n^2\left(X_j'b, b\right) X_j X_j'.$$

It will be shown that by replacing μ with μ_n the estimating equation does not have asymptotic variance equal to W, hence, the sample estimators σ_n^2 and W_n are not used to derive a feasible estimator. However, σ_n^2 still provides information about any possible heteroskedasticity in the data, as it represents the conditional asymptotic variance of the error term not corrupted by the fact that μ is being replaced by μ_n. Hence, it can be used for data analysis. Consistency of the above statistics rests on regularity conditions. The following are sufficient for the present purposes.

Condition 1 *The sequence $(Y_i, X_i)_{i\in\mathbb{N}}$ is i.i.d. with Y_i having density conditional on $X_i = x$ equal to (1) with $\eta = \eta(x'\beta) = x'\beta$.*

Condition 2 *Let $\mathcal{B}$ be a compact Euclidean set such that the moment condition (2) is uniquely satisfied by $b = \lambda\beta$, for $\lambda\beta$ in the interior of $\mathcal{B}$.*

Condition 3 *$X_1 \in \mathcal{X}$, where $\mathcal{X}$ is a Euclidean subset such that*

$$\max_{b\in\mathcal{B}} \sup_{x\in\mathcal{X}} \left|x'b\right| \leq T < \infty$$

and $\mathbb{E}X_1X_1'$ has full rank.

Condition 4 *There is an $\epsilon_1 > 0$ such that*

$$\mathbb{E}\exp\left\{(4+\epsilon_1) T\left|Y_1\right|\right\} < \infty.$$

Remarks on the conditions are deferred to Section 3. The fact that one is using μ_n instead of μ leads to extra terms in addition to V in the variance of $g_n(b)$. Let $\mathbb{E}^i$ be expectation w.r.t. to the variables with index i only. For X and μ, define

$$\bar{p}_i(X;b) = \mathbb{E}^j \frac{X_j \exp\left\{Y_i (X_j - X_i)' b\right\}}{\mathbb{E}^i \exp\left\{Y_i (X_j - X_i)' b\right\}},\ \bar{p}_i(\mu X;b) = \mathbb{E}^j \frac{\mu\left(X_j' b, b\right) X_j \exp\left\{Y_i (X_j - X_i)' b\right\}}{\mathbb{E}^i \exp\left\{Y_i (X_j - X_i)' b\right\}}. \tag{7}$$

The following gives the variance matrix of the feasible estimating equation.

$$R(b) := Var\left(\left(Y_1 - \mu\left(X_1' b, b\right)\right) X_1 + Y_1 \bar{p}_1(X;b) + \bar{p}_1(\mu X;b)\right), \tag{8}$$

so that the first term in $R(\lambda\beta)$ is W as in (4) (here, for any random column vector Z, $Var(Z) = \mathbb{E}\left[(1-\mathbb{E})Z\right]\left[(1-\mathbb{E})Z\right]'$). Though $R(b)$ is unknown, it can be replaced by the sample estimator

$$\begin{aligned} R_n(b) : &= \frac{1}{n}\sum_{i=1}^{n}\left\{\left(1-\frac{1}{n}\sum_{i=1}^{n}\right)\left[\left(Y_i - \mu_n\left(X_i' b, b\right)\right) X_i + Y_i \frac{1}{n}\sum_{j=1}^{n} X_j\left(1+\mu_n\left(X_j' b, b\right)\right) p_{in}\left(X_j' b, b\right)\right]\right\} \\ &\times \left\{\left(1-\frac{1}{n}\sum_{i=1}^{n}\right)\left[\left(Y_i - \mu_n\left(X_i' b, b\right)\right) X_i + Y_i \frac{1}{n}\sum_{j=1}^{n} X_j\left(1+\mu_n\left(X_j' b, b\right)\right) p_{in}\left(X_j' b, b\right)\right]\right\}'. \end{aligned}$$

In Section 4 (Lemma 7) it is shown that $R_n(b)$ is consistent for $R(b)$ (Lemma 6 also shows that $W_n(\lambda\beta)$ is consistent for W). Recall that the present procedure only allows to identify $\lambda\beta$. The feasible estimator is obtained as follows:

$$\hat{b}_0 := \arg\min_{b\in\mathcal{B}} g_n(b)' W_{n0}^{-1} g_n(b), \tag{9}$$

and then using $\hat{b}_0$ to obtain the estimator

$$\hat{b} = \arg\min_{b\in\mathcal{B}} g_n(b)' \left[R_n\left(\hat{b}_0\right)\right]^{-1} g_n(b), \tag{10}$$

where $\mathcal{B}$ is as in Condition 2. Amongst estimators derived from $g_n(b)' A^{-1} g_n(b)$ with some full rank matrix A, the one derived from (10) has smallest asymptotic variance (e.g. Hansen, 1982, Godambe and Heyde, 1987).

Under regularity conditions, the consistency and asymptotic normality of the unfeasible estimator follow from standard results on M-estimators (e.g. Hansen, 1982, for the original proof in the GMM context). Hence, the goal is to show that the same holds for the feasible estimator under regularity conditions.

