Profile Likelihood and Conditionally Parametric Models

Thomas A. Severini, Wing Hung Wong

Annals of Statistics, Volume 20, Issue 4 (Dec., 1992), 1768-1802.

Stable URL:
http://links.jstor.org/sici?sici=0090-5364%28199212%2920%3A4%3C1768%3APLACPM%3E2.0.CO%3B2-H

The Annals of Statistics
1992, Vol. 20, No. 4, 1768–1802

PROFILE LIKELIHOOD AND CONDITIONALLY PARAMETRIC MODELS

BY THOMAS A. SEVERINI AND WING HUNG WONG

Northwestern University and University of Chicago

In this paper, we outline a general approach to estimating the parametric component of a semiparametric model. For the case of a scalar parametric component, the method is based on the idea of first estimating a one-dimensional subproblem of the original problem that is least favorable in the sense of Stein. The likelihood function for the scalar parameter along this estimated subproblem may be viewed as a generalization of the profile likelihood for that parameter. The scalar parameter is then estimated by maximizing this "generalized profile likelihood." This method of estimation is applied to a particular class of semiparametric models, where it is shown that the resulting estimator is asymptotically efficient.

1. Introduction. Semiparametric models are models containing both parametric and nonparametric components, the nonparametric component playing the role of a nuisance parameter. More precisely, a semiparametric model is parameterized by a parameter of interest taking values in finite-dimensional Euclidean space and a nuisance parameter taking values in an infinite-dimensional space. The goal is then to estimate the parameter of interest in the presence of the infinite-dimensional nuisance parameter; see Bickel, Klaasen, Ritov and Wellner (1991) for a general discussion of estimation in semiparametric models. In this paper, we outline a general approach to estimating the parametric component of a semiparametric model; this general method is then applied to a particular class of semiparametric models, where it is shown that the resulting estimates are asymptotically efficient.

Stein (1956) proposed a method for obtaining a lower bound to the asymptotic variance of an estimator of the parametric component of a semiparametric model which generalizes the bound provided by the inverse of the marginal Fisher information in parametric models. This result is based on the observation that a nonparametric problem is at least "as difficult" as any one-dimensional subproblem. That is, the Fisher information for estimating the parameter of interest in a semiparametric problem is no greater than the Fisher information for estimating that parameter in any one-dimensional subproblem. Hence, by looking at the "least favorable" subproblem, we may obtain a bound on the asymptotic variance of an estimator of the parameter of interest in the original, semiparametric problem. This approach has been further developed by Levit (1974), Koshevnik and Levit (1976), Lindsay (1980), Pfanzagl (1982) and Begun, Hall, Huang and Wellner (1983).

Received February 1991; revised November 1991.

AMS 1980 *subject classifications*. 62F35, 62G07, 62F10.

Key words and phrases. Efficiency, estimation, nonparametric, semiparametric.

For the case of a scalar parameter of interest, the method of estimation proposed here is based on the idea of first estimating a one dimensional subproblem of the original problem which is least favorable in the sense of Stein (1956) and then proceeding to estimate the scalar parameter as if it was known to lie on this curve. The method may be viewed as generalization of maximum likelihood to the semiparametric setting and the proposed estimator reduces to the maximum likelihood estimator in the case of a finite-dimensional nuisance parameter.

The main condition required for this method of estimation is the existence of an estimator of a "curve" in the nuisance parameter space which corresponds to a least favorable subproblem. In this paper we present a general method for obtaining such an estimator for the following class of models. Suppose Y and X are random variables such that the conditional distribution of Y given $X = x$ depends on parameters θ and η, each of which takes values in finite-dimensional Euclidean space, where the value of η depends on the value of x, so that $\eta \equiv \eta_x$. Furthermore, suppose that η_x is a smooth function of x, $\eta_x = \lambda(x)$. Given a random sample from the distribution of (Y, X), our goal is to estimate the parameter of interest θ in the presence of the infinite-dimensional nuisance parameter λ. Since conditional on a particular value of x the model is parameterized by a finite-dimensional parameter, we have called such a semiparametric model a *conditionally parametric* model. In this paper it is shown that the method of estimation described above leads to an asymptotically efficient estimator of the parameter of interest when applied to a conditionally parametric model.

Estimation of a finite dimensional parameter in the presence of an infinite-dimensional nuisance parameter has been considered by a number of authors. Levit (1974, 1975) considered estimation of a functional defined on an infinite-dimensional family of distribution functions. In some cases, the parameter of interest θ can be estimated as well asymptotically when λ is unknown as when λ is known; this situation was studied by Bickel (1982) and his results were extended to the general case by Schick (1986). Another approach to estimation in semiparametric models is given by van der Vaart (1986); see Bickel, Klaasen, Ritov and Wellner (1991) for further references.

The outline of the paper is as follows. Section 2 reviews the marginal Fisher information bound for the asymptotic variance of a regular estimator in parametric problems. In Section 3, the generalization of this bound and the marginal Fisher information to semiparametric models is discussed. The ideas behind the proposed method of estimation are presented in Section 4. In Section 5 conditionally parametric models and their properties are discussed. The large sample properties of the proposed estimator when applied to a conditionally parametric model are presented in Section 6. In Section 7 a general method of obtaining the required estimator of a curve corresponding to a least favorable subproblem is given. In Section 8 an alternative method of estimating such a curve is given for a particular type of conditionally parametric model. Section 9 contains several examples. Technical proofs are given in Section 10.

2. Fisher's bound for asymptotic variances. Let $Y_1, \dots, Y_n$ denote independent observations from a common density $p(\cdot, \phi)$, where ϕ is a p-dimensional parameter; we will use $l(\phi)$ to denote $\log p(Y_1; \phi)$. If a sequence of estimators $\{T_n\}$ satisfies

$$\sqrt{n}(T_n - \phi) \to_{\mathscr{D}} N(0, V),$$

then V is called the asymptotic variance of T_n. Following Fisher and Pitman, the asymptotic variance will be used in this paper for the comparison of estimators. It is well known that, under some regularity conditions, the asymptotic variance of the maximum likelihood estimator (MLE) of ϕ is I_ϕ^{-1}, where I_ϕ denotes the Fisher information matrix for ϕ.

According to Bahadur (1964) and Hájek (1970) any regular estimator (see their papers for detailed regularity conditions) will have asymptotic variance at least as large as that of the MLE. Furthermore, if $\theta = g(\phi)$ is a smooth scalar function of ϕ, then the asymptotic variance of any regular estimator of θ will be at least as large as that of $\hat{\theta} = g(\hat{\phi})$, where $\hat{\phi}$ denotes the MLE of ϕ. That is,

$$i_\theta^{-1} = (\nabla g)' I_\phi^{-1} (\nabla g)$$

provides a lower bound to the asymptotic variance of regular estimators of θ, where (∇g) denotes the gradient of g evaluated at ϕ. We will refer to i_θ as the *marginal Fisher information* for θ, although the terms *effective Fisher information* and *efficient Fisher information* are sometimes also used.

For simplicity of discussion, assume now that ϕ is three-dimensional and that ϕ may be parameterized so that $\phi = (\theta, \lambda_1, \lambda_2)$, where $\theta \in \Theta$ is the real-valued parameter of interest and $\lambda = (\lambda_1, \lambda_2) \in \Lambda$ is a two-dimensional nuisance parameter in the problem of estimating θ. Let $U_\theta, U_{\lambda_1}, U_{\lambda_2}$ denote the partial score functions given by

$$U_\theta = \frac{\partial l}{\partial \theta}, \qquad U_{\lambda_j} = \frac{\partial l}{\partial \lambda_j}, \qquad j = 1, 2.$$

Under standard regularity conditions, the partial scores are mean-zero random variables with finite variance. The Fisher information matrix I_ϕ is the 3×3 matrix with elements

$$E_\phi\left(\frac{\partial l}{\partial \phi_i} \frac{\partial l}{\partial \phi_j}\right),$$

that is, I_ϕ is the matrix whose associated quadratic form determines the inner product in the L_2 space spanned by the partial scores. The marginal Fisher information for θ, i_θ, admits a useful geometric interpretation in this L_2 space: i_θ is the squared length of the residual of U_θ after projection onto $\operatorname{span}\{U_{\lambda_1}, U_{\lambda_2}\}$.

Let $\gamma \mapsto \phi(\gamma) = (\theta(\gamma), \lambda_1(\gamma), \lambda_2(\gamma))$ denote a smooth mapping from a real interval (a, b) to the parameter space $\Theta \times \Lambda$. Then $\{\phi(\gamma): a < \gamma < b\}$ describes a smooth curve in $\Theta \times \Lambda$. Let $\phi_0 = (\theta_0, \lambda_{10}, \lambda_{20})$ denote the true parameter point; the subscript 0 will always be used to denote evaluation at the true state. Assume that the curve may be parameterized so that $\theta(\gamma) = \gamma$, that is, the curve is $\theta \mapsto (\theta, \lambda_\theta)$ with $\lambda_{\theta_0} = \lambda_0$. For each curve λ_θ there is an associated total score function

$$\frac{d}{d\theta} l(\theta, \lambda_\theta)\bigg|_{\theta=\theta_0} = \frac{\partial l}{\partial \theta}(\theta_0, \lambda_0) + \frac{\partial l}{\partial \lambda}(\theta_0, \lambda_0)\left(\frac{d}{d\theta}\lambda_\theta\bigg|_{\theta=\theta_0}\right).$$

The Fisher information for estimating θ along the subproblem defined by this curve is given by

$$E_0\left(\frac{d}{d\theta} l(\theta, \lambda_\theta)\bigg|_{\theta=\theta_0}\right)^2 = E_0(U_\theta + U)^2,$$

where

$$U = \left(\frac{d}{d\theta}\lambda_{1\theta}\bigg|_{\theta=\theta_0}\right)U_{\lambda_1} + \left(\frac{d}{d\theta}\lambda_{2\theta}\bigg|_{\theta=\theta_0}\right)U_{\lambda_2} \in \operatorname{span}\{U_{\lambda_1}, U_{\lambda_2}\}.$$

The minimum Fisher information for θ over all possible one-dimensional subproblems is given by

$$\inf\left\{E_0(U_\theta + U)^2 : U \in \operatorname{span}\{U_{\lambda_1}, U_{\lambda_2}\}\right\} = E_0(U_\theta + U^*)^2,$$

where $U^* = U_{\lambda_1} v_1^* + U_{\lambda_2} v_2^*$ and $-U^*$ is the projection of U_θ onto $\operatorname{span}\{U_{\lambda_1}, U_{\lambda_2}\}$. It follows from the previous characterization of i_θ that $i_\theta = E_0(U_\theta + U^*)^2$.

This provides another useful interpretation of i_θ, as the minimum Fisher information over all possible smooth one-dimensional subproblems. Note that the Fisher information of θ associated with a curve λ_θ depends on the curve only through the tangent vector at the true point,

$$\lambda'_{\theta_0} = \frac{d}{d\theta}\lambda_\theta\bigg|_{\theta=\theta_0}.$$

Any curve with $\lambda'_{\theta_0} = v^* \equiv (v_1^*, v_2^*)$ satisfies

$$i_\theta = E_0\left(\frac{d}{d\theta} l(\theta, \lambda_\theta)\bigg|_{\theta=\theta_0}\right)^2$$

and hence, the Fisher information for θ in this subproblem is minimal among all possible one-dimensional subproblems. We will call such a curve λ_θ a *least favorable curve* and we will call the tangent vector v^* the *least favorable direction*. From the characterization of U^* as a projection, it follows immediately that a necessary and sufficient condition for v^* to be the least favorable direction is that

$$E_0\left(\frac{\partial l}{\partial \theta}(\theta_0, \lambda_0) + \frac{\partial l}{\partial \lambda}(\theta_0, \lambda_0)v^*\right)\left(\frac{\partial l}{\partial \lambda}(\theta_0, \lambda_0)v\right) = 0 \quad \text{for all } v \in \Re^2.$$

3. Extension to the semiparametric case. We now consider the extension of the above results to the case in which the nuisance parameter is infinite-dimensional. Hence, we will now assume that $\phi = (\theta, \lambda)$, where $\theta \in \Theta$ is the real-valued parameter of interest and $\lambda \in \Lambda$, where Λ is an open subset of a normed linear space Λ_0, which may be infinite-dimensional. Stein (1956) generalized the concept of the marginal Fisher information for a scalar parameter of interest in the presence of a finite-dimensional nuisance parameter to the case of an infinite-dimensional nuisance parameter by taking the concept of a least favorable curve as the basis for the generalization. A curve $\theta \mapsto \lambda_\theta$ with $\lambda_{\theta_0} = \lambda_0$ is said to be a least favorable curve if

$$E_0\left(\frac{d}{d\theta}l(\theta, \lambda_\theta)\bigg|_{\theta=\theta_0}\right)^2 \le E_0\left(\frac{d}{d\theta}l(\theta, \lambda_{1\theta})\bigg|_{\theta=\theta_0}\right)^2$$

for any other smooth curve $\theta \mapsto \lambda_{1\theta}$ in Λ with $\lambda_{1\theta_0} = \lambda_0$.

