SCHOOL OF ECONOMICS AND MANAGEMENT
FACULTY OF SOCIAL SCIENCES
AARHUS UNIVERSITY

CREATES
Center for Research in Econometric Analysis of Time Series

CREATES Research Paper 2009-44

Semiparametric Modelling and Estimation: A Selective Overview

Dennis Kristensen

School of Economics and Management
Aarhus University
Bartholins Allé 10, Building 1322, DK-8000 Aarhus C
Denmark

Semiparametric Modelling and Estimation: A Selective Overview*

Dennis Kristensen†
Columbia University and CREATES‡

September, 2009

Abstract

Semiparametric models are characterized by a finite- and infinite-dimensional (functional) component. As such they allow for added flexibility over fully parametric models, and at the same time estimators of parametric components can be developed that exhibit standard parametric convergence rates. These two features have made semiparametric models and estimators increasingly popular in applied economics. We give a partial overview over the literature on semiparametric modelling and estimation with particular emphasis on semiparametric regression models. The main focus is on developing two-step semiparametric estimators and deriving their asymptotic properties. We do however also briefly discuss sieve-based estimators and semiparametric efficiency.

Keywords: efficiency, kernel estimation, regression, semiparametric, sieve, two-step estimation.

JEL-Classification: C13, C14, C51.

*Prepared for *Quantile* 7. The author would like to thank Stanislav Anatolyev, Bruno Giovannetti and Shin Kanaya for helpful comments and suggestions.

†E-mail: dk2313@columbia.edu. Mailing address: Economics Department, Columbia University, 1018 International Affairs Building, 420 West 118th Street, New York, NY 10027, USA.

‡Center for Research in Econometric Analysis of TimE Series, University of Aarhus, is funded by the Danish National Science Foundation.

1 Introduction

Semiparametric modelling and estimation of economic processes have received a lot of attention over the past three decades. The main reason for the popularity of this approach is that it works as a compromise between two extremes, fully parametric and fully nonparametric modelling. In the former case, a fully parameterized model is used to explain data and a natural estimator is the maximum-likelihood estimator (MLE). If correctly specified, the MLE enjoys the usual good properties such as maximum efficiency. But if some parts of the model are misspecified, the MLE will suffer from asymptotic biases and conclusions drawn from the estimated model may be severely misleading. Situated at the other end of the spectrum, fully nonparametric models allow for maximum flexibility and therefore carry no risk of misspecification. On the other hand, nonparametric estimators require a lot of data, and tend to be rather imprecise in small samples; this is in particular the case in large-dimensional models where the precision of nonparametric estimators tend to deteriorate as more conditioning variables are included; this is normally referred to as the "curse of dimensionality."

Semiparametric models are situated between the nonparametric and parametric extremes in the sense that they contain both a nonparametric and parametric component. Thus, semiparametric models maintain, to some extent, the flexibility of the fully nonparametric model, and as such better safeguard against misspecification compared to a fully parametric model. At the same time, parametric components of the semiparametric model can in general be estimated with a precision comparable to what we would obtain by using a (correctly specified) fully parametric model.

We will here try to give a brief introduction to and overview of semiparametric modelling and estimation with special focus on regression models. We here introduce the main concepts in semiparametric modelling and estimation within the framework of regression models for two reasons: Firstly, these models are widely used in economics and as such should be familiar to the average reader. Second, regression models are fairly simple to work with, thereby allowing for a relatively straightforward introduction of the major semiparametric conventions and techniques. Secondly, many of the techniques that we will introduce in the regression framework can be carried over to many other settings, so the interested reader should be able to apply these tools to other types of models. To illustrate the last point, we will briefly touch on semiparametric copulas and demonstrate how the same ideas introduced in a regression framework can be utilized in this setting.

After having introduced estimators of some leading semiparametric regression models, we set up a general framework within which we can analyze the asymptotic properties of these. The general class of estimators that we consider are so-called two-step semiparametric estimators, where in the first step a nonparametric component of the model is estimated, which in turn is used to estimate the parametric part. We derive a set of high-level conditions under

which the semiparametric estimator is consistent and asymptotically normally distributed, and discuss in further details how these conditions can be verified for specific models.

As an alternative estimation strategy, we give a brief introduction to a class of semiparametric estimators based on so-called sieve methods. We will however not cover the underlying theory of such estimators in any detail. Finally, we devote some time to discuss the issue of semiparametric efficiency, and its uses in developing estimators. Again, this part of the paper is non-technical and we only try to convey the intuition behind the various concepts in this part of the literature.

This survey has no ambition of being exhaustive, and it should be noted that many other, excellent reviews of the literature on semiparametric modelling and estimation are available. These include, amongst others, Ichimura and Todd (2007), Härdle et al (2004), Horowitz (2009), Li and Racine (2007), Pagan and Ullah (1999), Powell (1994), and Robinson (1988) which complement and extend our survey in a number of directions.

The remains of the paper are organized as follows: In Sections 2-4, we start by giving a number of examples of semiparametric models, and discuss the estimation of these. In Section 5, we analyze the properties of a fairly general class of semiparametric two-step estimators that include the specific estimators presented in the previous sections. We focus on estimators based on kernel smoothing since these are relatively easy to analyze, and are popular in applied work. In Section 6, we briefly introduce simultaneous estimation of both components using so-called sieve-methods to handle the nonparametric component, while semiparametric efficiency is discussed in Section 7. We conclude in Section 8 by pointing to more detailed works on the different topics covered in the survey. All proofs have been relegated to Appendix A.

While Sections 2-4 and 6-7 can be read without any strong knowledge of econometric theory, Section 5 may be somewhat more challenging for the less technical-minded reader. In order to keep the technicalities at a reasonable level, some mathematical arguments are only sketched. Furthermore, some relevant papers containing more precise results and rigorous proofs are listed in Section 8.

2 Semiparametric Regression

In its most general form, a regression model can be formulated as

$$Y = m(X) + \varepsilon, \quad E[\varepsilon|X] = 0, \tag{1}$$

where $Y \in \mathbb{R}$ is the response (or dependent) variable, $X \in \mathbb{R}^d$ is a set of $d \geq 1$ regressors (or independent variables), and $\varepsilon \in \mathbb{R}$ is the error term. The regression function $m : \mathbb{R}^d \mapsto \mathbb{R}$ explains how the conditional mean of Y changes with X:

$$E[Y|X = x] = m(x).$$

Also, let $f_{\varepsilon|X}(e|x)$ denote the conditional density of ε given $X = x$.[1] Suppose we have observed a random sample, (Y_i, X_i) for $i = 1, ..., n$, from the model. We are then interested in drawing inference regarding the functions $m(x)$ and $f_{\varepsilon|X}(e|x)$.

In the fully parametric case, both the regression function, m, and the (conditional) error distribution, $f_{\varepsilon|X}$, are assumed to be known up to some finite-dimensional parameter. That is, we have specified parametric functions $m(x;\beta)$ and $f_{\varepsilon|X}(e|x;\sigma)$ where $\beta \in \mathcal{B}$ contains the regression coefficients characterizing the shape of m, while $\sigma \in \Sigma$ is a parameter capturing the shape of the (conditional) error distribution. Assuming that the model is correctly specified, that is, $m(x) = m(x;\beta_0)$ and $f_{\varepsilon|X}(e|x) = f_{\varepsilon|X}(e|x;\sigma_0)$ for some $\theta_0 = (\beta_0, \sigma_0)$, a natural estimator of the model would be the MLE,

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta\in\Theta} \sum_{i=1}^{n} \log f_{\varepsilon|X}(Y_i - m(X_i;\beta)|X_i;\sigma).$$

A popular specification is the Gaussian regression model: The error term is assumed to be independent of X and normally distributed $N(0,\sigma^2)$. In this case, the MLE's of $\theta = (\beta, \sigma^2)$ are the least-squares estimators: $\hat{\beta}_{\text{MLE}} = \hat{\beta}_{\text{LS}}$ and $\hat{\sigma}^2_{\text{MLE}} = \hat{\sigma}^2_{\text{LS}}$ where

$$\hat{\beta}_{\text{LS}} = \arg\min_{\beta\in\mathcal{B}} \sum_{i=1}^{n} (Y_i - m(X_i;\beta))^2, \quad \hat{\sigma}^2_{\text{LS}} = \frac{1}{n}\sum_{i=1}^{n} \left(Y_i - m\left(X_i;\hat{\beta}\right)\right)^2.$$

Regarding the specification of the regression function, a linear regression function is widely used, $m(x;\beta) = \beta_1 x_1 + ... + \beta_d x_d$, and the MLE collapses to the ordinary least-squares estimator,

$$\hat{\beta}_{\text{OLS}} = \left(\sum_{i=1}^{n} X_i X_i'\right)^{-1} \left(\sum_{i=1}^{n} X_i Y_i\right). \tag{2}$$

Under regularity conditions, the estimator $\hat{\beta}_{\text{MLE}}$ is $\sqrt{n}$-consistent and asymptotically normally distributed. For example, with Gaussian errors, the MLE satisfies

$$\sqrt{n}(\hat{\beta}_{\text{MLE}} - \beta_0) \to^d N(0, V), \quad V = \sigma^2 E\left[\dot{m}(x;\beta)\dot{m}(x;\beta)'\right]^{-1},$$

where $\dot{m}(x;\beta) = \partial m(x;\beta)/(\partial\beta)$ (see, for example, Amemiya, 1985). This in turn implies that the regression function can be estimated by $\hat{m}_{\text{MLE}}(x) = m(x;\hat{\beta}_{\text{MLE}})$.

However, the parametric model may be misspecified meaning that $m(x;\beta) \neq m(x)$ for all $\beta \in \mathcal{B}$ and/or $f_{\varepsilon|X}(e|x;\sigma) \neq f_{\varepsilon|X}(e|x)$ for all values of $\sigma \in \Sigma$. In this case, the estimated regression function $\hat{m}_{\text{MLE}}(x)$ is in general inconsistent and will give a misleading picture of how X impacts Y. To remove the risk of misspecification, one can instead use fully nonparametric estimators of m such as kernel estimators or series/sieve estimators. We will

[1] We here assume, as is standard in the non- and semiparametric literature, that all variables in question have continuous distributions.

here focus on kernel estimators and give a brief overview of these; we refer the reader to Härdle (1992) and Silverman (1986) for further details. Sieve estimators are briefly discussed in Section 6. Kernel estimators form a particular class of nonparametric estimators which use local information in data to draw inference about characteristics of the distribution. Suppose that X has a continuous distribution described by a density $f(x)$. This density can then be estimated nonparametrically by a kernel density estimator: For any given value $x \in \mathbb{R}^d$, this is computed as

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(X_i - x), \tag{3}$$

where $K_h(x) = K(x/h)/h^d$, $K : \mathbb{R}^d \mapsto \mathbb{R}$ is a so-called kernel function and $h > 0$ is a so-called bandwidth. Both K and h are chosen by the researcher. The kernel density estimator is similar to the histogram estimator of a distribution where the bandwidth determines the width of each cell in the histogram and the kernel how much weight individual observations within a cell should be given. Most weight is given to observations close to x while observations far from x play little, if no role. Similarly, the kernel regression estimator of $m(x) = E[Y|X = x]$ at a given value of $x \in \mathbb{R}^d$ takes the form of a weighted sample average,

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} Y_i K_h(X_i - x)}{\sum_{i=1}^{n} K_h(X_i - x)}. \tag{4}$$

Again, this is a local estimator that uses those observations, X_i, that are close to x to extract information regarding the shape of $m(\cdot)$ at x.

Kernel regression estimators are very robust: The estimator $\hat{m}(x)$ is consistent as $h \to 0$ and $nh^d \to \infty$, no matter what the shape of the true regression function m is. But one pays a price in terms of precision with the kernel estimator exhibiting more finite-sample variation compared to parametric estimators. On a theoretical level, this shows up in the fact that the optimal rate of kernel estimators are $\sqrt{n^{4/(4+d)}}$, which is slower than the $\sqrt{n}$-rate of parametric estimators.[2] We note that the precision of the nonparametric estimator is influenced by the dimension of X, $d \geq 1$: As d increases the convergence rate of the nonparametric estimator deteriorates (this is the aformentioned "curse of dimensionality"). In addition to these issues, even if precise nonparametric estimates can be obtained, it can be difficult to present and interpret the kernel estimators $\hat{f}(x)$ and $\hat{m}(x)$ when d is large.

Thus, when choosing between different modelling and estimation techniques, we face a trade-off between risk of misspecification and degree of precision of estimators. The MLE of a fully parametric model has maximum precision but a very high risk of suffering from misspecification biases. In contrast, the fully nonparametric estimator has no risk of misspecification but on the other hand can have very low precision. This motivates the use of semiparametric models and estimators. These still allows for a relatively high degree of

[2] We have here assumed that the function of interest is twice continuously differentiable.

flexibility of the model while improving on the convergence rate for the certain components of the model. We now present two semiparametric regression models as illustrations.

2.1 The Single Index Model

A popular semiparametric regression model is the so-called single-index model which takes the following form:

$$Y = g\left(\beta' X\right) + \varepsilon, \quad E\left[\varepsilon | X\right] = 0, \tag{5}$$

where the function $g : \mathbb{R} \mapsto \mathbb{R}$ and the parameter $\beta \in \mathbb{R}^d$ are unknown. We make no assumptions regarding the (conditional) distribution of ε, and treat g and $f_{\varepsilon|X}$ as nonparametric objects. Thus, in this case, our infinite-dimensional parameter is $\gamma = \left(g, f_{\varepsilon|X}\right)$ while the parametric component is β.

The name single-index comes from the fact that g here is a function of the *index* $\beta' X \in \mathbb{R}$ instead of the full vector $X \in \mathbb{R}^d$. Thus, we assume that X only influences Y through the index $\beta' X$ which is a restriction relative to the fully general regression function m given in eq. (1). Thus, in contrast to the fully nonparametric setting, we now face a risk of misspecification.

On the other hand, the model has a nice interpretation with the impact of X on Y described by the finite-dimensional parameter β and the univariate function g. In this regard, observe that g here has $\mathbb{R}$ as its domain in contrast to the function m appearing in (1) which has domain $\mathbb{R}^d$. Thus, regardless of the dimension of X, the estimation of g remains a univariate problem, and as such the curse of dimensionality has been removed.

The above framework accommodates for certain types of transformation models. Suppose that the random variable Y^* satisfies

$$Y^* = \beta_0' X + \eta,$$

where $\eta \sim F_\eta$ is independent of X. We do not observe Y^* however, but only

$$Y = t\left(Y^*\right),$$

for some transformation t which may be known or unknown. We see that

$$E\left[Y | X = x\right] = E\left[t\left(\beta_0' X + \eta\right) | X = x\right] = \int t\left(\beta_0' x + v\right) dF_\eta\left(v\right).$$

By defining g and ε as

$$g\left(z\right) := \int t\left(z + v\right) dF_\eta\left(v\right), \quad \varepsilon := Y - E\left[Y | X = x\right],$$

the class of transformation models can be written on the form of eq. (5). The transformation models include limited dependent variable models such as censored regression models and

duration models. For example, with $t(y) = 1\{y > 0\}$ the transformation model is a binary choice model such as the probit and logit models. If the transformation is $t(y) = y \cdot 1\{y > 0\}$, we obtain a Tobit model. The advantage of the semiparametric approach is that we can still estimate β, without having to take a stand on the precise form of t and F_η.

