An Efficient Nonparametric Estimator
for Models with Nonlinear Dependence

Patrick Gagliardini*

(University of Lugano and University of St. Gallen, Switzerland)

Christian Gouriéroux

(CREST, CEPREMAP, France, and University of Toronto, Canada)

First Version: June 2002

This Version: March 2005

Abstract

We provide a convenient econometric framework for the analysis of nonlinear dependence in financial applications. We introduce models with constrained nonparametric dependence, which specify the conditional distribution or the copula in terms of a one-dimensional functional parameter. Our approach is intermediate between standard parametric specifications (which are in general too restrictive) and the fully unrestricted approach (which suffers from the curse of dimensionality). We introduce a nonparametric estimator defined by minimizing a chi-square distance between the constrained densities in the family and an unconstrained kernel estimator of the density. We derive the nonparametric efficiency bound for linear forms and show that the minimum chi-square estimator is nonparametrically efficient for linear forms.

Keywords: Nonlinear Dependence, Copula, Nonparametric Estimation, Efficiency.

JEL classification: C14, C51.

---

*Corresponding author: Patrick Gagliardini, Swiss Institute of Banking and Finance, Rosenbergstrasse 52, CH-9000 St. Gallen, Switzerland. Tel.: 0041 71 224 7010. Fax: 0041 71 224 7088. Email: patrick.gagliardini@unisg.ch.

1 Introduction

Recent developments in risk management emphasize the need to carefully assess nonlinear dependence between risks. Typical examples are the dependence:

i) between the default risk of different firms to capture the so-called default correlation, that is some clustering in corporate failure (see Duffie and Singleton, 1999; Li, 2000; Schönbucher and Schubert, 2001; Gouriéroux and Monfort, 2002a), and more generally the joint migration between rating categories (Gagliardini and Gouriéroux, 2005);

ii) between the extreme risks in different budget lines of a bank's balance sheet, in order to aggregate the Value at Risk (VaR), and the required capital, computed per line (Embrechts et al., 2003);

iii) between intertrade durations to detect clustering effects in trading activity and analyse the liquidity risk (Engle and Russell, 1998; Ghysels et al., 2004).

In most of these problems the nonlinear dependence relates to the whole joint distribution of the variables (not only the first conditional moments) and the main concern is often about the tail of the joint distribution, as when the required capital is introduced to hedge extreme risks. Moreover, these problems generally involve a rather large number of variables. Indeed in example i) above the number of firms may run well over hundred, and in example ii) the number of budget lines is typically between ten and twenty.

Different approaches have been proposed in the econometric and statistical literature to describe nonlinear dependence. A considerable attention has been recently devoted to methods based on the joint distribution of the risk variables, such as copulas (see Nelsen, 1999, for a textbook introduction and Chen and Fan, 2003, for a recent application in time series modeling), especially in the framework of financial risk management. In this context parametric specifications are typically adopted. However, the dependence between financial variables such as times to default, returns, or intertrade durations is not well-captured by the standard parametric families of densities proposed in the literature (see Joe, 1997, for a survey). Indeed, these parametric specifications are often excessively constrained (implying a poor data fit) and not very appropriate for a separate analysis of dependence distinguishing between the risks at low, medium and high level [as required in example ii)]. Finally, their parameters often do not admit a clear structural interpretation for financial applications.

The alternative approach for modeling nonlinear dependence consists in estimating nonparametrically the unrestricted joint density; see Silverman (1986) and Scott (1992) for surveys on density estimation and Deheuvels (1981) and Fermanian and Scaillet (2003) for bivariate copulas. The weakness of this approach is that the absence of any structure complicates the interpretation of the patterns of nonlinear dependence, especially when more than 2 variables are considered, since the joint density is hard to visualize. Moreover

2

this approach suffers from the curse of dimensionality when the number of variables of interest is larger than 4 or 5. Finally, even in the bivariate case, it can provide very inaccurate and erratic results for the VaR [see example ii)], which is evaluated by considering the rather extreme observations.