Theorem 1 *Under Conditions 1, 2, 3, and 4,*

$$\sqrt{n}\left(\hat{b} - \lambda\beta\right) \to N\left(0_K, \frac{\Delta' R^{-1} \Delta}{\lambda}\right)$$

where $R = R(\lambda\beta)$ *is as in (8), and* $\Delta = \lim_n dm_n(\lambda\beta)/d(\lambda\beta)$. *Moreover,*

$$\left.\frac{dg_n(b)}{db}\right]_{b=\hat{b}} = \Delta + o_p(1)$$

and

$$R_n\left(\hat{b}\right) = R + o_p(1).$$

3 Discussion

It is important to understand the implications of the regularity conditions. Lemma 1 critically relies on Condition 1. Condition 1 rules out any form of endogeneity. It is possible to extend the moment condition using an instrument in place of X. However, in this case Lemma 1 does not hold and the procedure is not fully justified. Condition 2 is high level. For some problems, identification is not necessarily straightforward. This is particularly so for binary response (e.g. Manski, 1988). Condition 3 requires X to be bounded. This condition has non trivial implications on Condition 2 for the case of binary response. If the predictors have bounded support, Chamberlain (2010) shows that, for binary response, identification in Condition 2 is only satisfied in the logistic case. As mentioned in the introduction, Condition 1 implies that, for binary data, the link function is the logistic with the possibility of heteroskedasticity of unknown form. Condition 4 requires Y to have tails as thin as an exponential density with mean parameter less than

$[(4+\epsilon_1)\,T]^{-1}$. Since T would be rarely known in practice, essentially this requires Y to have super exponential tails. Unfortunately, the fact that μ_n is based on exponential functions with infinite support makes the control of the estimation error more difficult than for methods that are based on kernel smoothers, where the kernel is assumed to be bounded (e.g. Assumption 1 in Hansen, 2008). A bounded support for the explanatory variables seems to be required. The latter is needed to show that using μ_n rather then μ is asymptotically equivalent for estimation of $\lambda\beta$. At present, the author has been unable to show that this condition could be dispensed by truncation of X and successfully controlling the resulting error. Besides this strong condition, all other conditions appear to be relatively standard within the exponential model set-up. It might be possible that using some clever argument based on the fact that μ_n and σ_n^2 are expectations w.r.t. to the empirical measure (6), one could weaken the conditions used.

The conclusions of Theorem 1 are (1.) that the resulting estimator is consistent as the unfeasible estimator based on knowledge of μ (up to the unknown parameter $\lambda\beta$ to be estimated), and (2.) that confidence intervals can be constructed using weakly consistent estimators of the covariance matrix. The theoretical result does not guarantee that the estimator might perform well in finite samples. Section 3.3 provides some numerical evidence to complement the theoretical one.

The fact that σ_n and W_n are consistent estimators of σ and W respectively (Lemma 6) can be used for model diagnostic. A cross-plot of $\left\{t, \sigma_n^2\left(t, \hat{b}_0\right)\right\}$ can show any possible dependence of the variance of the model on the regressors, as $\sup_t \left|\sigma_n^2\left(t, \hat{b}_0\right) - \sigma^2\left(t, \lambda\beta\right)\right| = o_p\left(1\right)$. The Frobenius norm$\left|R_n\left(\hat{b}_0\right) - W_n\left(\hat{b}_0\right)\right|_2$ (for any $K \times K$ dimensional matrix A, $Trace\left(AA'/K\right)$) can be used to evaluate the loss incurred in estimating μ via μ_n using $loss := \left|R_n\left(\hat{b}_0\right) - W_n\left(\hat{b}_0\right)\right|_2^2 / \left|W_n\left(\hat{b}_0\right)\right|_2^2$, as both $R_n\left(\hat{b}_0\right)$ and $W_n\left(\hat{b}_0\right)$ are consistent for R and W, respectively . Clearly, $loss \simeq 0$ means that the sample is very informative and the semiparametric model holds.

3.1 Relation with More General Methods

The method discussed here falls in between fully parametric Generalized Linear Model estimation and non-parametric estimation of the single index model. Despite the differences, to put the current approach into prospective, it is instructive to highlight common assumptions and results for nonparametric estimation of the single index model. For definiteness, consider the main assumptions of the EFM in Cui et al. (2010). There, identification is assumed. The mean and variance function need to have two continuous derivatives in order to control the approximation error. Discrete regressors are allowed and only a second moment condition is needed for the regressors. The link function, the dependent variable, and the regressors are also constrained implicitly via the expectation of the supremum of the square of the first order conditions in the estimation (Condition (e) in that paper). That condition requires either the regressors to have bounded support, or the link function to be bounded. Cui et al. (2010) show that their estimator for the single index is root-n consistent and efficient.

For binary response variables, other nonparametric procedures have also been studied under even weaker conditions, though at the cost of not achieving root-n consistency. Only imposing a conditional median assumption, allowing for general forms of heteroskedasticity, the method of Maximum-Score (Manski, 1975) attains the cube-root convergence (Kim and Pollard, 1991) and, under additional smoothing restrictions, the smoothed version improves the rate to $n^{-2/5}$, where n is the sample size (Horowitz, 1992). Recently, Khan (2013) has proposed a sieve type estimator for such problems which attains the optimal rate for such sieve estimators. See Gerfin (1996) for a comparison of some of these methods.

3.2 Remarks on Optimization

The solution to (9) and (10) requires non-linear optimization of a function that may not be strictly convex. In consequence, gradient based methods can lead to a local minimum rather than a global one. The usual suggestion is to try different initial solution. In practice, one should attempt to derive an initial solution for (9) based on global optimizers such as genetic algorithms (e.g. Langdon and Poli, 2002, for a textbook reference). Such optimizers are easy to code and routinely available in some computer packages (e.g. Matlab). This initial solution can then be used in gradient based methods such as the Levenberg–Marquardt algorithm or similar trust region algorithms (e.g. Byrd et al., 1987). These remarks are valid not just for the current estimator but for most nonlinear least squares estimators.