The interpretations concerning the geometry of the partial scores and the least favorable direction remain the same as in the finite-dimensional case. The Fisher information along a particular subproblem parameterized by (θ, λ_θ) is given by

$$E_0\left(\frac{\partial l}{\partial\theta}(\theta_0, \lambda_0) + \frac{\partial l}{\partial\lambda}(\theta_0, \lambda_0)(v)\right)^2,$$

where $v \in \Lambda$ represents the tangent vector to the curve $\theta \mapsto \lambda_\theta$ at θ_0 and $\partial l/\partial\lambda$ represents the Fréchet derivative of l with respect to λ. A vector $v^* \in \Lambda$ is the least favorable direction if

$$E_0\left(\frac{\partial l}{\partial\theta}(\theta_0, \lambda_0) + \frac{\partial l}{\partial\lambda}(\theta_0, \lambda_0)(v^*)\right)^2 = \inf_{v\in\Lambda} E_0\left(\frac{\partial l}{\partial\theta}(\theta_0, \lambda_0) + \frac{\partial l}{\partial\lambda}(\theta_0, \lambda_0)(v)\right)^2.$$

Clearly, as in the finite-dimensional case,

$$-\frac{\partial l}{\partial\lambda}(\theta_0, \lambda_0)(v^*) = \text{projection of } \frac{\partial l}{\partial\theta}(\theta_0, \lambda_0) \text{ onto } \left\{\frac{\partial l}{\partial\lambda}(\theta_0, \lambda_0)(v) \colon v \in \Lambda\right\}$$

and v^* is the least favorable direction if and only if

$$E_0\left(\frac{\partial l}{\partial\theta}(\theta_0, \lambda_0) + \frac{\partial l}{\partial\lambda}(\theta_0, \lambda_0)(v^*)\right)\left(\frac{\partial l}{\partial\lambda}(\theta_0, \lambda_0)(v)\right) = 0 \quad \text{for all } v \in \Lambda.$$

Furthermore, any curve $\theta \mapsto \lambda_\theta$ with $\lambda_{\theta_0} = \lambda_0$ and $\lambda'_{\theta_0} = v^*$ will be called a least favorable curve. Therefore, we define the marginal Fisher information for θ in a semiparametric model by

$$i_\theta = E_0\left(\frac{\partial l}{\partial\theta}(\theta_0, \lambda_0) + \frac{\partial l}{\partial\lambda}(\theta_0, \lambda_0)(v^*)\right)^2.$$

It is clear that this definition reduces to the usual definition of marginal Fisher information in the case of a finite-dimensional nuisance parameter. Furthermore, i_θ^{-1} plays the same role in providing a lower bound to the asymptotic variance of a regular estimator of θ in a semiparametric model as the inverse of the marginal Fisher information for θ plays in a parametric model. See Levit (1974), Lindsay (1980), Begun, Hall, Huang and Wellner (1983), Wong

(1986) and Bickel, Klaasen, Ritov and Wellner (1991) for further discussion of this bound and technical details.

In the parametric case, it is well known that under fairly general conditions, there exist many estimators with asymptotic variance equal to i_θ^{-1}, so that i_θ^{-1} is a valid criterion for efficiency. In the semiparametric case, i_θ^{-1} remains a lower bound to the asymptotic variance of a regular estimator. However, there are few results on the construction of estimators which achieve this lower bound. Wong and Severini (1991) show that under strong regularity conditions, the maximum likelihood estimator of θ is one such estimator; Bickel, Klaasen, Ritov and Wellner (1991) describe the construction of efficient estimators in several specific examples and give further references. In this paper, we present an approach to estimating the parametric component of a semiparametric model that can be applied generally. Furthermore, it is shown that for the rich class of conditionally parametric models, the proposed estimator has variance equal to i_θ^{-1} and hence, for these models, the estimator is asymptotically efficient.

4. Least-favorable curves and generalized profile likelihood.

Suppose that we are able to identify a curve λ_θ in Λ satisfying $\lambda_{\theta_0} = \theta_0$. Then we may use the log-likelihood

$$L_n(\theta, \lambda_\theta) = \sum \log p(Y_j; \theta, \lambda_\theta)$$

for the estimation of θ. The MLE $\tilde{\theta}$ based on $L_n(\theta, \lambda_\theta)$ will have asymptotic variance

$$\left[E_0\left(\frac{d}{d\theta} l(\theta, \lambda_0)\bigg|_{\theta=\theta_0}\right)^2\right]^{-1},$$

which we have seen, is less than or equal to i_θ^{-1}. Of course in practice such a curve λ_θ is not available. However, suppose we are able to obtain an estimate $\hat{\lambda}_\theta$ (based on $Y_1, \dots, Y_n$) of a curve λ_θ satisfying $\lambda_{\theta_0} = \lambda_0$; we may then obtain an estimate of θ by maximizing $L_n(\theta, \hat{\lambda}_\theta)$. To carry out this approach we need to determine which curve λ_θ to estimate and also which estimator $\hat{\lambda}_\theta$ to use. Clearly, we want to make these choices so that the resulting estimator is asymptotically efficient, if possible. To gain some insight into these choices consider the situation in which λ is finite-dimensional.

For fixed θ, let $\hat{\lambda}_\theta$ denote a $\lambda \in \Lambda$ that maximizes $L_n(\theta, \lambda)$. Then $L_n(\theta, \hat{\lambda}_\theta)$ is called the *profile log-likelihood* for θ. Maximizing $L_n(\theta, \hat{\lambda}_\theta)$ leads to an estimator $\hat{\theta}$ with asymptotic variance i_θ^{-1}; in fact, $\hat{\theta}$ is exactly the MLE of θ. To understand how we may generalize this procedure, consider the following.

Let λ_θ denote the value of $\lambda \in \Lambda$ that minimizes $K(\theta, \lambda)$ for fixed θ, where $K(\theta, \lambda)$ is given by

$$K(\theta, \lambda) = -\int \log p(y; \theta, \lambda) p(y; \theta_0, \lambda_0)\, dy.$$

Let $\hat{\lambda}_\theta$ denote the value of $\lambda \in \Lambda$ that maximizes $L_n(\theta, \lambda)$. Then, under standard regularity conditions, $\hat{\lambda}_\theta$ is a consistent estimator of λ_θ [Huber (1967)]. Furthermore, if $\lambda_{1\theta}$ is another curve in Λ satisfying $\lambda_{1\theta_0} = \lambda_0$, then by the definition of λ_θ,

$$\int (h(\theta) - h_1(\theta)) p(y; \theta_0, \lambda_0)\, dy \geq 0 \quad \text{for all } \theta,$$

where $h(\theta) = \log p(y; \theta, \lambda_\theta)$ and $h_1(\theta) = \log p(y; \theta, \lambda_{1\theta})$. Hence, expanding h and h_1 in terms of θ around θ_0 and using the fact that $E_0 h'(\theta_0) = E_0 h_1'(\theta_0) = 0$ we obtain that

$$\int h''(\theta_0) p(y; \theta_0, \lambda_0)\, dy \geq \int h_1''(\theta_0) p(y; \theta_0, \lambda_0)\, dy,$$

that is,

$$-E_0 \frac{d^2}{d\theta^2} \log p(Y_1; \theta, \lambda_\theta)\bigg|_{\theta=\theta_0} \leq -E_0 \frac{d^2}{d\theta^2} \log p(Y_1; \theta, \lambda_{1\theta})\bigg|_{\theta=\theta_0}.$$

Since, $\lambda_{1\theta}$ is arbitrary, this implies that λ_θ is a least favorable curve. The same argument holds when λ is an infinite-dimensional parameter provided that $h(\theta)$ and $h_1(\theta)$ are twice-differentiable functions of θ and that $E_0 h'(\theta_0) = E_0 h_1'(\theta_0) = 0$; once attention is restricted to a curve λ_θ the problem is essentially finite-dimensional. Furthermore, the results of Bahadur (1971) on the consistency of the MLE suggest that $\hat{\lambda}_\theta$ is still a consistent estimator of λ_θ.

Note that maximizing $L_n(\theta, \hat{\lambda}_\theta)$ yields the MLE, an estimator with the same asymptotic distribution as the estimator obtained by maximizing $L_n(\theta, \lambda_\theta)$, where λ_θ is the curve previously described. However, why should we choose $\hat{\lambda}_\theta$ to be the estimator of a least favorable curve, which by definition, maximizes the asymptotic variance of the resulting estimator. Of course, we know that using another curve will not result in an estimator with smaller variance, since i_θ^{-1} is a lower bound to the asymptotic variance of a regular estimator. The following argument gives insight into why this is so and is useful in understanding the proofs of the later sections.

If λ_θ is a curve in Λ and $\hat{\lambda}_\theta$ is an estimator of λ_θ, then we would like the estimator $\hat{\theta}$ obtained by maximizing $L_n(\theta, \hat{\lambda}_\theta)$ to have the same asymptotic distribution as the estimator obtained by maximizing $L_n(\theta, \lambda_\theta)$; in that case $\sqrt{n}(\hat{\theta} - \theta)$ will be asymptotically normally distributed with mean 0 and variance depending on the curve λ_θ. For this to occur $L_n(\theta, \hat{\lambda}_\theta)$ and $L_n(\theta, \lambda_\theta)$ must have the same local behavior, as functions of θ, near $\theta = \theta_0$. In particular, we need that

$$\frac{1}{\sqrt{n}} \frac{dL_n(\theta, \hat{\lambda}_\theta)}{d\theta}\bigg|_{\theta=\theta_0} = \frac{1}{\sqrt{n}} \frac{dL_n(\theta, \lambda_\theta)}{d\theta}\bigg|_{\theta=\theta_0} + o_p(1).$$

Now,

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\frac{dL_n(\theta,\hat{\lambda}_\theta)}{d\theta}\bigg|_{\theta=\theta_0}-\frac{1}{\sqrt{n}}\frac{dL_n(\theta,\lambda_\theta)}{d\theta}\bigg|_{\theta=\theta_0}\\
&\quad\doteq\frac{1}{\sqrt{n}}\frac{d}{d\theta}\left(\frac{\partial L_n}{\partial\lambda}(\theta,\lambda_\theta)(\hat{\lambda}_\theta-\lambda_\theta)\right)\bigg|_{\theta=\theta_0}\\
(1)\qquad&\quad=\frac{1}{\sqrt{n}}\frac{d}{d\theta}\frac{\partial L_n}{\partial\lambda}(\theta,\lambda_\theta)\bigg|_{\theta=\theta_0}(\hat{\lambda}_0-\lambda_0)\\
&\quad\quad+\frac{1}{\sqrt{n}}\frac{\partial L_n}{\partial\lambda}(\theta,\lambda_\theta)\bigg|_{\theta=\theta_0}(\hat{\lambda}'_0-\lambda'_0).
\end{aligned}$$

Under regularity conditions,

$$\frac{\partial L_n}{\partial\lambda}(\theta,\lambda_\theta)\bigg|_{\theta=\theta_0}$$

is the sum of n mean-zero random variables, so the second term of (1) will be of order $o_p(1)$ [provided that $(\hat{\lambda}'_0-\lambda'_0)\to 0$]. However, in general,

$$\frac{d}{d\theta}\frac{\partial L_n}{\partial\lambda}(\theta,\lambda_\theta)\bigg|_{\theta=\theta_0}$$

is a sum of n random variables with nonzero means. Hence, the first term in (1) is, in general, not of order $o_p(1)$. In this case,

$$\frac{1}{\sqrt{n}}\frac{dL_n(\theta,\hat{\lambda}_\theta)}{d\theta}\bigg|_{\theta=\theta_0}=\frac{1}{\sqrt{n}}\frac{dL_n(\theta,\lambda_\theta)}{d\theta}\bigg|_{\theta=\theta_0}+B_n+o_p(1),$$

where B_n is a bias term that is of greater order than $o_p(1)$. However, when λ_θ is a least favorable curve, then for any tangent vector v

$$\begin{aligned}
&E_0\left(\frac{d}{d\theta}\frac{\partial l}{\partial\lambda}(\theta,\lambda_\theta)\bigg|_{\theta=\theta_0}(v)\right)\\
&\quad=E_0\left(\frac{\partial^2 l}{\partial\theta\,\partial\lambda}(\theta_0,\lambda_0)(v)+\frac{\partial^2 l}{\partial\lambda}(\theta_0,\lambda_0)(v,v^*)\right)\\
&\quad=-E_0\left(\frac{\partial l}{\partial\theta}(\theta_0,\lambda_0)+\frac{\partial l}{\partial\lambda}(\theta_0,\lambda_0)(v^*)\right)\left(\frac{\partial l}{\partial\lambda}(\theta_0,\lambda_0)(v)\right)=0
\end{aligned}$$

by the properties of the least favorable direction described in Section 3.

Hence, when λ_θ is a least favorable curve, the first term in (1) will also be of order $o_p(1)$ and under additional regularity conditions, the estimators obtained by maximizing $L_n(\theta,\hat{\lambda}_\theta)$ and $L_n(\theta,\lambda_\theta)$ will have the same asymptotic distribution. That is, if $\hat{\theta}$ denotes the maximizer of $L_n(\theta,\hat{\lambda}_\theta)$, then

$$\sqrt{n}\left(\hat{\theta}-\theta_0\right)\to_{\mathscr{D}} N\left(0,i_\theta^{-1}\right).$$

This illustrates the importance of obtaining an estimator of a least favorable curve; if λ_θ is not a least favorable curve, the estimator $\hat{\theta}$ will be asymptotically biased. When λ_θ is a least favorable curve this bias term disappears due to the orthogonality of the score functions, as discussed in Section 3.

When $\hat{\lambda}_\theta$ is the MLE of λ (for fixed θ), it is well known that maximizing $L_n(\theta, \hat{\lambda}_\theta)$, the profile log-likelihood for θ, leads to an asymptotically efficient estimator of θ. However, it is clear from the above argument that this property of the profile likelihood is shared by other functions. Hence, if λ_θ is any least favorable curve and $\hat{\lambda}_\theta$ is a consistent estimator of λ_θ for each θ, then $L_n(\theta, \hat{\lambda}_\theta)$ will be called a *generalized profile likelihood* for θ. An estimator of θ may then be obtained by maximizing $L_n(\theta, \hat{\lambda}_\theta)$; under some regularity conditions, this estimator is asymptotically efficient.