In this paper we explore the intermediate approach in which the joint density is constrained and depends on a small number of one-dimensional functional parameters, that is to say functions of one variable. In this case the density is parameterized by means of a (vector-valued) function A defined on a subset of $\mathbb{R}$. The aim of this paper is to provide efficient nonparametric estimators for the one-dimensional functional parameters A that characterize nonlinear dependence.

Such a constrained nonparametric approach has three major advantages for modeling nonlinear dependence. First, by using functional parameters instead of scalar parameters, we achieve a higher flexibility and a better data fit. A clear structural interpretation of nonlinear dependence is maintained, since in applied examples the functional parameters generally correspond to the distribution of a random variable such as a latent factor, an omitted heterogeneity, or an innovation, or they represent for instance the response function of a lagged variable. Second, the graphical representation of the one-dimensional functional parameters highlights the patterns of nonlinear dependence. To give an example, in the case of dynamic proportional hazard models used for the analysis of liquidity risk [see example iii) above, and example iii) in Section 2.2], the serial dependence in the whole sample depends on the elasticity of the functional parameter, while tail dependence is revealed by its behaviour close to the boundary points of its support. Thus, compared to the fully unrestricted approach, the interpretation of nonlinear dependence is made easier. At the same time, functional parameters allow for a much richer description of nonlinear dependence compared to standard finite-dimensional parameters. Finally, as a third advantage, we show in the paper that the rate of convergence of appropriate estimators, both for functional parameters and joint density, is the standard one-dimensional nonparametric rate, and is independent of the number of underlying variables of interest.

Constrained nonparametric densities have already been analyzed in the literature, under various restrictions. A typical example is the transformation model, in which an unknown transformation of the endogenous variable satisfies a linear regression model with iid errors (see Han, 1987a, b; Horowitz, 1996), or the location-scale model in which the mean and the volatility are unrestricted functions of a set of regressors (Härdle and Tsybakov, 1997). To avoid the curse of dimensionality when the number of regressors is high, these models typically adopt additivity assumptions (see Hastie and Tibshirani, 1990), or assume an index structure (see Powell et al., 1989; Ichimura, 1993). These constrained nonparametric regressions are suitable for describing dependence between an endogenous variable and a set of regressors, but not for modeling dependence between several endogenous variables, such as times to default for several borrowers [as in example i)]. Moreover, they have been introduced as extensions of the standard linear model, which

explains the form of the index function, which is often linear and neglects cross effects. Finally they assume that the same index matters for the extreme and standard values of the endogenous variable. Our purpose is to consider other types of nonparametric constraints better suited for financial or duration analysis, and admitting structural interpretations, for instance in terms of factors, or omitted heterogeneity.

The paper is organized as follows. In Section 2 we consider the specification of constrained nonparametric families. In particular, we discuss the approaches where the conditional distribution, or an equivalent representation of the joint density called copula, is characterized by a one-dimensional functional parameter A. We provide several examples of constrained nonparametric families which are useful for applications. In Section 3 we introduce the derivative of the log-density with respect to the functional parameter and the corresponding information operator. We discuss the choice of functional parameter A to ensure the one-dimensional nonparametric rate of convergence of appropriate estimators. In Section 4, we consider a nonparametric estimator of functional parameter A. In a cross-sectional framework the idea is to minimize a chi-square distance between the constrained densities in the family and an unconstrained kernel estimator of the density. In a time series framework the transition densities are used. We derive the asymptotic properties of the estimator and of its linear forms. The nonparametric efficiency of the minimum chi-square estimator for linear forms is proved in Section 5, where the nonparametric efficiency bounds are derived. In many examples the functional parameter A is subject to restrictions, which are due either to the natural constraint on the marginal density to sum up to 1, or to identification restrictions. The extension of the results to these cases is discussed in Section 6. Section 7 concludes. Proofs are gathered in Appendices and on the website http://www.istituti.usilu.net/gagliarp/web/proofs.htm.