3.3 Numerical Experiment

Following Friedman (2001), amongst others, simulations are carried out from a random model. This is done to reduce the dependence of the results on the Monte Carlo set up. The simulation setup is as follows:

$$\begin{aligned} Y_i &= G\left(X_i'\beta\right) + \sigma Z_i \\ G\left(X_i'\beta\right) &= a_0 + a_1 X_i'\beta + a_2 \cos\left(2\pi\left(a_3 X_i'\beta + a_4\right)\right) \end{aligned}$$

where $a_0, a_1, a_2, a_3, a_4 \in [0,1]$, $\beta = (\beta_1, \beta_2, ..., \beta_K)'$, with $\beta_k \in [-1,1]/\sqrt{K}$, $K \in \{2, 3, ..., 50\}$, $X_{ik}, Z_i \sim N(0,1)$, $\sigma \in \left[0, .5\sqrt{Var\left(G\left(X_i'\beta\right)\right)}\right]$. The parameters $a's$, $\beta's$, K and σ are sampled from a uniform distribution in their respective range and the sample size is $n = 400$. The number of simulations is 250. Note that the above model does not belong to (1), as a conditional Gaussian distribution with canonical link always imply a linear G, moreover, even if another conditional distribution were used, G is always monotonic for the model in (1).

For simplicity, the semiparametric estimator (SP) is estimated using (9) only. The estimator of Cui et al. (2010) (EMF) is also computed and used as benchmark. The initial value of b for the estimation of β is set equal to $(1, 1, ..., 1)/\sqrt{K}$ and estimation is carried out using a trust region algorithm for SP and the algorithm in Cui et al. (2010) for EFM.

Figure 1 reports the box plot for the l_1 error

$$\sum_{k=1}^{K}\left|\frac{\hat{b_k}}{\sqrt{\hat{b}'\hat{b}}} - \frac{\beta_k}{\sqrt{\beta'\beta}}\right|$$

where $\hat{b}$ is the estimator from SP, EMF, EMF when the starting value for b is the output of SP (hence SP_EMF), and from the naive linear regression estimation of β (OLS). The EMF estimator requires to tune the smoothing parameter, say h, in the kernel estimation. Here,

$$h = c\frac{\sqrt{\sigma_X \sigma_Y}}{0.6745}\left(\frac{4}{3n}\right)^{\frac{1}{5}}$$

where σ_X is the (in sample) median absolute deviation of X and similarly for σ_Y. The above is just a regression version of Silverman's rule of thumb to find the order of magnitude of smoothing. The constant $c \in \{.05, .25, .5, .75, 1, 1.25, 1.75, 2.25, 2.75, 3.5\}$ is then chosen in each simulation to minimize the ex post l_1 error of the EFM estimator.

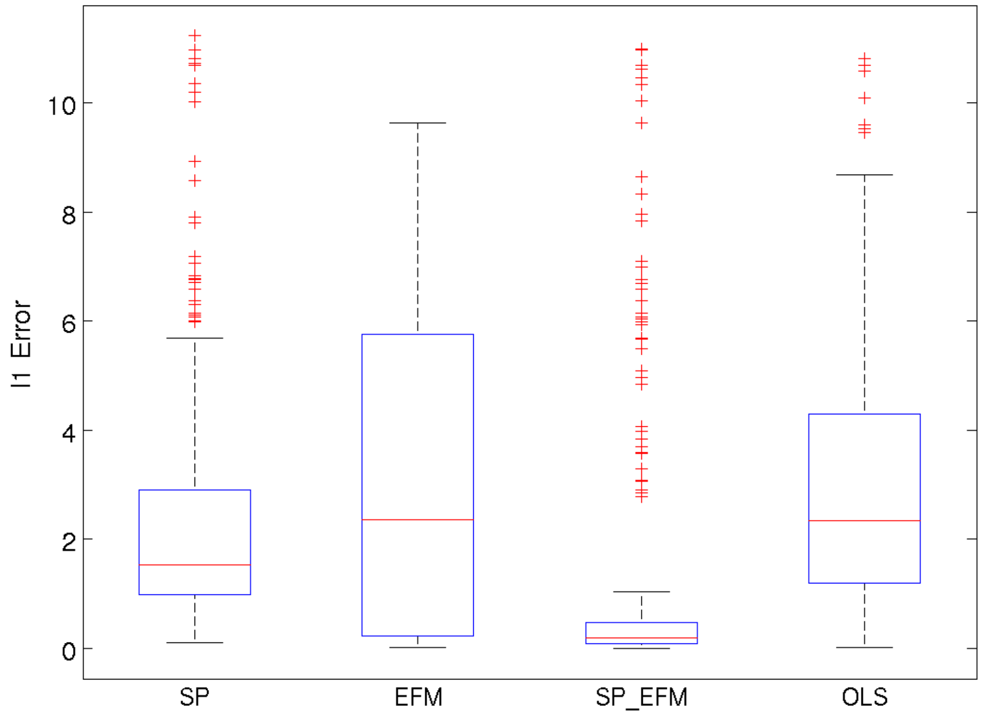

Figure 1. Simulation Results

The goal of the simulation is not to show that the current estimator outperforms the EMF, as this is also quite dependent on the above simulation setup and the optimization algorithm used. Nevertheless, the simulation framework is relatively general and shows that the performance is acceptable. The goal is to show that SP is a viable estimator that does not require fine tuning of smoothing. As the simulation shows, once SP is available, one can then use this as a starting value for more general non-parametric techniques as the EFM, in the present case. It is worth mentioning that the SP estimator was compared to EFM using the Monte Carlo set up in Cui et al. (2010), see also Xia (2006). In that specific case, with the initial guess given by Cui et al. (2010), the performance of EFM was considerably better.