These observations form the basis of our approach to the estimation of the parameter of interest in a semiparametric model. Suppose that λ is a infinite-dimensional nuisance parameter and let λ_θ denote a least favorable curve. If $\hat{\lambda}_\theta$ is an estimator of λ_θ, then an estimator of θ can be obtained by maximizing the generalized profile log-likelihood $L_n(\theta, \hat{\lambda}_\theta)$. We expect that, under some regularity conditions, the resulting estimator is asymptotically efficient. Note that in many semiparametric problems the use of a *generalized* profile likelihood is absolutely necessary since the profile likelihood requires maximization over an infinite-dimensional space and hence, may not exist or may lead to an inconsistent estimator of θ.

In the remaining sections of the paper, we apply the approach described above to a particular class of semiparametric models. For these models, it is shown that an estimator of a least favorable curve may be obtained and the estimator of θ obtained by maximizing the generalized profile likelihood is asymptotically efficient.

5. Conditionally parametric models. Let $\{p(\cdot; \theta, \eta): \theta \in \Theta, \eta \in H\}$ denote a family of density functions indexed by parameters θ, η. Assume that Θ is a compact subset of $\mathfrak{R}$, H is a compact subset of $\mathfrak{R}$ and that $p(y; \theta, \eta)$ is a measurable function of y for each θ, η. Suppose we observe random variables Y, X, $X \in [0, 1]$, such that the conditional distribution of Y given $X = x$ has density $p(y; \theta, \eta_x)$, where η_x depends on the value of x. Assume that $\eta_x = \lambda(x)$ for some smooth function $\lambda: [0, 1] \mapsto H$, taking values in a set Λ,

$$\Lambda = \{h \in C^2[0, 1]: h(x) \in \operatorname{int}(H) \text{ for all } x \in [0, 1]\}.$$

Suppose we observe independent replicates of (X, Y), $(X_1, Y_1), \dots, (X_n, Y_n)$; our goal is to estimate θ in the presence of the nuisance parameter λ. It is worth noting that the results below can be extended to the case in which θ, η, X are each multidimensional.

We now introduce some notation and terminology that will be used in the remainder of the paper.

Let

$$l(y; \theta, \eta) = \log p(y; \theta, \eta)$$

and

$$l_j(\theta, \lambda) = l(Y_j; \theta, \eta_j), \qquad \eta_j = \lambda(x_j).$$

We can easily obtain partial derivatives of l with respect to λ: For any functions $v_1, \dots, v_s$ on [0, 1],

$$\frac{\partial^{r+s} l_j}{\partial\theta^r \, \partial\lambda^s}(\theta, \lambda)(v_1, \dots, v_s) = \frac{\partial^{r+s} l}{\partial\theta^r \, \partial\eta^s}(Y_j; \theta, \eta_j) v_1(X_j) \cdots v_s(X_j).$$

Let $L_n(\theta, \lambda) = \Sigma l_j(\theta, \lambda)$ so that

$$\begin{aligned}\frac{\partial^{r+s} L_n}{\partial\theta^r \, \partial\lambda^s}(\theta, \lambda)(v_1, \dots, v_s) &= \sum \frac{\partial^{r+s} l_j}{\partial\theta^r \, \partial\lambda^s}(\theta, \lambda)(v_1, \dots, v_s) \\ &= \sum \frac{\partial^{r+s} l}{\partial\theta^r \, \partial\eta^s}(Y_j; \theta, \eta_j) v_1(X_j) \cdots v_s(X_j).\end{aligned}$$

Let (θ_0, λ_0) denote the true parameter values; we assume that $\theta_0 \in \operatorname{int}(\Theta)$. Let E_0 denote expectation under (θ_0, λ_0).

We require that the family of density functions $\{p(\cdot; \theta, \eta)\colon \theta \in \Theta,\ \eta \in H\}$ satisfies the following identifiability (I) and smoothness (S) conditions.

CONDITIONS I. (a) For fixed but arbitrary θ_1, η_1, where $\theta_1 \in \Theta$ and $\eta_1 \in H$, let

$$\rho(\theta, \eta) = \int \log p(y; \theta, \eta) p(y; \theta_1, \eta_1)\, dy, \qquad \theta \in \Theta,\ \eta \in H.$$

If $\theta \neq \theta_1$, then

$$\rho(\theta, \eta) < \rho(\theta_1, \eta_1).$$

(b) Let $\tilde{i}_\theta(\theta, \eta)$ denote the marginal Fisher information for θ in the parametric model, that is,

$$\begin{aligned}\tilde{i}_\theta(\theta, \eta) &= E_{\theta,\eta}\left(\frac{\partial l}{\partial\theta}(Y; \theta, \eta)^2\right) \\ &\quad - E_{\theta,\eta}\left(\frac{\partial l}{\partial\theta}(Y; \theta, \eta)\frac{\partial l}{\partial\eta}(Y; \theta, \eta)\right)^2 E_{\theta,\eta}\left(\frac{\partial l}{\partial\eta}(Y; \theta, \eta)^2\right)^{-1}.\end{aligned}$$

Then assume that $\tilde{i}_\theta(\theta, \eta) > 0$ for all $\theta \in \Theta$, $\eta \in H$.

CONDITIONS S. Assume that for all $r, s = 0, \dots, 4$, $r + s \leq 4$, the derivative

$$\frac{\partial^{r+s} l}{\partial\theta^r \, \partial\eta^s}(y; \theta, \eta)$$

exists for almost all y and that

$$E_0\left\{\sup_{\theta\in\Theta}\sup_{\eta\in H}\left|\frac{\partial^{r+s} l}{\partial\theta^r \, \partial\eta^s}(Y; \theta, \eta)\right|^2\right\} < \infty.$$

Note that under Conditions S it is permissible to interchange differentiation and integration when differentiating

$$\int \log p(y;\theta,\eta)p(y;\theta,\eta)\,dy.$$

For models of this form we may derive explicit expressions for v^* and i_θ.

LEMMA 1. *For the model described above, the least favorable direction v^* is given by*

$$v^*(x) = -\frac{E_0((\partial^2 l/\partial\theta\,\partial\eta)(Y;\theta_0,\lambda_0(X))|X=x)}{E_0((\partial^2 l/\partial\eta^2)(Y;\theta_0,\lambda_0(X))|X=x)}.$$

The marginal Fisher information for θ is given by

$$i_\theta \equiv i_\theta(\theta_0,\lambda_0) = E_0\{\tilde{i}_\theta(\theta_0,\lambda_0(X))\}.$$

PROOF. v^* must satisfy

$$E_0\left\{\left[\frac{\partial l}{\partial\theta}(Y;\theta_0,\lambda_0(X)) + \frac{\partial l}{\partial\eta}(Y;\theta_0,\lambda_0(X))v^*(X)\right]\frac{\partial l}{\partial\eta}(Y;\theta_0,\lambda_0(X))v(X)\right\}$$
$$= 0$$

for all continuous functions v [Begun, Hall, Huang and Wellner (1983)]. Hence, a sufficient condition for v^* to be the least favorable direction is that

$$E_0\left\{\frac{\partial l}{\partial\theta}(Y;\theta_0,\lambda_0(X))\frac{\partial l}{\partial\eta}(Y;\theta_0,\lambda_0(X))\middle|X=x\right\}v(x)$$
$$= -E_0\left\{\left[\frac{\partial l}{\partial\eta}(Y;\theta_0,\lambda_0(X))\right]^2\middle|X=x\right\}v^*(x)v(x)$$

for all continuous functions v and all $x \in [0,1]$. The result follows. Note that Conditions S and the fact that $\lambda_0 \in C^2[0,1]$ imply that $v^* \in C^2[0,1]$.

The expression for i_θ is then obtained by calculating

$$E_0\left\{\left[\frac{\partial l}{\partial\theta}(Y;\theta_0,\lambda_0(X)) + \frac{\partial l}{\partial\eta}(Y;\theta_0,\lambda_0(X))v^*(X)\right]^2\right\}. \qquad \Box$$

We can now define a least favorable curve for a conditionally parametric model. Call a function λ_θ: $\Theta \to \Lambda$ a least favorable curve if:

(a) $\lambda_{\theta_0} = \lambda_0$.

(b) For each $x \in [0,1]$

$$\lambda'_\theta(x) \equiv \frac{\partial}{\partial\theta}\lambda_\theta(x) \quad\text{and}\quad \lambda''_\theta(x) \equiv \frac{\partial^2}{\partial\theta^2}\lambda_\theta(x)$$

exist and $\|\lambda'_\theta\|$ and $\|\lambda''_\theta\|$ are finite, where

$$\|h\| = \sup_{x\in[0,1]} |h(x)|.$$

(c) For each $x \in [0,1]$,

$$\lambda_0'(x) \equiv \lambda_\theta'(x)|_{\theta=\theta_0} = v^*(x).$$

The results given below can also be applied to the following generalization of the conditionally parametric model described above. Suppose Y consists of two components $Y = (Y_1, Y_2)$ such that the conditional density of Y_1 given X and Y_2 is of the form $p(\cdot|y_2; \theta, \eta_x)$ and the conditional density of Y_2 given X does not depend on θ or η. Then the log-likelihood for a single observation can be written

$$\log p(Y_1|Y_2; \theta, \eta_X) + \log p_{Y_2|X}(Y_2|X) + \log p_X(X)$$

and since only the first term in this expression depends on the parameters, we can take

$$l(y; \theta, \eta) = \log p(y_1|y_2; \theta, \eta_x).$$

The results regarding the estimation of θ apply directly to this case as well; note, however, that certain expressions, like the ones given in Lemma 1, must be modified.

We now consider the large sample properties of the estimator of θ described in the previous section as applied to a conditionally parametric model.

6. Large sample properties of the estimator. We will use the following method to estimate the parameter of interest θ. Let λ_θ denote a least favorable curve in Λ and let $\hat{\lambda}_\theta$ denote an estimator of λ_θ. Let $\hat{\theta}$ denote the value of $\theta \in \Theta$ that maximizes the log-likelihood along the estimated curve, given by $L_n(\theta, \hat{\lambda}_\theta)$. In this section, it will be shown that $\hat{\theta}$ is an asymptotically efficient estimator of θ.

For this method to be successful, the estimator $\hat{\lambda}_\theta$ must satisfy the following conditions; methods for constructing an estimator satisfying these conditions will be considered in Sections 7 and 8.

CONDITIONS NP (Nuisance parameter). (a) For each $x \in [0,1]$ and each $\theta \in \Theta$, $\hat{\lambda}_\theta(x)$ converges in probability to some constant as $n \to \infty$; denote that constant by $\tilde{\lambda}_\theta(x)$. Assume that for each $\theta \in \Theta$, $\tilde{\lambda}_\theta \in \Lambda$ and that for all $r, s = 0, 1, 2$, $r + s \leq 2$,

$$\frac{\partial^{r+s}}{\partial x^r \partial \theta^s} \tilde{\lambda}_\theta(x) \quad \text{and} \quad \frac{\partial^{r+s}}{\partial x^r \partial \theta^s} \hat{\lambda}_\theta(x)$$

exist. Let

$$\tilde{\lambda}_0 = \tilde{\lambda}_\theta \bigg|_{\theta=\theta_0} \quad \text{and} \quad \tilde{\lambda}_0' = \frac{d}{d\theta} \tilde{\lambda}_\theta \bigg|_{\theta=\theta_0}.$$

Then suppose

$$\left\| \hat{\lambda}_0 - \tilde{\lambda}_0 \right\| = o_p(n^{-\alpha})$$

and

$$\left\|\hat{\lambda}'_0 - \tilde{\lambda}'_0\right\| = o_p(n^{-\beta}),$$

where $\alpha + \beta \geq 1/2$ and $\alpha \geq 1/4$.

Furthermore, suppose that $\sup_{\theta\in\Theta}\|\hat{\lambda}_\theta - \tilde{\lambda}_\theta\|$, $\sup_{\theta\in\Theta}\|\hat{\lambda}'_\theta - \tilde{\lambda}'_\theta\|$ and $\sup_{\theta\in\Theta}\|\hat{\lambda}''_\theta - \tilde{\lambda}''_\theta\|$ are all of order $o_p(1)$ as $n \to \infty$.

For some $\delta > 0$, assume that

$$\left\|\frac{\partial}{\partial x}\hat{\lambda}_0 - \frac{\partial}{\partial x}\tilde{\lambda}_0\right\| = o_p(n^{-\delta})$$

and

$$\left\|\frac{\partial}{\partial x}\hat{\lambda}'_0 - \frac{\partial}{\partial x}\tilde{\lambda}'_0\right\| = o_p(n^{-\delta}).$$

(b) The curve $\tilde{\lambda}_\theta$ is a least favorable curve as defined in the previous section.

We now establish the consistency of $\hat{\theta}$; the proof of the proposition is in Section 10.

PROPOSITION 1. *For each n define $\hat{\theta} \equiv \hat{\theta}_n$ to be any element of Θ satisfying*

$$L_n\left(\hat{\theta}, \hat{\lambda}_{\hat{\theta}}\right) = \sup_{\theta\in\Theta} L_n\left(\theta, \hat{\lambda}_\theta\right).$$

Then, under the above regularity conditions,

$$\hat{\theta} \to_p \theta_0 \quad \text{as } n \to \infty.$$

The following proposition establishes that $\hat{\theta}$ is asymptotically normally distributed with asymptotic variance equal to the marginal Fisher information for θ.

PROPOSITION 2. *Under the above regularity conditions,*

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \to_{\mathscr{D}} N\left(0, i_\theta^{-1}\right).$$

PROOF. Using a Taylor's series expansion,

$$\begin{aligned} 0 &= \left.\frac{dL_n\left(\theta, \hat{\lambda}_\theta\right)}{d\theta}\right|_{\theta=\hat{\theta}} \\ &= \left.\frac{dL_n\left(\theta, \hat{\lambda}_\theta\right)}{d\theta}\right|_{\theta=\theta_0} + \left.\frac{d^2L_n}{d\theta^2}\left(\theta, \hat{\lambda}_\theta\right)\right|_{\theta=\hat{\theta}^*}\left(\hat{\theta} - \theta_0\right), \end{aligned}$$

where $\hat{\theta}^*$ lies between $\hat{\theta}$ and θ_0 and hence, by Proposition 1, $\hat{\theta}^* \to_p \theta_0$.

