

Gutierrez, Roberto G.; Carroll, Raymond J.

Working Paper

Plug-in semiparametric estimating equations

SFB 373 Discussion Paper, No. 1997,13

Provided in Cooperation with:

Collaborative Research Center 373: Quantification and Simulation of Economic Processes,
Humboldt University Berlin

Suggested Citation: Gutierrez, Roberto G.; Carroll, Raymond J. (1995) : Plug-in semiparametric estimating equations, SFB 373 Discussion Paper, No. 1997,13, Humboldt University of Berlin, Interdisciplinary Research Project 373: Quantification and Simulation of Economic Processes, Berlin, <https://nbn-resolving.de/urn:nbn:de:kobv:11-10063736>

This Version is available at:

<https://hdl.handle.net/10419/66244>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

PLUG-IN SEMIPARAMETRIC ESTIMATING EQUATIONS

Roberto G. Gutierrez
Department of Statistical Science
Southern Methodist University
Dallas, Texas 77275-0332

Raymond J. Carroll
Department of Statistics
Texas A&M University
College Station, Texas 77843-3143

December 15, 1995

Abstract

In parametric regression problems, estimation of the parameter of interest is typically achieved via the solution of a set of unbiased estimating equations. We are interested in problems where in addition to this parameter, the estimating equations consist of an unknown nuisance function which does not depend on the parameter. We study the effects of using a plug-in nonparametric estimator of the nuisance function (for example, a local-linear regression estimator) on the estimability of the parameter. In particular, we specify conditions on the functional estimator which ensure that the parametric rate of consistency for estimating the parameter of interest is preserved, and we give a general asymptotic covariance formula. We apply this theory to three examples.

Some Key Words: Generalized Linear Models; Local Linear Regression; Logistic Regression; Missing Data; Nonparametric Regression; Partially Linear Models; Semiparametric Regression.

1. INTRODUCTION

In many estimation problems a nuisance function is present. For example, consider the linear regression setting where observations on some regressors are sometimes missing. In this particular case, the estimating scheme of Horvitz and Thompson (1952) estimates the regression parameters by using weighted least squares with weights set equal to the inverse of the missingness probability. In situations where the data are not missing by design, this probability of missingness would have to be estimated as well.

One solution would be to also specify a parametric model for the missingness probability. This brings up the issue of lack of fit, in that misspecification of such a model could result in inconsistent estimates of the regression parameters. An alternative strategy would be to model the missingness process using a more flexible semiparametric model, or to estimate the missingness probabilities using nonparametric regression. Under such a strategy, estimation of the regression parameters becomes a semiparametric scheme, consisting of a parametric part (the regression parameters), and a nonparametric or semiparametric part (the missingness process).

Semiparametric problems arise in numerous other settings, such as measurement–error models, partial spline models, and partially linear models. In this paper, particular attention is given to the parametric part of the problem, and the semiparametric or nonparametric part is considered to be a nuisance quantity for which a plug–in estimator is used. Once this plug–in estimator is in place, estimation of the parametric portion may be done using conventional techniques such as maximum likelihood, maximum quaslikelihood, etc.

Because of the relaxation of model assumption, estimates of the semiparametric part of the estimation scheme converge at a rate slower than the $n^{1/2}$ –rate, and thus may have an effect in terms of bias and variance on the estimates of the parametric part. It is the focus of this paper to study such effects in a general setting.

In Section 2 we formulate the plug–in semiparametrics setting in general terms. In Section 3 we outline a general asymptotic theory for parameter estimation in terms of the plug–in estimators of the semiparametric part of the estimation scheme. In Section 4 we apply the theory to three examples motivated by missing data and give some concluding remarks in Section 5.

2. GENERAL FORMULATION

Parametric regression problems which consist of using independent and identically distributed data \mathcal{Y}_i ($i = 1, \dots, n$) to estimate the parameter Θ_0 vary to the extent of what is assumed about

the probability distribution of \mathcal{Y} . Given such an assumption, an estimate $\hat{\Theta}$ of Θ_0 is obtained by solving a set of estimating equations

$$0 = \sum_{i=1}^n \Psi(\mathcal{Y}_i, \Theta) \quad (2.1)$$

in Θ . For example when the probability distribution of \mathcal{Y} is fully specified up to Θ_0 , Ψ will correspond to a loglikelihood score function. In the case where \mathcal{Y} consists of a response and a set of regressors, one could specify the mean and variance of the response as functions in Θ_0 of the regressors. In such a case Ψ will correspond to a *quasilielihood* score function (McCullagh & Nelder, 1989).

In this paper, we generalize (2.1) to allow for the presence of an unknown nuisance function which does not depend on Θ , and in doing so we allow for two additional nuisance parameters. Consider the case where finding $\hat{\Theta}$ requires solving in Θ

$$0 = \sum_{i=1}^n \Psi\{\mathcal{Y}_i, \Theta, \eta_0(U_i^t \alpha_0), \gamma_0\}, \quad (2.2)$$

where Ψ is a vector of the same dimension of Θ , U is some component of \mathcal{Y} , $\eta_0(\cdot)$ is an unknown function, $U^t \alpha_0$ is a linear projection of U (often called a “single-index”), and γ_0 is a parameter of secondary interest which arises from the inclusion of $\eta_0(\cdot)$.

Semiparametric estimating schemes commonly occur in problems involving missing data. In this context, (2.2) is widely applicable, and in this paper we consider three missing data examples which fit the framework of it; see Section 4.

We wish to ascertain the effect of using *plug-in* estimators of the unknown nuisance quantities in (2.2) on the estimation of Θ_0 . We are interested in the properties of $\hat{\Theta}$ which solves

$$0 = \sum_{i=1}^n \Psi\{\mathcal{Y}_i, \Theta, \hat{\eta}(U_i^t \hat{\alpha}, \hat{\alpha}, \hat{\gamma}), \hat{\gamma}\}, \quad (2.3)$$

where $(\hat{\alpha}, \hat{\gamma})$ estimates (α_0, γ_0) and $\hat{\eta}(\cdot)$ is a nonparametric regression estimate of the univariate function $\eta_0(\cdot)$ evaluated at $U^t \hat{\alpha}$ and allowed to depend on $(\hat{\alpha}, \hat{\gamma})$. It is assumed that Θ is of dimension p , γ is of dimension q , and α is of dimension d , whereupon it follows that (2.3) represents a system of p equations to solve in Θ . The critical distinction here is that estimation of the nuisance quantities takes place independently of the estimation of Θ , and thus we can estimate Θ using regular parametric techniques once the plug-in estimators are in place.