In the present simulations, the author believes that the EFM might get stuck to a local minimum with higher probability than the SP due to the fact that one needs to control smoothing. As shown as shown in Figure 1, once an initial good choice of b and of bandwidth is available, a non-parametric method should outperform the SP estimator, unless the sample size is quite small and the bias of SP is also small.

4 Proofs

Throughout, $b_0 = \lambda\beta$, and

$$\begin{aligned} \mu_n^{(k)}(t,b) &= \frac{1}{n}\sum_{i=1}^{n} Y_i^k \exp\left\{Y_i\left(t - X_i'b\right)\right\} \\ \mu^{(k)}(t,b) &= \mathbb{E}Y_i^k \exp\left\{Y_i\left(t - X_i'b\right)\right\} \end{aligned}$$

so that $\mu_n = \mu_n^{(1)}/\mu_n^{(0)}$ and similarly for μ. For any K dimensional vector x, $|x|_1 = \sum_{k=1}^{K}|x_k|$ is the $l_1$norm where x_k denotes the k^{th} entry in x. The method of proof exploits properties of U-statistics together with uniform convergence. The following is useful in deriving uniform convergence rates.

Lemma 3 *For any finite constants k_l, $l = 0, 1, ..., K$, define*

$$f(x, y; b, t) := y^{k_0}\prod_{l=1}^{K} x_l^{k_l} \exp\left\{y\left(t - x'b\right)\right\},$$

where for a vector x, x_l denotes the l^{th} entry. The class of functions

$$\mathfrak{F} := \{f(\bullet, \bullet; b, t) : b \in \mathcal{B}, |t| \leq T\}$$

has finite envelope function under the L_2 norm and δ-bracketing number w.r.t. the L_2 norm equal to $N(\delta) = O(\delta^{-p})$ for some finite p depending on k_l, $l = 0, 1, ..., K$.

Proof. By the Mean Value Theorem, infer that

$$\begin{aligned} & y^{k_0}\prod_{l=1}^{K} x_l^{k_l} \exp\left\{y\left(t - x'b\right)\right\} - y^{k_0}\prod_{l=1}^{K} x_l^{k_l} \exp\left\{y\left(s - x'a\right)\right\} \\ \leq & \left(y^{k_0}\prod_{l=1}^{K} x_l^{k_l} \sup_{b\in\mathcal{B},|t|\leq T} \exp\left\{y\left(t - x'b\right)\right\}\right)\left(|y|\,|s-t| + \sum_{l=1}^{K}|x_l|\,|a_l - b_l|\right) \end{aligned}$$

$$\leq \quad C\left(\exp\left\{(2T+\epsilon)\,y\right\}\right)\left(|s-t|+\sum_{l=1}^{K}|a_l-b_l|\right) \qquad (11)$$

for some finite absolute constant C that depends on k_l and any $\epsilon > 0$. By Condition 4 $\mathbb{E}\exp\left\{(4T+2\epsilon)\,Y_1\right\} < \infty$ taking $\epsilon_1 = 2\epsilon$. Hence, Theorem 2.7.11 in van der Vaart and Wellner (2000) says that $\mathfrak{F}$ has finite δ-bracketing number under the L_2 norm which is as stated in the lemma because $T < \infty$ and $\mathcal{B}$ is a compact Euclidean set. The fact that the envelope function is finite under the L_2 norm also follows from (11). ■

Lemma 4 *Under Conditions 1, 3 and 4, for finite constants k_l, $l = 0, 1, ..., K$,*

$$\sup_{b\in\mathcal{B}}\sup_{|t|\leq T}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(1-\mathbb{E}^i\right)Y_i^{k_0}\prod_{l=1}^{K}X_{il}^{k_l}\exp\left\{Y_i\left(t-X_i'b\right)\right\}\right| = O_p(1),$$

where X_{il} is the l^{th} entry in the vector X_i. In consequence, for any finite k,

$$\sup_{b\in\mathcal{B}}\sup_{|t|\leq T}\left|\mu_n^{(k)}(t,b)-\mu^{(k)}(t,b)\right| = O_p\left(n^{-1/2}\right)$$

and

$$\sup_{b\in\mathcal{B}}\sup_{|t|\leq T}\left|\frac{1}{\mu_n^{(0)}(t,b)}-\frac{1}{\mu^{(0)}(t,b)}\right| = O_p\left(n^{-1/2}\right).$$

Proof. By Lemma 3, the class of functions $\mathfrak{F}$ has bracketing number under the L_2 norm satisfying $\int_0^\infty\sqrt{\ln N(\delta)}d\delta < \infty$ and a finite envelope function under the L_2 norm. Hence, Theorem 2.5.6 in van der Vaart and Wellner (2000) implies that $\mathfrak{F}$ is Donsker, i.e.