Therefore,

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) = \frac{(1/\sqrt{n})dL_n\left(\theta, \hat{\lambda}_\theta\right)/d\theta\big|_{\theta=\theta_0}}{-(1/n)\left(d^2L_n/d\theta^2\right)\left(\theta, \hat{\lambda}_\theta\right)\big|_{\theta=\hat{\theta}^*}}.$$

The result now follows provided that

$$(2) \qquad \frac{1}{\sqrt{n}} \frac{dL_n\left(\theta, \hat{\lambda}_\theta\right)}{d\theta}\Bigg|_{\theta=\theta_0} = \frac{1}{\sqrt{n}} \frac{dL_n\left(\theta, \lambda_\theta\right)}{d\theta}\Bigg|_{\theta=\theta_0} + o_p(1)$$

and

$$(3) \qquad \sup_\theta \left| \frac{1}{n} \frac{d^2L_n}{d\theta^2}\left(\theta, \hat{\lambda}_\theta\right) - \frac{1}{n} \frac{d^2L_n}{d\theta^2}\left(\theta, \lambda_\theta\right) \right| = o_p(1)$$

hold. To verify (2) and (3) we will use the following lemmas; the proofs are given in Section 10.

LEMMA 2. *Under the above regularity conditions*:

$$\text{(i) } \frac{1}{\sqrt{n}} \frac{d}{d\theta} \frac{\partial L_n}{\partial \lambda}(\theta, \lambda_\theta)\Bigg|_{\theta=\theta_0} \left(\hat{\lambda}_0 - \lambda_0\right) = o_p(1).$$

$$\text{(ii) } \frac{1}{\sqrt{n}} \frac{\partial L_n}{\partial \lambda}(\theta, \lambda_\theta)\Bigg|_{\theta=\theta_0} \left(\hat{\lambda}'_0 - \lambda'_0\right) = o_p(1).$$

LEMMA 3. *Under the above regularity conditions*:

(i) $L_n(\theta, \hat{\lambda}_\theta) - L_n(\theta, \lambda_\theta) = r_n^{(1)}(\theta)$, *where*

$$\sup_\theta \left| n^{-1} \frac{d^2}{d\theta^2} r_n^{(1)}(\theta) \right| = o_p(1).$$

(ii) $L_n(\theta, \hat{\lambda}_\theta) = L_n(\theta, \lambda_\theta) + (\partial L_n/\partial\lambda)(\theta, \lambda_\theta)(\hat{\lambda}_\theta - \lambda_\theta) + r_n^{(2)}(\theta)$, *where*

$$n^{-1/2} \frac{d}{d\theta} r_n^{(2)}(\theta)\Bigg|_{\theta=\theta_0} = o_p(1).$$

Using these lemmas we can now verify (2) and (3) above. Consider (2):

$$\frac{1}{\sqrt{n}} \frac{d}{d\theta}\left(L_n\left(\theta, \hat{\lambda}_\theta\right) - L_n\left(\theta, \lambda_\theta\right)\right) = \frac{1}{\sqrt{n}} \frac{d}{d\theta}\left(\frac{\partial L_n}{\partial \lambda}(\theta, \lambda_\theta)\left(\hat{\lambda}_\theta - \lambda_\theta\right) + r_n^{(2)}(\theta)\right).$$

Therefore,

$$\begin{aligned}
&\frac{1}{\sqrt{n}} \frac{d}{d\theta} L_n\left(\theta, \hat{\lambda}_\theta\right)\Bigg|_{\theta=\theta_0} - \frac{1}{\sqrt{n}} \frac{d}{d\theta} L_n\left(\theta, \lambda_\theta\right)\Bigg|_{\theta=\theta_0} \\
&\quad = \frac{1}{\sqrt{n}} \frac{d}{d\theta} \frac{\partial L_n}{\partial \lambda}(\theta, \lambda_\theta)\Bigg|_{\theta=\theta_0} \left(\hat{\lambda}_0 - \lambda_0\right) \\
&\qquad + \frac{1}{\sqrt{n}} \frac{\partial L_n}{\partial \lambda}(\theta_0, \lambda_0)\left(\hat{\lambda}'_0 - \lambda'_0\right) + \frac{1}{\sqrt{n}} \frac{d}{d\theta} r_n^{(2)}(\theta)\Bigg|_{\theta=\theta_0} = o_p(1)
\end{aligned}$$

by Lemma 2 and Lemma 3(ii).

Equation (3) follows immediately from Lemma 3(i), completing the proof of Proposition 2. □

The generalized profile log-likelihood function $L_n(\theta, \hat{\lambda}_\theta)$ may also be used to obtain an estimate of i_θ, the asymptotic variance of $\hat{\theta}$. Hence, for inference based on first-order asymptotic theory, the generalized profile likelihood function may be treated as a standard likelihood function for θ.

PROPOSITION 3. *Let*

$$\hat{i}_\theta = -\frac{1}{n}\frac{d^2L_n}{d\theta^2}(\theta, \hat{\lambda}_\theta)\bigg|_{\theta=\hat{\theta}}.$$

Then, under the regularity conditions,

$$\hat{i}_\theta \to_p i_\theta \quad \textit{as } n \to \infty.$$

PROOF. The proof follows from (3) and the fact that $\hat{\theta} \to_p \theta_0$. □

7. Estimation of a least favorable curve. In order to carry out the estimation procedure outlined in the previous sections, an estimator of a least favorable curve must be available. Hence, in this section we present a general method for estimating a least favorable curve.

We have seen that if λ is a one-dimensional parameter and $\hat{\lambda}_\theta$ represents the MLE of λ for fixed θ, then $\hat{\lambda}_\theta$ is an estimator of a least favorable curve. Hence, for the semiparametric model under consideration, one approach to estimating $\lambda_\theta(x)$ is to use maximum likelihood. However, instead of maximizing the likelihood itself we maximize an empirical version of

$$E_0\{\log p(Y; \theta, \eta)|X = x\}$$

given by

$$\sum \log p(Y_j; \theta, \eta) K\left(\frac{x - X_j}{h_n}\right),$$

where $K(\cdot)$ is a kernel on the real line [Staniswalis (1989)]. Carrying out this procedure for each x yields an estimator $\hat{\lambda}_\theta$. In Lemma 4 it is shown that if $\hat{\lambda}_\theta$ has the required convergence properties, then $\hat{\lambda}_\theta$ is an estimator of a least favorable curve.

LEMMA 4. *Suppose that for each $x \in [0,1]$ and $\theta \in \Theta$, $\hat{\lambda}_\theta(x)$ is obtained by the following procedure*:

$$\textit{maximize } \sum l(Y_j; \theta, \eta) K\left(\frac{x - X_j}{h_n}\right) \quad \textit{with respect to } \eta,$$

where $K(\cdot)$ satisfies the following properties

$$K(u) = 0 \quad \textit{for } |u| > 1, \qquad \sup_u |K(u)| < \infty,$$

$$\int K(u)\,du = 1, \qquad \int uK(u)\,du = 0, \qquad \int u^2K(u)\,du < \infty$$

and h_n is a sequence of constants satisfying

$$h_n \to 0 \quad \textit{and} \quad nh_n \to \infty \quad \textit{as } n \to \infty.$$

Then, if $\hat{\lambda}_\theta$ satisfies Condition NP(a), *then Condition NP*(b) *is satisfied as well, that is, $\hat{\lambda}_\theta$ is an estimator of a least favorable curve.*

Proof. Let $\tilde{\lambda}_\theta(x)$ denote the limit in probability of $\hat{\lambda}_\theta$ as $n \to \infty$, which exists by condition NP(a). We now show that $\tilde{\lambda}_\theta$ is a least favorable curve.

From Staniswalis (1989) it follows that

$$\hat{\lambda}_0(x) \to_p \lambda_0(x) \quad \text{as } n \to \infty$$

for each x so that $\tilde{\lambda}_0 = \lambda_0$. The estimator $\hat{\lambda}_\theta(x)$ satisfies

$$\sum \frac{\partial l}{\partial \eta}\big(Y_j; \theta, \hat{\lambda}_\theta(x)\big) K\left(\frac{x - X_j}{h_n}\right) = 0$$

for each x and θ so that

$$\begin{aligned}
&\frac{d}{d\theta} \sum \frac{\partial l}{\partial \eta}\big(Y_j; \theta, \hat{\lambda}_\theta(x)\big) K\left(\frac{x - X_j}{h_n}\right)\bigg|_{\theta = \theta_0} \\
&\quad = \sum \frac{\partial^2 l}{\partial \theta\, \partial \eta}\big(Y_j; \theta_0, \hat{\lambda}_0(x)\big) K\left(\frac{x - X_j}{h_n}\right) \\
&\qquad + \sum \frac{\partial^2 l}{\partial \eta^2}\big(Y_j; \theta_0, \hat{\lambda}_0(x)\big) K\left(\frac{x - X_j}{h_n}\right) \hat{\lambda}_0'(x) \\
&\quad = 0 \quad \text{for each } x.
\end{aligned}$$

Therefore,

$$\hat{\lambda}_0'(x) = -\frac{\Sigma(\partial^2 l/\partial\theta\, \partial\eta)\big(Y_j; \theta_0, \hat{\lambda}_0(x)\big) K\big((x - X_j)/h_n\big)}{\Sigma(\partial^2 l/\partial \eta^2)\big(Y_j; \theta_0, \hat{\lambda}_0(x)\big) K\big((x - X_j)/h_n\big)}.$$

Note that $\tilde{\lambda}_0 = \lambda_0$ implies that $\|\hat{\lambda}_0 - \lambda_0\| = o_p(1)$. Using this fact together with Conditions S, it follows that

$$\begin{aligned}
&\frac{\Sigma(\partial^2 l/\partial\theta\, \partial\eta)\big(Y_j; \theta_0, \hat{\lambda}_0(x)\big) K\big((x - X_j)/h_n\big)}{\Sigma K\big((x - X_j)/h_n\big)} \\
&\quad = \frac{\Sigma(\partial^2 l/\partial\theta\, \partial\eta)\big(Y_j, \theta_0, \lambda_0(x)\big) K\big((x - X_j)/h_n\big)}{\Sigma K\big((x - X_j)/h_n\big)} + o_p(1)
\end{aligned}$$

and

$$\frac{\Sigma(\partial^2 l/\partial\eta^2)\big(Y_j;\theta_0,\hat{\lambda}_0(x)\big)K\big((x-X_j)/h_n\big)}{\Sigma K\big((x-X_j)/h_n\big)}$$

$$= \frac{\Sigma(\partial^2 l/\partial\eta^2)\big(Y_j,\theta_0,\lambda_0(x)\big)K\big((x-X_j)/h_n\big)}{\Sigma K\big((x-X_j)/h_n\big)} + o_p(1).$$

Hence,

$$\hat{\lambda}'_0(x) \to_p -\frac{E_0\{(\partial^2 l/\partial\theta\,\partial\eta)(Y;\theta_0,\lambda_0(x))|X=x\}}{E_0\{(\partial^2 l/\partial\eta^2)(Y;\theta_0,\lambda_0(x))|X=x\}}$$

for each x [e.g., Nadaraya (1964)] so that, by Lemma 1, $\tilde{\lambda}'_0(x) = v^*(x)$ for each x proving the lemma. □

In order to use the approach of Lemma 4 in estimating a least favorable curve, the convergence requirements of Conditions NP(a) must be verified. However it is clear that if an estimator $\hat{h}_n(\theta,\eta,x)$ of

$$h(\theta,\eta,x) \equiv E_0\{\log p(Y;\theta,\eta)|X=x\}$$

is available such that $\hat{h}_n(\theta,\eta,x)$ and its derivatives converge to $h(\theta,\eta,x)$ and its derivatives at a suitable rate, then the estimator of $\lambda_\theta(x)$ given by maximizing $\hat{h}_n(\theta,\eta,x)$ with respect to η will satisfy Conditions NP(a). This idea is considered in Lemma 5 using a nonparametric regression estimate of $h(\theta,\eta,x)$; the proof is given in Section 10.

LEMMA 5. *For each $\theta\in\Theta$, $x\in[0,1]$, let $h(\theta,\eta,x) = E_0\{\log p(Y;\theta,\eta)|X = x\}$. Assume that*

$$\sup_{\theta,\eta,x}\left|\frac{\partial^k}{\partial\theta^k}\frac{\partial^l}{\partial x^l}h^{(j)}(\theta,\eta,x)\right| < \infty$$

for $j = 2,3,4$, $k = 0,1,2$, $l = 0,1$, $j+k+l \le 4$; here $h^{(j)}(\theta,\eta,x) = \partial^j h(\theta,\eta,x)/\partial\eta^j$. Let $\lambda_\theta(x)$ denote a solution to

$$\partial h(\theta,\eta,x)/\partial\eta = 0$$

with respect to η for each fixed θ and x. Assume that $\lambda_\theta(x)$ is unique and that for any $\varepsilon > 0$ there exists a $\delta > 0$ such that

$$\sup_\theta\sup_x\big|h'\big(\theta,\bar{\lambda}_\theta(x),x\big)\big| \le \delta$$

implies that

$$\sup_\theta\sup_x\big|\bar{\lambda}_\theta(x) - \lambda_\theta(x)\big| \le \varepsilon.$$

Let

$$T_{\theta,\eta}^{(j,k)}(Y) = \frac{\partial^j}{\partial\theta^j}\frac{\partial^k}{\partial\eta^k}\log p(Y;\theta,\eta)$$

and let $f_\theta^{(j,k)}(y|x)$ *denote the conditional density of* $T_{\theta,\eta}^{(j,k)}(Y)$ *given* $X = x$. *Let* $f(x)$ *denote the marginal density of* x. *Assume the following conditions are satisfied*:

(a) $E\{\sup_\theta \sup_\eta |T_{\theta,\eta}^{(j,k)}(Y)|\} < \infty$, $j = 0,\dots,3$; $k = 0,\dots,5$.