We now wish to develop an estimator of the single-index model. To this end, we first need to discuss identification of the parameters of interest. That is, can we uniquely identify the parameters β and the functions g and $f_{\varepsilon|X}$ given data? We first note that the parameter β cannot be identified if $P(\delta' X = c) = 1$ for some constants $c \in \mathbb{R}$ and $\delta \in \mathbb{R}^d$. Furthermore, we need to normalise β to be able to identify g. To see why this is the case, define $\tilde{g}(z) = g(a + bz)$ for any constants $(a, b) \in \mathbb{R}^2$, which is equivalent to $\tilde{g}(-a + 1/bz) = g(z)$. It then holds that the two specifications are observationally equivalent:

$$g\left(\beta' x\right) = \tilde{g}\left(-a + (1/b)\beta' x\right) = \tilde{g}(\tilde{\beta}' x),$$

where $\tilde{\beta} = (1/b)\beta$. That is, given data, we will not be able to distinguish between $\tilde{g}$ and g. We therefore will require that X does not to contain any constants, and we also set one of the coefficients β equal to one; we choose $\beta_1 = 1$ (one can always rearrange the order of the components of X). Finally, we note that if g is linear then we cannot identify β (unless we assume that g is known).

Under the identifying restrictions, we are able to develop estimators of β and g: Suppose first that the function g is known; then a natural estimator of $\hat{\beta}$ would be the least-squares estimator,

$$\hat{\beta}_g = \arg\min_{\beta \in \mathcal{B}} \frac{1}{n} \sum_{i=1}^n \left[Y_i - g\left(\beta' X_i\right)\right]^2. \tag{6}$$

Conversely, suppose that $\beta \in \mathbb{R}^d$ was known; then a natural nonparametric estimator of g would be the standard kernel regression estimator,

$$\hat{g}(z; \beta) = \frac{\sum_{i=1}^n Y_i K_h(\beta' X_i - z)}{\sum_{i=1}^n K_h(\beta' X_i - z)}.$$

However, since both β and g are unknown, neither of these are feasible estimators. Instead, we propose to combine them as follows: By substituting the nonparametric estimator $\hat{g}(z; \beta)$ into the least-squares criterion in eq. (6), a feasible estimator of β is obtained as:

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^n \left[Y_i - \hat{g}\left(\beta' X_i; \beta\right)\right]^2.$$

Once $\hat{\beta}$ has been obtained, the obvious estimator of $g(z)$ is $\hat{g}(z; \hat{\beta})$.

An alternative strategy is the *average-derivative* estimation method as proposed in Powell et al (1989). Assuming g is differentiable, the following identity holds:

$$\frac{\partial E[Y|X = x]}{\partial x} = \beta g'\left(\beta' x\right),$$

where $g'(x) = \frac{\partial g(z)}{\partial z}$. Thus, for any bounded function w,

$$E\left[\frac{\partial E[Y|X]}{\partial x} w(X)\right] = \beta E\left[w(X) g'\left(\theta' X\right)\right].$$

This shows that the parameter δ defined as

$$\delta := E\left[\frac{\partial E[Y|X]}{\partial x} w(X)\right]$$

is observationally equivalent to β up to a scale normalization ($E[w(X) g'(\theta' X)]$). We now develop an estimator of δ with the weight function w chosen as $w(x) = f(x)$, where f denotes the density of X: First, observe

$$\begin{aligned} E\left[\frac{\partial E[Y|X]}{\partial x} f(X)\right] &= \int_{\mathbb{R}^d} \frac{\partial E[Y|X=x]}{\partial x} f^2(x)\, dx \\ &= -2\int_{\mathbb{R}^d} E[Y|X=x] f(x) \frac{\partial f(x)}{\partial x} dx \\ &= -2E\left[E[Y|X] \frac{\partial f(X)}{\partial x}\right] \\ &= -2E\left[Y \frac{\partial f(X)}{\partial x}\right]. \end{aligned}$$

The last expression on the right hand side will form the basis for our estimator of δ: Replacing population expectations with sample expectations and the density, f, with its kernel estimator, $\hat{f}$, as given in eq. (3), we obtain:

$$\hat{\delta} = \frac{1}{n}\sum_{i=1}^{n} Y_i \frac{\partial \hat{f}(X_i)}{\partial x}.$$

An advantage of $\hat{\delta}$ over $\hat{\beta}$ is that the former is on closed-form and requires no numerical optimization.

One can extend the single index model to the following more general class of models,

$$Y = g(v(X;\beta_0)) + \varepsilon, \quad E[\varepsilon|X] = 0,$$

for some function $v : \mathbb{R} \times \mathcal{B} \mapsto \mathbb{R}$ which is known up to β_0. The estimation strategy outlined above carries through to this more general setting.

2.2 The Partially Linear Model

An alternative specification is obtained by assuming that m in (1) is linear in some of its arguments. Suppose $X = (X_1, X_2)$ where $X_i \in \mathbb{R}^{d_i}$, $i = 1, 2$, and $d = d_1 + d_2$, such that

$$Y = \beta_0' X_1 + g(X_2) + \varepsilon, \quad E[\varepsilon|X] = 0, \tag{7}$$

for some $g : \mathbb{R}^{d_2} \mapsto \mathbb{R}$ and $\beta \in \mathbb{R}^{d_1}$. As before, we leave the distribution of $\varepsilon|X$ unspecified.

Compared to the fully general regression model in eq. (1), the following restriction has been imposed on the shape of the regression function, $m(x) = \beta_0' x_1 + g(x_2)$. That is, Y is additive in X_1 and X_2, and X_1 impacts Y in a linear fashion. Our model consists of a parametric component, β_0, and two nonparametric components, g and $f_{\varepsilon|X}$, and as such it is semiparametric.

Again, we need to impose restrictions on the model for g and β to be identified. We cannot allow any of the components of X to be constant since with $\tilde{g}(x_2) = g(x_2) - a$, $a \in \mathbb{R}$, we cannot distinguish between $\beta' x_1 + g(x_2)$ and $\{a + \beta' x_1\} + \tilde{g}(x_2)$. In fact, we have to assume that

$$\Omega = E\left[(X_1 - E[X_1|X_2])(X_1 - E[X_1|X_2])'\right]$$

is nonsingular. If this does not hold, we cannot distinguish between the linear and the nonlinear term. To see this, observe that

$$E[Y|X_2] = \beta_0' E[X_1|X_2] + g(X_2) + E[\varepsilon|X_2],$$

implying

$$Y - E[Y|X_2] = \beta_0'(X_1 - E[X_1|X_2]) + \eta, \tag{8}$$

where $\eta = \varepsilon - E[\varepsilon|X_2]$ satisfies $E[\eta|X] = 0$. So in order to identify β_0, we need Ω to be nonsingular.

The equation (8) forms the basis of the following "residual-based" estimator: Construct kernel estimators of $m_Y(x_2) = E[Y|X_2 = x_2]$ and $m_{X_1}(x_2) = E[X_1|X_2 = x_2]$,

$$\hat{m}_Y(x_2) = \frac{\sum_{i=1}^n Y_i K_h(X_{2,i} - x_2)}{\sum_{i=1}^n K_h(X_{2,i} - x_2)}, \quad \hat{m}_{X_1}(x_2) = \frac{\sum_{i=1}^n X_{1,i} K_h(X_{2,i} - x_2)}{\sum_{i=1}^n K_h(X_{2,i} - x_2)},$$

and substitute these into (8). We can then estimate β by OLS,

$$\hat{\beta} = \left(\sum_{i=1}^n \hat{Z}_i \hat{Z}_i'\right)^{-1} \sum_{i=1}^n \hat{Z}_i \left(Y_i - \hat{m}_Y(X_{2,i})\right)'. \tag{9}$$

where $\hat{Z}_i = X_{1,i} - \hat{m}_{\mathrm{NP},X_1}(X_{2,i})$.

The estimation method can be extended to the following, more general model,

$$Y = v(X_1; \beta) + g(X_2) + \varepsilon, \quad E[\varepsilon|X] = 0, \tag{10}$$

where $v : \mathbb{R}^{d_1} \times \mathcal{B} \mapsto \mathbb{R}$ is known up to $\theta \in \Theta$. The resulting estimator is however no longer on closed form, and numerical optimization techniques now have to be employed.

3 Specification of Error Distribution

So far, we have only discussed how the functional form of m in the general regression model can be modelled and estimated using semiparametric techniques. In this section, we focus on the error term, ε, and discuss how different assumptions regarding the error terms lead to different (semiparametric) estimation strategies for the regression function. In some situations, one can derive an estimator of the parameter of interest without having to estimate infinite-dimensional objects. These estimators however tend to be inefficient though, and semiparametric estimation techniques can be used to improve on the efficiency.

3.1 The Linear Regression Model

Consider the standard linear regression model:

$$Y = \beta' X + \varepsilon, \tag{11}$$

where $E[\varepsilon|X] = 0$. This is normally seen as a fully parametric model, but in our terminology this is a semiparametric model if the distribution of $\varepsilon|X$, $f_{\varepsilon|X}$, is not fully specified. If $f_{\varepsilon|X}$ has not been specified, we have a parametric component, θ, and a nonparametric one, $f_{\varepsilon|X}$.

If we assume that the error term follows a normal distribution, we saw in the previous section that the MLE takes the form of the standard OLS estimator as given in eq. (2). However, the OLS estimator can also be interpreted as a semiparametric estimator of θ since it remains $\sqrt{n}$-consistent regardless of the precise specification of $f_{\varepsilon|X}$. Moreover, an attractive feature of OLS is that we do not need to estimate $f_{\varepsilon|X}$ in order to compute it. This is in contrast to the semiparametric estimators considered in the previous section, where we had to obtain a preliminary estimator of a nonparametric component in order to estimate the parametric one.

However, one may wonder whether other, better estimators are available? Obviously, if we impose a (correct) parametric structure on $f_{\varepsilon|X}$, we can compute the MLE which in general is more efficient than OLS. But even without imposing a parametric form on the distribution, we shall in the following see that OLS is in general not efficient within the class of semiparametric estimators.

3.2 Heteroskedasticity of Unknown Form

We maintain the linear model in eq. (11), but now assume that the errors are heteroskedastic,

$$E\left[\varepsilon^2|X = x\right] = \sigma^2(x), \tag{12}$$

with the form of the conditional variance function, $\sigma^2(\cdot)$, being *unknown*.

The standard OLS estimator given in Eq. (2) is still consistent and asymptotically normally distributed but now the asymptotic distribution is:

$$\sqrt{n}(\hat{\theta}_{\text{OLS}} - \theta) \to^d N\left(0, E\left[XX'\right]^{-1} E\left[\sigma^2(X) XX'\right]^{-1} E\left[XX'\right]^{-1}\right).$$

In particular, it is no longer efficient as we shall now see: Consider first the case where the conditional variance function $\sigma^2(x)$ is *known*. Then we can do weighted least squares (WLS),

$$\tilde{\theta}_{\text{WLS}} = \left(\sum_{i=1}^{n} \sigma^{-2}(X_i) X_i X_i'\right)^{-1} \left(\sum_{i=1}^{n} \sigma^{-2}(X_i) X_i Y_i\right), \tag{13}$$

which improves on the asymptotic variance relative to the OLS estimator:

$$\sqrt{n}(\tilde{\theta}_{\text{WLS}} - \theta) \to^d N\left(0, E\left[\sigma^{-2}(X) XX'\right]^{-1}\right),$$

where

$$E\left[\sigma^{-2}(X) XX'\right]^{-1} \leq E\left[XX'\right]^{-1} E\left[\sigma^2(X) XX'\right]^{-1} E\left[XX'\right]^{-1}$$

with "=" if and only if $\sigma^2(X) = \sigma^2 = E\left[\varepsilon^2\right]$ is constant almost surely.

If the conditional variance function $\sigma^2(x)$ is *unknown*, $\tilde{\theta}_{\text{WLS}}$ is not feasible. One could then impose a parametric form on $\sigma^2(x)$ and estimate the unknown parameters using standard methods. This procedure requires that the functional form of the conditional variance is correctly specified however. In order to avoid the risk of working with a misspecificied model, a nonparametric estimator of $\sigma^2(x)$ should instead be used. To motivate our estimator, first note that $\sigma^2(x)$ by definition is simply the conditional mean of ε^2, c.f. eq. (12). A natural estimator of a conditional mean is the kernel regression estimator as introduced in eq. (4). Thus, ideally we would like to compute $\hat{\sigma}^2(x) = \sum_{i=1}^{n} \varepsilon_i^2 K_h(X_i - x) / (\sum_{i=1}^{n} K_h(X_i - x))$. However, since ε_i, $i = 1, ..., n$, are not observed, we replace these by the residuals. This leads to the following three-step procedure:

1. Compute the OLS estimator, $\hat{\theta}_{\text{OLS}}$, as given in eq. (2).

2. Compute the associated residuals, $\hat{\varepsilon}_i = Y_i - \hat{\theta}_{\text{OLS}}' X_i$, $i = 1, ..., n$, and use those to estimate the conditional variance nonparametrically,

$$\hat{\sigma}^2(x) = \frac{\sum_{i=1}^{n} \hat{\varepsilon}_i^2 K_h(X_i - x)}{\sum_{i=1}^{n} K_h(X_i - x)}.$$

3. Obtain the WLS estimator as given in (13), but with $\sigma^2(x)$ substituted for $\hat{\sigma}^2(x)$,

$$\hat{\theta}_{\text{WLS}} = \left(\sum_{i=1}^{n} \hat{\sigma}^{-2}(X_i) X_i X_i'\right)^{-1} \left(\sum_{i=1}^{n} \hat{\sigma}^{-2}(X_i) X_i Y_i\right), \tag{14}$$

Again, the above estimation method can be generalised to allow for more complicated parametric forms,

$$Y = g(X; \theta) + \varepsilon, \quad E[\varepsilon | X] = 0,$$

where $g : \mathbb{R}^d \times \Theta \mapsto \mathbb{R}$ is known up to $\theta \in \Theta$.

3.3 Independence Assumption

The above idea can be adapted to obtain ML-type estimators when the distribution of errors are of unknown form. We maintain the linear specification in eq. (11), but now assume that

$$\varepsilon \text{ and } X \text{ are independent}$$

such that $f_{\varepsilon|X}(\varepsilon|X) = f_\varepsilon(\varepsilon)$ where

$$E[\varepsilon] = \int_{\mathbb{R}} z f_\varepsilon(z)\, dz = 0, \quad \sigma^2 = \int_{\mathbb{R}} z^2 f_\varepsilon(z)\, dz < \infty.$$

Compared to the previous sections, we have here imposed an additional assumption of independence between regressors and errors. However, we do not assume that the distribution of ε is known, and as such the model remains semiparametric.