$$\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(1-\mathbb{E}^i\right)Y_i^{k_0}\prod_{l=1}^{K}X_{il}^{k_l}\exp\left\{Y_i\left(t-X_i'b\right)\right\} : |t|\leq T, b\in\mathcal{B}\right\}$$

converges weakly to a Gaussian process with a.s. continuous sample paths. Hence, the first display in the lemma holds by compactness of the parameter space. The second display in the statement is a special case of the first by setting $k_0 = k$ and $k_l = 0$, $l = 1, 2, ..., K$, and then dividing by $\sqrt{n}$. For the the last part note that

$$\begin{aligned} &= \sup_{t,b}\left|\frac{1}{\mu_n^{(0)}(t,b)}-\frac{1}{\mu^{(0)}(t,b)}\right| \\ &\leq \frac{1}{\inf_{t,b}\mu_n^{(0)}(t,b)\,\mu^{(0)}(t,b)}\sup_{t,b}\left|\mu_n^{(0)}(t,b)-\mu^{(0)}(t,b)\right| \end{aligned}$$

$$
\begin{aligned}
&\leq \frac{1}{\inf_{t,b}\mu^{(0)}(t,b)\left(\mu^{(0)}(t,b)+\mu_n^{(0)}(t,b)-\mu^{(0)}(t,b)\right)}O_p\left(n^{-1/2}\right)\\
&= O_p\left(n^{-1/2}\right)
\end{aligned}
$$

using the fact that

$$
\begin{aligned}
\inf_{b\in\mathcal{B},|t|\leq T}\mu^{(0)}(t,b) &\geq \mathbb{E}\exp\{-2T|Y_1|\} \\
&> \epsilon
\end{aligned}
\qquad (12)
$$

/for some $\epsilon>0$, because Y_1 is tight by Condition (4), and using the fact that $\mu_n^{(0)}-\mu^{(0)}$ converges uniformly to zero. ■

The following provides the basic ingredients for asymptotic normality of the estimator.

Lemma 5 *Under Conditions 1, 2, 3 and 4,*

$$\sqrt{n}g_n(b_0)\to N(0,R)$$

in distribution where $R=R(b_0)$ with $R(b)$ as in (8). Moreover,

$$\sup_{b\in\mathcal{B}}|g_n(b)-m_n(b)|=o_p(1)$$

and

$$\sup_{b\in\mathcal{B}}\left|\frac{dg_n(b)}{db}-\frac{dm_n(b)}{db}\right|_1=o_p(1)$$

Proof. Adding and subtracting μ,

$$\sqrt{n}g_n(b)=\sqrt{n}m_n(b)+\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\left(\mu_n\left(X_j'b,b\right)-\mu\left(X_j'b,b\right)\right)X_j.$$

It is convenient to deal with the two terms separately. First note that $\sqrt{n}m_n(b_0)$ is a root-n standardized partial sum of mean zero random with finite variance, hence, it is mean zero with asymptotic variance given by (5). To control the second term in the previous display, note that, omitting arguments for convenience,

$$\mu_n\left(X_j'b,b\right)-\mu\left(X_j'b,b\right) = \frac{\mu_n^{(1)}-\mu^{(1)}}{\mu_n^{(0)}}+\left(\frac{\mu^{(1)}}{\mu^{(0)}}\right)\frac{\mu^{(0)}-\mu_n^{(0)}}{\mu_n^{(0)}}$$

$$
\begin{aligned}
&= \frac{\mu_n^{(1)} - \mu^{(1)}}{\mu^{(0)}} + \left(\frac{\mu^{(1)}}{\mu^{(0)}}\right)\frac{\mu^{(0)} - \mu_n^{(0)}}{\mu^{(0)}} \\
&\quad + \left[\frac{\mu_n^{(1)} - \mu^{(1)}}{\mu^{(0)}} + \left(\frac{\mu^{(1)}}{\mu^{(0)}}\right)\frac{\mu^{(0)} - \mu_n^{(0)}}{\mu^{(0)}}\right] \\
&\quad \times \left(\frac{\mu^{(0)} - \mu_n^{(0)}}{\mu_n^{(0)}}\right) \\
&=: I_n\left(X_j'b, b\right) + II_n\left(X_j'b, b\right) + III_n\left(X_j'b, b\right)
\end{aligned}
$$

with obvious notation in the last definition. Then, for every $b \in \mathcal{B}$,

$$
\begin{aligned}
\sqrt{n}U_n(b) &:= \frac{1}{\sqrt{n}}\sum_{j=1}^{n} X_j\left[I_n\left(X_j'b, b\right) + II_n\left(X_j'b, b\right)\right] \\
&= \frac{1}{n^{3/2}}\sum_{i,j=1}^{n} X_j\left[\frac{\left(1-\mathbb{E}^i\right)Y_i\exp\left\{Y_i\left(X_j - X_i\right)b\right\}}{\mathbb{E}^i\exp\left\{Y_i\left(X_j - X_i\right)b\right\}}\right. \\
&\quad \left.+\frac{\mathbb{E}^iY_i\exp\left\{Y_i\left(X_j - X_i\right)b\right\}}{\mathbb{E}^i\exp\left\{Y_i\left(X_j - X_i\right)b\right\}}\frac{\left(1-\mathbb{E}^i\right)\exp\left\{Y_i\left(X_j - X_i\right)b\right\}}{\mathbb{E}^i\exp\left\{Y_i\left(X_j - X_i\right)b\right\}}\right] \\
&= \frac{1}{n^{3/2}}\sum_{i\neq j} X_j\left[\frac{\left(1-\mathbb{E}^i\right)Y_i\exp\left\{Y_i\left(X_j - X_i\right)b\right\}}{\mathbb{E}^i\exp\left\{Y_i\left(X_j - X_i\right)b\right\}}\right. \\
&\quad \left.+\frac{\mathbb{E}^iY_i\exp\left\{Y_i\left(X_j - X_i\right)b\right\}}{\mathbb{E}^i\exp\left\{Y_i\left(X_j - X_i\right)b\right\}}\frac{\left(1-\mathbb{E}^i\right)\exp\left\{Y_i\left(X_j - X_i\right)b\right\}}{\mathbb{E}^i\exp\left\{Y_i\left(X_j - X_i\right)b\right\}}\right] + o_p(1)
\end{aligned}
$$