To clarify the ideas introduced above, consider again the example given in Section 1 where we now introduce some notation. Consider a linear regression of Y on (X, W) , so that $E(Y|X, W) =$

$\theta_0 + X^t\theta_1 + W^t\theta_2$, and $\Theta = (\theta_0, \theta_1^t, \theta_2^t)^t$. Suppose that (Y, W) are observable for all sample units, but that X is missing at random (Little & Rubin, 1987) and observable only with probability $\pi(Y_i, W_i)$ for $i = 1, \dots, n$. If $\Delta_i = 1$ when X_i is observed and $\Delta_i = 0$ otherwise, the Horvitz and Thompson (1952) estimator is obtained by solving

$$0 = \sum_{i=1}^n \frac{\Delta_i}{\pi(Y_i, W_i)} (Y_i - \theta_0 - X_i^t\theta_1 - W_i^t\theta_2) \begin{pmatrix} 1 \\ X_i \\ W_i \end{pmatrix}.$$

This is a weighted least squares estimate based on the “complete” data with (Y, X, W) all observed, and weighted inversely to the sample probabilities. Assuming that the sampling probabilities $\pi(Y_i, W_i)$ are known, this estimator fits into the framework (2.1) with $\mathcal{Y} = (Y, X, W, \Delta)$.

Now suppose though that this is observational data and the sampling weights are unknown, in which case one would be forced to estimate them. One could use logistic regression with higher order polynomial terms to insure model robustness (Robins, Rotnitzky & Zhao, 1994), or one could estimate the sampling probabilities by nonparametric regression of Δ on (Y, W) ; this latter approach is taken by Wang, Wang, Zhao & Ou (1995). The difficulty with the nonparametric approach is that if W is multivariate, the nonparametric regression may suffer from the curse of dimensionality; this shows up in the Wang et al. theory in the need for progressively higher order kernels. Instead, one might propose a compromise between linear logistic regression and nonparametric regression. For example, if Y were the primary determiner of missingness, a natural model is

$$\Pr(\Delta = 1|Y, W) = \text{Logistic} \left\{ \eta_0(Y) + W^t\gamma_0 \right\}, \quad (2.4)$$

while an alternative is

$$\Pr(\Delta = 1|Y, W) = \text{Logistic} \left\{ \eta_0(W^t\alpha_0) + \gamma_0 Y \right\}. \quad (2.5)$$

Model (2.4) is called a partially linear model (Severini & Staniswalis, 1994) and is a special case of generalized additive models (Hastie & Tibshirani, 1990), while model (2.5) with $\gamma_0 = 0$ is a single-index model (Härdle, Hall & Ichimura, 1993); model (2.5) itself is a partially linear, single-index model (Carroll, Fan, Gijbels & Wand, 1995). If $H(\cdot)$ is the logistic distribution function and (2.5) is used, then the resulting estimator fits into the framework (2.2) with $U = W$ and

$$\Psi(\mathcal{Y}, \Theta, \eta_0(U^t\alpha_0), \gamma_0) = \frac{\Delta}{H\{\eta_0(U^t\alpha_0) + \gamma_0 Y\}} (Y - \theta_0 - X^t\theta_1 - W^t\theta_2) \begin{pmatrix} 1 \\ X \\ W \end{pmatrix}.$$

We develop a general asymptotic theory for $\hat{\Theta}$ which solves (2.3). The theory will consist of two parts. The first part will outline a set of sufficient conditions on the estimating function $\Psi(\cdot)$

and the plug-in estimators $(\hat{\eta}(\cdot), \hat{\alpha}, \hat{\gamma})$ which ensure that the parametric rate of consistency for $\hat{\Theta}$ is retained. The second part will give a general formula for the asymptotic covariance of $\hat{\Theta}$.

Newey (1994) refers to the estimating scheme given by (2.3) as a two-step estimator, where the first step is the estimation of the plug-in quantities and the second is the estimation of Θ_0 . Newey considers the situation where $\hat{\eta}(\cdot)$ is a series estimator, such as a spline or polynomial regression, and where α_0 is known. We consider the case where $\hat{\eta}(\cdot)$ is a kernel regression estimator such as that of Nadaraya (1964) and Watson (1964) or a local polynomial regression with kernel weights, and thus our results are presented in terms of smoothing parameters associated with these smoothing methods. We also consider the effects of estimating α_0 and γ_0 .

3. ASYMPTOTIC THEORY

3.1. Asymptotic Expansions of the Plug-in Quantities

The large sample bias and variance properties of $\hat{\Theta}$ will depend mainly on the properties of $\hat{\alpha}$, $\hat{\gamma}$ and $\hat{\eta}(\cdot)$. The nonparametric estimator $\hat{\eta}(\cdot)$ used here is of a kernel regression form (ordinary or local linear). We will state our conditions in general form, but Carroll et al. (1995) have shown that they hold for the class of partially linear single-index models using either backfitting as commonly employed (see Weisberg & Welsh, 1994 for a recent example) or for nonparametric likelihood (Severini & Wong, 1992). As a technical matter the sums in (2.3) must be constrained to a compact subset of the support of the U 's. Our assumption is that uniformly over u in a compact set interior to the support of U ,

$$\begin{aligned} \hat{\eta}(u^t \hat{\alpha}, \hat{\alpha}, \hat{\gamma}) - \eta_0(u^t \alpha_0) &= h^2 \zeta_0(u) + a_\gamma^t(u)(\hat{\gamma} - \gamma_0) + a_\alpha^t(u)(\hat{\alpha} - \alpha_0) \\ &\quad + n^{-1} \sum_{j=1}^n K_h(U_j^t \alpha_0 - u^t \alpha_0) \rho_0(\mathcal{Y}_j, u^t \alpha_0) + O_p(h^3) + o_p(n^{-1/2}), \end{aligned} \quad (3.1)$$

for some functions $\zeta_0(u)$, $a_\alpha(u)$, $a_\gamma(u)$, and $\rho_0(\cdot)$. In (3.1) K is a symmetric probability density function and h is the bandwidth, with rescaling $K_h(t) = K(t/h)/h$. It is also assumed that $E\{\rho_0(\mathcal{Y}, u^t \alpha_0) | U^t \alpha_0\} = 0$, for all u in the compact set described above.

We also suppose that the following expansions exist:

$$\begin{aligned} \hat{\alpha} - \alpha_0 &= h^2 b_\alpha + n^{-1} \sum_{j=1}^n V_{j\alpha} + O_p(h^3) + o_p(n^{-1/2}); \\ \hat{\gamma} - \gamma_0 &= h^2 b_\gamma + n^{-1} \sum_{j=1}^n V_{j\gamma} + O_p(h^3) + o_p(n^{-1/2}), \end{aligned} \quad (3.2)$$

for some constants b_α and b_γ and random variables $V_{j\alpha}$ and $V_{j\gamma}$ ($j = 1, \dots, n$) which are independent and identically distributed observations on the random vectors V_α and V_γ respectively, having mean

zero and finite covariance matrix. It should be noted that the above expansions allow for estimation (α_0, γ_0) at a rate of convergence slower than that of the parametric rate, which in some cases may be a direct result of their estimation occurring in conjunction with that of $\eta_0(\cdot)$. When in fact we can (for example) estimate α_0 at the $n^{1/2}$ -rate, the expansion reduces accordingly and $b_\alpha = 0$. For example, in partially linear single-index models, in backfitting it is typical that $b_\alpha \neq 0$ and $b_\gamma \neq 0$, while for nonparametric likelihood $b_\alpha = b_\gamma = 0$.