The independence assumption makes it possible to estimate the parametric component by semiparametric MLE: Suppose that the density f_ε was *known*; then we could do MLE,

$$\tilde{\theta}_{\text{MLE}} = \arg\max_{\theta\in\Theta} \frac{1}{n}\sum_i \log f_\varepsilon\left(Y_i - \theta' X_i\right), \tag{15}$$

which under regularity conditions will satisfy:

$$\sqrt{n}(\tilde{\theta}_{\text{MLE}} - \theta) \to^d N\left(0, H_0^{-1}\right),$$

where

$$H_0 = E\left[\frac{\partial \log f_\varepsilon(Y_i - \theta' X_i)}{\partial\theta}\frac{\partial \log f_\varepsilon(Y_i - \theta' X_i)}{\partial\theta'}\right] = \int \frac{f_\varepsilon'(z)^2}{f_\varepsilon(z)} dz E\left[XX'\right].$$

However, the density f_ε is *unknown*, and $\tilde{\theta}_{\text{MLE}}$ is therefore not feasible. On the other hand, observe that OLS is still a feasible option and will yield a consistent estimator. The OLS estimator will however not be as efficient as the MLE since $\int f_\varepsilon'(z)^2 / f_\varepsilon(z)\, dz \leq \sigma^2$ with "=" if and only if f_ε is the $N\left(0, \sigma^2\right)$ density.

To improve on the efficiency of the OLS estimator, we therefore propose to obtain a semiparametric version of the MLE by the following 3-step procedure:

1. Compute the OLS estimator, $\hat{\theta}_{\text{OLS}}$, as given in eq. (2).

2. Compute the associated residuals, $\hat{\varepsilon}_i = Y_i - \hat{\theta}_{\text{OLS}}' X_i$, $i = 1, ..., n$, and use these to estimate the marginal density f_ε nonparametrically, e.g.

$$\hat{f}_\varepsilon(x) = \frac{1}{n}\sum_{i=1}^n K_h(\hat{\varepsilon}_i - x). \tag{16}$$

3. Obtain the MLE estimator as given in (15), but with f_ε substituted for $\hat{f}_\varepsilon$,

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \log \hat{f}_\varepsilon \left(Y_i - \theta' X_i\right). \tag{17}$$

Again, the above estimation method can be generalised to allow for more complicated parametric forms. Suppose for example

$$Y = g\left(X;\theta\right) + \sigma\left(X;\theta\right)\varepsilon,$$

where $g, \sigma : \mathbb{R}^d \times \Theta \mapsto \mathbb{R}$ are known up to $\theta \in \Theta$, and ε and X are independent with

$$E\left[\varepsilon\right] = \int_{\mathbb{R}} z f_\varepsilon\left(z\right) dz = 0, \quad E\left[\varepsilon^2\right] = \int_{\mathbb{R}} z^2 f_\varepsilon\left(z\right) dz = 1.$$

Suppose that we have obtained a preliminary estimator of θ, e.g. the MLE based on normal errors, $\hat{\theta}_{\text{QMLE}}$ which will remain consistent even if the errors are not normally distributed. We can then compute the corresponding residuals

$$\hat{\varepsilon}_i = \frac{Y_i - g(X_i;\hat{\theta}_{\text{QMLE}})}{\sigma(X_i;\hat{\theta}_{\text{QMLE}})}, \quad i = 1, ..., n,$$

and then estimate the density nonparametrically as in eq. (16). In the final step, we then define

$$\hat{\theta} = \arg\max_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \left\{\log \hat{f}_\varepsilon \left(\frac{Y - g\left(X;\theta\right)}{\sigma\left(X;\theta\right)}\right) + \log\left(\sigma\left(X;\theta\right)\right)\right\},$$

We would expect that while $\hat{\theta}_{\text{QMLE}}$ will not enjoy full efficiency, $\hat{\theta}$ will.

4 Copulas

To show that semiparametric modelling have applications outside of a regression framework, we give a last example involving copulas. Copulas have proved to be a useful tool in the modelling of multivariate dependence structures; they have in particular found use in finance, see e.g. Genest et al (2009). We here present a semiparametric family of copulas and associated estimators.

Let $Z = (Z_1, Z_2) \in \mathbb{R}^2$ be a bivariate continuous random variable and denote the joint probability density function (pdf) and cumulative distribution function (cdf) by f and F respectively,

$$P\left(Z_1 \leq z_1, Z_2 \leq z_2\right) = F\left(z_1, z_2\right) = \int_{-\infty}^{z_2}\int_{-\infty}^{z_1} f\left(v_1, v_2\right) dv_1 dv_2.$$

Also let f_k and F_k denote the marginal pdf and cdf respectively of Z_k,

$$P\left(Z_k \leq z_k\right) = F_k\left(x\right) = \int_{-\infty}^{z_k} f_k\left(u\right) du, \quad k = 1, 2.$$

The so-called copulas are then used to model the dependence structure between Z_1 and Z_2 based on the following standard result: There exists a unique function $C : [0,1]^2 \mapsto [0,1]$ such that

$$F(z_1, z_2) = C(F_1(z_1), F_2(z_2)),$$

c.f. Joe (1997). The function C is referred to as the *copula* of Z. We easily see that C is the cdf of the uniformly distributed random variable $U := (F_1(Z_1), F_2(Z_2))$:

$$C(u_1, u_2) = P(F_1(Z_1) \leq u_1, F_2(Z_2) \leq u_2).$$

Furthermore, the joint density of Z can be expressed by

$$f(z_1, z_2) = c(F_1(z_1), F_2(z_2)) f_1(z_1) f_2(z_2),$$

where $c : [0,1] \times [0,1] \mapsto \mathbb{R}_+$ is the pdf of U.

One can now model the joint distribution of Z by specifying the two marginal distributions and the copula. In a fully parametric framework, this could for example be done by

$$f(z_1, z_2; \xi) = c(F_1(z_1; \alpha_1), F_2(z_2; \alpha_1); \theta) f_1(z_1; \alpha_1) f_2(z_2; \alpha_1),$$

where $\xi = (\theta', \alpha_1', \alpha_2')' \in \Theta \times \mathcal{A}_1 \times \mathcal{A}_1$ is a finite dimensional parameter. We could then proceed to estimate ξ by MLE,

$$\hat{\xi} = \arg\max_{\xi \in \Xi} \frac{1}{n} \sum_{i=1}^{n} \{\log c(F_1(Z_{1,i}; \alpha_1), F_2(Z_{2,i}; \alpha_1); \theta) + \log f_1(Z_{1,i}; \alpha_1) + \log f_2(Z_{2,i}; \alpha_1)\}.$$

This may potentially be a difficult problem to solve numerically if the dimension of ξ is large. Instead, one could instead estimate the parameters using a 2-step estimation procedure: First estimate $(\alpha_1', \alpha_2')'$ by

$$\hat{\alpha}_k = \arg\max_{\alpha_k \in \mathcal{A}_k} \frac{1}{n} \sum_{i=1}^{n} \log f_k(Z_{k,i}; \alpha_k), \quad k = 1, 2,$$

and then

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \log c(F_1(Z_{1,i}; \hat{\alpha}_1), F_2(Z_{2,i}; \hat{\alpha}_2); \theta).$$

This two-step estimator may have reduced efficiency compared to the full MLE above, but is easier to implement.

An obvious semiparametric copula model is the following: We still specify a parametric family of copulas, $c(u_1, u_2; \theta)$, but now leave the two marginal distributions unspecified. We then wish to estimate the marginal distributions nonparametrically, and use these to draw inference about θ. Let

$$\hat{F}_k(z_k) = \frac{1}{n} \sum_{i=1}^{n} 1\{Z_{k,i} \leq z_k\}, \quad k = 1, 2,$$

be the empirical cdf's. A natural estimator of θ would then be

$$\hat{\theta} = \arg\max_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \log c\left(\hat{F}_1\left(Z_{1,i}\right), \hat{F}_2\left(Z_{2,i}\right); \theta\right).$$

5 A Class of Two-Step Estimators

In the previous two sections, we presented a number of examples of semiparametric models, and derived estimators of the parameters of interest. In this section, we wish to develop a framework within which we can analyze the asymptotic properties of these estimators. In particular, we will give conditions for the estimators to be $\sqrt{n}$-consistent and with an asymptotically normal distribution.

We start out by introducing a general class of semiparametric two-step estimators: In the first step, a preliminary nonparametric estimator is computed. In the second step, this nonparametric estimator is plugged into a criterion function which is then minimized in order to obtain an estimator of the parametric component. The class is sufficiently general to contain all of the estimators defined in the previous two sections. We give general conditions for consistency and asymptotic normality of the estimator of the parametric component under suitable regularity conditions.

Our estimation problem has a lot in common with standard parametric two-step estimation problems where a preliminary estimator of a nuisance parameter is used to obtain an estimator of the parameter of interest. The only difference is that in our case the preliminary estimator is a function and not a finite-dimensional parameter. However, the strategy of proof for parametric two-step estimators can after suitable modifications still be used.

5.1 The Framework

We are interested in estimating a finite dimensional parameter $\theta \in \Theta \subseteq \mathbb{R}^k$ by a random objective function $Q_n(\theta, \gamma)$ where $\gamma \in \Gamma$ is some infinite-dimensional parameter, in most cases a function. The objective function will in most situations be a function of available data, (Y_i, X_i) for $i = 1, ..., n$, but we here suppress this dependence and only indicate it through the subscript n. We assume that the parameter space Γ is a linear space equipped with a norm $\|\cdot\|$. This norm could for example be the supremum norm, $\|\gamma\| = \sup_x |\gamma(x)|$, or the L_q-norm, $\|\gamma\| = \left(\int |\gamma(x)|^q w(x)\, dx\right)^{1/q}$ for some weighting function $w(x) \geq 0$.

If the true value of γ, which we denote γ_0, was known, we could estimate θ by

$$\tilde{\theta} = \arg\min_{\theta\in\Theta} Q_n(\theta, \gamma_0). \tag{18}$$

In this case, standard results for parametric estimators can be empoyed to derive the asymptotic properties of $\tilde{\theta}$, see e.g. Newey and McFadden (1994).

Here, we will consider the case where γ_0 is unknown, hence $\tilde{\theta}$ is not feasible. However, suppose a preliminary estimator, $\hat{\gamma}$, of it is available. Then, but by substituting γ_0 for $\hat{\gamma}$, we can instead use

$$\hat{\theta} = \arg\min_{\theta \in \Theta} Q_n(\theta, \hat{\gamma}). \tag{19}$$

We will refer to $\hat{\theta}$ as a semiparametric two-step estimator.

Initially, we will make minimal assumptions regarding the form of $Q_n(\theta, \gamma)$ and $\hat{\gamma}$ and only require that it is a consistent estimator of γ_0, and converges with sufficiently fast rate. A leading case is where the objective function takes the form

$$Q_n(\theta, \gamma) = \frac{1}{n} \sum_{i=1}^{n} q(Z_i; \theta, \gamma), \tag{20}$$

but we will not limit ourselves to this situation.

Before proceeding with the analysis of the general two-step estimator, we first demonstrate how the estimators presented in the previous sections can be written on the form of (19)-(20) by suitable choice of $q(z; \theta, \gamma)$:

Example 1: Single-Index Model. With $\gamma = g$, the estimator for this model can be written on the form of eqs. (19)-(20) with q given by

$$q(z; \theta, \gamma) = \left[y - \gamma\left(\theta' x\right)\right]^2,$$

and the estimator $\hat{\gamma}_\theta$ could be chosen as

$$\hat{\gamma}_\theta(z) = \frac{\sum_{i=1}^{n} Y_i K_h(\theta' X_i - z)}{\sum_{i=1}^{n} K_h(\theta' X_i - z)}.$$

Example 2: Partially Linear Model. Here, the estimator can be written on the desired form by defining

$$q(z; \theta, \gamma) = \left[y - \gamma_1(x_2) - \theta'(x_1 - \gamma_2(x_2))\right]^2,$$

where $\gamma_1(x_2) = E[Y|X_2 = x_2]$, $\gamma_2(x_2) = E[X_1|X_2 = x_2]$. The preliminary estimators were given as

$$\hat{\gamma}_1(x_2) = \frac{\sum_{i=1}^{n} Y_i K_h(X_{2,i} - x_2)}{\sum_{i=1}^{n} K_h(X_{2,i} - x_2)}, \quad \hat{\gamma}_2(x_2) = \frac{\sum_{i=1}^{n} X_{1,i} K_h(X_{2,i} - x_2)}{\sum_{i=1}^{n} K_h(X_{2,i} - x_2)}.$$

Example 3: Efficient Estimation in the Presence of Heteroskedasticity. By defining the function q by

$$q(z; \theta, \gamma) = \gamma^{-1}(x)\left[y - \theta' x\right]^2, \tag{21}$$

where $\gamma(x) = \sigma^2(x)$, the WLS estimator is seen to be a special case of the general two-step estimator. Here, the preliminary estimators are given by

$$\hat{\gamma}(x_2) = \frac{\sum_{i=1}^n \hat{\varepsilon}_i^2 K_h(X_i - x)}{\sum_{i=1}^n K_h(X_i - x)}. \tag{22}$$

Example 4: Semiparametric Copulas. The function q defining the copula estimator θ is given by

$$q(z; \theta, \gamma) = \log c(\gamma_1(z_1), \gamma_2(z_2); \theta),$$

where $\gamma_k(z) = F_k(z)$ is the marginal cdf of Z_k, $k = 1, 2$. These can be estimated by:

$$\hat{\gamma}_k(z) = \frac{1}{n} \sum_{i=1}^n 1\{Z_{k,i} \leq z\}, \quad k = 1, 2.$$

In two of the above examples, namely the partially linear model and the regression model with unknown heteroskedasticity, closed form expressions of $\hat{\theta}$ can be derived. Thus, a direct analysis of these particular estimators could be carried out, and would probably be more straightforward compared to the indirect analysis proposed here. But in general, explicit expressions of the estimators are not available, and analysis has to be centered around the properties of the objective function $Q_n(\theta, \gamma)$.

Within this general framework, we will first establish high-level conditions under which $\hat{\theta}$ is consistent and converges towards a normal distribution. Imposing more structure on the objective function $Q_n(\theta, \gamma)$ and the estimator $\hat{\gamma}$, we then sketch how these high-level conditions can be verified with particular emphasis on the case where $\hat{\gamma}$ is a kernel estimator. Finally, we establish the first-order asymptotic properties of the WLS estimator as defined in Section 3.2 by verifying the high-level conditions for this particular estimator.

5.2 Consistency

The proof of consistency is more or less identical to the one for parametric two-step estimators; the only difference is conceptual since we here work with an infinite-dimensional parameter. We will impose the following conditions on the objective function:

C.1 There exists a function $Q(\theta, \gamma)$ such that: $\sup_{\theta \in \Theta} |Q_n(\theta, \gamma_0) - Q(\theta, \gamma_0)| \to^P 0$.

C.2 For all $\varepsilon > 0$: $\inf_{\|\theta - \theta_0\| > \varepsilon} Q(\theta, \gamma_0) > Q(\theta_0, \gamma_0)$.

C.3 For some $\lambda > 0$ and $B_n = O_P(1)$:

$$\sup_{\theta \in \Theta} |Q_n(\theta, \gamma) - Q_n(\theta, \gamma_0)| \leq B_n \|\gamma - \gamma_0\|^{\lambda}$$

for γ in a neighbourhood of γ_0.