is a root-n standardized (mean zero) U-statistic of order 2; the r.h.s. of the first equality is a V-statistic, which is asymptotically equivalent to the U-statistic in the second equality. This U-statistic has non-degenerate kernel, hence by Hoeffding decomposition,

$$
\begin{aligned}
\sqrt{n}U_n(b) &= \frac{1}{n^{1/2}}\sum_{i=1}^{n}\left[\left(1-\mathbb{E}^i\right)Y_i\mathbb{E}^0\frac{X_0\exp\left\{Y_i\left(X_0 - X_i\right)b\right\}}{\mu^{(0)}\left(S_0(b), b\right)}\right. \\
&\quad \left.+\left(1-\mathbb{E}^i\right)\mathbb{E}^0\mu\left(X_0'b, b\right)\frac{X_0\exp\left\{Y_i\left(X_0 - X_i\right)b\right\}}{\mu^{(0)}\left(X_0'b, b\right)}\right] + o_p(1).
\end{aligned}
$$

The first term on the r.h.s. is asymptotically normal (e.g. van der Vaart, 1998, theorem 12.3) with asymptotic variance equal to

$$
\begin{aligned}
&Var\left(Y_1\mathbb{E}^0\frac{X_0\exp\left\{Y_1\left(X_0 - X_1\right)b\right\}}{\mu^{(0)}\left(X_0'b, b\right)} + \mathbb{E}^0\mu\left(X_0'b, b\right)\frac{X_0\exp\left\{Y_1\left(X_0 - X_1\right)b\right\}}{\mu^{(0)}\left(X_0'b, b\right)}\right) \\
&= Var\left(Y_1\bar{p}_1(X; b) + \bar{p}_1(\mu X; b)\right),
\end{aligned}
$$

with $\bar{p}_1$ as defined in (7).

Define

$$IV_n\left(X_j'b,b\right) := \left(\frac{\mu^{(0)}\left(X_j'b,b\right)-\mu_n^{(0)}\left(X_j'b,b\right)}{\mu_n^{(0)}\left(X_j'b,b\right)}\right)$$

Then,

$$\begin{aligned}
\sqrt{n}D_n\left(b\right) &:= \frac{1}{\sqrt{n}}\sum_{j=1}^{n}X_jIII_n\left(X_j'b,b\right)\\
&= \frac{1}{\sqrt{n}}\sum_{j=1}^{n}X_j\left[I_n\left(X_j'b,b\right)+II_n\left(X_j'b,b\right)\right]IV_n\left(X_j'b,b\right).
\end{aligned}$$

Consider each term in the brackets separately. Hence,

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{j=1}^{n}X_jI_n\left(X_j'b,b\right)IV_n\left(X_j'b,b\right)\\
=\;&\frac{1}{n^{5/2}}\sum_{i,j,l=1}^{n}X_j\frac{\left(1-\mathbb{E}^i\right)Y_i\exp\left\{Y_i\left(X_j-X_i\right)'b\right\}}{\mathbb{E}^0\exp\left\{Y_0\left(X_j-X_0\right)'b\right\}}\frac{\left(1-\mathbb{E}^l\right)\exp\left\{Y_l\left(X_j-X_l\right)'b\right\}}{\frac{1}{n}\sum_{l=1}^{n}\exp\left\{Y_l\left(X_j-X_l\right)'b\right\}}\\
=\;&\frac{\frac{1}{n}\sum_{j=1}^{n}\left|X_j\right|}{\inf_{|t|\leq T}\mathbb{E}^0\exp\left\{Y_0\left(t-X_0'b\right)\right\}\frac{1}{n}\sum_{l=1}^{n}\exp\left\{Y_l\left(t-X_l'b\right)\right\}}\\
&\times\frac{1}{n^{1/2}}\sup_{|t|\leq T}\left|\frac{1}{n^{1/2}}\sum_{i=1}^{n}\left(1-\mathbb{E}^i\right)Y_i\exp\left\{Y_i\left(t-X_i'b\right)\right\}\right|\sup_{|t|\leq T}\left|\frac{1}{n^{1/2}}\sum_{i=1}^{n}\left(1-\mathbb{E}^i\right)\exp\left\{Y_i\left(t-X_i'b\right)\right\}\right|\\
=\;&O_p\left(n^{-1/2}\right)
\end{aligned}$$

using Lemma 4 and (12). The second term in $\sqrt{n}D_n\left(b\right)$ is dealt with similarly. Hence, infer that $\sqrt{n}D_n\left(b\right)=O_p\left(n^{-1/2}\right)$ so that it does not contribute the the asymptotic distribution of g_n.