2 Modeling nonlinear dependence with one-dimensional functional parameters

Let us consider families of distributions which are parameterized in terms of a one-dimensional functional parameter A, that is a vector-valued function defined on a subset of $\mathbb{R}$. For expository purpose the results are presented in a bivariate framework, but the extension to any number d of observed variables is straightforward.

Assumption A.1: *The distributions of interest are continuous with respect to the Lebesgue measure λ_2, with p.d.f. $f(x, y; A)$, where A is a one-dimensional functional parameter. We denote by P_A the distribution associated with $f(x, y; A)$.*

The family of densities $f(x, y; A)$ defines a constrained nonparametric family. We discuss in Section 2.1

below how constrained nonparametric families can be appropriately specified, and their nonlinear dependence summarized in the functional parameter A. We provide in Section 2.2 several examples of constrained nonparametric families.

2.1 Characterizations of the joint density

A family of bivariate joint densities can be specified in various ways.

i) Conditional density and marginal density

One possibility is to parameterize the conditional distribution and one marginal distribution. Assume $f_{X|Y}(x \mid y; A)$ [resp. $f_Y(y; A)$] is a family of conditional densities of X given Y [resp. of marginal densities of Y], parameterized by function A. A family of bivariate densities is defined by:

$$f(x, y; A) = f_{X|Y}(x \mid y; A) f_Y(y; A).$$

ii) Copula and marginal distributions

In many applications, however, other equivalent representations of the joint density can be more appropriate to specify constrained nonparametric families. For instance, the distribution can be parameterized by the copula (see Nelsen, 1999) and the marginal distributions [1]. Let us recall that any joint c.d.f. can be written as (Sklar, 1959):

$$F(x, y) = C\left[F_X(x), F_Y(y)\right],$$

where F_X, F_Y are the marginal c.d.f. of the variables X and Y, and C is a c.d.f. on $[0, 1]^2$ with uniform marginal distributions, called copula. Such a decomposition allows to separate the marginal features (included in F_X, F_Y) and some dependence features (included in the copula). A family of bivariate densities for (X, Y) is defined by specifying the copula p.d.f. $c(u, v; A)$ and the marginal distributions $f_X(x; A)$, $f_Y(y; A)$:

$$f(x, y; A) = c\left[F_X(x; A), F_Y(y; A); A\right] f_X(x; A) f_Y(y; A).$$

In a cross-sectional framework the functional parameter A is often chosen as:

$$A = (f_X, f_Y, a)',$$

---

[1]Other approaches based on alternative equivalent representations of the joint density, such as real Laplace transforms, are discussed in Gagliardini and Gouriéroux (2002a).

where f_X, f_Y are the unconstrained densities of variables X, Y, and a is a one dimensional functional parameter which characterizes the copula. In this case, marginal features are separated from dependence features, which are summarized by functional parameter a. In a time series framework, such an approach can be used to study the risk dynamics. If (X_t) is a stationary homogeneous Markov process, the dynamics is fully characterized by the joint bivariate distribution of X_t, X_{t-1}, whose marginal distributions are identical because of stationarity. In this case the bivariate distribution $f(x_t, x_{t-1}; A)$ is parameterized by two one-dimensional functional parameters: $A = (f, a)$, where f is the p.d.f. of the stationary distribution and a the functional parameter which characterizes the copula of X_t, X_{t-1}, that is nonlinear serial dependence.

2.2 Examples

We consider below different constrained nonparametric families of bivariate densities, and discuss the parameterizations of the copula or of the conditional density.

i) Truncated model

Let us consider a latent variable X^* with p.d.f. f^*, $f^* > 0$, and assume that, for any value of $Y = y$, the value of X is drawn in the conditional distribution of X^* given $X^* < y$. This situation occurs in truncation models, where the truncation variable Y is independent of the latent variable X^* of interest. The parameter of interest is the latent p.d.f. f^*. The conditional p.d.f. of X given Y is:

$$f(x \mid y; A) = \frac{f^*(x)}{\int_{-\infty}^{y} f^*(z) dz} \mathbb{I}_{x \leq y},$$

where the functional parameter is $A = f^*$.