Condition (C.1) states that the infeasible finite-sample objective function, $Q_n(\theta, \gamma_0)$, has a well-defined limit, $Q(\theta, \gamma_0)$. Condition (C.2) is an identification condition saying that the limiting function uniquely identifies θ_0 as its minimum, $\theta_0 = \arg\min_{\theta \in \Theta} Q(\theta, \gamma_0)$. It can easily be shown that Condition (C.2) is implied by the following three conditions: Θ is compact, $\theta \mapsto Q(\theta, \gamma_0)$ is continuous, and $Q(\theta, \gamma_0) > Q(\theta_0, \gamma_0)$ for all $\theta \neq \theta_0$, while primitive conditions for C.1 can be found in Newey (1991). Conditions (C.1)-(C.2) imply that the infeasible estimator $\tilde{\theta}$ as defined in eq. (18) is consistent, $\tilde{\theta} \to^P \theta_0$; see e.g. Newey and McFadden (1994, Theorem 2.1).

The final condition, (C.3), states that the difference between the two objective functions, $Q_n(\theta, \hat{\gamma})$ and $Q_n(\theta, \gamma_0)$, is asymptotically negligible: $Q_n(\theta, \hat{\gamma}) \to^P Q_n(\theta, \gamma_0)$ as $\hat{\gamma} \to^P \gamma_0$. Note here that the norm $\|\gamma - \gamma_0\|$ is a functional norm as discussed in the beginning of this section. Under this assumption, the feasible estimator converges towards the infeasible one, $\hat{\theta} = \tilde{\theta} + o_P(1)$.

Conditions (C.1) and (C.3) could be exchanged for the following two conditions: (C.1') $\sup_{\theta \in \Theta, \|\gamma - \gamma_0\| < \delta} |Q_n(\theta, \gamma) - Q(\theta, \gamma)| \to^P 0$, $\delta > 0$ and (C.3') $\sup_{\theta \in \Theta} |Q(\theta, \gamma) - Q(\theta, \gamma')| \to 0$ as $\gamma \to \gamma'$. Empirical process theory could be used to verify conditions (C.1') and (C.3'), c.f. Andrews (1994a,b), Chen, Linton and van Keilegom (2003), van der Vaart and Wellner (1996). This verification normally would involve a Lipschitz condition of the type stated in Condition C.3.

The formal consistency result is stated in the following theorem:

Theorem 1 *Assume that $Q_n(\theta, \gamma)$ satisfies (C.1)-(C.3). If $\hat{\gamma} \in \Gamma$ from a certain step with $\hat{\gamma} \to^P \gamma_0$, then $\hat{\theta} \to^P \theta_0$.*

Remark 2 *In the case where $\hat{\gamma}$ depends on θ, one needs to strengthen the consistency condition to $\sup_{\theta \in \Theta} \|\hat{\gamma}_\theta - \gamma_\theta\| \to^P 0$.*

We now verify conditions (C.1)-(C.3) for the WLS estimator:

Example 3 (cont.). Let σ_0^2 and θ_0 denote the true parameter values. We here assume that $X \in \mathcal{X}$, where $\mathcal{X} \subset \mathbb{R}^d$ is compact, and $E\left[Y^2\right] < \infty$. For identification we need that $E\left[XX'\sigma_0^{-2}(X)\right]$ is nonsingular. The assumption of compact support $\mathcal{X}$ can be dispensed of, but one then has to introduce trimming of the estimator, c.f. Robinson (1987).

Also, assume that $\sigma_0^2(x), f(x) > 0$ are twice continuously differentiable. In particular, $\underline{\sigma}^2 := \inf_{x\in\mathcal{X}} \sigma_0^2(x) > 0$. We restrict Θ to be compact so there exists constant c such that $|\theta' x| \leq c$ for any $\theta \in \Theta$ and $x \in \mathcal{X}$. We define the norm of σ^2 as $\left\|\sigma^2\right\|_\infty = \sup_{x\in\mathcal{X}} \left|\sigma^2(x)\right|$, and assume that we have established

$$\left\|\hat{\sigma}^2 - \sigma_0^2\right\|_\infty \to^P 0, \tag{23}$$

where $\hat{\sigma}^2$ is the kernel estimator given in (22); this could for example be done using the results of Kristensen (2009b).

First, we show (C.3): The criterion function takes the form in eq. (21). By a first order Taylor expansion of the function $a \mapsto 1/a$,

$$\begin{aligned} q(z;\theta,\hat{\sigma}^2) - q\left(z;\theta,\sigma_0^2\right) &= \left[y-\theta' x\right]^2 \left\{\frac{1}{\hat{\sigma}^2(x)} - \frac{1}{\sigma_0^2(x)}\right\} \\ &= \left[y-\theta' x\right]^2 \frac{-1}{\left[\lambda_x \hat{\sigma}^2(x) + (1-\lambda_x)\sigma_0^2(x)\right]^2} \left\{\hat{\sigma}^2(x) - \sigma_0^2(x)\right\}, \end{aligned}$$

for some $\lambda_x \in [0,1]$. Because of (23), $\inf_{x\in\mathcal{X}} \hat{\sigma}^2(x) \geq \underline{\sigma}^2/2$ almost surely from a certain step as $n \to \infty$. Thus,

$$\frac{1}{\left[\lambda_x \hat{\sigma}^2(x) + (1-\lambda_x)\sigma_0^2(x)\right]^2} \leq \frac{1}{\left[\lambda_x \underline{\sigma}^2/2 + (1-\lambda_x)\underline{\sigma}^2/2\right]^2} = \frac{4}{\underline{\sigma}^4} < \infty.$$

Also, since $\theta \mapsto (Y-\theta' X)^2$ is continuous and $(Y-\theta' X)^2 \leq 3Y^2 + 3c^2$ where $E\left[Y^2\right] < \infty$, it follows from standard uniform convergence results (see e.g. Newey, 1991) that

$$\sup_{\theta\in\Theta} \left| \frac{1}{n}\sum_{i=1}^n \left(Y_i - \theta' X_i\right)^2 - E[\left(Y_i - \theta' X_i\right)^2] \right| \to^P 0,$$

and $\sup_{\theta\in\Theta} E[(Y-\theta' X)^2] < \infty$. Thus, $\sup_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^n (Y_i - \theta' X_i)^2 = O_P(1)$. Next, write:

$$\begin{aligned} \sup_{\theta\in\Theta} \left|Q_n\left(\theta,\hat{\sigma}^2\right) - Q_n\left(\theta,\sigma_0^2\right)\right| &\leq \sup_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^n \left|q(Z_i;\theta,\hat{\sigma}^2) - q\left(Z_i;\theta,\sigma_0^2\right)\right| \\ &\leq \sup_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^n \left[Y_i - \theta' X_i\right]^2 \times \sup_x \left| \frac{1}{\hat{\sigma}^2(x)} - \frac{1}{\sigma_0^2(x)} \right|, \end{aligned}$$

where

$$\sup_x \left| \frac{1}{\hat{\sigma}^2(x)} - \frac{1}{\sigma_0^2(x)} \right| \leq \frac{4\left\|\hat{\sigma}^2 - \sigma_0^2\right\|_\infty}{\underline{\sigma}^4} = o_P(1).$$

Thus, (C.3) is satisfied with

$$B_n = \frac{4}{\underline{\sigma}^2} \times \sup_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^n \left[Y_i - \theta' X_i\right]^2,$$

and and $\lambda = 1$.

Next, we verify (C.1): By yet another application of a uniform Law of Large Numbers (LLN), one easily shows that

$$\sup_{\theta \in \Theta} \left| Q_n\left(\theta, \sigma_0^2\right) - Q\left(\theta, \sigma_0^2\right) \right| \to^P 0,$$

where $\theta \mapsto Q\left(\theta, \sigma_0^2\right) = E\left[\left[Y - \theta' X\right]^2 \sigma_0^{-2}\left(X\right)\right]$ is continuous.

Finally, to verify (C.2), observe that for any $\theta \neq \theta_0$,

$$\begin{aligned} Q\left(\theta, \sigma_0^2\right) &= E\left[\left[\left(\theta_0 - \theta\right)' X + \varepsilon\right]^2 \sigma_0^{-2}\left(X\right)\right] \\ &= \left(\theta_0 - \theta\right)' E\left[X X' \sigma_0^{-2}\left(X\right)\right]\left(\theta_0 - \theta\right) + E\left[\varepsilon^2 \sigma_0^{-2}\left(X\right)\right] \\ &> E\left[\varepsilon^2 \sigma_0^{-2}\left(X\right)\right] \\ &= Q\left(\theta_0, \sigma_0^2\right). \end{aligned}$$

Given the remarks following conditions (C.1)-(C.3), this implies (C.2). We have now verified these conditions and Theorem 1 now gives us consistency of $\hat{\theta}$.

5.3 Asymptotic Normality

To show asymptotic normality, we use the same strategy as for parametric two-step estimators: There, one normally would make a Taylor expansion w.r.t. the first-step estimator, thereby taking into account the additional sampling error due to the first-step estimator. However, in our setting the first-step estimator is a function, i.e. an infinite-dimensional parameter. Thus, in order to follow the strategy used for parametric two-step estimators, we first need to generalize the concept of derivatives from the standard finite-dimensional case to the infinite-dimensional one.

Let $T : \Gamma \mapsto \mathbb{R}^d$ be a functional taking any given $\gamma \in \Gamma$ into a Euclidean vector. For example , $T\left(\gamma\right) = \int \gamma\left(x\right) dx$, $T_x\left(\gamma\right) = \partial \gamma\left(x\right) / \partial x$, and $T_x\left(\gamma_1, \gamma_2\right) = \gamma_1\left(x\right) \gamma_2\left(x\right)$.

Definition 3 *We say that T is pathwise differentiable at $\gamma \in \Gamma$ if there exists a linear and continuous functional $\dot{T}\left(\gamma\right)\left[\cdot\right] : \Gamma \mapsto \mathbb{R}^d$ such that*

$$\dot{T}\left(\gamma\right)\left[h\right] = \lim_{t \to 0} \frac{T\left(\gamma + th\right) - T\left(\gamma\right)}{t},$$

for all $h \in \Gamma$.

One normally refers to $\dot{T}$ as the *pathwise derivative* of T. In the finite-dimensional case, if T is differentiable with derivative $\partial T\left(\gamma\right) / \partial \gamma$ then $\dot{T}\left(\gamma\right)\left[h\right]$ is the differential of T,

$$\dot{T}\left(\gamma\right)\left[h\right] = \frac{\partial T\left(\gamma\right)}{\partial \gamma} h.$$

You can normally carry over results from the finite-dimensional case when deriving the pathwise derivative. In particular, the chain-rule is still valid.

Examples of Functionals. (i) $\Gamma = \{\gamma : \int |\gamma(x)|\, dx < \infty\}$ and $T(\gamma) = \int \gamma(x)\, dx$. Then $\dot{T}(\gamma)[h] = \int h(x)\, dx$: It's linear, continuous in the L_1-norm, and

$$T(\gamma + th) - T(\gamma) = \int (\gamma + th)(x)\, dx - \int \gamma(x)\, dx = t \int h(x)\, dx = t\dot{T}(\gamma)[h].$$

(ii) $\Gamma = \{\gamma | \partial\gamma(x)/\partial x \text{ exists}\}$ and $T(\gamma) = \partial\gamma(x)/\partial x$. Then $\dot{T}(\gamma)[h] = \partial h(x)/\partial x$:

$$T(\gamma + th) - T(\gamma) = \frac{\partial(\gamma + h)}{\partial x} - \frac{\partial\gamma}{\partial x} = t\frac{\partial h}{\partial x} = t\dot{T}(\gamma)[h]$$

These two examples are simple cases since in both T is a linear functional.

(iii) $T_x(\gamma) = F(\gamma(x))$ then $\dot{T}_x(\gamma)[h] = F'(\gamma(x))\, h(x)$.

(iv) $T(\gamma) = \int F(\gamma(x))\, dx$. Then $\dot{T}(\gamma)[h] = \int F'(\gamma(x))\, h(x)\, dx$ under suitable conditions on F and Γ.

We now wish to use pathwise derivatives to evaluate the additional sampling variation of our estimator $\hat{\theta}$ due to the presence of $\hat{\gamma}$. First, introduce the following functionals which are the score and the Hessian of the objective functions,

$$S_n(\theta, \gamma) = \frac{\partial Q_n(\theta, \gamma)}{\partial\theta}, \quad H_n(\theta, \gamma) = \frac{\partial^2 Q_n(\theta, \gamma)}{\partial\theta\partial\theta'}.$$

We then assume that the pathwise derivative of $S_n(\theta, \gamma)$ w.r.t. γ at $(\theta, \gamma) = (\theta_0, \gamma_0)$ exists in the direction $h \in \Gamma$ and denote this by $\dot{S}_n(\theta_0; \gamma_0)[h]$. We can then give conditions under which asymptotic normality holds:

N.1 $\|\hat{\gamma} - \gamma_0\| = o_P(n^{-1/4})$ and $\hat{\theta} \to^P \theta_0$.

N.2 $\theta_0 \in \text{int}\Theta$.

N.3 $Q_n(\theta, \gamma)$ is twice continuously differentiable w.r.t. θ in a neighbourhood $\mathcal{N}$ of θ_0.

N.4 The pathwise derivative $\dot{S}_n(\theta_0; \gamma_0)[h]$ exists and satisfies

$$\left\| S_n(\theta_0, \gamma) - S_n(\theta_0, \gamma_0) - \dot{S}_n(\theta_0; \gamma_0)[\gamma - \gamma_0] \right\| \leq B_n \|\gamma - \gamma_0\|^2$$

where $B_n = O_P(1)$.

N.5 $\sqrt{n}\left\{ S_n(\theta_0, \gamma_0) + \dot{S}_n(\theta_0; \gamma_0)[\hat{\gamma} - \gamma_0] \right\} \to^d N(0, \Omega_0)$.

N.6 $||H_n(\theta, \gamma) - H_n(\theta, \gamma_0)|| \leq B_n \|\gamma - \gamma_0\|^\lambda$ where $B_n = O_P(1)$.

N.7 $\sup_{\theta\in\mathcal{N}} \|H_n(\theta,\gamma_0) - H(\theta,\gamma_0)\| \to^P 0$, where $H_0 = H(\theta_0,\gamma_0)$ is non-singular.

As was the case with the consistency result, our conditions consist of two parts: The first set of conditions, (N.2), (N.3), (N.5) (setting $\dot{S}_n(\theta_0;\hat{\gamma}-\gamma_0)=0$) and (N.7), imply that the infeasible estimator assuming γ_0 known, $\tilde{\theta}$, is $\sqrt{n}$-asymptotically normally distributed; see e.g. Newey and McFadden (1994, Theorem 3.1). The remaining conditions, (N.1), (N.4) and (N.6), then enable us to show that the feasible estimator is also $\sqrt{n}$-asymptotically normally distributed.

Theorem 4 *Assume that $\hat{\theta} \to^P \theta_0$, and that (N.1)-(N.7) hold. Then:*

$$\sqrt{n}(\hat{\theta}-\theta_0) \to^d N\left(0, H_0^{-1}\Omega_0 H_0^{-1}\right),$$

where Ω_0 and H_0 are given in (N.5) and (N.7).