(b) *For some even integer* $q \geq 10$, $\sup_\theta \sup_\eta E\{|T_{\theta,\eta}^{(j,k)}(Y)|^q\} < \infty$, $j = 0,\dots,3$; $k = 0,\dots,4$.

(c) $\sup_\theta \sup_\eta \sup_{y,x}|f_{\theta,\eta}^{(j,k)}(y|x)| < \infty$, $j = 0,\dots,3$; $k = 0,\dots,4$; $r = 0,\dots,4$.

(d) $\sup_x|f^{(r)}(x)| < \infty$, $r = 0,\dots,4$.

(e) $0 < \inf_x f(x) \leq \sup_x f(x) < \infty$.

Let

$$\hat{h}_n(\theta,\eta,x) = \frac{\Sigma \log p(Y_j;\theta,\eta)K\big((x - X_j)/h_n\big)}{\Sigma K\big((x - X_j)/h_n\big)},$$

where $K(\cdot)$ *satisfies*

$$\int K(u)\,du = 1, \qquad \int uK(u)\,du = 0, \qquad \int u^2K(u)\,du < \infty,$$

$$\sup_u |K^{(r)}(u)| < \infty, \qquad r = 0,\dots,4$$

and h_n *is a sequence of constants satisfying* $h_n = O(n^{-\alpha})$,

$$\frac{1}{8} < \alpha < \frac{1}{4}\frac{(q+3)(q-2)}{4(q+6)(q+2)}.$$

Let $\hat{\lambda}_\theta(x)$ *denote an estimator of* $\lambda_\theta(x)$ *obtained by solving*

$$\partial\hat{h}_n(\theta,\eta,x)/\partial\eta = 0$$

with respect to η *for each fixed* θ *and* x.

Then, under the regularity conditions in effect, $\hat{\lambda}_\theta(x)$ *satisfies Conditions NP taking* $\tilde{\lambda}_\theta(x) = \lambda_\theta(x)$ *for each* θ *and* x.

Although the approach of Lemma 5 may be used to obtain an estimator of a least favorable curve, this method may be computationally intensive since each evaluation of $L_n(\theta, \hat{\lambda}_\theta)$ requires a separate maximization of $\hat{h}_n(\theta, \eta, x)$ for each $x = X_j$, $j = 1,\dots,n$. However, for a particular type of model a simpler method of obtaining an estimator of a least favorable curve is available.

8. Conditionally exponential families. Suppose that for each fixed θ there exists a real-valued function $\psi_\theta(\cdot)$ such that the distribution of $\psi_\theta(Y)$ does not depend on θ and the density of the conditional distribution of $w = \psi_\theta(Y)$ given $X = x$ forms an exponential family; without loss of generality we may take the density to be of the form

$$\exp\{wT(\eta) - A(\eta) + S_1(w)\}$$

for some functions T, A, S_1, where $\eta = \eta_x$ and $T''(\eta)$ and $A''(\eta)$ exist for each $\eta \in H$. Then the conditional distribution of Y given $X = x$ is of the form

$$(4) \qquad p(y;\theta,\eta) = \exp\{\psi_\theta(y)T(\eta) - A(\eta) + S(y;\theta)\}$$

for some function S not depending on η. The following lemma gives an expression for the least favorable direction in this setting.

LEMMA 6. *Suppose $p(y;\theta,\eta)$ is of the form* (4) *above. Assume that for almost all y*

$$\psi'_\theta(y) \equiv \frac{\partial}{\partial\theta}\psi_\theta(y)$$

exists for $\theta = \theta_0$. Then the least favorable direction v^* is given by

$$v^*(x) = \left(\frac{A'(\eta_0)T''(\eta_0)}{T'(\eta_0)} - A''(\eta_0)\right)^{-1} E_0\{\psi'_0(Y)|X = x\}T'(\eta_0),$$

where $\psi'_0 = \psi'_\theta$ at $\theta = \theta_0$ and $\eta_0 = \lambda_0(x)$.

PROOF. Using (4) we have that

$$\frac{\partial^2 l}{\partial\eta^2}(\theta,\eta) = \psi_\theta(Y)T''(\eta) - A''(\eta)$$

and

$$\frac{\partial^2 l}{\partial\theta\,\partial\eta}(\theta,\eta) = \psi'_\theta(Y)T'(\eta).$$

The result now follows from Lemma 1 using the fact that

$$E_{\theta,\eta}\{\psi_\theta(Y)|X = x\} = \frac{A'(\eta)}{T'(\eta)},$$

which follows from the properties of an exponential family. □

Lemma 7 gives a method for determining a least favorable curve in this setting.

LEMMA 7. *Suppose the conditional density of Y given $X = x$ is of the form* (4) *above. Let φ: $\mathfrak{R} \mapsto \mathfrak{R}$ denote a one-to-one differentiable function satisfying*

$$\eta = \varphi\left(E_{\theta,\eta}\{\psi_\theta(Y)\}\right) \quad \textit{for each } \theta, \eta.$$

Then

$$\lambda_\theta(x) = \varphi\big(E_0\{\psi_\theta(Y)|X=x\}\big)$$

is a least favorable curve.

PROOF. By definition of φ,

$$\lambda_0(x) = \varphi\big(E_0\{\psi_0(Y)|X=x\}\big),$$

where $\psi_0(Y) = \psi_{\theta_0}(Y)$. Furthermore,

$$\lambda'_\theta(x)|_{\theta=\theta_0} = \varphi'\big(E_0\{\psi_0(Y)|X=x\}\big)E_0\{\psi'_0(Y)|X=x\}.$$

By the properties of an exponential family,

$$E_{\theta,\eta}\{\psi_\theta(Y)\} = \varphi^{-1}(\eta), \qquad \varphi^{-1}(\eta) = \frac{A'(\eta)}{T'(\eta)}$$

and

$$\varphi'(\varphi^{-1}(\eta)) = \left(\frac{A'(\eta)T''(\eta)}{T'(\eta)} - A''(\eta)\right)^{-1} T'(\eta).$$

It follows that

$$\lambda'_{\theta_0}(x) = \varphi'(\varphi^{-1}(\lambda_0(x)))E_0\{\psi'_0(Y)|X=x\} = v^*(x).$$

The remaining requirements of a least favorable curve follow immediately from Conditions S. □

Since the nonparametric regression estimate $\Sigma\psi_\theta(Y_j)K((x-X_j)/h_n)/\Sigma K((x-X_j)/h_n)$ is an estimate of $E_0(\psi_\theta(Y)|x=x)$ Lemma 7 suggests that nonparametric regression may be used to estimate a least favorable curve. In order to carry out this approach, it is convenient to use the following result on nonparametric regression; the proof is given in Section 10.

LEMMA 8. *Let $T_\theta(Y)$ denote a scalar function of a random variable Y depending on a scalar parameter θ and let $f_{\theta j}(y|x)$, $j=0,1,2$ denote the conditional density of*

$$T_\theta^{(j)}(Y) \equiv \frac{\partial^j}{\partial\theta^j}T_\theta(Y)$$

given $X=x$. Let $f(x)$ denote the marginal density of X.

Assume the following conditions hold:

(a) $E\{\sup_\theta |T_\theta^{(j)}(Y)|\} < \infty$, $j=0,1,2,3$.

(b) *For some even integer* $q \geq 2$, $\sup_\theta E\{|T_\theta^{(j)}(Y)|^q\} < \infty$, $j=0,1,2$.

(c) $\sup_\theta \sup_{y,x} |f_{\theta j}^{(r)}(y|x)| < \infty$, $j=0,1,2$; $r=0,\dots,4$.

(d) $\sup_x |f^{(r)}(x)| < \infty$, $r=0,\dots,4$.

(e) $0 < \inf_x f(x) \leq \sup_x f(x) < \infty$.

Let $m_\theta(x) = E\{T_\theta(Y)|X = x\}$ *and*

$$\hat{m}_\theta(x) = \frac{\Sigma T_\theta(Y_j)K\big((x - X_j)/h_n\big)}{\Sigma K\big((x - X_j)/h_n\big)},$$

where $Y_1, \dots,$ *denotes a sequence of independent random variables each distributed according to the distribution of* Y, $K(\cdot)$ *satisfies*

$$\int K(u)\,du = 1, \qquad \int uK(u)\,du = 0, \qquad \int u^2 K(u)\,du < \infty,$$

$$\sup_u |K^{(r)}(u)| < \infty, \qquad r = 0, \dots, 4$$

and h_n *is a sequence of constants satisfying* $h_n \to 0$ *as* $n \to \infty$.

Then, for any $\gamma > 0$,

$$\sup_\theta \left\| \frac{\partial^k}{\partial x^k}\frac{\partial^j}{\partial \theta^j}\hat{m}_\theta - \frac{\partial^k}{\partial x^k}\frac{\partial^j}{\partial \theta^j} m_\theta \right\|$$
$$= O_p\big(n^{-q/(2q+4)} h_n^{-(k+(q+4)/(q+2))} n^\gamma + h_n^2\big) \quad \text{as } n \to \infty$$

for $j = 0, 1, 2$ *and* $k = 0, 1$.

Lemma 9 shows that nonparametric regression can be used to obtain an estimate of a least favorable curve.

LEMMA 9. *Suppose the conditions of Lemma* 7 *are satisfied and the conditions of Lemma* 8 *are satisfied with* $T_\theta(Y) = \psi_\theta(Y)$ *and* $q \geq 10$. *Let*

$$m_\theta(x) = E_0(\psi_\theta(Y)|X = x)$$

and let M *denote a compact subset of the real line such that*

$$m_\theta(x) \in \operatorname{int}(M) \quad \textit{for all } x, \theta.$$

Assume that

$$\sup_{m \in M} |\varphi^{(j)}(m)| < \infty \quad \textit{for } j = 0, \dots, 3,$$

where $\varphi(\cdot)$ *is the function defined in Lemma* 7. *Let*

$$\hat{\lambda}_\theta(x) = \varphi\left(\frac{\Sigma \psi_\theta(Y_j)K\big((x - X_j)/h_n\big)}{\Sigma K\big((x - X_j)/h_n\big)} \right),$$

where $K(\cdot)$ *satisfies the conditions of Lemma* 8 *and*

$$h_n = O(n^{-\alpha}) \quad \textit{with } 1/8 < \alpha < (q - 2)/(4(q + 4)).$$

Then $\hat{\lambda}_\theta$ *is an estimator of a least favorable curve satisfying Conditions NP.*

PROOF. Let

$$\hat{m}_\theta(x) = \frac{\Sigma\psi_\theta(Y_j)K((x - X_j)/h_n)}{\Sigma K((x - X_j)/h_n)}.$$

By Lemma 8,

$$\sup_\theta \left\| \frac{\partial^j}{\partial\theta^j}\hat{m}_\theta - \frac{\partial^j}{\partial\theta^j}m_\theta \right\| = o_p(n^{-1/4}), \qquad j = 0, 1, 2$$

and

$$\sup_\theta \left\| \frac{\partial}{\partial x}\frac{\partial^j}{\partial\theta^j}\hat{m}_\theta - \frac{\partial}{\partial x}\frac{\partial^j}{\partial\theta^j}m_\theta \right\| = o_p(n^{-1/4}h_n^{-1}), \qquad j = 0, 1.$$

Note that $n^{-1/4}h_n^{-1} = o(n^{-\varepsilon})$ for some $\varepsilon > 0$.

By Lemma 6, $\lambda_\theta = \varphi(m_\theta)$ is a least favorable curve. Since

$$\begin{aligned}\sup_x |\hat{\lambda}_\theta(x) - \lambda_\theta(x)| &= \sup_x |\varphi(\hat{m}_\theta(x)) - \varphi(m_\theta(x))| \\ &\leq \sup_{m\in M} |\varphi'(m)|\, \|\hat{m}_\theta - m_\theta\|\end{aligned}$$

it follows that

$$\sup_\theta \|\hat{\lambda}_\theta - \lambda_\theta\| = o_p(n^{-1/4}).$$

Using the same approach it can be shown that the remaining conditions of Conditions NP are satisfied. □

9. Examples.

EXAMPLE 1. Suppose that the conditional distribution of Y given $X = x$ is a two-parameter exponential family with parameters θ, η, $\eta = \lambda(x)$.

(a) Let

$$p(y; \theta, \eta) = (2\pi\eta)^{-1/2}\exp\{-(y - \theta)^2/(2\eta)\}, \qquad -\infty < y < \infty.$$

Assume that Θ is a compact subset of $\Re$ and that H is a compact subset of $(0, \infty)$. Take $\psi_\theta(Y) = (Y - \theta)^2$ and $\varphi(t) = t$. Then

$$m_\theta(x) = E_0(\psi_0(Y)|X = x) = \lambda_0(x) + (\theta - \theta_0)^2,$$

so we may take M to be any compact subset of $(0, \infty)$ containing $\eta + 4\theta^2$ for all $\eta \in H$, $\theta \in \Theta$.

It is easy to verify that the conditions of Lemma 8 are satisfied with arbitrarily large q so we may estimate θ by maximizing

$$L_n(\theta, \hat{\lambda}_\theta) = -\frac{1}{2}\sum\frac{(Y_j - \theta)^2}{\hat{\lambda}_\theta(X_j)} - \frac{1}{2}\sum\log\hat{\lambda}_\theta(X_j),$$

where

$$\hat{\lambda}_\theta(x) = \frac{\Sigma(Y_j - \theta)^2 K\big((x - X_j)/h_n\big)}{\Sigma K\big((x - X_j)/h_n\big)},$$

$K(\cdot)$ satisfies the conditions of Lemma 8 and $h_n = O(n^{-1/5})$.