3.2. Statement of the Main Result

Define the following:

$$\begin{aligned} S_\Theta(\mathcal{Y}, u^t \alpha_0, \Theta_*, \gamma_*) &= \frac{\partial}{\partial \Theta^t} \Psi(\mathcal{Y}, \Theta, v, \gamma); \\ S_v(\mathcal{Y}, u^t \alpha_0, \Theta_*, \gamma_*) &= \frac{\partial}{\partial v} \Psi(\mathcal{Y}, \Theta, v, \gamma); \\ S_\gamma(\mathcal{Y}, u^t \alpha_0, \Theta_*, \gamma_*) &= \frac{\partial}{\partial \gamma^t} \Psi(\mathcal{Y}, \Theta, v, \gamma), \end{aligned} \tag{3.3}$$

each evaluated at $\Theta = \Theta_*$, $v = \eta_0(u^t \alpha_0)$, and $\gamma = \gamma_*$, noting that $S_\Theta(\cdot)$ is a $(p \times p)$ matrix, $S_v(\cdot)$ is $(p \times 1)$, and $S_\gamma(\cdot)$ is $(p \times q)$. Let $R = U^t \alpha_0$ and $f_R(r)$ be the density function of R evaluated at r . Also define $\mathbf{C} = E\{S_\Theta(\mathcal{Y}, R, \Theta_0, \gamma_0)\}$ and $\Sigma = E\Omega\Omega^t$, where

$$\begin{aligned} G(\mathcal{Y}, u^t \alpha_0) &= f_R(u^t \alpha_0) E\{S_v(\mathcal{Y}, u^t \alpha_0, \Theta_0, \gamma_0) | R = u^t \alpha_0\} \rho_0(\mathcal{Y}, u^t \alpha_0); \\ \Omega &= \Psi\{\mathcal{Y}, \Theta_0, \eta_0(R), \gamma_0\} + G(\mathcal{Y}, R) + E\{S_\gamma(\mathcal{Y}, R, \Theta_0, \gamma_0)\} V_\gamma \\ &\quad + E\{S_v(\mathcal{Y}, R, \Theta_0, \gamma_0) a_\gamma^t(U)\} V_\gamma + E\{S_v(\mathcal{Y}, R, \Theta_0, \gamma_0) a_\alpha^t(U)\} V_\alpha. \end{aligned} \tag{3.4}$$

THEOREM 1: Suppose that $\widehat{\Theta}$ solves (2.3) and the expansions (3.1) and (3.2) hold. Given the assumptions listed in Section 6.1, suppose also that as $n \rightarrow \infty$, $h \rightarrow 0$, $nh^2 \rightarrow \infty$, and either of the following two conditions holds:

Condition 1: $nh^4 \rightarrow 0$.

Condition 2:

$$\begin{aligned} nh^6 &\rightarrow 0; \\ E\{S_v(\mathcal{Y}, R, \Theta_0, \gamma_0) \zeta_0(U)\} &= 0; \quad E\{S_v(\mathcal{Y}, R, \Theta_0, \gamma_0) a_\gamma^t(U)\} b_\gamma = 0; \\ E\{S_v(\mathcal{Y}, R, \Theta_0, \gamma_0) a_\alpha^t(U)\} b_\alpha &= 0; \quad E\{S_\gamma(\mathcal{Y}, R, \Theta_0, \gamma_0)\} b_\gamma = 0. \end{aligned}$$

Then,

$$n^{1/2}(\widehat{\Theta} - \Theta_0) \xrightarrow{D} \text{Normal}(0, \mathbf{C}^{-1} \Sigma \mathbf{C}^{-t}). \tag{3.5}$$

Some comments on this main result are appropriate here:

- (a) Conditions 1 and 2 are provided as a tool to determine what sort of bandwidths h may be used in the nonparametric estimate of $\eta_0(\cdot)$ while preserving the $n^{1/2}$ -consistency of $\hat{\Theta}$. Nominally, optimal bandwidths in kernel regression are of order $O(n^{-1/5})$. In general cases where a kernel regression estimate is used in the context of a parametric problem, one requires an *undersmoothed* version of the nonparametric regression, and bandwidths of the optimal order are specifically excluded (Condition 1). However, if the structure of $\Psi(\cdot)$, and specifically the context in which $\hat{\eta}(\cdot)$ is used within $\Psi(\cdot)$, possesses certain properties (Condition 2), then bandwidths of the nominally optimal order are permitted.
- (b) Equation (3.5) gives a general asymptotic covariance calculation. Recalling that $\Sigma = E\Omega\Omega^t$, the matrix Ω defined in (3.4) consists of five terms, the latter four of which correspond to the asymptotic cost in efficiency due to using the plug-in estimates of the nuisance quantities. Again, in certain cases $\Psi(\cdot)$ may be structured so that one or more of these terms will equal zero.
- (c) It is worth noting that none of the terms in (3.5) depend on the kernel or the bandwidth. This does not mean, however, that the type of smoothing one does to estimate $(\eta_0, \alpha_0, \gamma_0)$ is immaterial, because the type of smoother determines $(\rho_0, a_\alpha, a_\gamma)$ in (3.1) and $(V_{j_\alpha}, V_{j_\gamma})$ in (3.2). Carroll et al. (1995) exhibit different estimation methods in which these terms differ in partially linear and single-index models.

4. EXAMPLES

In this section we consider three examples where we apply the semiparametric methods discussed in this paper. Each example arises from a parametric regression problem in which one of the regressors is sometimes missing. We thus have the common notation $\mathcal{Y} = (Y, X, W, \Delta)$, where Y is a univariate response, X is a univariate regressor which is missing on a subset of the data, W is a possibly multivariate set of regressors which are always observed, and Δ is an indicator variable equal to 1 when X is observed and 0 otherwise. In the absence of missing data, we would estimate the crucial parameter Θ_0 by solving in Θ

$$0 = \sum_{i=1}^n \Psi_c(Y_i, X_i, W_i, \Theta),$$

where $\Psi_c(\cdot)$ is the complete-data estimating function. Define $L = (Y, W)$ and $\pi(L) = \Pr(\Delta = 1|L)$, the probability that X is observed given L . We assume missingness at random, so that $\pi(L) = \Pr(\Delta = 1|L, X)$, and any parameters in semiparametric models do not involve Θ .

4.1. Adaptive Efficient Semiparametric Regression

Suppose that $E(Y|X, W) = m(X, W, \Theta_0)$. For $\Psi_c(Y, X, W, \Theta) = a(X, W)\{Y - m(X, W, \Theta)\}$, Robins et al. (1994) consider estimating Θ_0 via solving the equation

$$0 = \sum_{i=1}^n \left[\frac{\Delta_i}{\pi(L_i)} \Psi_c(Y_i, X_i, W_i, \Theta) - \frac{\Delta_i - \pi(L_i)}{\pi(L_i)} E\{\Psi_c(Y, X, W, \Theta)|L = L_i\} \right], \quad (4.1)$$

where the expectation in the second term is dependent upon the distribution of X given L . Gutierrez (1995) proposed fitting this conditional distribution using a generalized partially linear single-index model. In such a model it is supposed that for a specified partition $L = (L_1, L_2)$, the conditional density of X given L is of the form $f_{X|L}\{x, \eta_0(L_1^t \alpha_0) + L_2^t \gamma_0\}$, where $\eta_0(\cdot)$, α_0 , and γ_0 are estimated using an iterative backfitting routine based upon local linear regression. Carroll et al. (1995) show that for these estimates $(\hat{\eta}(\cdot, \hat{\alpha}, \hat{\gamma}), \hat{\alpha}$, and $\hat{\gamma})$ the asymptotic expansions of (3.1) and (3.2) do exist, yet we will not require their explicit forms in order to apply Theorem 1.