Since $\sqrt{n}m_n\left(b_0\right)$ plus $\sqrt{n}U_n\left(b_0\right)$ forms a sum of i.i.d. random variables, the Central Limit Theorem applies with variance

$$R := Var\left(\left(Y_1-\mu\left(X_1'b_0,b_0\right)\right)X_1+Y_1\bar{p}_1\left(X;b_0\right)+\bar{p}_1\left(\mu X;b_0\right)\right).$$

For the last part of the lemma, note that

$$\begin{aligned}\sup_{b\in\mathcal{B}}\left|g_n(b)-m_n(b)\right| &= \sup_{b\in\mathcal{B}}\left|\frac{1}{n}\sum_{j=1}^{n}\left(\mu_n\left(X_j'b,b\right)-\mu\left(X_j'b,b\right)\right)X_j\right| \\ &\leq \sup_{b\in\mathcal{B}}\left|U_n(b)\right|+\sup_{b\in\mathcal{B}}\left|D_n(b)\right|,\end{aligned}$$

for U_n and D_n as defined above in this proof. The uniform convergence of the above terms follow along the lines of the previous parts of the proof using Lemma 4. Similarly,

$$\sup_{b\in\mathcal{B}}\left|\frac{dg_n(b)}{db}-\frac{dm_n(b)}{db}\right|_1 \leq \sup_{b\in\mathcal{B}}\left|\frac{dU_n(b)}{db}\right|_1+\sup_{b\in\mathcal{B}}\left|\frac{dD_n(b)}{db}\right|_1.$$

From the above proof, it is clear that convergence of the r.h.s. in the above display requires uniform convergence of terms such as

$$\frac{1}{n}\sum_{i=1}^{n}\left(1-\mathbb{E}^i\right)Y_iX_{il}\exp\left\{Y_i\left(t-X_i'b\right)\right\},$$

which again follows by Lemma 4. ■

Lemma 6 *Under Conditions 1, 2, 3 and 4,*

$$\sup_{b\in\mathcal{B}}\left|W_n(b)-W(b)\right|\to 0_{K\times K},$$

where

$$W(b):=\lim_n\frac{1}{n}\sum_{j=1}^{n}\sigma^2\left(X_j'b,b\right)X_jX_j'$$

exists and is elementwise finite, $0_{K\times K}$ *is the* K*-dimensional square matrix of zeros, and* $|\bullet|$ *is understood as elementwise absolute norm. Note that* $W=W(\lambda\beta)$ *is as in (4).*

Proof. At first one needs to show that

$$\sup_{b\in\mathcal{B},|t|\leq T}\left|\sigma_n^2(t,b)-\sigma^2(t,b)\right|=o_p(1). \tag{13}$$

To this end note that

$$\sigma_n^2(t,b):=\frac{\mu_n^{(2)}(t,b)}{\mu_n^{(0)}(t,b)}-\left[\frac{\mu_n^{(1)}(t,b)}{\mu_n^{(0)}(t,b)}\right]^2$$

Then (13) follows from the convergence of the above sample quantities to the population ones using Lemma 4. Convergence of $W_n(b)$ uniform in b follows by ergodicity of $(X_j)_{j\in\mathbb{N}}$ and (13). ■

Lemma 7 *Under Conditions 1, 2, 3 and 4,*

where $|\bullet|$ is understood as elementwise absolute norm.

Proof. Consider the following heuristic steps for two typical terms in the definition of $R_n(b)$,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\left[Y_iX_i\mu_n\left(X_i'b,b\right)\right]^2 &= \frac{1}{n}\sum_{i=1}^{n}\left[Y_iX_i\mu\left(X_i'b,b\right)\right]^2+o_p(1)\\ &= \mathbb{E}\left[Y_1X_1\mu\left(X_1'b,b\right)\right]^2+o_p(1)\end{aligned}$$

and using the definition of p_{in},

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\left[Y_i\frac{1}{n}\sum_{j=1}^{n}X_j\mu_n\left(X_j'b,b\right)np_{in}\left(X_j'b,b\right)\right]^2\\ =\;&\frac{1}{n}\sum_{i=1}^{n}\left[Y_i\frac{1}{n}\sum_{j=1}^{n}X_j\mu\left(X_j'b,b\right)\frac{\exp\left\{Y_i\left(X_j-X_i\right)'b\right\}}{\mu^{(0)}\left(X_j'b,b\right)}+o_p(1)\right]^2\\ =\;&\frac{1}{n^3}\sum_{i,j,l}Y_i^2X_j\mu\left(X_j'b,b\right)\frac{\exp\left\{Y_i\left(X_j-X_i\right)'b\right\}}{\mu^{(0)}\left(X_j'b,b\right)}X_l\mu\left(X_l'b,b\right)\frac{\exp\left\{Y_i\left(X_l-X_i\right)'b\right\}}{\mu^{(0)}\left(X_l'b,b\right)}+o_p(1)\\ =\;&\frac{1}{n^2}\sum_{j,l}\mathbb{E}^0Y_0^2X_j\mu\left(X_j'b,b\right)\frac{\exp\left\{Y_0\left(X_j-X_0\right)'b\right\}}{\mu^{(0)}\left(X_j'b,b\right)}X_l\mu\left(X_l'b,b\right)\frac{\exp\left\{Y_0\left(X_l-X_0\right)'b\right\}}{\mu^{(0)}\left(X_l'b,b\right)}+o_p(1)\\ =\;&\mathbb{E}^0Y_0^2\left[\mathbb{E}^1X_1\mu\left(X_1'b,b\right)\frac{\exp\left\{Y_0\left(X_1-X_0\right)'b\right\}}{\mu^{(0)}\left(X_1'b,b\right)}\right]^2+o_p(1),\end{aligned}$$

where the last display required extra care, as $\sum_{j=1}^{n}X_j\mu\left(X_j'b,b\right)\frac{\exp\left\{Y_i\left(X_j-X_i\right)'b\right\}}{\mu^{(0)}\left(X_j'b,b\right)}$ could not be bounded uniformly in Y_i, but only in $X'b$ (i.e. it was necessary to expand the square and take limit with respect to the sum with index i first). Using Lemma 4, one can make the above arguments precise and uniform in b as well. In the interest of conciseness, the details are left to the reader. ■