ii) Stochastic unit root

The stochastic unit root model has been introduced by Gouriéroux and Robert (2005) to study the links between long memory, endogenous switching regimes and heavy tails, often encountered in financial time series. The process is defined by:

$$X_t = \begin{cases} X_{t-1} + \varepsilon_t & \text{, with prob. } \pi(X_{t-1}), \\ \varepsilon_t & \text{, with prob. } 1 - \pi(X_{t-1}), \end{cases}$$

where the ε_t are i.i.d. errors independent from $\underline{X_{t-1}}$, with density g, $g > 0$, and π is a function with values in $]0, 1]$. This is a Markov process with two stochastic regimes, corresponding to either a random walk, or

a white noise[2]. Such a specification underlies the analysis of purchasing power parity (PPP), when it is assumed that unit roots can exist within a band for the PPP equilibrium, whereas mean-reverting effects exist outside the band (e.g., Bec et al., 2004; Rahbek and Shephard, 2002). Function π characterizes the nonlinear serial dependence properties of Markov process (X_t). For instance, tail behaviour of π when $y \to \infty$ characterizes the durations spent by the process in the random walk regime. Under appropriate conditions on functions π and g, process (X_t) is stationary (see Gouriéroux and Robert, 2005), and its conditional density is given by:

$$f(x \mid y; A) = \pi(y)g(x - y) + [1 - \pi(y)]\, g(x).$$

The model is parameterized by $A = (\pi, g)'$.

iii) Dynamic models with proportional hazard

This specification concerns time series (X_t) of duration variables, where the lagged values are explanatory variables with proportional hazard effect. Such models are used for analyzing liquidity risk from intertrade duration data (see Gagliardini and Gouriéroux, 2002b). Since the proportional hazard condition is invariant by increasing transformation, it only concerns the copula of the process, and any stationary distribution can be imposed by an appropriate marginal transformation. The distribution of the Markov process (U_t) with proportional hazard and uniform marginal distribution can be written as:

$$P[U_t \geq u \mid U_{t-1} = v] = \exp\left[-a(v)H_0(u)\right],$$

where a is a positive function on $[0, 1]$, and H_0 is a baseline cumulated hazard on $[0, 1]$. Functions a and H_0 are restricted by the condition of uniform margins:

$$1 - u = E\left[P[U_t \geq u \mid U_{t-1}]\right], \forall u \in [0, 1],$$

that is:

$$H_0^{-1}(z) = 1 - \int_0^1 \exp\left[-za(v)\right] dv, \ z \geq 0. \tag{1}$$

Thus the proportional hazard copula of (U_t, U_{t-1}) is characterized by the functional parameter a only and it is given by:

$$c(u, v; a) = a(v)h_0(u; a) \exp\left[-a(v)H_0(u; a)\right],$$

where $H_0(u; a)$ is defined by (1), and $h_0 = dH_0/du$. The distribution of the Markov process $X_t = F^{-1}(U_t)$

---

[2]The specification is easily extended to a second regime which is a stationary autoregression.

with proportional hazard and marginal c.d.f. F (resp. p.d.f. f) is characterized by two one-dimensional functional parameters (f, a). In Gagliardini and Gouriéroux (2002b) it is shown that the strength of serial dependence is related to the elasticity of function a, whereas the behaviour of the latter close to the boundary points $v = 0$, $v = 1$ characterizes the tail dependence properties of the process. Note finally that two functional parameters differing by a multiplicative constant, a and ka (say), define the same proportional hazard copula, which creates an identifiability problem.