For $\ell = (y, w)$ with partition $\ell = (\ell_1, \ell_2)$ corresponding to the above, define $g(\ell, a, \Theta) = \int \Psi_c(y, x, w, \Theta) f_{X|L}(x, a) dx$. It follows that $\hat{\Theta}$ solves

$$0 = \sum_{i=1}^n \Psi\{\mathcal{Y}_i, \Theta, \hat{\eta}(L_1^t \hat{\alpha}, \hat{\alpha}, \hat{\gamma}), \hat{\gamma}\},$$

where

$$\Psi(\mathcal{Y}, \Theta, v, \gamma) = \frac{\Delta}{\pi(L)} \Psi_c(Y, X, W, \Theta) - \frac{\Delta - \pi(L)}{\pi(L)} g(L, v + L_2^t \gamma, \Theta).$$

Note that in this context L_1 plays the role of U in the statement of Theorem 1 and let $R = L_1^t \alpha_0$.

Since $E(\Delta|\mathcal{Y}) = \pi(L)$, it is easily seen that

$$\begin{aligned} E\{S_v(\mathcal{Y}, R, \Theta_0, \gamma_0)|L_1\} &= E[E\{S_v(\mathcal{Y}, R, \Theta_0, \gamma_0)|\mathcal{Y}\}] = 0 \quad \text{and} \\ E\{S_\gamma(\mathcal{Y}, R, \Theta_0, \gamma_0)\} &= E[E\{S_\gamma(\mathcal{Y}, R, \Theta_0, \gamma_0)|\mathcal{Y}\}] = 0. \end{aligned}$$

Thus, we only require $nh^6 \rightarrow 0$ to ensure that Condition 2 is satisfied.

Furthermore, the above arguments confirm that $G(\mathcal{Y}, r) = 0$ for each choice of r , and thus the last four terms of Ω are equal to 0. Hence by Theorem 1, the asymptotic covariance of $\hat{\Theta}$ is $\mathbf{C}^{-1}(E\Omega\Omega^t)\mathbf{C}^{-t}$, where $\Omega = \Psi\{\mathcal{Y}, \Theta_0, \eta_0(R), \gamma_0\}$ and

$$\mathbf{C} = E\{S_\Theta(\mathcal{Y}, R, \Theta_0, \gamma_0)\} = E[E\{S_\Theta(\mathcal{Y}, R, \Theta_0, \gamma_0)|L\}] = E\left\{ \frac{\partial}{\partial \Theta^t} \Psi_c(Y, X, W, \Theta) |_{\Theta=\Theta_0} \right\}.$$

In particular two things should be noted. First, since the $n^{1/2}$ -consistency of $\hat{\Theta}$ follows from Condition 2, optimal bandwidths h of order $O(n^{-1/5})$ are allowed and thus may be obtained via any number of standard data-driven bandwidth selection routines. Second, the asymptotic covariance of $\hat{\Theta}$ is the same to that as if $E\{\Psi_c(Y, X, W, \Theta)|L\}$ were known.

4.2. Estimation of Weights in Efficient Semiparametric Regression

Again considering the estimating scheme of Robins et al. (1994), we estimate Θ_0 via (4.1). For $\phi(L, \Theta) = E\{\Psi_c(Y, X, W, \Theta)|L\}$, we showed in the previous section that in certain cases Θ_0 may be estimated as well as if $\phi(L, \Theta)$ were known. We now assume that $\phi(L, \Theta)$ is indeed known and that the probability of observing X given L , $\pi(L)$, is unknown.

In order to estimate $\pi(L)$ nonparametrically, it is convenient to reduce the dimension of L and thus assume that $\pi(L) \equiv \pi(L^t \alpha_0)$ for some unknown “direction” α_0 . We assign to $\pi(L)$ a logistic single-index model (Carroll et al., 1995), so that $\pi(L^t \alpha_0) = H\{\eta_0(L^t \alpha_0)\}$, where $H(t) = (1 + e^{-t})^{-1}$ and $\eta_0(\cdot)$ is an unknown link function. Simultaneous estimation of α_0 and $\eta_0(\cdot)$ is achieved using nonparametric likelihood (Carroll et al., 1995), whereupon it follows that (3.2) holds with $b_\alpha = 0$, and

$$\begin{aligned} \hat{\eta}(\ell^t \hat{\alpha}, \hat{\alpha}) - \eta_0(\ell^t \alpha_0) &= \frac{1}{2} h^2 \mu_2 \eta_0^{(2)}(\ell^t \alpha_0) + \eta_0^{(1)}(\ell^t \alpha_0) \{\ell - E(L|L^T \alpha_0 = \ell^t \alpha_0)\}^t (\hat{\alpha} - \alpha_0) + \\ & n^{-1} \sum_{j=1}^n K_h(L_j^t \alpha_0 - \ell^t \alpha_0) \frac{\Delta_j - H\{\eta_0(L_j^t \alpha_0)\}}{f_{L^t \alpha_0}(\ell^t \alpha_0) \dot{H}(\ell^t \alpha_0)} + O_p(h^3) + o_p(n^{-1/2}), \end{aligned} \quad (4.2)$$

where $f_{L^t \alpha_0}(\cdot)$ is the density of $L^t \alpha_0$, $\mu_2 = \int u^2 K(u) du$, $\eta_0^{(k)}(\cdot)$ is the k th derivative of $\eta_0(\cdot)$, and $\dot{H}(\cdot) = H(\cdot)\{1 - H(\cdot)\}$.

Simple calculations show that

$$S_v(\mathcal{Y}, r, \Theta_0) = -\dot{H}(r) \left\{ \frac{\Delta}{H^2(r)} \Psi_c(Y, X, W, \Theta_0) - \frac{\Delta}{H^2(r)} \phi(L, \Theta_0) \right\}.$$

Since $E\{S_v(\mathcal{Y}, L^t \alpha_0, \Theta_0)|L\} = E[E\{S_v(\mathcal{Y}, L^t \alpha_0, \Theta_0)|L, \Delta\}] = 0$, by definition of $\phi(\cdot)$, we only require $nh^6 \rightarrow 0$ to ensure that Condition 2 of Theorem 1 is satisfied. Further calculations show that the last four terms of Ω are all equal to zero and therefore as in the previous section, the asymptotic covariance of $\hat{\Theta}$ is the same to that as if $\pi(L)$ were a known function.

4.3. Estimation of Weights in a Complete Data Scheme

For purposes of computational simplicity, it is sometimes preferable to estimate Θ_0 using only those data for which X is observed. Consider the complete data scheme of Horvitz and Thompson (1952) where $\hat{\Theta}$ solves the weighted estimating equation

$$0 = \sum_{i=1}^n \frac{\Delta_i}{\pi(L_i)} \Psi_c(Y_i, X_i, W_i, \Theta).$$

In cases where $\pi(L)$ is unknown, a fully parametric model for $\pi(L)$ can be problematic because misspecification would result in a $\hat{\Theta}$ which is inconsistent.