It is easy to show that conditions S are satisfied so that by Proposition 2 the resulting estimator is asymptotically efficient.

Note that similar results could be obtained for the model in which θ is a regression parameter, that is, for some random variable Z_j, Y_j has mean θZ_j and variance $\lambda(X_j)$ given X_j.

(b) Let

$$p(y;\theta,\eta) = \frac{1}{\eta^\theta \Gamma(\eta)} y^{\theta - 1} \exp\left\{-\frac{y}{\eta}\right\}, \qquad y > 0,$$

where Θ and H are compact subsets of $(0, \infty)$.

In this example Lemma 9 cannot be applied; instead we will use the approach of Lemma 4. Let $\hat{\lambda}_\theta(x)$ denote the value of η that maximizes

$$\begin{aligned}\sum l(Y_j;\theta,\eta) K\left(\frac{x - X_j}{h_n}\right) \\ = \sum\left[-\theta \log \eta - \log\Gamma(\theta) + (\theta - 1)\log Y_j - \eta^{-1}Y_j\right] K\left(\frac{x - X_j}{h_n}\right),\end{aligned}$$

where $K(\cdot)$ satisfies the conditions of Lemma 8 and $h_n = O(n^{-1/5})$. Then

$$\hat{\lambda}_\theta(x) = \theta^{-1}\frac{\Sigma Y_j K\big((x - X_j)/h_n\big)}{\Sigma K\big((x - X_j)/h_n\big)}$$

and Conditions NP follow from Lemmas 4 and 8. Conditions S are easily satisfied so that the estimator obtained by maximizing

$$\begin{aligned}L_n\big(\theta, \hat{\lambda}_\theta\big) = -\theta \sum \log \hat{\lambda}_\theta(X_j) - n \log(\theta) \\ + (\theta - 1)\sum \log(Y_j) - \sum\big(\hat{\lambda}_\theta(X_j)\big)^{-1} Y_j\end{aligned}$$

is asymptotically efficient.

Example 2. Suppose that X, Y, Z are random variables such that the conditional distribution of Y given $X = x$ and $Z = z$ is an exponential family distribution with parameter $\theta z + \eta$, $\eta = \lambda(x)$.

(a) Let

$$p(y, z;\theta,\eta) = (2\pi)^{-1/2}\exp\{-(y - \theta z - \eta)^2/2\} f_Z(z), \qquad -\infty < y < \infty,$$

where $f_Z(\cdot)$ represents the marginal density of Z. Assume that Θ and H are compact subsets of $\mathfrak{R}$; assume that the distribution of Z has support on [0, 1].

Take $\psi_\theta(Y, Z) = Y - \theta Z$ and $\varphi(t) = t$. Let

$$\hat{\lambda}_\theta(x) = \frac{\Sigma(Y_j - \theta Z_j)K\big((x - X_j)/h_n\big)}{\Sigma K\big((x - X_j)/h_n\big)},$$

where $K(\cdot)$ satisfies the conditions of Lemma 8 and $h_n = O(n^{-1/5})$. It is easy to show that the remaining conditions of Lemma 8 are satisfied with arbitrarily large q so that $\hat{\lambda}_\theta$ satisfies Conditions NP.

Conditions S are satisfied so that the estimator obtained by maximizing

$$L_n(\theta, \hat{\lambda}_\theta) = -\tfrac{1}{2}\sum\big(Y_j - \theta Z_j - \hat{\lambda}_\theta(X_j)\big)^2$$

is asymptotically efficient. For this example we can derive an explicit expression of $\hat{\theta}$,

$$\hat{\theta} = \frac{\Sigma\big(Z_j - \hat{E}(Z|X_j)\big)\big(Y_j - \hat{E}(Y|X_j)\big)}{\Sigma\big(Z_j - \hat{E}(Z|X_j)\big)^2},$$

where

$$\hat{E}(Z|x) = \frac{\Sigma Z_j K\big((x - X_j)/h_n\big)}{\Sigma K\big((x - X_j)/h_n\big)}$$

and

$$\hat{E}(Y|x) = \frac{\Sigma Y_j K\big((x - X_j)/h_n\big)}{\Sigma K\big((x - X_j)/h_n\big)}.$$

(b) Let

$$p(y, z; \theta, \eta) = \exp\{-(\theta z + \eta)\}\exp\{-y\exp\{-(\theta z + \eta)\}\} f_Z(z),$$
$$y > 0, 0 \le z \le 1.$$

Assume that Θ and H are compact subsets of $\Re$. Take $\psi_\theta(Y, Z) = Y\exp\{-\theta Z\}$ and $\varphi(t) = \log t$. Then

$$m_\theta(x) = \exp\{\lambda_0(x)\}E\big(\exp\{(\theta_0 - \theta)Z\}|X = x\big)$$

so we may take M to be a sufficiently large compact subset of $(0, \infty)$. It follows that

$$\sup_m |\varphi^{(j)}(m)| < \infty, \qquad j = 1, 2, 3.$$

Let

$$\hat{\lambda}_\theta(x) = \log\left(\frac{\Sigma Y_j \exp\{-\theta Z_j\}K\big((x - X_j)/h_n\big)}{\Sigma K\big((x - X_j)/h_n\big)}\right),$$

where $K(\cdot)$ satisfies the conditions of Lemma 8 and $h_n = O(n^{-1/5})$. It is easy to show that the remaining conditions of Lemma 8 are satisfied with arbitrarily large q so that $\hat{\lambda}_\theta$ satisfies Conditions NP.

Conditions S are satisfied so that by Proposition 2 the estimator obtained by maximizing

$$L_n(\theta, \hat{\lambda}_\theta) = -\theta \sum Z_j - \sum \hat{\lambda}_\theta(X_j) - \sum Y_j \exp\{-(\theta Z_j + \hat{\lambda}_\theta(X_j))\}$$

is asympotically efficient.

(c) Suppose that

$$p(y, z; \theta, \eta) = \exp\{(\theta z + \eta) y\} \exp\{-\exp\{\theta z + \eta\}\}/y! F_Z z, \qquad y = 0, 1, \dots, \; 0 \le z \le 1.$$

Assume that Θ and H are compact subsets of $\mathfrak{R}$.

In this example Lemma 9 cannot be applied; instead we will use the approach of Lemma 4. Let $\hat{\lambda}_\theta(x)$ denote the value of η that maximizes

$$\sum l(Y_j, Z_j; \theta, \eta) K\left(\frac{x - X_j}{h_n}\right) = \sum \left[Y_j(\theta Z_j + \eta) - \exp\{\theta Z_j + \eta\}\right] K\left(\frac{x - X_j}{h_n}\right),$$

where $K(\cdot)$ satisfies the conditions of Lemma 8 and $h_n = O(n^{-1/5})$. Then

$$\hat{\lambda}_\theta(x) = \log\left(\frac{\Sigma Y_j K((x - X_j)/h_n)}{\Sigma \exp\{\theta Z_j K((x - X_j)/h_n)\}}\right).$$

By Lemma 8,

$$\frac{\Sigma Y_j K((x - X_j)/h_n)}{\Sigma K((x - X_j)/h_n)} = E_0(Y|X = x) + o_p(n^{-1/4})$$

uniformly in x and

$$\frac{\Sigma \exp\{\theta Z_j\} K((x - X_j)/h_n)}{\Sigma K((x - X_j)/h_n)} = E(\exp\{\theta Z\}|X = x) + o_p(n^{-1/4})$$

uniformly in θ and x. Since $E_0(\exp\{\theta Z\}|X = x)$ is bounded away from 0 for $\theta \in \Theta$, it can be shown that

$$\frac{\Sigma Y_j K((x - X_j)/h_n)}{\Sigma \exp\{\theta Z_j\} K((x - X_j)/h_n)} = \frac{E_0(Y|X = x)}{E_0(\exp\{\theta Z\}|X = x)} + o_p(n^{-1/4})$$

uniformly in x and θ. It now follows that

$$\|\hat{\lambda}_\theta - \lambda_\theta\| = o_p(n^{-1/4}).$$

The remaining conditions of Conditions NP may be verified in a similar manner.

Conditions S are satisfied so that by Proposition 2 the estimator obtained by maximizing

$$L_n(\theta, \hat{\lambda}_\theta) = \theta \sum Y_j Z_j + \sum Y_j \hat{\lambda}_\theta(X_j) - \sum \exp\{\theta Z_j + \hat{\lambda}_\theta(X_j)\}$$

is asymptotically efficient.

10. Technical proofs.

PROOF OF PROPOSITION 1. Note that, under the regularity conditions in effect, $L_n(\theta, \hat{\lambda}_\theta)$ is continuous in θ and a measurable function of $Y_1, \dots, Y_n$; $X_1, \dots, X_n$ for each θ; it follows that $\hat{\theta}$ is measurable.

Let

$$\gamma(\theta) = E_0\{l(Y; \theta, \lambda_\theta(X))\}.$$

By Conditions I, $\gamma(\theta)$ satisfies the following condition:

$$\text{for all } \theta \in \Theta, \qquad \theta \neq \theta_0, \qquad \gamma(\theta) < \gamma(\theta_0).$$

By the weak law of large numbers,

$$n^{-1}L_n(\theta, \lambda_\theta) \to_p \gamma(\theta) \quad \text{for each } \theta \in \Theta.$$

Furthermore, for $\theta_1, \theta_2 \in \Theta$,

$$\begin{aligned} &n^{-1}\left|L_n(\theta_1, \lambda_{\theta_1}) - L_n(\theta_2, \lambda_{\theta_2})\right| \\ &\quad \leq n^{-1} \sum \left|l_j(\theta_1, \lambda_{\theta_1}) - l_j(\theta_2, \lambda_{\theta_2})\right| \\ &\quad \leq n^{-1} \sum \left\{ \sup_{\theta, \eta} \left| \frac{\partial l_j}{\partial \theta}(\theta, \eta) \right| |\theta_1 - \theta_2| + \sup_{\theta, \eta} \left| \frac{\partial l_j}{\partial \eta}(\theta, \eta) \right| \left\| \lambda_{\theta_1} - \lambda_{\theta_2} \right\| \right\} \\ &\quad \leq n^{-1} \sum \left\{ \sup_{\theta, \eta} \left| \frac{\partial l_j}{\partial \theta}(\theta, \eta) \right| |\theta_1 - \theta_2| + \sup_{\theta, \eta} \left| \frac{\partial l_j}{\partial \eta}(\theta, \eta) \right| \sup_\theta \|\lambda'_\theta\| \, |\theta_1 - \theta_2| \right\} \\ &\quad \equiv A_n |\theta_1 - \theta_2|. \end{aligned}$$

Since, by Conditions S, A_n is bounded in probability, it follows that

$$\{n^{-1}L_n(\theta, \lambda_\theta) \colon \theta \in \Theta\}$$

is tight and hence,

$$n^{-1}L_n(\theta, \lambda_\theta) \to_{\mathscr{D}} \gamma(\theta) \quad \text{in } C(\Theta).$$

For each θ,

$$\begin{aligned} \frac{1}{n}\left|L_n(\theta, \hat{\lambda}_\theta) - L_n(\theta, \lambda_\theta)\right| &\leq \frac{1}{n} \sum \left|l_j(\theta, \hat{\lambda}_\theta(X_j)) - l(\theta, \lambda_\theta(X_j)\right| \\ &\leq \frac{1}{n} \sum \sup_\theta \sup_\eta \left| \frac{\partial l}{\partial \eta}(\theta, \eta) \right| \sup_\theta \left\| \hat{\lambda}_\theta - \lambda_\theta \right\|. \end{aligned}$$

Therefore,

$$\sup_\theta \frac{1}{n}\left|L_n(\theta, \hat{\lambda}_\theta) - L_n(\theta, \lambda_\theta)\right| \to_p 0 \quad \text{as } n \to \infty$$

and hence,

$$\sup_\theta \left| \frac{1}{n} L_n(\theta, \hat{\lambda}_\theta) - \gamma(\theta) \right| \to_p 0 \quad \text{as } n \to \infty.$$

Furthermore, since

$$\sup_{\theta} \frac{1}{n} L_n(\theta, \hat{\lambda}_\theta) \to_p \sup_{\theta} \gamma(\theta) = \gamma(\theta_0)$$

it follows that

$$\gamma(\hat{\theta}) \to_p \gamma(\theta_0) \quad \text{as } n \to \infty.$$

For a given $\theta \in \Theta$, there exists an $\varepsilon > 0$ and an open neighborhood N_θ of θ such that

$$\inf_{\theta_1 \in N_\theta} |\gamma(\theta_1) - \gamma(\theta_0)| > \varepsilon.$$

Therefore,

$$P_0\big(\hat{\theta} \in N_\theta\big) \leq P_0\big(|\gamma(\hat{\theta}) - \gamma(\theta_0)| > \varepsilon\big) \to 0 \quad \text{as } n \to \infty.$$

Let N_0 denote an open neighborhood of θ_0 and consider the compact set $\Theta_0 = \Theta \setminus N_0$. Let $\{N_\theta \colon \theta \in \Theta, \theta \neq \theta_0\}$ denote the open cover of Θ_0 constructed by the preceding procedure. By compactness of Θ_0 there exists a finite subcover $\{N_{\theta_1}, \dots, N_{\theta_k}\}$. Then

$$P_0\big(\hat{\theta} \notin N_0\big) = P_0\big(\hat{\theta} \in \Theta_0\big) \leq \sum_1^k P_0\big(\hat{\theta} \in N_{\theta_j}\big) \to 0 \quad \text{as } n \to \infty.$$

Therefore,

$$\hat{\theta} \to_p \theta_0 \quad \text{as } n \to \infty. \qquad \square$$

Proof of Lemma 2. Consider (i). Let

$$\Lambda_0 = \left\{ h \in C^2[0,1] \colon \|h\| \leq 1, \left\| \frac{\partial}{\partial x} h \right\| \leq 1 \right\}.$$

Λ_0 may be viewed as a metric space with metric

$$\rho(h_1, h_2) = \|h_1 - h_2\|.$$

By Conditions NP,

$$P_0\big(n^\delta(\hat{\lambda}_0 - \lambda_0) \in \Lambda_0\big) \to 1 \quad \text{as } n \to \infty.$$

Hence, assume that, for sufficiently large n, $n^\delta(\hat{\lambda}_0 - \lambda_0) \in \Lambda_0$ since the probability that this does not occur can be made arbitrarily small.