As mentioned before the theorem, the infeasible estimator, $\tilde{\theta}$, is also $\sqrt{n}$-asymptotically normally distributed under (N.1)-(N.7). However, $\tilde{\theta}$ will in general have a smaller asymptotic variance and as such be more efficient than $\hat{\theta}$. The two estimators asymptotic variances are only equal if the adjustment term vanishes asymptotically. That is, $\sqrt{n}\dot{S}_n(\theta_0;\gamma_0)[\hat{\gamma}-\gamma_0]) = o_P(1)$ in which case, $\tilde{\theta}$ and $\hat{\theta}$ are first order equivalent. In most cases however, one pays a price for not knowing γ_0 in which case $\text{Var}(\hat{\theta}) > \text{Var}(\tilde{\theta})$.

An alternative set of conditions which in some cases might be easier to verify can be used instead of (N.3)-(N.7):

N.3' $Q_n(\theta,\gamma)$ is continuously differentiable w.r.t. θ in a neighbourhood $\mathcal{N}$ of θ_0.

N.4' There exists a functional $S(\theta,\gamma)$ such that $\nu_n(\theta,\gamma) := S_n(\theta,\gamma) - S(\theta,\gamma)$ satisfies:

$$\sup_{\|\theta-\theta_0\|<\delta, \|\gamma-\gamma_0\|<\delta} \|\nu_n(\theta,\gamma) - \nu_n(\theta_0,\gamma_0)\| = o_P\left(1/\sqrt{n}\right),$$

and $S(\theta_0,\gamma_0)=0$.

N.5' The pathwise derivative $\dot{S}(\theta,\gamma)[h]$ of $S(\theta,\gamma_0)$ exists and satisfies

$$\left\| S(\theta_0,\gamma) - S(\theta_0,\gamma_0) - \dot{S}(\theta_0,\gamma_0)[\gamma-\gamma_0] \right\| \leq B\|\gamma-\gamma_0\|^2$$

for a constant $B<\infty$.

N.6' $\sqrt{n}\{S_n(\theta_0,\gamma_0) + (\theta_0;\gamma_0)[\hat{\gamma}-\gamma_0]\} \to^d N(0,\Omega_0)$.

N.7' The function $S(\theta,\gamma)$ is continuously differentiable w.r.t. θ in a neighbourhood $\mathcal{N}$ of θ_0 with continuous derivative $H(\theta,\gamma)$ which satisfies $\sup_{\theta\in\mathcal{N}} \|H(\theta,\gamma) - H(\theta,\gamma_0)\| \leq B\|\gamma-\gamma_0\|^{\lambda}$, where $H_0 = H(\theta_0,\gamma_0)$ is non-singular.

We note that (N.3) has been weakened to only require $Q_n(\theta,\gamma)$ having one derivative. The condition (N.4') is rather high-level, but can be verified by empirical process techniques; see, for example, Chen, Linton and van Keilegom (2003) for a set of sufficient conditions. In most cases, $S(\theta,\gamma)$ can be chosen as $S(\theta,\gamma) = \partial Q(\theta,\gamma)/(\partial\theta)$, in which case the identification condition given in (C.2) will normally ensure that $S(\theta_0,\gamma_0) = 0$.

Also note that the conditions in (N.5') and (N.6') involve the limiting score function $S(\theta,\gamma)$ instead of, as in (N.5) and (N.6), the sample version, $S_n(\theta,\gamma)$.

Theorem 5 *Assume that $\hat{\theta} \to^P \theta_0$, and (N.1)-(N.2) and (N.3')-(N.7') hold. Then the conclusion of Theorem 4 remains true.*

While in Theorem 4 we require a CLT to hold for $\sqrt{n}\{S_n(\theta_0,\gamma_0) + \dot{S}_n(\theta_0;\gamma_0)[\hat{\gamma}-\gamma_0]\}$, in Theorem 5 we now require $\sqrt{n}\{S_n(\theta_0,\gamma_0) + \dot{S}(\theta_0;\gamma_0)[\hat{\gamma}-\gamma_0]\}$ to satisfy one. To apply either of these theorems, the major challenge lies in establishing a CLT for either of the two terms. At a first glance, this might seem impossible due to the presence of $\dot{S}_n(\theta_0;\gamma_0)[\hat{\gamma}-\gamma_0]$ and $\dot{S}(\theta_0;\gamma_0)[\hat{\gamma}-\gamma_0]$ since both terms involve a nonparametric estimator, $\hat{\gamma}$, which in general converges with a rate slower than $\sqrt{n}$. However, additional smoothing of the nonparametric estimator is implicitly taking place when plugging $\hat{\gamma}$ into the the pathwise derivative. As we shall see, this smoothing will in general increase the convergence rate and make it possible to verify (N.5) or (N.6').

To demonstrate how $\sqrt{n}$-convergence can be verified, we restrict ourselves to the case where the score can be written as

$$S_n(\theta,\gamma) = \frac{1}{n}\sum_{i=1}^{n} s(Z_i;\theta,\gamma) + o_P\left(n^{-1/2}\right),$$

where $Z_i \in \mathbb{R}^d$, $i = 1,...,n$, are i.i.d. data. This restriction holds, for example, when $Q_n(\theta,\gamma)$ is given by eq. (20) in which case $s(z;\theta,\gamma) = \partial q(z;\theta,\gamma)/(\partial\theta)$. Under this restriction, the pathwise derivatives of $S_n(\theta,\gamma)$ and $S(\theta,\gamma)$ are given by

$$\dot{S}_n(\theta,\gamma)[h] = \frac{1}{n}\sum_{i=1}^{n} \dot{s}(Z_i;\theta,\gamma)[h], \quad \dot{S}(\theta,\gamma)[h] = E[\dot{s}(Z;\theta,\gamma)[h]],$$

where $\dot{s}(z;\theta,\gamma)[h]$ is the pathwise derivative of $s(z;\theta,\gamma)[h]$ w.r.t. γ in the direction h.

We first give sufficient conditions for Assumption (N.4)-(N.5) to hold:

N.4.i $\|s(z;\theta_0,\gamma) - s(z;\theta_0,\gamma_0) - \dot{s}(z;\theta_0,\gamma_0)[\gamma-\gamma_0]\| \leq b(z)\|\gamma-\gamma_0\|^2$ with $E[b(Z)] < \infty$.

N.5.i $n^{-1}\sum_{i=1}^{n} \dot{s}(Z_i;\theta_0,\gamma_0)[\hat{\gamma}-\gamma_0] = \dot{S}(\theta_0,\gamma_0)[\hat{\gamma}-\gamma_0] + o_P(1/\sqrt{n})$.

N.5.ii There exists a function $\delta : \mathbb{R}^d \mapsto \mathbb{R}^k$ with $E[\delta(Z)] = 0$ and $E[\|\delta(Z)\|^2] < \infty$ such that

$$\dot{S}(\theta_0, \gamma_0)[\hat{\gamma} - \gamma_0] = \frac{1}{n}\sum_{i=1}^{n} \delta(Z_i) + o_P\left(1/\sqrt{n}\right).$$

N.5.iii $E[s(Z; \theta_0, \gamma_0)] = 0$ and $E[\|s(Z; \theta_0, \gamma_0)\|^2] < \infty$.

Lemma 6 *Assume that $\dot{s}(z; \theta, \gamma)[h]$ exists and satisfies (N.4.i)-(N.5.iii). Then Assumptions (N.4) and (N.5) hold with*

$$\Omega = E\left[[s(Z; \theta_0, \gamma_0) + \delta(Z)][s(Z; \theta_0, \gamma_0) + \delta(Z)]'\right].$$

While (N.4.i)-(N.5.iii) are more primitive conditions, it is in many cases still not so obvious how to actually verify them. In particular, assumptions (N.5.i)-(N.5.ii) are not straightforwardly shown to hold. To make any further progress, we assume that $\hat{\gamma}$ can be written on the form

$$\hat{\gamma}(x) = \frac{1}{n}\sum_{i=1}^{n} w_n(x, Z_i) + o_P\left(n^{-1/4}\right),$$

for some function w_n which is allowed to depend on sample size n. This restriction is for example satisfied for kernel estimators by defining $w_n(x, Z_i) = Y_i K_h(X_i - x)$. One can easily check that series estimators also fall within this framework, see e.g. Newey (1997). In the following, we suppress the dependence on (θ_0, γ_0), and for example write $\dot{S}[\hat{\gamma} - \gamma_0]$ for $\dot{S}(\theta_0, \gamma_0)[\hat{\gamma} - \gamma_0]$.

Verification of (N.5.i). First, observe that since $\dot{s}$ is a linear functional, we can write

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} \dot{s}(Z_i)[\hat{\gamma} - \gamma_0] - \dot{S}[\hat{\gamma} - \gamma_0] \\
&= \left\{\frac{1}{n}\sum_{i=1}^{n} \dot{s}(Z_i)[\hat{\gamma} - \bar{\gamma}] - \dot{S}[\hat{\gamma} - \bar{\gamma}]\right\} \\
&+ \left\{\frac{1}{n}\sum_{i=1}^{n} \dot{s}(Z_i)[\bar{\gamma} - \gamma_0] - \dot{S}[\bar{\gamma} - \gamma_0]\right\} \\
&=: I_{n,1} + I_{n,2},
\end{aligned}$$

where $\bar{\gamma}(x) = E[\hat{\gamma}(x)]$. Defining

$$\begin{aligned}
V_n(x, x') = \dot{s}(z)[w_n(\cdot, x')], \quad V_n = E[V_n(Z_1, Z_2)], \\
V_{n,1}(x) = E[V_n(x, Z)], \quad V_{n,2}(x) = E[V_n(Z, x)],
\end{aligned}$$

and again using that $\dot{s}$ is a linear functional, we can write

$$\begin{aligned} I_{n,1} &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\dot{s}\left(Z_i\right)\left[w_n\left(\cdot,Z_j\right)\right] - \frac{1}{n}\sum_{i=1}^{n}\dot{s}\left(Z_i\right)\left[E[w_n\left(\cdot,Z\right)]\right] \\ &\quad - \frac{1}{n}\sum_{i=1}^{n}\dot{S}[w_n\left(\cdot,Z_j\right)] + \dot{S}[E[w_n\left(\cdot,Z_2\right)]] \\ &= \frac{1}{n}\sum_{i,j=1}^{n}V_n\left(Z_i,Z_j\right) - \frac{1}{n}\sum_{i=1}^{n}V_{n,1}\left(Z_i\right) - \frac{1}{n}\sum_{j=1}^{n}V_{n,2}\left(Z_j\right) - V_n. \end{aligned}$$

One can then use results for so-called U-statistics (see, for example, Lee, 1990 for an introduction) to show that the RHS is $o_P\left(n^{-1/2}\right)$ in great generality. This takes care of the first term. To deal with the second term, one can normally show that

$$\left\|\dot{s}\left(z;\theta_0,\gamma_0\right)[h]\right\| \leq b\left(z\right)\left\|h\right\|,$$

in which case,

$$E\left[\left\|I_{n,2}\right\|\right] \leq E\left[b\left(Z\right)\right]\left\|\left[E[\hat{\gamma}]-\gamma_0\right]\right\|$$

and one then needs to show that the bias vanishes sufficiently fast, $\left\|E[\hat{\gamma}]-\gamma_0\right\| = o_P\left(n^{-1/2}\right)$. In the case where $\hat{\gamma}$ is a kernel estimator, this can be verified in great generality by combining so-called higher order kernels with undersmoothing.

Verification of (N.5.ii). This is normally established by first showing that there exists a function d such that

$$\dot{S}[h] = \int d\left(x\right)h\left(x\right)dx.$$

Often one can establish this directly if one has an explicit expression for $\dot{S}$, see Newey (1994b). Alternatively, one can use Riesz' Representation Theorem to establish this as utilized in Aït-Sahalia (1993).

Given this representation, one can normally find the function δ. Suppose, for example, that $\gamma_0\left(x\right) = f_X\left(x\right)E\left[Y|X=x\right]$, $\hat{\gamma}\left(x\right) = 1/n\sum_{i=1}^{n}Y_iK\left(\left(X_i-x\right)/h\right)/h^d$, and $\dot{S}[h] = \int d\left(x\right)h\left(x\right)dx$. Then, we first write

$$\dot{S}[\hat{\gamma}-\gamma_0] = \dot{S}[\hat{\gamma}] - \dot{S}[\gamma_0].$$

The first term satisfies

$$\dot{S}[\hat{\gamma}] = \int d\left(x\right)\hat{\gamma}\left(x\right)dx = \frac{1}{n}\sum_{i=1}^{n}Y_i\frac{1}{h^d}\int d\left(x\right)K\left(\frac{X_i-x}{h}\right)dx = \frac{1}{n}\sum_{i=1}^{n}Y_id\left(X_i\right) + o_P\left(1/\sqrt{n}\right),$$

where the last equality follows under suitable regularity conditions since

$$h^{-d}\int d\left(x\right)K\left(\frac{X_i-x}{h}\right)dx \to d\left(X_i\right),$$

as $h \to 0$. The second term can be written as

$$\dot{S}[\gamma_0] = \int d(x)\,\gamma_0(x)\,dx = \int d(x)\,f_X(x)\,E[Y|X=x]\,dx = E[Y d(X)].$$

So by defining $\delta(z)$ by

$$\delta(z) := y d(x) - E[Y d(X)],$$

we obtain as desired that

$$\dot{S}[\hat{\gamma} - \gamma_0] = \frac{1}{n}\sum_{i=1}^{n} \delta(Z_i) + o_P\left(1/\sqrt{n}\right).$$

Further techniques for verification of (N.5.ii) for kernel estimators can be found in Newey (1994b).

Example 3 (cont.). We assume that we have already verified that

$$\left\|\hat{\sigma}^2 - \sigma_0^2\right\|_\infty = o_P\left(n^{-1/4}\right),$$

for some bandwidth sequence (see e.g. Kristensen, 2009b).