Wang et al. (1995) consider using a Nadaraya (1964) and Watson (1964) nonparametric regression estimate of $\pi(L)$. Because we use dimension reduction, our models coincide with theirs only in the case that $\pi(L)$ depends only on a single component of L , say R . Routine calculations show that our general theory yields the result of Wang et al. in this case. The only detail worth mentioning here is that they estimate $\pi(R)$ by ordinary kernel regression. If $K(\cdot)$ is the kernel and $K_h(t) = K(t/h)/h$, we use the standard kernel expansion

$$\hat{\pi}(r) - \pi(r) = \frac{1}{2}h^2\mu_2s^{(2)}(r) + n^{-1}\sum_{j=1}^n K_h(R_j - r)\frac{\Delta_j - \pi(R_j)}{f_R(r)} + O_p(h^3) + o_p(n^{-1/2}),$$

where $f_R(\cdot)$ is the probability density of R , and $s(r) = \pi(r)f_R(r)$.

Suppose now that $\pi(L)$ depends on more than one known component of L . In this case Wang et al. (1995) propose the use of high dimensional kernel regression with higher order kernels to control the bias. We propose instead fitting $\pi(L)$ using a logistic partially linear model with

$$\pi(L) = \Pr(\Delta = 1|L) = H\{\eta_0(L_1) + L_2^t\gamma_0\},$$

where $H(\cdot)$ is the logistic function defined in Section 4.2, L_1 is a single component of L , and $\eta_0(\cdot)$ and γ_0 are unknown. Using nonparametric likelihood estimation, letting $R = U = L_1$ and defining $\xi = \eta_0(R) + L_2^t\gamma_0$, Carroll et al. (1995) show that

$$\hat{\gamma} - \gamma_0 = h^2b_\gamma + n^{-1}\sum_{j=1}^n\{\Delta_j - H(\xi_j)\}B_1^{-1}\left[L_{2j} - \frac{E\{L_2\dot{H}(\xi)|R_j\}}{E\{\dot{H}(\xi)|R_j\}}\right] + O_p(h^3) + o_p(n^{-1/2}),$$

and

$$\begin{aligned}\hat{\eta}(r, \hat{\gamma}) - \eta_0(r) &= \frac{1}{2}h^2\mu_2\eta_0^{(2)}(r) - \frac{E\{L_2^t\dot{H}(\xi)|R=r\}}{E\{\dot{H}(\xi)|R=r\}}(\hat{\gamma} - \gamma_0) + \\ & n^{-1}\sum_{j=1}^n K_h(R_j - r)\frac{\Delta_j - H(\xi_j)}{f_R(r)E\{\dot{H}(\xi)|R=r\}} + O_p(h^3) + o_p(n^{-1/2}),\end{aligned}\quad (4.3)$$

where

$$B_1 = E\{L_2L_2^t\dot{H}(\xi)\} - E\left[\frac{E\{L_2\dot{H}(\xi)|R\}E\{L_2^t\dot{H}(\xi)|R\}}{E\{\dot{H}(\xi)\}}\right].$$

Define $\hat{\Theta}(\hat{\gamma})$ to be the estimate which solves $0 = \sum_{i=1}^n \Psi\{\mathcal{Y}_i, \Theta, \hat{\eta}(R_i, \hat{\gamma}), \hat{\gamma}\}$ where

$$\Psi(\mathcal{Y}, \Theta, v, \gamma) = \frac{\Delta}{H(v + L_2^t\gamma)}\Psi_c(Y, X, W, \Theta).$$

It follows then that

$$\begin{aligned}S_v(\mathcal{Y}, r, \Theta_0, \gamma_0) &= -\frac{\Delta[1 - H\{\eta_0(r) + L_2^t\gamma_0\}]}{H\{\eta_0(r) + L_2^t\gamma_0\}}\Psi_c(Y, X, W, \Theta_0); \\ S_\gamma(\mathcal{Y}, r, \Theta_0, \gamma_0) &= -\frac{\Delta[1 - H\{\eta_0(r) + L_2^t\gamma_0\}]}{H\{\eta_0(r) + L_2^t\gamma_0\}}\Psi_c(Y, X, W, \Theta_0)L_2^t.\end{aligned}$$

Since for example $E\{S_v(\mathcal{Y}, R, \Theta_0, \gamma_0)\zeta_0(R)\}$ is not necessarily zero, we must assume $nh^4 \rightarrow 0$ to satisfy Condition 1.

Calculating the asymptotic covariance of $n^{1/2}\{\hat{\Theta}(\hat{\gamma}) - \Theta_0\}$, we first obtain

$$a_\gamma(r) = -E\{L_2\dot{H}(\xi)|R=r\} \left[E\{\dot{H}(\xi)|R=r\} \right]^{-1}.$$

from the expansion given by (4.3). It is easily seen that

$$G(\mathcal{Y}, R) = -\{\Delta - H(\xi)\}E\{[1 - H(\xi)]\Psi_c|R\} \left[E\{\dot{H}(\xi)|R\} \right]^{-1},$$

where we have suppressed the arguments of $\Psi_c(\cdot)$. Also, conditioning on \mathcal{Y} will show that

$$E\{S_v(\mathcal{Y}, R, \Theta_0, \gamma_0)a_\gamma^t(R)\} = E \left[\{1 - H(\xi)\}\Psi_c \frac{E\{L_2^t\dot{H}(\xi)|R=r\}}{E\{\dot{H}(\xi)|R=r\}} \right],$$

and

$$E\{S_\gamma(\mathcal{Y}, R, \Theta_0, \gamma_0)\} = -E \left[\{1 - H(\xi)\}\Psi_c L_2^t \right].$$

Defining

$$B_2 = E \left(\{1 - H(\xi)\}\Psi_c \left[L_2^t - \frac{E\{L_2^t\dot{H}(\xi)|R=r\}}{E\{\dot{H}(\xi)|R=r\}} \right] \right),$$

and calculating Ω from Theorem 1 will yield $\Omega = \Omega_1 + \Omega_2 + \Omega_3$, where $\Omega_1 = \Delta\{H(\xi)\}^{-1}\Psi_c$, $\Omega_2 = G(\mathcal{Y}, R)$, and

$$\Omega_3 = -\{\Delta - H(\xi)\}B_2B_1^{-1} \left[L_2 - \frac{E\{L_2\dot{H}(\xi)|R\}}{E\{\dot{H}(\xi)|R\}} \right].$$

A fairly lengthy covariance calculation which is sketched in Section 6.2 shows that for

$$\mathbf{C} = E\{S_\Theta(\mathcal{Y}, R, \Theta_0, \gamma_0)\},$$

$$n^{1/2}\{\Theta(\hat{\gamma}) - \Theta_0\} \xrightarrow{D} \text{Normal}\{0, \mathbf{C}^{-1}(\Sigma_1 - \Sigma_2 - \Sigma_3)\mathbf{C}^{-t}\}, \quad (4.4)$$

where $\Sigma_1 = E[\{H(\xi)\}^{-1}\Psi_c\Psi_c^t]$, $\Sigma_3 = B_2B_1^{-1}B_2^t$, and

$$\Sigma_2 = E \left[\frac{E[\{1 - H(\xi)\}\Psi_c|R]E[\{1 - H(\xi)\}\Psi_c^t|R]}{E\{\dot{H}(\xi)|R\}} \right].$$

Since Σ_2 is positive definite, the term $\mathbf{C}^{-1}\Sigma_2\mathbf{C}^{-t}$ actually represents a gain in asymptotic efficiency due to the data adjustment of $\pi(L)$. This phenomenon is quite common, and occurs in a related context in the theory of Wang et al. (1995). Likewise, the term $\mathbf{C}^{-1}\Sigma_3\mathbf{C}^{-t}$ represents the gain in efficiency due to data adjustment of γ_0 . Both gains diminish as $H(\cdot)$ tends to 1 and hence when X is observed on a larger proportion of the data.