iv) Archimedean copula

The family is usually defined by (see Genest and Mc Kay, 1986):

$$C(u, v) = \phi \left[\phi^{-1}(u) + \phi^{-1}(v) \right], \tag{2}$$

where the (strict) generator ϕ^{-1} is a convex, decreasing function defined on $(0, 1]$, such that $\phi^{-1}(1) = 0$, and $\phi^{-1}(0) = +\infty$ [3]. The most well-known Archimedean copulas are derived from factor models. Typically they correspond to duration models, where the duration variables X and Y are independent identically distributed conditional on an omitted factor Z, and the factor Z has identical proportional hazard effects on the duration distributions (see e.g. Van der Berg, 2001). More precisely, up to some increasing transformations on X and Y, we can assume that the variables X and Y are independent conditional on Z, with identical survivor function $P[X > x \mid Z] = \exp(-Zx)$, $P[Y > y \mid Z] = \exp(-Zy)$ [4]. Then the joint survivor function is $P[X > x, Y > y] = E \exp[-Z(x+y)] = \phi(x+y)$, where $\phi(s) = E[\exp(-sZ)]$. A similar computation provides the marginal survivor function $P[X > x] = \phi(x)$ and the Archimedean expression of the (survivor) copula in equation (2) . In this case ϕ is the Laplace transform of the positive heterogeneity factor Z:

$$\phi(s) = E[\exp(-sZ)], \ s \geq 0. \tag{3}$$

This specification is useful for analyzing credit risk and especially for modeling default correlation, with heterogeneity Z being a latent economic factor with a common proportional hazard effect on the times to default X, Y of two firms. The pattern of the nonlinear dependence between X and Y is characterized by the Laplace transform ϕ of the omitted factor Z. Function ϕ is directly related to the age structure of default correlation in the distribution of X and Y and, as a consequence, is linked to the term structure of spread

---

[3] The Archimedean copula admits a direct extension to multidimensional framework as $\phi \left[\sum_{i=1}^{d} \phi^{-1}(u_i) \right]$. This is a typical example of symmetric copula with large dimension depending on a single one-dimensional functional parameter. Clearly such symmetric copulas, useful in default correlation analysis, do not belong to the class of index models.

[4] These models correspond to the so-called frailty models introduced in the multivariate failure time literature (see e.g. Oakes, 1989).

of interest rates for any derivative written on the two firms [5]. This provides a direct interpretation of ϕ in terms of prices, and the flexibility of the model with respect to ϕ implies the flexibility of the associated interest rate spread model.

When the generator ϕ is twice continuously differentiable, the copula p.d.f. is given by:

$$c(u,v) = \frac{\phi'' \left[\phi^{-1}(u) + \phi^{-1}(v)\right]}{\phi' \left[\phi^{-1}(u)\right] \phi' \left[\phi^{-1}(v)\right]},$$

and is parameterized by the generator ϕ (or by ϕ^{-1}).

v) Markov process with finite dimensional canonical decomposition

Nonlinear canonical analysis provides a decomposition of a stationary Markov process X_t, $t \in \mathbb{N}$, into orthogonal functional directions $\varphi_j(X_t)$, $\psi_j(X_{t-1})$, $j \in \mathbb{N}$ varying, of decreasing serial dependence[6]. Functions φ_j, ψ_j, j varying, are called canonical directions, and $\lambda_j = corr\left[\varphi_j(X_t), \psi_j(X_{t-1})\right]$, j varying, are the associated canonical correlations. The canonical decomposition of Markov process (X_t) is characterized, up to increasing transformations of the canonical directions, by the canonical decomposition of the copula.

A stationary Markov process with one dimensional canonical decomposition is obtained when $\lambda_j = 0$, $j \geq 2$, and $\lambda_1 = \lambda > 0$ (see Gouriéroux and Jasiak, 2001). Its copula is given by:

$$c(u,v) = 1 + \lambda \varphi(u) \psi(v),$$

where the canonical directions φ and ψ satisfy the conditions:

$$\int_0^1 \varphi(u) \, du = \int_0^1 \psi(v) \, dv = 0, \quad \int_0^1 \varphi(u)^2 \, du = \int_0^1 \psi(v)^2 \, dv = 1,$$

and are such that the copula density is positive. Let us for simplicity consider the case of reversible Markov processes, that is $\varphi = \psi$. Then the copula density can be parameterized by $a = \sqrt{\lambda}\varphi$, and we get:

$$c(u,v;a) = 1 + a(u)a(v),$$

where the functional parameter a satisfies the constraint:

$$\int_0^1 a(v)dv = 0.$$

---

[5] See Gagliardini and Gouriéroux (2002a) and Gouriéroux and Monfort (2002b).