Proof. [Theorem 1] At first one establishes consistency of $\hat{b}$. Since

$$|g_n(b) - \mathbb{E}m_n(b)| \leq |g_n(b) - m_n(b)| + |(1 - \mathbb{E})\, m_n(b)|,$$

the second part of Lemma 5 gives uniform convergence to zero of the first term on the r.h.s. The convergence of the second term on the r.h.s. can also be easily established. Then, by Conditions 1 and 2, the consistency follows by standard arguments (e.g. Corollary 3.2.3 in van der Vaart and Wellner, 2000) because $R_n\left(\hat{b}_0\right)$ has a non-stochastic limit by Lemma 7. For this to hold, one needs $\hat{b}_0$ to be consistent, which is the case by the aforementioned arguments and the fact that this estimator is derived from W_{n0} which does not depend on b and has a non-stochastic limit by the law of large numbers and Condition 3. Once consistency is established, Theorem 1 can be explicitly shown using Theorem 3.1 in Hansen (1981). Lemma 5 and 7 give sufficient conditions for the assumptions in Theorem 3.1 of Hansen (1981) to hold, proving the theorem. ■

References

[1] Byrd, R.H., R.B. Schnabel, and G.A. Schultz (1987) A Trust Region Algorithm for Nonlinearly Constrained Optimization. SIAM Journal of Numerical Analysis 24, 1152-1170.

[2] Carroll, R. J., J. Fan, I. Gijbels and M.P. Wand (1997). Generalized Partially Linear Single-Index Models. Journal of the American Statistical Association 92, 447–489.

[3] Chamberlain, G. (2010), Binary Response Models for Panel Data: Identification and Information. Econometrica 78, 159-168.

[4] Cui, X., W.K. Härdle and L. Zhu (2011). The EFM Approach for Single-Index Models. Annals of Statistics 39, 1658-1688.

[5] Fan, Y, W.K. Härdle, W. Wang and L. Zhu (2013) Composite Quantile Regression for the Single-Index Model. SFB 649 Discussion Paper 2013-010, URL: <http://sfb649.wiwi.hu-berlin.de/papers/pdf/SFB649DP2013-010.pdf>

[6] Friedman, J.H. (2001) Greedy Function Approximation: A Gradient Boosting Machine. Annals of Statistics 29, 1189-1232.

[7] Godambe, V.P. and C.C. Heyde (1987) Quasi-likelihood and Optimal Estimation. International Statistical Review 55, 231-244.

[8] Hansen, B.E. (2008). Uniform Convergence Rates for Kernel Estimation with Dependent Data. Econometric Theory 24, 726–748.

[9] Hansen, L.P. (1982). Large Sample Properties of Generalized Method of Moments Estimators. Econometrica 50, 1029-1054.

[10] Härdle, W.K., P. Hall and H. Ichimura (1993). Optimal Smoothing in Single-Index Models. Annals of Statistics 21, 157–178.

[11] Härdle, W.K. and T.M. Stoker (1989). Investigating Smooth Multiple Regression by Method of Average Derivatives. Journal of the American Statistical Association 84, 986–995.

[12] Horowitz, J.L. (1992) A Smoothed Maximum Score Estimator for the Binary Response Model. Econometrica 60, 505-531.

[13] Horowitz, J. L. and W.K. Härdle (1996). Direct Semiparametric Estimation of a Single-Index Model with Discrete Covariates. Journal of the American Statistical Association 91, 1632-1640.

[14] Hristache, M., A. Juditski and V. Spokoiny (2001). Direct Estimation of the Index Coefficients in a Single-Index Model. Annals of Statistics 29, 595-623.

[15] Langdon, W.B. and R. Poli (2002) Foundations of Genetic Programming. New York: Springer.

[16] Jørgensen , B. (1986). Some Properties of Exponential Dispersion Models. Scandinavian Journal of Statistics 13, 187-197.

[17] Jørgensen , B. (1987). Exponential Dispersion Models. Journal of the Royal Statistical Society. Series B 49, 127-162.

[18] Khan, S. (2013) Distribution Free Estimation of Heteroskedastic Binary Response Models Using Probit/Logit Criterion Functions. Journal of Econometrics 172, 168-182.

[19] Kim J. and D. Pollard (1990) Cube Root Asymptotics. Annals of Statistics 18, 191-219.

[20] Manski, C.F. (1975) Maximum Score Estimation of the Stochastic Utility Model of Choice. Journal of Econometrics 3, 205-228.

[21] Manski, C.F. (1988) Identification of Binary Response Models Charles. JASA 83, 729-738.

[22] McCullagh, P. and J.A. Nelder (1989). Generalzied Linear Models. London: Champman and Hall.

[23] Nelder, J.A. and R.W.M. Wedderburn (1972) Generalized Linear Models. Journal of the Royal Statistical Society. Series A 135, 370-384.

[24] Newey, W.K. (1991). Uniform Convergence in Probability and Stochastic Equicontinuity. Econometrica 59, 1161-1167.

[25] Powell, J. L., J.H. Stock, and T.M. Stoker (1989). Semiparametric Estimation of Index Coefficients. Econometrica 57, 1403-1430.

[26] Xia, Y. (2006). Asymptotic Distributions for Two Estimators of the Single-Index Model. Econometric Theory 22, 1112-1137.

[27] Van der Vaart, A. (1998) Asymptotic Statistics. Cambridge: Cambridge University Press.

[28] Van der Vaart, A. and J.A. Wellner (2000) Weak Convergence and Emprical Processes. New York: Springer.