5. DISCUSSION

The purpose of this paper was to introduce a class of semiparametric estimating functions which are general enough to be widely applicable, giving a simple and direct method of ensuring $n^{1/2}$ -consistent estimates of crucial parameters and a covariance form which is easily administered. The methods introduced were applied to three examples.

In estimating $\eta_0(\cdot)$ for a given $(\hat{\alpha}, \hat{\gamma})$, most kernel local averages and local linear regressions envision bandwidths of order $h \sim n^{-1/5}$. These bandwidths have been considered global in our calculations, but the same results apply for local bandwidths. When estimating Θ_0 using bandwidths of the usual order, when Condition 2 fails we have assumed Condition 1, namely that $nh^4 \rightarrow 0$, a contradiction. What is happening here is that while the variance of $n^{1/2}(\hat{\Theta} - \Theta_0)$ is of order $O(1)$, the bias is of order $O\{(nh^4)^{1/2}\}$. The natural question is what one should do if Condition 2 does not apply. We suggest here four possible approaches.

The first approach is to ignore the issue of bias. This is in fact a typical approach. In generalized additive models, it is known that backfitting has the same difficulty with bias (Hastie & Tibshirani, 1990, pp. 154–155), but this fact is often ignored.

There are three ad hoc methods to eliminate bias. The first is to multiply one's favorite bandwidth so that condition 1 is satisfied, i.e., use $hn^{-2/15}$. The second, suggested by Weisberg & Welsh (1994), is to use a standard bandwidth (local or global) but at the final estimate of $\eta_0(\cdot)$ use a third order kernel, i.e., one for which $\int K(v)dv = 1$, $\int v^3K(v)dv \neq 0$, and $\int v^jK(v)dv = 0$ for $j = 1, 2$. Finally, and somewhat similar in spirit to the previous suggestion, one can use local polynomial fits of order ≥ 2 applied with the candidate bandwidth.

Estimating the asymptotic covariance matrix of $\hat{\Theta}$ can be done in one of two ways. First, one can estimate each of the terms in (3.5) directly. To implement this, the form of (3.5) requires that one estimate additionally a number of conditional regressions, such as $E\{\dot{H}(\xi)|R\}$, which is easily accomplished.

An alternative covariance matrix estimate can be obtained by use of the so-called “ m out of n ” bootstrap studied by Politis & Romano (1994). Their remarkably general work shows that if a statistic is asymptotically normally distributed, then the m out of n bootstrap provides asymptotically valid standard error estimates and inferences.

ACKNOWLEDGMENTS

Gutierrez was supported by a Minority Merit Fellowship at Texas A&M University. Gutierrez and Carroll are supported by a grant from the National Cancer Institute (CA-57030). This research was completed in part while the first author was at the Australian Graduate School of Management, University of New South Wales. Carroll's research was partially completed while visiting the Institut für Statistik und Ökonometrie, Sonderforschungsbereich 373, Humboldt Universität zu Berlin, with partial support from a senior Alexander von Humboldt Foundation research award. The authors are grateful to M. P. Wand for his insightful suggestions.

REFERENCES

- Carroll, R. J., Fan, J., Gijbels, I. & Wand, M. P. (1995). Generalized partially linear single-index models. *Journal of the American Statistical Association*, under review.
- Gutierrez, R. G. (1995). Adaptive efficient estimation of regression coefficients when some regressors are missing. *Statistica Sinica*, under review.
- Härdle, W., Hall, P. & Ichimura, H. (1993). Optimal smoothing in single-index models. *Annals of Statistics*, **21**, 157–178.
- Hastie, T. J. & Tibshirani, R. J. (1990). *Generalized Additive Models*. London: Chapman & Hall.
- Horvitz, D. G. & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**, 663–685.
- Li, K. C. (1991). Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association*, **86**, 337–342.
- Little, R. J. A. & Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. New York: John Wiley.
- McCullagh, P. & Nelder, J. A. (1989). *Generalized Linear Models*. London: Chapman & Hall.
- Nadaraya, E. A. (1964). On estimating regression. *Theory Probab. Appl.*, **10**, 186–190.
- Newey, W. K. (1994). Series Estimation of Regression Functionals. *Econometric Theory*, **10**, 1–28.
- Politis, D. N. & Romano, J. P. (1994). Large sample confidence regions based on subsamples under minimal assumptions. *Annals of Statistics*, **22**, 2031–2050.
- Robins, J. M., Rotnitzky, A. & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, **89**, 846–866.
- Severini, T. A. & Staniswalis, J. G. (1994). Quasi-likelihood Estimation in Semiparametric Models. *Journal of the American Statistical Association*, **89**, 501–511.
- Severini, T. A. & Wong, W. H. (1992). Generalized Profile Likelihood and Conditionally Parametric Models. *Annals of Statistics*, **20**, 1768–1802.

Wang, C. Y., Wang, S., Zhao, L.-P. & Ou, S.-T. (1995). Weighted semiparametric estimation in regression analysis with missing covariate data. *Journal of the American Statistical Association*, to appear.

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

Weisberg, S. & Welsh, A. H. (1994). Adapting for the missing link. *Annals of Statistics*, **22**, 1674–1700.

6. PROOFS AND CALCULATIONS

6.1. Proof of Theorem 1

Definitions

- (i) \mathcal{U} is a fixed compact set interior to the range of U , and \mathcal{C} is the set generated as the product space of the set \mathcal{U} and the range of the remaining components of \mathcal{Y} .
- (ii) $\mathcal{R} = \{u^t \alpha_0 : u \in \mathcal{U}\}$.
- (iii) The notation $\langle k \rangle$ used as a superscript such as $A^{\langle k \rangle}$ denotes the k th row of A .
- (iv) $\Lambda_k(\mathcal{Y}, \Theta^*, v^*, \gamma^*)$ is the matrix of second order partial derivatives of $\Psi^{\langle k \rangle}(\mathcal{Y}, \Theta, v, \gamma)$ with respect to $(\Theta^t, v, \gamma^t)^t$ evaluated at $(\mathcal{Y}, \Theta^*, v^*, \gamma^*)$.