Let

$$B(y,x) = \frac{\partial^2 l}{\partial\theta\,\partial\eta}(y; \theta_0, \lambda_0(x)) + \frac{\partial^2 l}{\partial\eta^2}(y; \theta_0, \lambda_0(x)) v^*(x).$$

Note that, by the definition of $v^*(x)$,

$$E_0\{B(Y,X) | X = x\} = 0$$

for each x and hence, $E_0\{B(Y,X)\} = 0$.

For any $v \in \Lambda_0$,

$$\frac{d}{d\theta}\frac{\partial L_n}{\partial \lambda}(\theta, \lambda_\theta)\bigg|_{\theta=\theta_0}(v) = \sum B(Y_j, X_j)(v).$$

We will view the function $v \mapsto B(y, x)v(x)$ as an element of the space of continuous functions on Λ_0 together with the sup norm, which will be denoted by $C(\Lambda_0)$.

For any $v_1, v_2 \in \Lambda_0$,

$$\begin{aligned} |B(y,x)v_1(x) - B(y,x)v_2(x)| &= |B(y,x)(v_1 - v_2)(x)| \\ &\le 2|B(y,x)|\,\|v_1 - v_2\| \end{aligned}$$

and by conditions S,

$$E_0\big(|B(Y,X)|^2\big) < \infty.$$

Let $H(\cdot, \Lambda_0)$ denote the metric entropy of the set Λ_0 with respect to the metric ρ. Then

$$H(\varepsilon, \Lambda_0) \le A_0 \varepsilon^{-1}$$

for some constant A_0.

Jain and Marcus (1975) have shown that if $\mathscr{T}$ is a metric space with metric d and Z is a random element taking values in $C(\mathscr{T})$ satisfying $E(Z) = 0$ and

$$|Z(s) - Z(t)| \le Vd(s,t),$$

where V is a random variable with $E(V^2) < \infty$, then the distribution of

$$n^{-1/2}(Z_1 + \cdots + Z_n),$$

where $Z_1, \dots, Z_n$ are independent replicates of Z, converges weakly to an appropriately defined Gaussian measure provided that

$$\int_0^1 H(\varepsilon, \mathscr{T})\, d\varepsilon < \infty.$$

Hence, it follows from Jain and Marcus (1975) that

$$\frac{1}{\sqrt{n}}\frac{d}{d\theta}\frac{\partial L_n}{\partial \lambda}(\theta, \lambda_\theta)\bigg|_{\theta=\theta_0}$$

satisfies the central limit theorem as an element of $C(\Lambda_0)$ and hence,

$$\frac{1}{\sqrt{n}}\frac{d}{d\theta}\frac{\partial L_n}{\partial \lambda}(\theta, \lambda_\theta)\bigg|_{\theta=\theta_0}\big(n^\delta(\hat{\lambda}_0 - \lambda_0)\big) = O_p(1),$$

which implies that

$$\frac{1}{\sqrt{n}}\frac{d}{d\theta}\frac{\partial L_n}{\partial \lambda}(\theta, \lambda_\theta)\bigg|_{\theta=\theta_0}\big(\hat{\lambda}_0 - \lambda_0\big) = o_p(1)$$

proving (i).

The proof of (ii) follows along similar lines and hence, is omitted. □

PROOF OF LEMMA 3. By Taylor's theorem,

$$l(y;\theta,\hat{\lambda}_\theta(x)) = l(y;\theta,\lambda_\theta(x)) + r(y,x;\theta),$$

where

$$\begin{aligned} r(y,x;\theta) &= \int_0^1 \frac{\partial l}{\partial\eta}\big(y;\lambda_\theta(x) + t(\hat{\lambda}_\theta(x) - \lambda_\theta(x))\big)\,dt\big(\hat{\lambda}_\theta(x) - \lambda_\theta(x)\big) \\ &\equiv Q_\theta^{(1)}(y,x)\big(\hat{\lambda}_\theta(x) - \lambda_\theta(x)\big). \end{aligned}$$

Hence, to prove (i) it suffices to show that

$$\text{(5)} \qquad \sup_\theta \sup_x \left| n^{-1} \sum_i \frac{\partial^j}{\partial\theta^j} Q_\theta^{(1)}(y_i,x_i) \right| = O_p(1), \qquad j = 0,1,2.$$

For $j = 0$,

$$|Q_\theta^{(1)}(y,x)| \le \sup_\theta \sup_{\eta\in H} \left| \frac{\partial l}{\partial\eta}(y;\theta,\eta) \right| \quad \text{for all } x.$$

Hence, (5) for $j = 0$ follows from Conditions S; cases $j = 1,2$ can be established in a similar manner, proving (i).

We now consider (ii). By Taylor's theorem,

$$\begin{aligned} l(y;\theta,\hat{\lambda}_\theta(x)) &= l(y;\theta,\lambda_\theta(x)) + \frac{\partial l}{\partial\eta}(y;\theta,\lambda_\theta(x))\big(\hat{\lambda}_\theta(x) - \lambda_\theta(x)\big) \\ &\quad + r(y,x;\theta), \end{aligned}$$

where

$$\begin{aligned} r(y,x;\theta) &= \frac{1}{2}\int_0^1 \frac{\partial^2 l}{\partial\eta^2}\big(y;\theta,\lambda_\theta(x) + t(\hat{\lambda}_\theta(x) - \lambda_\theta(x))\big)\,dt\big(\hat{\lambda}_\theta(x) - \lambda_\theta(x)\big)^2 \\ &\equiv \frac{1}{2} Q_\theta^{(2)}(y,x)\big(\hat{\lambda}_\theta(x) - \lambda_\theta(x)\big)^2. \end{aligned}$$

Hence, to prove (ii) it suffices to show that

$$\sup_x \left| n^{-1} \sum_i \frac{\partial^j}{\partial\theta^j} Q_\theta^{(2)}(y_i,x_i) \Big|_{\theta=\theta_0} \right| = O_p(1) \quad \text{for } j = 0,1.$$

This can be shown to follow from Conditions S using the same approach as in the proof of (i). □

PROOF OF LEMMA 5. Using the same approach as in the proof of Lemma 8 below it can be shown that, under the conditions of the lemma, there exist constants $\alpha_0 \geq 1/4$, $\alpha_1 > 0$, $\alpha_0 \geq \alpha_1$, such that for all $k = 0, 1$, $j = 0, 1, 2$:

$$\text{(i)} \qquad \sup_{\theta, \eta, x} \left| \frac{\partial^k}{\partial x^k} \frac{\partial^j}{\partial \theta^j} \left(\hat{h}'_n(\theta, \eta, x) - h'(\theta, \eta, x) \right) \right| = o_p(n^{-\alpha_k}),$$

$$\text{(ii)} \qquad \sup_{\theta, \eta, x} \left| \frac{\partial^k}{\partial x^k} \frac{\partial^j}{\partial \theta^j} \left(\hat{h}''_n(\theta, \eta, x) - h''(\theta, \eta, x) \right) \right| = o_p(n^{-\alpha_k}),$$

$$\text{(iii)} \qquad \sup_{\theta, \eta, x} \left| \frac{\partial^k}{\partial x^k} \frac{\partial^j}{\partial \theta^j} \left(\hat{h}'''_n(\theta, \eta, x) - h'''(\theta, \eta, x) \right) \right| = o_p(n^{-\alpha_k}),$$

$$\text{(iv)} \qquad \sup_{\theta, \eta, x} \left| \frac{\partial^k}{\partial x^k} \frac{\partial^j}{\partial \theta^j} \left(\hat{h}^{(4)}_n(\theta, \eta, x) - h^{(4)}(\theta, \eta, x) \right) \right| = o_p(1);$$

here $\hat{h}^{(j)}_n(\theta, \eta, x) = \partial^j \hat{h}_n(\theta, \eta, x)/\partial \eta^j$.

To show that conditions NP are satisfied, it is enough to show that:

$$1. \qquad \sup_{\theta} \left\| \hat{\lambda}_\theta - \lambda_\theta \right\| = o_p(n^{-\alpha_0}),$$

$$2. \qquad \sup_{\theta} \left\| \hat{\lambda}'_\theta - \lambda'_\theta \right\| = o_p(n^{-\alpha_0}),$$

$$3. \qquad \sup_{\theta} \left\| \hat{\lambda}''_\theta - \lambda''_\theta \right\| = o_p(n^{-\alpha_0}),$$

$$4. \qquad \sup_{\theta} \left\| \frac{\partial}{\partial x} \left(\hat{\lambda}_\theta - \lambda_\theta \right) \right\| = o_p(n^{-\alpha_1}),$$

$$5. \qquad \sup_{\theta} \left\| \frac{\partial}{\partial x} \left(\hat{\lambda}'_\theta - \lambda'_\theta \right) \right\| = o_p(n^{-\alpha_1});$$

note that the differentiability of $\hat{\lambda}_\theta(x)$ with respect to θ and x follows immediately from the implicit function theorem [e.g., Saaty and Bram (1964)] and Conditions S.

Under the regularity conditions in effect, for any $\varepsilon > 0$ there exists a $\delta > 0$ such that

$$\begin{aligned} P\left\{ \sup_{\theta} \sup_{x} \left| \hat{\lambda}_\theta(x) - \lambda(x) \right| > \varepsilon \right\} & \\ \leq P\left\{ \sup_{\theta} \sup_{x} \left| h'\left(\theta, \hat{\lambda}_\theta(x), x\right) \right| > \delta \right\} & \\ = P\left\{ \sup_{\theta} \sup_{x} \left| \hat{h}'_n\left(\theta, \hat{\lambda}_\theta(x), x\right) - h'\left(\theta, \hat{\lambda}_\theta(x), x\right) \right| > \delta \right\} & \\ \leq P\left\{ \sup_{\theta} \sup_{x} \sup_{\eta} \left| \hat{h}'_n(\theta, \eta, x) - h'(\theta, \eta, x) \right| > \delta \right\} & \\ \to 0 \quad \text{as } n \to \infty. & \end{aligned}$$

Hence, $\sup_\theta \|\hat{\lambda}_\theta - \lambda_\theta\| = o_p(1)$.

By Conditions I,

$$\inf_\theta \inf_x -h''(\theta, \lambda_\theta(x), x) > 0$$

and by Conditions S, for every $\delta > 0$ there exists an $\varepsilon > 0$ such that

$$\sup_\theta \sup_x \sup_{\eta_1, \eta_2 : |\eta_1 - \eta_2| \le \varepsilon} |h''(\theta, \eta_2, x) - h''(\theta, \eta_1, x)| < \delta.$$

Hence, there exists an $\varepsilon > 0$ such that

$$\inf_\theta \inf_x \inf_{\eta : |\eta - \lambda_\theta(x)| \le \varepsilon} |h''(\theta, \eta, x)| > 0. \tag{6}$$

Since $\hat{\lambda}_\theta(x)$ must satisfy

$$\hat{h}'_n(\theta, \hat{\lambda}_\theta(x), x) = 0$$

for all θ, x and $\lambda_\theta(x)$ must satisfy

$$h'(\theta, \lambda_\theta(x), x) = 0$$

for all θ, x it follows that

$$\begin{aligned} 0 &= \hat{h}'_n(\theta, \hat{\lambda}_\theta(x), x) - h'(\theta, \lambda_\theta(x), x) \\ &= \hat{h}'_n(\theta, \hat{\lambda}_\theta(x), x) - h'(\theta, \hat{\lambda}_\theta(x), x) + h'(\theta, \hat{\lambda}_\theta(x), x) - h'(\theta, \lambda_\theta(x), x) \\ &= \hat{h}'_n(\theta, \hat{\lambda}_\theta(x), x) - h'(\theta, \hat{\lambda}_\theta(x), x) + d_n(\theta, x)(\hat{\lambda}_\theta(x) - \lambda_\theta(x)) \end{aligned}$$

for each θ, x where

$$d_n(\theta, x) = \int_0^1 h''(\theta, s\lambda_\theta(x) + (1-s)\hat{\lambda}_\theta(x), x)\, ds.$$

Note that (6) and the fact that $\sup_\theta \|\hat{\lambda}_\theta - \lambda_\theta\| = o_p(1)$ imply that

$$\begin{aligned} &\liminf \inf_x \inf_\theta \left|\hat{h}''_n(\theta, \hat{\lambda}_\theta(x), x)\right| > 0 \quad \text{and} \\ &\liminf \inf_x \inf_\theta |d_n(\theta, x)| > 0 \quad \text{as } n \to \infty. \end{aligned} \tag{7}$$

Since

$$\hat{h}'_n(\theta, \hat{\lambda}_\theta(x), x) = 0$$

for all θ, x,

$$\hat{h}''_n(\theta, \hat{\lambda}_\theta(x), x) \frac{\partial}{\partial\theta} \hat{\lambda}_\theta(x) + \frac{\partial}{\partial\theta} \hat{h}'_n(\theta, \hat{\lambda}_\theta(x), x) = 0$$

so that by (7) and the preceding conditions (i) and (ii)

$$\sup_x \sup_\theta \left| \frac{\partial}{\partial\theta} \hat{\lambda}_\theta(x) \right| = O_p(1).$$

Similarly,

$$\sup_x \sup_\theta \left| \frac{\partial^k}{\partial x^k} \frac{\partial^j}{\partial \theta^j} \hat{\lambda}_\theta(x) \right| = O_p(1), \qquad k = 0,1, j = 0,1,2. \tag{8}$$