[6] See Lancaster (1968) and Dunford and Schwartz (1968) for the definition of canonical analysis and Gouriéroux and Jasiak (2002) for an application to intertrade durations.

The canonical correlation λ and the canonical direction φ are deduced from functional parameter a by the equations:

$$\lambda = \int_0^1 a(v)^2 dv, \quad \varphi(u) = \frac{1}{\sqrt{\lambda}} a(u).$$

3 The information operator

Let $f(x, y; A)$ be a constrained nonparametric family of bivariate densities. The discussion of the nonparametric estimation of function A follows the same lines as in the standard finite-dimensional parametric framework, but the main ideas are generalized to take into account the functional nature of the parameter. In particular, we introduce in Section 3.1 the derivative $D \log f(x, y; A)$ of the log-density with respect to A, that is the functional score, and the information operator I, which is the functional counterpart of the usual information matrix. In Section 3.2 we provide the expressions of the functional score and of the information operator in some of the examples presented above.

3.1 Hadamard derivative and information operator

Let $f(x, y; A)$ be a family of bivariate densities, where the functional parameter A belongs to an open set $\mathcal{A}$ of $\mathbb{R}^q$-valued functions defined on a subset of $\mathbb{R}$. Set $\mathcal{A}$ is endowed with the $L^2(\lambda)$-norm $\|.\|_{L^2(\lambda)}$ corresponding to the Lebesgue measure λ. Function $A_0 \in \mathcal{A}$ denotes the true value of the functional parameter, and $f(.,.) = f(.,.; A_0)$ the corresponding true p.d.f.

i) The differentiability assumption

Let us first define the derivative of the log-density with respect to A.

Assumption A.2 *The Hadamard derivative of* $\log f(x, y; A)$ *with respect to* A, *denoted by* $D \log f(x, y; A)$, *exists:*

$$\log f(x, y; A + h) - \log f(x, y; A) = \langle D \log f(x, y; A), h \rangle + R(x, y; A, h),$$

for $A, A + h \in \mathcal{A}$, *where* $D \log f(x, y; A)$ *is a linear mapping from* $L^2(\lambda)$ *to* $\mathbb{R}$ *which associates to* $h \in L^2(\lambda)$ *the quantity* $\langle D \log f(x, y; A), h \rangle \in \mathbb{R}$. *The Hadamard derivative can be considered stochastic when* x, y *are replaced by* X, Y *with distribution* P_A. *Then* $D \log f(X, Y; A)$ *is a linear operator from* $L^2(\lambda)$ *to* $L^2(P_A)$ *which associates to* $h \in L^2(\lambda)$ *the random variable* $\langle D \log f(X, Y; A), h \rangle \in L^2(P_A)$. *We also assume that:*

   i) *operator* $D \log f(X, Y; A) : L^2(\lambda) \to L^2(P_A)$ *is bounded, for any* $A \in \mathcal{A}$,

ii) *the stochastic residual term $R(X, Y; A, h)$ is such that $\|R(X, Y; A, h)\|_{L^2(P_A)} = o\left(\|h\|_{L^2(\lambda)}\right)$, uniformly on h in the class of compact sets, for any $A \in \mathcal{A}$ [7].*