Assumptions

- (i) The asymptotic expansions in (3.2) hold, and the expansion in (3.1) holds uniformly over \mathcal{U} .
- (ii) $\hat{\Theta} - \Theta_0 = o_p(1)$ and $\|\hat{\Theta} - \Theta_0\|^2 = o_p(n^{-1/2})$.
- (iii) $\Lambda_k(\mathcal{Y}, \Theta, v, \gamma)$ is uniformly bounded in $\mathcal{Y} \in \mathcal{C}$ and v in a neighborhood of $\eta_0(U^t \alpha)$ for $U \in \mathcal{U}$ and (α, Θ, γ) in a neighborhood of $(\alpha_0, \Theta_0, \gamma_0)$.
- (iv) $E\{S_{\Theta}(\mathcal{Y}, R, \Theta_0, \gamma_0) S_{\Theta}^t(\mathcal{Y}, R, \Theta_0, \gamma_0)\}$ is positive definite.
- (v) $\Psi\{\mathcal{Y}, \Theta_0, \eta_0(R), \gamma_0\}$ has mean zero and positive definite covariance matrix.
- (vi) The random matrices $S_v(\mathcal{Y}, R, \Theta_0, \gamma_0) \zeta_0(U)$, $S_v(\mathcal{Y}, R, \Theta_0, \gamma_0) a_{\gamma}^t(U)$, $S_v(\mathcal{Y}, R, \Theta_0, \gamma_0) a_{\alpha}^t(U)$, and $S_{\gamma}(\mathcal{Y}, R, \Theta_0, \gamma_0)$ are such that for each of these matrices M , EMM^t is positive definite.
- (vii) The random variables V_{α} and V_{γ} each have mean 0 and positive definite covariance matrix.
- (viii) $E\{\rho_0(\mathcal{Y}, r) | R\} = 0$ for each $r \in \mathcal{R}$.

(ix) The function $G(\mathcal{Y}, r)$ has first two derivatives with respect to r which are uniformly bounded, for all $\mathcal{Y} \in \mathcal{C}$.

(x) K is a second order kernel function symmetric at 0 with compact support.

(xi) $f_R(r)$ is positive and continuous for all $r \in \mathcal{R}$.

Letting $\hat{\Theta}$ be a one-step Newton-Raphson solution to (2.3) and the presence of a $n^{1/2}$ -consistent starting value for $\hat{\Theta}$ will ensure that Assumption (ii) is satisfied.

Proof

By (2.3) and a Taylor expansion,

$$\begin{aligned}
0 &= n^{-1/2} \sum_{i=1}^n \Psi^{<k>} \{\mathcal{Y}_i, \Theta, \hat{\eta}(U_i^t \hat{\alpha}, \hat{\alpha}, \hat{\gamma}), \hat{\gamma}\} \\
&= n^{-1/2} \sum_{i=1}^n \Psi^{<k>} \{\mathcal{Y}_i, \Theta_0, \eta_0(R_i), \gamma_0\} + n^{-1/2} \sum_{i=1}^n S_{\Theta}^{<k>}(\mathcal{Y}_i, R_i, \Theta_0, \gamma_0)(\hat{\Theta} - \Theta_0) + \\
&\quad n^{-1/2} \sum_{i=1}^n S_v^{<k>}(\mathcal{Y}_i, R_i, \Theta_0, \gamma_0) \{\hat{\eta}(U_i^t \hat{\alpha}, \hat{\alpha}, \hat{\gamma}) - \eta_0(R_i)\} + n^{-1/2} \sum_{i=1}^n S_{\gamma}^{<k>}(\mathcal{Y}_i, R_i)(\hat{\gamma} - \gamma_0) + \\
&\quad n^{-1/2} \sum_{i=1}^n \frac{1}{2} \begin{bmatrix} \hat{\Theta} - \Theta_0 \\ \hat{\eta}(U_i^t \hat{\alpha}, \hat{\alpha}, \hat{\gamma}) - \eta_0(R_i) \\ \hat{\gamma} - \gamma_0 \end{bmatrix}^t \Lambda_k(\mathcal{Y}_i, \Theta^*, v^*, \gamma^*) \begin{bmatrix} \hat{\Theta} - \Theta_0 \\ \hat{\eta}(U_i^t \hat{\alpha}, \hat{\alpha}, \hat{\gamma}) - \eta_0(R_i) \\ \hat{\gamma} - \gamma_0 \end{bmatrix}, \tag{6.1}
\end{aligned}$$

for Θ^* in between $\hat{\Theta}$ and Θ_0 , v^* in between $\hat{\eta}(U_i^t \hat{\alpha}, \hat{\alpha}, \hat{\gamma})$ and $\eta_0(R_i)$, and γ^* in between $\hat{\gamma}$ and γ_0 . Given assumptions (i) – (iii) it is easily seen that the last term in (6.1) is $o_p(1)$ for $k = 1, \dots, p$. Combining all the components of $\Psi(\cdot)$ we have that

$$\begin{aligned}
0 &= n^{-1/2} \sum_{i=1}^n \Psi\{\mathcal{Y}_i, \Theta_0, \eta_0(R_i), \gamma_0\} + n^{-1/2} \sum_{i=1}^n S_v(\mathcal{Y}_i, R_i, \Theta_0, \gamma_0) \{\hat{\eta}(U_i^t \hat{\alpha}, \hat{\alpha}, \hat{\gamma}) - \eta_0(R_i)\} + \\
&\quad n^{-1/2} \sum_{i=1}^n S_{\Theta}(\mathcal{Y}_i, R_i, \Theta_0, \gamma_0)(\hat{\Theta} - \Theta_0) + n^{-1/2} \sum_{i=1}^n S_{\gamma}(\mathcal{Y}_i, R_i, \Theta_0, \gamma_0)(\hat{\gamma} - \gamma_0) + o_p(1).
\end{aligned}$$

By Assumption (iv) and the law of large numbers, $n^{-1} \sum_{i=1}^n S_{\Theta}(\mathcal{Y}_i, R_i, \Theta_0, \gamma_0) = \mathbf{C} + O_p(n^{-1/2})$.

Making this substitution and applying Assumption (ii) we find that

$$\begin{aligned}
-\mathbf{C}n^{1/2}(\hat{\Theta} - \Theta_0) &= n^{-1/2} \sum_{i=1}^n \Psi\{\mathcal{Y}_i, \Theta_0, \eta_0(R_i), \gamma_0\} + n^{-1/2} \sum_{i=1}^n S_{\gamma}(\mathcal{Y}_i, R_i, \Theta_0, \gamma_0)(\hat{\gamma} - \gamma_0) + \\
&\quad n^{-1/2} \sum_{i=1}^n S_v(\mathcal{Y}_i, R_i, \Theta_0, \gamma_0) \{\hat{\eta}(U_i^t \hat{\alpha}, \hat{\alpha}, \hat{\gamma}) - \eta_0(R_i)\} + o_p(1).
\end{aligned}$$