To derive the asymptotic distribution of θ, we first find the score function and the Hessian,

$$S_n\left(\theta,\sigma^2\right) = \frac{\partial Q_n\left(\theta,\sigma^2\right)}{\partial\theta} = -\frac{2}{n}\sum_{i=1}^{n}\sigma^{-2}(X_i)\left[Y_i - \theta' X_i\right]X_i,$$
$$H_n\left(\theta,\sigma^2\right) = \frac{\partial^2 Q_n\left(\theta,\sigma^2\right)}{\partial\theta\partial\theta'} = \frac{2}{n}\sum_{i=1}^{n}\sigma^{-2}(X_i)\,X_i X_i'.$$

We make the following guess for the pathwise derivative of the score,

$$\dot{S}_n[h] = \frac{1}{n}\sum_{i=1}^{n}\dot{s}(Z_i)[h],$$

where

$$\dot{s}(z)[h] = \left[y - \theta_0' x\right] x'\sigma_0^{-4}(x)\,h(x) = d(y,x)\,h(x),$$
$$d(y,x) = \left[y - \theta_0' x\right] x'\sigma_0^{-4}(x),$$

and h is the direction w.r.t. σ^2. We have here suppressed the dependence of $\dot{S}_n[h]$ on θ_0 and ${}'\sigma_0^2(x)$. Observe that

$$d(Y,X) = \left[Y - \theta_0' X\right] X\sigma_0^{-4}(X)\,h(X) = \varepsilon X\sigma_0^{-4}(X)\,h(X).$$

We verify that this satisfies the necessary conditions: First, $h \mapsto \dot{S}_n[h]$ is linear; second, a second order Taylor expansion of the function $a \mapsto 1/a$ yields:

$$\frac{1}{a} - \frac{1}{a_0} = -\frac{1}{a_0^2}(a - a_0) + \frac{2}{\left(\lambda a + (1-\lambda)a_0\right)^3}(a - a_0)^2,$$

for some $\lambda \in [0,1]$. Use this equality with $a = \hat{\sigma}^2(x)$, $a_0 = \sigma_0^2(x)$ to obtain

$$\hat{\sigma}^{-2}(x) - \sigma_0^{-2}(x) = -\sigma_0^{-4}(x)\left(\hat{\sigma}^2(x) - \sigma_0^2(x)\right) + \frac{2}{\left(\lambda_x \hat{\sigma}^2(x) + (1-\lambda_x)\sigma_0^2(x)\right)^3}\left(\hat{\sigma}^2(x) - \sigma_0^2(x)\right)^2,$$

where

$$\left|\frac{2}{\left(\lambda_x \hat{\sigma}^2(x) + (1-\lambda_x)\sigma_0^2(x)\right)^3}\right| \leq \frac{16}{\underline{\sigma}^6},$$

by the same arguments as in the proof of consistency. Thus,

$$\begin{aligned}
&\left\| S_n\left(\theta_0, \hat{\sigma}^2\right) - S_n\left(\theta_0, \sigma_0^2\right) - \dot{S}_n\left[\hat{\sigma}^2 - \sigma_0^2\right] \right\| \\
&\leq \frac{2}{n}\sum_{i=1}^n \left\|\varepsilon_i X_i'\right\| \left|\hat{\sigma}^{-2}(X_i) - \sigma_0^{-2}(X_i) + \sigma_0^{-4}(X_i)\left(\hat{\sigma}^2(X_i) - \sigma_0^2(X_i)\right)\right| \\
&\leq \left\{\frac{2}{n}\sum_{i=1}^n \left\|\varepsilon_i X_i'\right\|\right\} \sup_{x\in\mathcal{X}} \left|\hat{\sigma}^{-2}(x) - \sigma_0^{-2}(x) + \sigma_0^{-4}(x)\left(\hat{\sigma}^2(x) - \sigma_0^2(x)\right)\right|,
\end{aligned}$$

where $n^{-1}\sum_{i=1}^n \|\varepsilon X'\| = E[\|\varepsilon X'\|] + o_P(1)$ by the LLN, while

$$\begin{aligned}
&\sup_{x\in\mathcal{X}} \left|\hat{\sigma}^{-2}(x) - \sigma_0^{-2}(x) + \sigma_0^{-4}(x)\left(\hat{\sigma}^2(x) - \sigma_0^2(x)\right)\right| \\
&\leq \frac{16}{\underline{\sigma}^6} \sup_{x\in\mathcal{X}} \left|\hat{\sigma}^2(x) - \sigma_0^2(x)\right|^2 \\
&= \frac{16}{\underline{\sigma}^6} \left\|\hat{\sigma}^2 - \sigma_0^2\right\|_\infty^2 \\
&= o_P\left(n^{-1/2}\right).
\end{aligned}$$

We conclude

$$\sqrt{n} S_n\left(\theta_0, \sigma^2\right) = \sqrt{n} S_n\left(\theta_0, \sigma_0^2\right) + \sqrt{n}\dot{S}_n\left[\hat{\sigma}^2 - \sigma_0^2\right] + o_P(1).$$

Next, for any h,

$$\begin{aligned}
\dot{S}[h] &= E[\dot{s}(Z)[h]] = E[d(Y,X)h(X)] \\
&= E\left[\varepsilon \sigma_0^{-4}(X) h(X) X'\right] = E\left[E[\varepsilon|X]\sigma_0^{-4}(X) h(X) X'\right] \\
&= 0,
\end{aligned}$$

since $E[\varepsilon|X] = 0$. This implies that the adjustment term is $\delta(Z) \equiv 0$. So if we can verify that

$$\sqrt{n}\left(\dot{S}_n\left[\hat{\sigma}^2 - \sigma_0^2\right] - \dot{S}\left[\hat{\sigma}^2 - \sigma_0^2\right]\right) \to^P 0,$$

we are able to conclude that

$$\sqrt{n}\dot{S}_n\left[\hat{\sigma}^2 - \sigma_0^2\right] = o_P(1).$$

The Hessian satisfies

$$\begin{aligned}\left\|H_n\left(\hat{\sigma}^2\right)-H_n\left(\sigma_0^2\right)\right\| &\leq \frac{2}{n}\sum_{i=1}^{n}\left\|X_i\right\|^2\left|\hat{\sigma}^{-2}\left(X_i\right)-\sigma_0^{-2}\left(X_i\right)\right| \\ &\leq \left\{\frac{2}{n}\sum_{i=1}^{n}\left\|X_i\right\|^2\right\}\sup_{x\in\mathcal{X}}\left|\hat{\sigma}^{-2}\left(x\right)-\sigma_0^{-2}\left(x\right)\right| \\ &= O_P\left(1\right)\times o_P\left(1\right),\end{aligned}$$

and $H_n\left(\sigma_0^2\right)\to^P H_0=E\left[\sigma_0^{-2}\left(X\right)XX'\right]$ by the LLN.

Collecting the above results,

$$\begin{aligned}\sqrt{n}(\hat{\theta}-\theta_0) &= H_n^{-1}\left(\bar{\theta},\hat{\sigma}^2\right)\sqrt{n}S_n\left(\theta_0,\hat{\sigma}^2\right) \\ &= H_n^{-1}\left(\bar{\theta},\hat{\sigma}^2\right)\left\{\sqrt{n}S_n\left(\theta_0,\sigma_0^2\right)+\sqrt{n}\dot{S}_n\left[\hat{\sigma}^2-\sigma_0^2\right]+o_P\left(1\right)\right\} \\ &= H_n^{-1}\left(\bar{\theta},\hat{\sigma}^2\right)\left\{\sqrt{n}S_n\left(\theta_0,\sigma_0^2\right)+o_P\left(1\right)\right\} \\ &\to^d N\left(0,H_0^{-1}\Omega_0H_0^{-1}\right),\end{aligned}$$

where

$$\Omega_0=E\left[\sigma^{-4}\left(X\right)\varepsilon^2XX'\right]=E\left[\sigma^{-2}\left(X\right)XX'\right],$$

such that

$$H_0^{-1}\Omega_0H_0^{-1}=E\left[\sigma^{-2}\left(X\right)XX'\right]^{-1}.$$

Observe that the infeasible WLS,

$$\tilde{\theta}=\arg\min_{\theta\in\Theta}Q_n\left(\theta,\sigma_0^2\right),$$

has the same asymptotic distribution as $\hat{\theta}$. Thus, the feasible WLS estimator based on the nonparametric estimator $\hat{\sigma}^2\left(x\right)$ is asymptotically equivalent to the unfeasible WLS. So one looses no efficiency in substituting σ_0^2 for $\hat{\sigma}^2$ in the estimation of θ. However, note that this does not mean that $\hat{\theta}$ is necessarily the most efficient estimator available, since it does not reach the Cramer-Rao bound (except in the case where the rescaled error term, $\sigma^{-1}\left(X\right)\varepsilon$, is i.i.d. standard Normally distributed).

5.4 Estimation of Variance

The estimation of H_0 can be done by $\hat{H}=H_n(\hat{\theta},\hat{\gamma})$ and is consistent under the conditions given in Theorem 4. If an estimator $\hat{\delta}$ of δ is available then under regularity conditions

$$\hat{\Omega}=\frac{1}{n}\sum_{i=1}^{n}\left[s(Z_i;\hat{\theta},\hat{\gamma})+\hat{\delta}(Z_i)\right]\left[s(Z_i;\hat{\theta},\hat{\gamma})+\hat{\delta}(Z_i)\right]'$$

will be a consistent estimator of Ω_0. Normally, one is able to derive an explicit expression of δ as $\delta\left(z\right)=\delta\left(z;\theta_0,\gamma_0\right)$ in which case a natural choice is $\hat{\delta}(z)=\delta(z;\hat{\theta},\hat{\gamma})$. Using standard

techniques, one can show that the variance estimator will be consistent if s and δ satisfy Lipschitz conditions in (θ, γ), see e.g. Newey and McFadden (1994, Theorem 8.13).

In complicated models however a closed form expression of δ is not available (see, for example, Kristensen, 2009a). One can then either do bootstrapping (Chen et al, 2003) or use numerical methods (Newey, 1994a).

6 Sieve Estimation

While in many cases, semiparametric models can be estimated by a two-step procedure, an alternative approach is to estimate both the parametric and nonparametric component simultaneously. We discuss how this can be done in the context of sieve-estimators.

As in the previous section, we wish to estimate a parameter $\theta \in \Theta \subseteq \mathbb{R}^k$ using a criterion function $Q_n(\theta, \gamma)$ where $\gamma \in \Gamma$ is an infinite-dimensional parameter. Instead of relying on a preliminary estimator (if one such is available at all), and developing a two-step procedure, one can try to estimate θ and γ simultaneously using so-called sieves. The method of sieves is a general nonparametric method where infinite-dimensional function spaces are replaced by approximating, finite-dimensional spaces (a so-called sieve) in finite-samples. The approximation error due to use of a finite-dimensional space vanishes asymptotically by letting the dimension of the sieve increase with sample size.

In order to define the semiparametric sieve estimator, we first need to introduce some additional notation. Suppose we have chosen a sequence of approximating, finite-dimensional spaces $\{\Gamma_J\}$ such that $\Gamma_J \subseteq \Gamma$, $J \geq 1$, and $\overline{\bigcup_{J=1}^{\infty} \Gamma_J} = \Gamma$. We then define

$$(\hat{\theta}, \hat{\gamma}) = \arg \min_{\theta \in \Theta, \gamma \in \Gamma_{J_n}} Q_n(\theta, \gamma), \tag{24}$$

for some sequence $J_n \to \infty$ as $n \to \infty$. Here, we estimate θ and γ simultaneously using the same objective function, Q_n. In contrast, the two-step estimators considered in the previous section used two different objective functions to obtain estimates of θ and γ respectively.

General results establishing consistency and convergence rates for the case where $Q_n(\theta, \gamma)$ takes the form of a sample-average as in eq. (20) can be found in Shen and Wong (1994) and Shen (1997). Moreover, conditions for $\hat{\theta}$ to be $\sqrt{n}$-asymptotically normally distributed are derived in these two papers. GMM-type sieve estimators for models defined through conditional moment conditions are developed and analyzed in Ai and Chen (2003); see also Blundell et al (2007). The conditions to obtain these results are fairly technical however (and so are the proofs) so we will not go into further details here.

One of the disadvantage of the above sieve-approach is the practical implementation. In the two-step estimation, $\hat{\gamma}$ is given as a preliminary estimator, and one therefore only have solve the ("low"-dimensional) optimization problem, $\hat{\theta} = \arg \min_{\theta \in \Theta} Q_n(\theta, \hat{\gamma})$. In contrast, sieve estimators require simultaneous optimization over both θ and γ. In particular, the

dimension of γ can be quite large in standard problems and grows exponentially with the number of variables that it is a function of. Thus, the numerical problem of solving for $(\hat{\theta}, \hat{\gamma})$ in eq. (24) is "high"-dimensional and can be computationally infeasible. However, in many cases, a closed-form solution is available, thereby reducing the numerical problems.

We here give two examples demonstrating how a semiparametric sieve estimator can be implemented.

Example 1 (cont.). With $\gamma = g$, the criterion function for the single-index model is on the form of eq. (20) with q given by

$$q(z; \theta, \gamma) = \left[y - \gamma\left(\beta' x\right)\right]^2.$$

Suppose that Γ is some function space for which there exists a sieve on the form

$$\Gamma_J = \left\{ \gamma_J(z) = \sum_{j=1}^{J} \alpha_j \varphi_j(z) : \alpha_j \in \mathbb{R},\, j = 1, ..., J \right\} \tag{25}$$

where $\varphi_1(z), \varphi_2(z), ...$ are *known* basis functions. The sieve estimator defined in eq. (24) then takes the form

$$(\hat{\beta}, \hat{\gamma}) = \arg\min_{\beta, A_{J_n}} \sum_{i=1}^{n} \left[Y_i - A'_{J_n} \Phi_{J_n}\left(\beta' X_i\right)\right]^2,$$

where $A_{J_n} = (\alpha_1, ..., \alpha_{J_n})'$ and $\Phi_{J_n}(z) = (\varphi_1(z), ..., \varphi_{J_n}(z))'$. For any given value of β, the first order condition w.r.t. A_{J_n} is

$$\sum_{i=1}^{n} \left[Y_i - A'_{J_n} \varphi_{J_n}\left(\theta' X_i\right)\right] \Phi_{J_n}\left(\beta' X_i\right) = 0,$$

which yields the solution

$$\hat{A}_{J_n}(\beta) = \left(\sum_{i=1}^{n} \Phi_{J_n}\left(\beta' X_i\right) \Phi_{J_n}\left(\beta' X_i\right)' \right)^{-1} \sum_{i=1}^{n} \Phi_{J_n}\left(\beta' X_i\right) Y_i.$$

Substituting this in, we then get a profiled estimator:

$$\hat{\beta} = \arg\min_{\theta} \sum_{i=1}^{n} \left[Y_i - \hat{A}_{J_n}(\beta)' \Phi_{J_n}\left(\beta' X_i\right) \right]^2.$$

So here the computational burden is restricted to numerical optimization over θ. Observe that the simultaneous estimator in this case is identical to the two-step estimator where a series estimator is used as a preliminary estimator of γ.

Example 4 (cont.). In the case of the semiparametric copula model, $\gamma = (f_1, f_2)$ and the objective function can again be written as in eq. (20) with q given by

$$q(z; \theta, \gamma) = \left\{ \log c\left(F_1(z_1), F_2(z_2); \theta\right) + \log f_1(z_1) + \log f_2(z_2) \right\}.$$

This estimator was proposed by Chen et al (2006) who also explain how a sieve space for the two densities can be constructed. The sieve estimator can in general not be written on closed form and has to be found by numerical optimization which makes it maybe less attractive. The resulting sieve estimator based on this criterion will in general be more efficient than the two-step estimator proposed in Section 4.

While in the first of the above two examples, the sieve and two-step kernel estimator are very similar, the sieve estimator in general will lead to different estimators. In particular, sieve estimators will in general be more efficient than two-step kernel estimators due to its construction where the parametric and nonparametric component are estimated simultaneously. This leads us to the issue of efficiency of semiparametric estimators:

7 Semiparametric Efficiency

Recall that any semiparametric model is completely characterized by a parametric component, θ_0, and a nonparametric one, $\gamma_0(\cdot)$. The parameter of interest is θ_0, and one may ask how efficiently this parameter can be estimated without any prior knowledge of the nonparametric component, $\gamma_0(\cdot)$. This is in general a hard question to answer, but a constructive approach has been to compute bounds on the level of efficiency for θ_0.