Let

$$r_n(\theta, x) = \hat{h}'_n(\theta, \hat{\lambda}_\theta(x), x) - h'(\theta, \hat{\lambda}_\theta(x), x).$$

By (8), together with conditions (i)–(iv),

$$\sup_x \sup_\theta \left| \frac{\partial^k}{\partial x^k} \frac{\partial^j}{\partial \theta^j} r_n(\theta, x) \right| = o_p(n^{-\alpha_k}), \qquad k = 0,1, j = 0,1,2 \tag{9}$$

and

$$\sup_x \sup_\theta \left| \frac{\partial^k}{\partial x^k} \frac{\partial^j}{\partial \theta^j} d_n(\theta, x) \right| = O_p(1), \qquad k = 0,1, j = 0,1,2. \tag{10}$$

Now, for each x and θ

$$r_n(\theta, x) + d_n(\theta, x)\big(\hat{\lambda}_\theta(x) - \lambda_\theta(x)\big) = 0$$

so that by (7)–(10).

$$\sup_\theta \left\| \hat{\lambda}_\theta - \lambda_\theta \right\| = o_p(n^{-\alpha_0}).$$

Differentiating with respect to θ yields

$$\frac{\partial}{\partial \theta} r_n(\theta, x) + \frac{\partial}{\partial \theta} d_n(\theta, x)\big(\hat{\lambda}_\theta(x) - \lambda_\theta(x)\big) + d_n(\theta, x)\left(\frac{\partial}{\partial \theta} \hat{\lambda}_\theta(x) - \frac{\partial}{\partial \theta} \lambda_\theta(x) \right) = 0$$

so that

$$\sup_\theta \left\| \hat{\lambda}'_\theta - \lambda'_\theta \right\| = o_p(n^{-\alpha_0}).$$

It can be shown, using the same approach, that (3)–(5) are satisfied as well; the lemma follows. □

PROOF OF LEMMA 8. Consider the case $j = 0$. Let

$$\hat{g}_\theta(x) = \frac{1}{nh_n} \sum T_\theta(T_j) K\left(\frac{x - X_j}{h_n} \right), \qquad g_\theta(x) = m_\theta(x) f(x),$$

$$\hat{f}(x) = \frac{1}{nh_n} \sum K\left(\frac{x - X_j}{h_n} \right),$$

$$f_\theta(y, x) = f_{\theta_0}(y|x) f(x),$$

$$\hat{g}_\theta^{(r)}(z) = \frac{\partial^r}{\partial z^r} \hat{g}_\theta(z), \qquad g_\theta^{(r)}(z) = \frac{\partial^r}{\partial z^r} g_\theta(z),$$

where $f_\theta(y)$ denotes the marginal density of $T_\theta(Y)$ and $f_\theta(\cdot|y)$ denotes the conditional density of X given $T_\theta(Y) = y$. Note that $m_\theta(x) = g_\theta(x)/f(x)$ and $\hat{m}_\theta(x) = \hat{g}_\theta(x)/\hat{f}(x)$.

First consider $E\{\hat{g}_\theta^{(r)}(z)\} - g_\theta^{(r)}(z)$.

$$E\{\hat{g}_\theta^{(r)}(z)\} = h_n^{-(r+1)} \int\int y K^{(r)}\left(\frac{x-z}{h_n}\right) f_\theta(y, x)\, dy\, dx$$

$$= \int y f_\theta(y) \int h_n^{-r} K^{(r)}(u)\, f_\theta(z - h_n u\,|y)\, du\, dy,$$

where $u = (z - x)/h_n$. By repeated application of integration by parts,

$$\int K^{(r)}(u)\, f_\theta(z - h_n u\,|y)\, du = h_n^r \int K(u)\, f_\theta^{(r)}(z - h_n u\,|y)\, du.$$

Hence,

$$\left|E\{\hat{g}_\theta^{(r)}(z)\} - g_\theta^{(r)}(z)\right|$$

$$= \left|\int y f_\theta(y) \int K(u)\left(f_\theta^{(r)}(z - h_n u\,|y) - f_\theta^{(r)}(z|y)\right) du\, dy\right|$$

$$= \left|\int y f_\theta(y) \int K(u)\left(f_\theta^{(r+1)}(z|y)\, h_n u + \tfrac{1}{2} f_\theta^{(r+2)}(z^*|y)\, h_n^2 u^2\right) du\, dy\right|$$

(where z^* lies between z and $z - h_n u$)

$$\leq \int |y| f_\theta(y)\, dy \int \tfrac{1}{2} u^2 K(u)\, du \sup_{z,y} \left| f_\theta^{(r+2)}(z|y)\right| h_n^2$$

$$= O(h_n^2)$$

uniformly in z and θ.

Now consider $\tilde{g}^{(r)}(z) \equiv \hat{g}_\theta^{(r)}(z) - E\{\hat{g}_\theta^{(r)}(z)\}$.

$$\tilde{g}^{(r)}(z) = \frac{1}{n} h_n^{-(r+1)} \sum \left[T_\theta(Y_j) K^{(r)}\left(\frac{z - X_j}{h_n}\right) - E\left\{T_\theta(Y_1) K^{(r)}\left(\frac{z - X_1}{h_n}\right)\right\}\right].$$

Note that if $W_1, \dots, W_n$ are independent, identically distributed random variables with $EW_1 = 0$ and $E\{W_1^q\} < \infty$ for some $q = 2, 4, \dots,$ then for $\varepsilon > 0$.

$$P\left\{\left|\frac{1}{n}\sum W_j\right| > \varepsilon\right\} \leq \frac{E\{W_1^q\} c_q}{n^{q/2}\varepsilon^q}$$

for some constant c_q depending only on q. Hence, for each $\theta \in \Theta$, $z \in [0, 1]$ and $\varepsilon > 0$,

$$P\left\{\left|\tilde{g}_\theta^{(r)}(z)\right| > \varepsilon\right\} \leq \frac{c}{n^{q/2}(\varepsilon h_n)^q}$$

for some constant c not depending on θ or z. This follows from the fact that

$$E\left\{\left|T_\theta(Y_1)K^{(r)}\left(\frac{z-X_1}{h_n}\right)\right|^q\right\} \le \sup_z |K^{(r)}(z)|^q \sup_\theta E\{|T_\theta(Y_1)|^q\} < \infty.$$

Let θ_1, θ_2 be elements of Θ. Since $E\{\sup_\theta |T_\theta'(Y_1)|\} < \infty$, there exist i.i.d. random variables $M_1^{(1)}, M_2^{(1)}, \dots,$ not depending on θ_1 or θ_2, such that $E\{|M_j^{(1)}|\} < \infty$ and

$$\sup_z \left|\tilde{g}_{\theta_1}^{(r)}(z) - \tilde{g}_{\theta_2}^{(r)}(z)\right| \le \sup_z |K^{(r)}(z)| \frac{|\theta_2 - \theta_1|}{h_n^{(r+1)}} \frac{1}{n} \sum M_j^{(1)}.$$

Similarly, for any z_1, $z_2 \in [0, 1]$, there exist i.i.d. random variables $M_1^{(2)}, M_2^{(2)}, \dots,$ not depending on z_1, z_2, such that $E\{|M_j^{(2)}|\} < \infty$ and

$$\sup_\theta \left|\tilde{g}_\theta^{(r)}(z_2) - \tilde{g}_\theta^{(r)}(z_1)\right| \le \sup_z |K^{(r+1)}(z)| \frac{|z_2 - z_1|}{h_n^{(r+2)}} \frac{1}{n} \sum M_j^{(2)}.$$

Hence, there exist i.i.d. random variables $M_1, M_2, \dots,$ such that $E\{|M_j|\} < \infty$ and

$$\begin{aligned}\sup_{|\theta_1-\theta_2|<\delta} \sup_{|z_1-z_2|<\delta} \left|\tilde{g}_{\theta_2}^{(r)}(z_2) - \tilde{g}_{\theta_1}^{(r)}(z_1)\right| &\le \left(\frac{1}{n}\sum M_j\right)\delta\left(h_n^{-(r+1)} + h_n^{-(r+2)}\right)\\ &\le 2\left(\frac{1}{n}\sum M_j\right)h_n^{-(r+2)}\delta\end{aligned}$$

for sufficiently large n.

Let δ_n be a sequence converging to 0 and let Θ_n and Z_n be δ_n-nets in Θ and $[0, 1]$, respectively. Then

$$\begin{aligned}P\left\{\sup_{\theta,z}\left|\tilde{g}_\theta^{(r)}(z)\right| > \varepsilon\right\} &\le P\left\{\max_{\theta\in\Theta}\max_{z\in Z_n}\left|\tilde{g}_\theta^{(r)}(z)\right| > \varepsilon/2\right\}\\ &\quad + P\left\{\sup_{|\theta_1-\theta_2|<\delta}\sup_{|z_1-z_2|<\delta}\left|\tilde{g}_{\theta_2}^{(r)}(z_2) - \tilde{g}_{\theta_1}^{(r)}(z_1)\right| > \varepsilon/2\right\}\\ &\le \frac{c_1\delta_n^{-2}}{n^{q/2}\varepsilon^q h_n^{q(r+1)}} + \frac{c_2\delta_n}{h_n^{(r+2)}\varepsilon}\end{aligned}$$

for some constants c_1, c_2. Hence, taking

$$\varepsilon_n = n^{-(q/(2q+4))}n^\gamma h_n^{-(r+(q+4)/(q+2))}$$

for some $\gamma > 0$ and $\delta_n = O(n^{-q/6}(\varepsilon_n h_n^{r+1})^{-(q-1)/3}h_n^{-1/3})$ it is easily shown that

$$P\left\{\sup_{\theta,z}\left|\tilde{g}_\theta^{(r)}(z)\right| > \varepsilon_n\right\} = o(1) \quad \text{as } n \to \infty.$$

It follows that

$$\sup_{\theta}\sup_{z}\left|\hat{g}_{\theta}^{(r)}(z)-g_{\theta}^{(r)}(z)\right|=O_p\left(h_n^2+n^{-(q/(2q+4))}n^{\gamma}h_n^{-(r+(q+4)/(q+2))}\right).$$

Using the same approach it can be shown that

$$\sup_{z}\left|\hat{f}^{(r)}(z)-f^{(r)}(z)\right|=O_p\left(h_n^2+n^{-(q/(2q+4))}n^{\gamma}h_n^{-(r+(q+4)/(q+2))}\right)$$

and since $f(\cdot)$ is bounded away from 0 the result for $j = 0$ follows; the proofs for $j = 1, 2$ follow along similar lines. □

REFERENCES

Bahadur, R. R. (1964). On Fisher's bound for asymptotic variances. *Ann. Math. Statist.* **35** 1545–1552.

Bahadur, R. R. (1971). *Some Limit Theorems in Statistics*. SIAM, Philadelphia.

Begun, J. M., Hall, W. J., Huang, W. M. and Wellner, J. A. (1983). Information and asymptotic efficiency in parametric-nonparametric models. *Ann. Statist.* **11** 432–452.

Bickel, P. J. (1982). On adaptive estimation. *Ann. Statist.* **10** 647–671.

Bickel, P. J., Klaasen, C. A. J., Ritov, Y. and Wellner, J. A. (1991). *Efficient and Adaptive Inference in Semiparametric Models*. Johns Hopkins Univ. Press. To appear.

Hájek, J. (1970). A characterization of limiting distributions of regular estimates. *Z. Wahrsch. Verw. Gebiete* **14** 323–330.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under non-standard conditions. *Proc. Fifth Berkeley Symp. Math. Statist. Probab.* **1** 221–234. Univ. California Press, Berkeley.

Jain, N. and Marcus, M. (1975). Central limit theorem for $C(S)$-valued random variables. *J. Funct. Anal.* **19** 216–231.

Koshevnik, Y. A. and Levit, B. Y. (1976). On a nonparametric analogue of the information matrix. *Theory Probab. Appl.* **21** 738–753.

Levit, B. Y. (1974). On optimality of some statistical estimates. In *Proceedings of the Prague Symposium on Asymptotic Statistics* (J. Hájek, ed.) **2** 215–238. Univ. Karlova, Prague.

Levit, B. Y. (1975). On the efficiency of a class of nonparametric estimates. *Theory Probab. Appl.* **20** 723–740.

Lindsay, B. G. (1980). Nuisance parameters, mixture models and the efficiency of partial likelihood estimators. *Philos. Trans. Roy. Soc. London Ser. A* **296** 639–665.

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

Pfanzagl, J. (1982). *Contributions to a General Asymptotic Statistical Theory*. Springer, New York.

Saaty, T. L. and Bram, J. (1964). *Nonlinear Mathematics*. McGraw-Hill, New York.

Schick, A. (1986). On asymptotically efficient estimation in semiparametric models. *Ann. Statist.* **14** 1132–1138.

Staniswalis, J. G. (1989). On the kernel estimate of a regression function in likelihood based models. *J. Amer. Statist. Assoc.* **84** 276–283.

Stein, C. (1956). Efficient nonparametric testing and estimation. *Proc. Third Berkeley Symp. Math. Statist. Probab.* **1** 187–196. Univ. California Press, Berkeley.

van der Vaart, A. W. (1988). Estimating a real parameter in a class of semiparametric models. *Ann. Statist.* **16** 1450–1474.

Wong, W. H. (1986). Theory of partial likelihood. *Ann. Statist.* **14** 88–123.

Wong, W. H. and Severini, T. A. (1991). On maximum likelihood estimation in infinite-dimensional parameter spaces. *Ann. Statist.* **19** 603–632.

Department of Statistics
Northwestern University
2006 Sheridan Road
Evanston, Illinois 60208-4070

Department of Statistics
University of Chicago
5734 University Avenue
Chicago, Illinois 60637