We now make the substitution given by the uniform expansion in (3.1). Noting that for both Condition 1 and Condition 2 $nh^6 \rightarrow 0$, it follows that

$$\begin{aligned} -\mathbf{C}n^{1/2}(\widehat{\Theta} - \Theta_0) &= n^{-1/2} \sum_{i=1}^n \Psi\{\mathcal{Y}_i, \Theta_0, \eta_0(R_i), \gamma_0\} + n^{-1/2} \sum_{i=1}^n S_\gamma(\mathcal{Y}_i, R_i, \Theta_0, \gamma_0)(\widehat{\gamma} - \gamma_0) \\ &\quad n^{-1/2} \sum_{i=1}^n S_v(\mathcal{Y}_i, R_i, \Theta_0, \gamma_0) \left\{ h^2 \zeta_0(U_i) + a_\gamma^t(U_i)(\widehat{\gamma} - \gamma_0) + \right. \\ &\quad \left. a_\alpha^t(U_i)(\widehat{\alpha} - \alpha_0) + n^{-1} \sum_{j=1}^n K_h(R_j - R_i) \rho_0(\mathcal{Y}_j, R_i) \right\} + o_p(1). \end{aligned}$$

Next we make the substitutions given by the expansions in (3.2) and find that

$$\begin{aligned} -\mathbf{C}n^{1/2}(\widehat{\Theta} - \Theta_0) &= n^{-1/2} \sum_{i=1}^n \Psi\{\mathcal{Y}_i, \Theta_0, \eta_0(R_i), \gamma_0\} + n^{-1/2} h^2 \sum_{i=1}^n S_v(\mathcal{Y}_i, R_i, \Theta_0, \gamma_0) \zeta_0(U_i) + \\ &\quad n^{-1/2} h^2 \sum_{i=1}^n S_v(\mathcal{Y}_i, R_i, \Theta_0, \gamma_0) a_\gamma^t(U_i) b_\gamma + n^{-1/2} h^2 \sum_{i=1}^n S_v(\mathcal{Y}_i, R_i, \Theta_0, \gamma_0) a_\alpha^t(U_i) b_\alpha + \\ &\quad n^{-1/2} h^2 \sum_{i=1}^n S_\gamma(\mathcal{Y}_i, R_i, \Theta_0, \gamma_0) b_\gamma + n^{-3/2} \sum_{i=1}^n \sum_{j=1}^n S_v(\mathcal{Y}_i, R_i, \Theta_0, \gamma_0) a_\gamma^t(U_i) V_{j\gamma} + \\ &\quad n^{-3/2} \sum_{i=1}^n \sum_{j=1}^n S_v(\mathcal{Y}_i, R_i, \Theta_0, \gamma_0) a_\alpha^t(U_i) V_{j\alpha} + n^{-3/2} \sum_{i=1}^n \sum_{j=1}^n S_\gamma(\mathcal{Y}_i, R_i, \Theta_0, \gamma_0) V_{j\gamma} + \\ &\quad n^{-3/2} \sum_{i=1}^n \sum_{j=1}^n S_v(\mathcal{Y}_i, R_i, \Theta_0, \gamma_0) \rho_0(\mathcal{Y}_j, R_i) K_h(R_j - R_i) + o_p(1) \\ &\equiv T_1 + T_2 + T_3 + T_4 + T_5 + T_6 + T_7 + T_8 + T_9 + o_p(1). \end{aligned}$$

The second term $T_2 = n^{1/2} h^2 \{n^{-1} \sum_{i=1}^n S_v(\mathcal{Y}_i, R_i, \Theta_0, \gamma_0) \zeta_0(U_i)\}$. If Condition 1 holds then $nh^4 \rightarrow 0$. If Condition 2 holds then $n^{-1} \sum_{i=1}^n S_v(\mathcal{Y}_i, R_i, \Theta_0, \gamma_0) \zeta_0(U_i) = O_p(n^{-1/2})$ given Assumption (vi). In either case, $T_2 = o_p(1)$. Similarly, it can be shown that $T_3 = o_p(1)$, $T_4 = o_p(1)$, and $T_5 = o_p(1)$.

The sixth term

$$\begin{aligned} T_6 &= n^{-3/2} \sum_{i=1}^n \sum_{j=1}^n S_v(\mathcal{Y}_i, R_i, \Theta_0, \gamma_0) a_\gamma^t(U_i) V_{j\gamma} = \left\{ n^{-1} \sum_{i=1}^n S_v(\mathcal{Y}_i, R_i, \Theta_0, \gamma_0) a_\gamma^t(U_i) \right\} n^{-1/2} \sum_{j=1}^n V_{j\gamma} \\ &= n^{-1/2} \sum_{j=1}^n E\{S_v(\mathcal{Y}, R, \Theta_0, \gamma_0) a_\gamma^t(U)\} V_{j\gamma} + o_p(1), \end{aligned}$$

since $n^{-1/2} \sum_{j=1}^n V_{j\gamma} = O_p(1)$. Similarly, it can be shown that

$$\begin{aligned} T_7 &= n^{-1/2} \sum_{j=1}^n E\{S_v(\mathcal{Y}, R, \Theta_0, \gamma_0) a_\alpha^t(U)\} V_{j\alpha} + o_p(1); \\ T_8 &= n^{-1/2} \sum_{j=1}^n E\{S_\gamma(\mathcal{Y}, R, \Theta_0, \gamma_0)\} V_{j\gamma} + o_p(1). \end{aligned}$$

This leaves T_9 . We show $T_9 = n^{-1/2} \sum_{i=1}^n G(\mathcal{Y}_i, R_i) + o_p(1)$ by showing that the first two moments of $T_9^* \equiv T_9 - n^{-1/2} \sum_{i=1}^n G(\mathcal{Y}_i, R_i)$ are $o(1)$. Details are available from the first author.

6.2. Calculations leading to (4.4)

We start by suppressing the dependence of H and \dot{H} on ξ . We must show that $E\Omega\Omega^t = \Sigma_1 - \Sigma_2 - \Sigma_3$. Since $\Omega = \Omega_1 + \Omega_2 + \Omega_3$, we have nine terms to consider. The first, $E\Omega_1\Omega_1^t = E\{(\Delta/H^2)\Psi_c\Psi_c^t\} = E[E\{(\Delta/H^2)\Psi_c\Psi_c^t\}|\mathcal{Y}] = E(H^{-1}\Psi_c\Psi_c^t) = \Sigma_1$. We next note that $-\Sigma_2 = E\Omega_1\Omega_2^t + E\Omega_2\Omega_1^t + E\Omega_2\Omega_2^t$, since it is easily seen that $E\Omega_1\Omega_2^t = -\Sigma_2 = E\Omega_2\Omega_1^t$. Further, direct calculations show that $E\Omega_2\Omega_2^t = \Sigma_2$ and it follows that $E\Omega_1\Omega_2^t + E\Omega_2\Omega_1^t + E\Omega_2\Omega_2^t = -2\Sigma_2 + \Sigma_2 = -\Sigma_2$. Studying the other terms in $E\Omega\Omega^t$ we see that $E\Omega_2\Omega_3^t = E\Omega_3\Omega_2^t = 0$ and $E\Omega_1\Omega_3^t = -\Sigma_3$. Finally, $E\Omega_3\Omega_3^t = \Sigma_3$. Combining all terms it then follows that $E\Omega\Omega^t = \Sigma_1 - \Sigma_2 - \Sigma_3$. Complete details are available from the first author.