The intuition behind the bounds that we are going to introduce is the following: Consider the estimation of two statistical models where the second model is contained (nested) within the first one. Clearly, we expect the estimation of the second model to be an easier problem than the estimation of the first one. In particular, if the two models share a common parameter, say θ, we expect this parameter to be estimated more precisely in the second model. Thus, if we can evaluate the efficiency of the estimation of θ in the second model, this will give us a bound for the efficiency of θ in the first model.

Stein (1956) used the above idea to construct efficiency bounds for semiparametric estimation problems. As the first, more complicated, model, he chose the semiparametric model of interest. This is characterized by $(\theta_0, \gamma_0(\cdot))$. As the second, simpler model, he then introduced a fully parametric submodel: Choose some parametric family of functions, $\gamma(\cdot;\alpha)$ where $\alpha \in \mathcal{A} \subseteq \mathbb{R}^l$ is the parameter, and suppose that the parametric submodel contains the true function $\gamma_0(\cdot)$, $\gamma_0(\cdot) = \gamma(\cdot;\alpha_0)$ for some $\alpha_0 \in \mathcal{A}$. Thus, the second model is characterized by (θ_0, α_0).

The estimation of the semiparametric model should clearly be at least as hard as the estimation of the fully parametric submodel. Thus, we cannot expect to be able to estimate θ_0 with higher precision in the semiparametric model. Since parametric submodel is fully specified in terms of (θ, α), we can write up the density of the model as a function of (θ, α), $(\theta, \alpha) \mapsto p(z; \theta, \gamma(\cdot;\alpha))$. A natural estimator is then the MLE, and the precision of the MLE

is determined by the associated Fisher information,

$$\mathcal{I} = E\left[\frac{\partial^2 \log p\left(Z;\theta,\gamma\left(\cdot;\alpha\right)\right)}{\partial\left(\theta,\alpha\right)\partial\left(\theta,\alpha\right)'}\right] = \left[\begin{array}{cc} \mathcal{I}_{\theta\theta} & \mathcal{I}_{\theta\alpha} \\ \mathcal{I}_{\theta\alpha} & \mathcal{I}_{\alpha\alpha} \end{array}\right].$$

For any given parametric specification, $\gamma\left(\cdot;\alpha\right)$, the efficiency level for θ is therefore given by the Cramer-Rao bound,

$$\mathcal{I}_{\mathrm{p}} = \mathcal{I}_{\theta\theta} - \mathcal{I}_{\theta\alpha}\mathcal{I}_{\alpha\alpha}^{-1}\mathcal{I}_{\theta\alpha}.$$

That is, the asymptotic variance of the MLE is $\mathcal{I}_{\mathrm{p}}^{-1}$. This variance expression quantifies the price we have to pay for not knowing $\gamma\left(\cdot\right)$ (corresponding to α in the parametric model): If α is known, θ can be estimated with asymptotic variance $\mathcal{I}_{\theta\theta}^{-1}$. If α is unknown and has to be estimated, the variance becomes $\mathcal{I}_{\mathrm{p}}^{-1} \geq \mathcal{I}_{\theta\theta}^{-1}$ with equality if and only if $\mathcal{I}_{\theta\alpha} = 0$.

Consider now an estimator not relying on any parametric information regarding $\gamma\left(\cdot\right)$, and let $\mathcal{I}_{\mathrm{sp}}^{-1}$ be its asymptotic variance. Then it must hold that $\mathcal{I}_{\mathrm{p}}^{-1} \leq \mathcal{I}_{\mathrm{sp}}^{-1}$. This will hold regardless of how the parametric submodel has been chosen such that $\sup_{\gamma(\cdot;\alpha)} \mathcal{I}_{\mathrm{p}}^{-1} \leq \mathcal{I}_{\mathrm{sp}}^{-1}$. This leads to the following definition of the semiparametric variance (or efficiency) bound as the asymptotic variance of the "least favourable" parametric submodel:

$$\text{semiparametric variance bound (SVB)} = \sup_{\gamma(\cdot;\alpha)} \mathcal{I}_{\mathrm{p}}^{-1}.$$

A particular attractive class of semiparametric estimators are those that perform as well as if we actually knew the nonparametric component: We call a (semiparametric) estimator $\hat{\theta}$ *adaptive* if

$$\sqrt{n}(\hat{\theta} - \theta_0) \rightarrow^d N\left(0, \mathcal{I}_{\theta\theta}^{-1}\right).$$

A necessary condition for this to hold is that $\mathcal{I}_{\theta\alpha} = 0$ for all parametric submodels, which in general is not satisfied. For example, none of the semiparametric estimators considered in Section 2 are adaptive. On the other hand, the MLE-type estimators introduced in Section 3.3. are indeed adaptive under certain regularity assumptions on the error distribution.

To see how the efficiency bound is linked to the asymptotic results of the previous section, recall that we found that the variance of the semiparametric estimator to be on the form $H_0^{-1}\Omega_0 H_0^{-1}$ where $H_0 = \mathcal{I}_{\theta\theta}$ and

$$\Omega_0 = E\left[\left\{s\left(Z;\theta_0,\gamma_0\right) + \delta\left(Z\right)\right\}\left\{s\left(Z;\theta_0,\gamma_0\right) + \delta\left(Z\right)\right\}'\right].$$

Here, $s\left(Z;\theta_0,\gamma_0\right) = \partial \log p\left(Z;\theta,\gamma_0\right)/\left(\partial\theta\right)$, while $\delta\left(Z\right)$ is the adjustment term due to the fact that we are using an estimator of γ_0 instead of the true value itself. The semiparametric efficiency bound is then roughly speaking a question of how you design the estimator to obtain the "smallest" possible δ in terms of variance. In particular, if $\delta = 0$, the estimator is *adaptive,* c.f. definition above, since it performs just as well as if we actually knew γ_0.

It should be noted here that there is no guarantee that there actually exists a semiparametric estimator that reaches the efficiency bound. As such the bound is not necessarily sharp. Examples of this situation can be found in Ritov and Bickel (1987) where the semiparametric efficiency bound is well-defined, but no $\sqrt{n}$-consistent semiparametric estimator exists.

While the efficiency bound makes intuitive sense, it is in general difficult to derive an explicit expression of it for general semiparametric problems. We will therefore not attempt to derive any efficiency bounds here. Instead, we will here try to give some more intuition for the efficiency bound by showing how this changes according to what assumptions the research is willing to impose on the model. As a simple example, consider the following semiparametric regression model,

$$Y = m(X;\theta) + \varepsilon,$$

where the conditional mean is fully parametrized, but the only restrictions on the errors are that $E[\varepsilon|X] = 0$ and $E\left[\varepsilon^2\right] < \infty$. In this case $\gamma = F_{\varepsilon|X}(e|x)$ is the nonparametric component. Depending on what additional assumptions the researcher is willing to make regarding $F_{\varepsilon|X}$, different efficiency bounds appears. If for example, we are only willing to impose the conditional mean restriction that $E[\varepsilon|X] = 0$, the SVB is given by

$$\text{SVB} = E\left[\sigma^{-2}(X)\,\dot{m}(X;\theta_0,\gamma_0)\,\dot{m}(X;\theta_0,\gamma_0)'\right]^{-1}, \tag{26}$$

where $\dot{m}(X;\theta,\gamma) = \partial m(X;\theta,\gamma)/(\partial\theta)$, and $\sigma^2(X) = E\left[\varepsilon^2|X\right]$ is the conditional variance. This can for example be reached using the general sieve-estimator of Ai and Chen (2003).

If on the other hand, one is willing to make the stronger assumption of independence between ε and X and that ε has a symmetric distribution, the efficiency bound becomes

$$\text{SVB} = E\left[\left(\frac{f_\varepsilon'(\varepsilon)}{f_\varepsilon(\varepsilon)}\right)^2 \dot{m}(X;\theta_0,\gamma_0)\,\dot{m}(X;\theta_0,\gamma_0)'\right]^{-1}, \tag{27}$$

where $f_\varepsilon(\varepsilon)$ is the density of ε. It is easily checked that this variance bound is smaller than the the one obtained in eq. (26). The intution behind this is that stronger restrictions on the model gives more information about the parameter of interest. In particular, the efficiency bound in eq. (27) is equal to the Cramer-Rao bound. An adaptive estimator can be developed along the same lines as done in Section 3.3.

8 Notes

The estimator of the single-index model was proposed in Ichimura (1993) who also derived its theoretical properties. The asymptotic theory for the average-derivative estimator were developed in Powel et al (1989); see also Hristache et al (2001). In the binary choice case, one

can alternatively use maximum likelihood methods to estimate β_0 in the single-index model, c.f. Klein and Spady (1993).

Robinson (1988b) and Speckman (1988) proposed the residual-based estimator of the partially linear model given in eq. (9) and derived its asymptotic distribution. Andrews (1994a) give results for the extended version in eq. (10).

Robinson (1987) derived the asymptotics of the WLS-estimator in the presence of heteroskedasticity of unknown form, and showed that his estimator reached the semiparametric efficiency bound. Ai and Chen (2003) propose semiparametrically efficient sieve estimators for a class of semiparametric models described by conditional moment restrictions.

For an introduction to and general results on copulas, see Joe (1997). The properties of the semiparametric copula estimator in Section 4 were derived in Genest et al. (1995), while the properties of the sieve estimator in Section 6 were analyzed in Chen et al (2008).

For further reading on functional derivatives, see e.g. Luenberger (1969) and Kantorovich and Akilov (1982).

Our asymptotic results for semiparametric two-step estimators are similar to those found in, amongst others, Andrews (1994a), Chen, Linton and van Keilegom (2003), Newey and McFadden (1994), Newey (1994b), Pakes and Olley (1995). These studies all give general conditions for consistency and asymptotic normality of two-step semiparametric estimators. For higher-order properties of semiparametric estimators, we refer to Linton (1995,1996).

For results on nonparametric sieve estimators, we refer to Andrews (1991), Fenton and Gallant (1996), Gallant and Nychka (1987), Newey (1997) and Shen and Wong (1994). For their use in semiparametric estimation, see Ai and Chen (2003), and Shen (1997). Chen (2007) give an overview of both non- and semiparametric estimation using sieve methods.

Newey (1990) gives a good introduction to semiparametric efficiency bounds, and how to derive these; general approaches to computing efficiency bounds can be found in Bickel et al (1993) and Severini and Tripathi (2001). Chamberlain (1987, 1992) derive efficiency bounds for conditional moment restrictions and semiparametric regressions. Manski (1984) develop efficiency bounds and adaptive estimators for nonlinear regression models under independence assumption; see Drost and Klaassen (1997) for similar results in the case of heteroskedastic time series models.

We have not discussed the practical implementation of semiparametric estimators. We refer to Ichimura and Todd (2007) for an overview. Cattaneo et al (2009) discuss in detail bandwidth selection for average derivative estimators, while Härdle et al (1993) propose a specific method for the single-index models.

We have throughout assumed that data was i.i.d. Most of the asymptotic results for the proposed estimators go through for stationary and mixing sequences; see e.g. Ang and Kristensen (2009), Chen et al. (2009), Hidalgo (1992), Kristensen (2009a) and Li and Wooldridge (2002). The issue of semiparametric efficiency bounds for time series models is however not

very well-developed once you leave the Markov setting and allow for general dependence; see Bickel and Kwon (2001) and Schick and Wefelmeyer (2005) for discussions and some results.

References

Ai, C. and X. Chen (2003) Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions. *Econometrica* 71, 1795-1844.

Aït-Sahalia, Y. (1993) The Delta Method for Nonparametric Kernel Functionals. Manuscript, University of Chicago.

Amemiya, T. (1985) Non-Linear Regression Models. In *Handbook of Econometrics,* Vol. 1 (eds. M.D. Intriligator and Z. Griliches), 333-389. Elsevier.

Amemiya, T. (1985) *Advanced Econometrics*. Harvard University Press.

Andrews, D.W.K. (1991) Asymptotic Normality of Series Estimators for Nonparametric and Semiparametric Regression Models. *Econometrica* 59, 307-45.

Andrews, D.W.K. (1994a) Asymptotics for Semiparametric Econometric Models via Stochastic Equicontinuity. *Econometrica* 62, 43-72.

Andrews, D.W.K. (1994b) Empirical Process Methods in Econometrics. In *Handbook of Econometrics,* Vol. 4 (eds. R.F. Engle and D.L. McFadden), 2246-2294. North-Holland.

Ang, A. and D. Kristensen (2009) Testing Conditional Factor Models. CREATES Research Papers 2009-09, University of Aarhus.

Bickel, P.J., C.A.J. Klaassen, Y. Ritov & J.A. Wellner (1993) *Efficient and Adaptive Estimation for Semiparametric Models*. The John Hopkins University Press.

Bickel, P.J. and J. Kwon (2001) Inference for Semiparametric Models: Some Questions and An Answer. *Statistica Sinica* 11, 863-960

Blundell, R., X. Chen and D. Kristensen (2007) Semi-Nonparametric IV Estimation of Shape Invariant Engel Curves. *Econometrica* 75, 1613-1670.

Cattaneo, M.D., R.K. Crump and M. Jansson (2009) Small Bandwidth Asymptotics for Density-Weighted Average Derivatives. Manuscript, Department of Economics, UC Berkeley.

Chamberlain, G. (1987) Asymptotic Efficiency in Estimation of Conditional Moment Restrictions. *Journal of Econometrics* 34, 305-334.

Chamberlain, G. (1992) Efficiency Bounds for Semiparametric Regression. *Econometrica* 60, 567-596.

Chen, X. (2007) Large Sample Sieve Estimation of Semi-nonparametric Models. In *Handbook of Econometrics*, Vol. 6B (eds. J.J. Heckman and E.E. Lleamer), 5549-5635. North-Holland.

Chen, X., Y. Fan and V. Tsyrennikov (2006) Efficient Estimation of Semiparametric Multivariate Copula Models. *Journal of the American Statistical Association* 101, 1228-1240.

Chen, X., O. Linton and I. van Keilegom (2003) Estimation of Semiparametric Models when the Criterion Function is not Smooth. *Econometrica* 71, 1591-1608.

Chen, X., W.B. Wu and Y. Yu (2009) Efficient Estimation of Copula-Based Semiparametric Markov Models. Cowles Foundation Discussion Papers, No. 1691.

Drost, F.C. and C.A.J. Klaassen (1997) Efficient estimation in semiparametric GARCH models. *Journal of Econometrics* 81, 193–221.

Fenton, V.M. & A.R. Gallant (1996) Convergence Rates of SNP Density Estimators. *Econometrica* 64, 719-727.

Gallant, A.R. & D.W. Nychka (1987) Semi-Nonparametric Maximum Likelihood Estimation. *Econometrica* 55, 363-390.

Genest, C., K. Ghoudi and L.-P. Rivest (1995) A Semiparametric Estimation Procedure of Dependence Parameters in Multivariate Families of Distributions. *Biometrika* 82, 543-552.

Genest, C., M. Gendron and M. Bourdeau-Brien (2009) The Advent of Copulas in Finance. Forthcoming in *European Journal of Finance* 15.

Härdle, W. (1990) *Applied Nonparametric Regression*. Cambridge: Cambridge University Press.

Härdle, W., P. Hall, H. Ichimura (1993) Optimal Smoothing in Single-index Models. *Annals of Statistics* 21, 157-178.

Härdle, W. and O. Linton (1994) Applied Nonparametric Methods. In *Handbook of Econometrics,* Vol. 4 (eds. R.F. Engle and D.L. McFadden), 2295-2339. North-Holland.

Härdle, W, M. Müller, S. Sperlich and A. Werwatz (2004) *Nonparametric and Semiparametric Models*. New York: Springer-Verlag.

Hidalgo, J. (1992) Adaptive Estimation in Time Series Regression Models with Heteroskedasticity of Unknown Form. *Econometric Theory* 8, 161-187.

Horowitz, J. (2009) *Semiparametric and Nonparametric Methods in Econometrics*. Springer-Verlag.

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

Ichimura, H. (1993) Semiparametic Least Squares (SLS) and Weighted SLS Estimation of Single-Index Models. *Journal of Econometrics* 58, 71-120.

Ichimura, H. and P.E. Todd (2007) Implementing Nonparametric and Semiparametric Estimators. In *Handbook of Econometrics*, Vol. 6B (eds. J.J. Heckman and E.E. Lleamer), 5369-5468. North-Holland.

Joe, H. (1997) *Multivariate Models and Dependence Concepts*. Chapman and Hall.

Kantorovich, L.V. and G.P. Akilov(1982) *Functional Analysis*. Pergamon Press, Oxford.

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

Kristensen, D. (2009a) Pseudo-Maximum-Likelihood Estimation in Two Classes of Semiparametric Diffusion Models. CREATES Research Papers 2009-41, University of Aarhus.

Kristensen, D. (2009b) Uniform Convergence Rates of Kernel Estimators with Heterogeneous, Dependent Data. *Econometric Theory* 25, 1433-1445.

Lee, A.J. (1990) *U-Statistics, Theory and Practice*. Marcel Dekker.

Li, Q. and J.S. Racine (2007) *Nonparametric Econometrics: Theory and Practice*. Princeton University Press.

Li, Q. and J.M. Wooldridge (2002) Semiparametric Estimation of Partially Linear Models for Dependent Data with Generated Regressors. *Econometric Theory* 18, 625-645.

Linton, O.B. (1995) Second Order Approximation in the Partially Linear Regression Model. *Econometrica* 63, 1079-1112.

Linton, O.B. (1996) Edgeworth Approximation for MINPIN Estimators in Semiparametric Regression Models. *Econometric Theory* 12, 30-60.

Luenberger, D. G. (1969) *Optimization by Vector Space Methods*. John Wiley.

Manski, C. (1984) Adaptive estimation of non-linear regression. *Econometric Reviews* 3, 145-194.

Newey, W.K. (1990) Semiparametric Efficiency Bounds. *Journal of Applied Econometrics* 5, 99-135.

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

Newey, W.K. (1994a) Kernel Estimation of Partial Means and a General Variance Estimator. *Econometric Theory* 10, 233-253.

Newey, W.K. (1994b) The Asymptotic Variance of Semiparametric Estimators. *Econometrica* 62, 1349-1362.

Newey, W.K. (1997) Convergence Rates and Asymptotic Normality for Series Estimators. *Journal of Econometrics* 79, 147-168.

Newey, W.K. and D.L. McFadden (1994) Large Sample Estimation and Hypothesis Testing. In *Handbook of Econometrics,* Vol. 4 (eds. R.F. Engle and D.L. McFadden), 2111-2245. North-Holland.

Pagan, A. and A. Ullah (1999) *Nonparametric Econometrics*. Cambridge University Press.

Pakes, A. and S. Olley (1995) A Limit Theorem for a Smooth Class of Semiparametric Estimators. *Journal of Econometrics* 65, 295-332.

Powell, J.L. (1994) Estimation of Semiparametric Models. In *Handbook of Econometrics,* Vol. 4 (eds. R.F. Engle and D.L. McFadden), 2443-2521. North-Holland.

Powell, J.L., J.H. Stock and T.M. Stoker (1989). Semiparametric Estimation of Index Coefficients. *Econometrica* 51, 1403-1430.

Robinson, P.M. (1987) Asymptotically Efficient Estimation in the Presence of Heteroskedasticity of Unknown Form. *Econometrica* 55, 875-891.

Robinson, P.M. (1988a) Semiparametric Econometrics: A Survey. *Journal of Applied Econometrics* 3, 35-51.

Robinson, P.M. (1988b) Root-N-Consistent Semiparametric Regression. *Econometrica* 56, 931-954.

Severini, T.A. & G. Tripathi (2001) A Simplified Approach to Computing Efficiency Bounds in Semiparametric Models. *Journal of Econometrics* 102, 23-66.

Shen, X. (1997) On Methods of Sieves and Penalization. *Annals of Statistics* 25, 2555-2591.

Shen, X. and W.H. Wong (1994) Convergence Rate of Sieve Estimates. *Annals of Statistics* 22, 580-615.

Schick, A. and W. Wefelmeyer (2006) Efficient Estimators for Time Series. In *Frontiers in Statistics* (eds. J. Fan and H. L. Koul), 45-62. Imperial College Press.

Silverman, B.W. (1986) *Density Estimation for Statistics and Data Analysis*. Chapman and Hall.

Speckman, P. (1988) Kernel Smoothing in Partial Linear Models. *Journal of the Royal Statistical Society, Series B* 50, 413-436.

van der Vaart, A & J. Wellner (1996) *Weak Convergence and Empirical Processes*. Springer-Verlag.

A Proofs

Proof of Theorem 1. We wish to show that for any $\varepsilon > 0$, $P(||\hat{\theta} - \theta_0|| > \varepsilon) \to 0$ as $n \to \infty$. Let $\varepsilon > 0$ be given; then by (C.3), there exists a $\delta > 0$ such that $\|\theta - \theta_0\| > \varepsilon$ implies $Q(\theta, \gamma_0) \geq Q(\theta_0, \gamma_0) + \delta$, which in turn implies $|Q(\theta, \gamma_0) - Q(\theta_0, \gamma_0)| \geq \delta$. Thus,

$$P(||\hat{\theta} - \theta_0|| > \varepsilon) \leq P(Q(\hat{\theta}, \gamma_0) \geq Q(\theta_0, \gamma_0) + \delta) \leq P(|Q(\hat{\theta}, \gamma_0) - Q(\theta_0, \gamma_0)| \geq \delta).$$

We then have to show that the RHS converges to zero which is equivalent to $Q(\hat{\theta}, \gamma_0) \to^P Q(\theta_0, \gamma_0)$. Since θ_0 is the unique minimiser of $Q(\theta, \gamma_0)$, we know that $Q(\theta_0, \gamma_0) \leq Q(\hat{\theta}, \gamma_0)$. Thus,

$$\begin{aligned} |Q(\hat{\theta}, \gamma_0) - Q(\theta_0, \gamma_0)| &= Q(\hat{\theta}, \gamma_0) - Q(\theta_0, \gamma_0) \\ &= \left\{Q(\hat{\theta}, \gamma_0) - Q_n(\hat{\theta}, \hat{\gamma})\right\} + \left\{Q_n(\hat{\theta}, \hat{\gamma}) - Q(\theta_0, \gamma_0)\right\} \\ &=: B_1 + B_2.. \end{aligned}$$

The first term on the right hand side of the last equation can be written as:

$$B_1 = \left\{Q_n(\hat{\theta}, \gamma_0) - Q_n(\hat{\theta}, \hat{\gamma})\right\} + \left\{Q(\hat{\theta}, \gamma_0) - Q_n(\hat{\theta}, \gamma_0)\right\},$$

while, using that $Q_n(\hat{\theta}, \hat{\gamma}) \leq Q_n(\theta_0, \hat{\gamma})$, the second term can is bounded by

$$B_2 \leq Q_n(\theta_0, \hat{\gamma}) - Q(\theta_0, \gamma_0) = \{Q_n(\theta_0, \hat{\gamma}) - Q_n(\theta_0, \gamma_0)\} + \{Q_n(\theta_0, \gamma_0) - Q(\theta_0, \gamma_0)\}.$$

Thus,

$$|B_i| \leq \sup_{\theta \in \Theta} |Q_n(\theta, \hat{\gamma}) - Q_n(\theta, \gamma_0)| + \sup_{\theta \in \Theta} |Q_n(\theta, \gamma_0) - Q(\theta, \gamma_0)|,$$

for $i = 1, 2$. It now follows from C.1 and C.2 that $|B_i| = o_P(1)$, $i = 1, 2$. In conclusion, $|Q(\hat{\theta}, \gamma_0) - Q(\theta_0, \gamma_0)| = o_P(1)$ as desired. ■

Proof of Theorem 4. A Taylor expansion of the score function $S_n(\hat{\theta}, \hat{\gamma})$ w.r.t. θ around θ_0 yields:

$$0 = S_n(\hat{\theta}, \hat{\gamma}) = S_n(\theta_0, \hat{\gamma}) + H_n(\bar{\theta}, \hat{\gamma})(\hat{\theta} - \theta_0),$$

where $\bar{\theta} \in [\hat{\theta}, \theta_0]$ is some intermediate point. Next, make a (functional) Taylor expansion of $S_n(\theta_0, \hat{\gamma})$ w.r.t. γ around γ_0,

$$S_n(\theta_0, \hat{\gamma}) = S_n(\theta_0, \gamma_0) + \dot{S}_n(\theta_0; \gamma_0)[\hat{\gamma} - \gamma_0] + R_n,$$

where, by Assumptions (N.1). and (N.4), the remainder term

$$R_n = O_P(B_n||\hat{\gamma} - \gamma_0||^2) = O_P(1) \times O_P(||\hat{\gamma} - \gamma_0||^2) = o_P(1/\sqrt{n}).$$

Combining this with Assumption (N.6)-(N.7), we obtain

$$\sqrt{n}(\tilde{\theta}-\theta_0)=H_0^{-1}\sqrt{n}\left\{S_n(\theta_0,\gamma_0)+\dot{S}_n(\theta_0,\gamma_0;\hat{\gamma}-\gamma_0)\right\}+o_P(1),$$

and the desired result follows from Assumption (N.5). ■

Proof of Theorem 5. First, by Condition N.4',

$$0=S_n(\hat{\theta},\hat{\gamma})=S_n(\theta_0,\gamma_0)+S(\hat{\theta},\hat{\gamma})+o_P\left(1/\sqrt{n}\right),$$

where, by a Taylor expansion w.r.t. θ,

$$S(\hat{\theta},\hat{\gamma})=S(\theta_0,\hat{\gamma})+H(\bar{\theta},\hat{\gamma})(\hat{\theta}-\theta_0),$$

for some intermediate point $\bar{\theta}\in[\tilde{\theta},\theta_0]$. Next, we expand $S(\theta_0,\hat{\gamma})$ w.r.t. γ around γ_0,

$$S(\theta_0,\hat{\gamma})=S(\theta_0,\gamma_0)+\dot{S}(\theta_0,\gamma_0)[\hat{\gamma}-\gamma_0]+R_n,$$

where the remainder term $R_n=O(||\hat{\gamma}-\gamma_0||^2)=o_P(1/\sqrt{n})$ by (N.5') and (N.1). Combining these results with (N.7'),

$$\sqrt{n}(\hat{\theta}-\theta_0)=H_0^{-1}\sqrt{n}\left\{S_n(\theta_0,\gamma_0)+\dot{S}(\theta_0,\gamma_0)[\hat{\gamma}-\gamma_0]\right\}+o_P(1),$$

and the desired result follows from Condition (N.6'). ■

Proof of Lemma 6. Define

$$\dot{S}_n(\theta_0;\gamma_0)[\gamma-\gamma_0]=\frac{1}{n}\sum_{i=1}^{n}\dot{s}(Z_i;\theta_0,\gamma_0)[\gamma-\gamma_0],$$

where $\dot{s}(z;\theta_0,\gamma_0)[\gamma-\gamma_0]$ is given in (N.4.i). Then,

$$\begin{aligned}&\left\|S_n(\theta_0,\gamma)-S_n(\theta_0,\gamma_0)-\dot{S}_n(\theta_0;\gamma_0)[\gamma-\gamma_0]\right\|\\&\leq\frac{1}{n}\sum_{i=1}^{n}\left\|s(Z_i;\theta_0,\gamma)-s(Z_i;\theta_0,\gamma_0)-\dot{s}(Z_i;\theta_0,\gamma_0)[\gamma-\gamma_0]\right\|\\&\leq B_n\left\|\gamma-\gamma_0\right\|^2,\end{aligned}$$

where $B_n:=\sum_{i=1}^{n}b(z_i)/n=E[b(z)]+o_P(1)$ by the Law of Large Numbers. This shows N.4.

Next, by (N.5.i)-(N.5.ii),

$$\begin{aligned}\sqrt{n}\left\{S_n(\theta_0,\gamma_0)+\dot{S}_n(\theta_0;\gamma_0)[\hat{\gamma}-\gamma_0]\right\}&=\sqrt{n}\left\{S_n(\theta_0,\gamma_0)+\dot{S}(\theta_0,\gamma_0)[\hat{\gamma}-\gamma_0]\right\}+o_P(1)\\&=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{s(Z_i;\theta_0,\gamma_0)+\delta(Z_i)\}+o_P(1),\end{aligned}$$

where, by the Central Limit Theorem,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{s(Z_i;\theta_0,\gamma_0)+\delta(Z_i)\}\to^d N(0,\Omega),$$

where Ω is given in the lemma. ■

Research Papers 2009

2009-31: Eduardo Rossi and Paolo Santucci de Magistris: A No Arbitrage Fractional Cointegration Analysis Of The Range Based Volatility

2009-32: Alessandro Palandri: The Effects of Interest Rate Movements on Assets' Conditional Second Moments

2009-33: Peter Christoffersen, Redouane Elkamhi, Bruno Feunou and Kris Jacobs: Option Valuation with Conditional Heteroskedasticity and Non-Normality

2009-34: Peter Christoffersen, Steven Heston and Kris Jacobs: The Shape and Term Structure of the Index Option Smirk: Why Multifactor Stochastic Volatility Models Work so Well

2009-35: Peter Christoffersen, Jeremy Berkowitz and Denis Pelletier: Evaluating Value-at-Risk Models with Desk-Level Data

2009-36: Tom Engsted and Thomas Q. Pedersen: The dividend-price ratio does predict dividend growth: International evidence

2009-37: Michael Jansson and Morten Ørregaard Nielsen: Nearly Efficient Likelihood Ratio Tests of the Unit Root Hypothesis

2009-38: Frank S. Nielsen: Local Whittle estimation of multivariate fractionally integrated processes

2009-39: Borus Jungbacker, Siem Jan Koopman and Michel van der Wel: Dynamic Factor Models with Smooth Loadings for Analyzing the Term Structure of Interest Rates

2009-40: Niels Haldrup, Antonio Montañés and Andreu Sansó: Detection of additive outliers in seasonal time series

2009-41: Dennis Kristensen: Pseudo-Maximum Likelihood Estimation in Two Classes of Semiparametric Diffusion Models

2009-42: Ole Eiler Barndorff-Nielsen and Robert Stelzer: The multivariate supOU stochastic volatility model

2009-43: Lasse Bork, Hans Dewachter and Romain Houssa: Identification of Macroeconomic Factors in Large Panels

2009-44: Dennis Kristensen: Semiparametric Modelling and Estimation: A Selective Overview