Under Assumption A.2 the information operator I can be defined by[8] :

$$(g, Ih)_{L^2(\lambda)} := \int g(v)' Ih(v) dv = E_0\left[\langle D \log f(X, Y; A_0), g \rangle \langle D \log f(X, Y; A_0), h \rangle\right], \tag{4}$$

for g, $h \in L^2(\lambda)$, where $E_0[.]$ denotes expectation w.r.t. variables X, Y with distribution P_{A_0}. The information operator I is the analogue of the information matrix defined for models with finite-dimensional parameters. It is the covariance operator of the functional score $D \log f(X, Y; A_0)$. Indeed from the definition of the adjoint operator $D \log f_0^* : L^2(P_{A_0}) \to L^2(\lambda)$ of the functional score $D \log f_0 := D \log f(X, Y; A_0)$, we get:

$$
\begin{aligned}
E_0\left[\langle D \log f(X, Y; A_0), g \rangle \langle D \log f(X, Y; A_0), h \rangle\right] &= \left(\langle D \log f_0, g \rangle, \langle D \log f_0, h \rangle\right)_{L^2\left(P_{A_0}\right)} \\
&= \left(g, D \log f_0^* \ D \log f_0 \ h\right)_{L^2(\lambda)}.
\end{aligned}
$$

We deduce that the information operator I can be written as $I = (D \log f_0)^* D \log f_0$. Thus the information operator I is a bounded, nonnegative, self-adjoint operator from $L^2(\lambda)$ into itself.

Similarly, a conditional information operator $I_{X|Y}$ and a copula information operator I_c can be defined from the differential of the conditional distribution $D \log f(X|Y; A_0)$ and the differential of the copula density $D \log c(U, V; A_0)$, respectively.

ii) Decomposition of the information operator

The differential and information operators admit simplified expressions in most of the applications. A particular decomposition of the information operator is generally encountered in examples (see Section 3.2 below). Specifically, we consider the following assumption:

Assumption A.3: *The information operator I admits the decomposition:*

$$(g, Ih)_{L^2(\lambda)} = \int g(w)' \alpha_0(w; A_0) h(w) dw + \int \int g(w)' \alpha_1(w, v; A_0) h(v) dv dw,$$

where α_0 and α_1 are matrix-valued functions, such that $\alpha_0(w; A_0) = \alpha_0(w; A_0)'$, $\alpha_1(v, w; A_0) = \alpha_1(w, v; A_0)'$, $\forall v, w$.

---

[7]Precisely: $\forall A \in \mathcal{A}$, $K \subset \mathcal{A}$ compact: $\|R(X, Y; A, h)\|_{L^2(P_A)} / \|h\|_{L^2(\lambda)} \to 0$, uniformly in $h \in K$ (see Aït-Sahalia, 1993; Van der Vaart and Wellner, 1996).

[8]See e.g. Koshevnik and Levit (1976); Begun et al. (1983); Bickel et al. (1993); Gill and Van der Vaart (1993); Holly (1995).

Under Assumption A.3 the information operator I is given by:

$$Ih(w) = \alpha_0(w; A_0)h(w) + \int \alpha_1(w, v; A_0)h(v)dv,$$

and is decomposed into a singular and an integral component, corresponding to functions α_0 and α_1, respectively. Such operators are at the core of the classical Fredholm theory of linear integral equations (see e.g. Yosida, 1995, Chapter 10; Rudin, 1973, Theorem 4.25 and Exercise 15 in Chapter 4). Assumption A.3 is needed to derive several asymptotic results about the minimum chi-square estimator defined in Section 4, in particular its one-dimensional nonparametric rate of convergence. Moreover under Assumption A.3 it is easy to derive sufficient conditions[9] for a bounded functional score (Assumption A.2) and for the invertibility of the information operator (Assumption A.4 and discussion below).

To illustrate the decomposition in Assumption A.3, let us consider the functional score:

$$\langle D \log f(x, y; A), h \rangle = \gamma_0(x, y; A)' h(x) + \gamma_1(x, y; A)' h(y) + \int \gamma_2(x, y, w; A)' h(w)dw, \tag{5}$$

where $\gamma_0, \gamma_1, \gamma_2$ are $\mathbb{R}^q$-valued functions[10]. In other words the joint density $f(x, y; A)$ depends on function A by means of values $A(x), A(y)$ at points x, y and of integrals of function A. The information operator I satisfies Assumption A.3 with component α_0 given by:

$$\alpha_0(w; A) = \int \gamma_0(w, y; A)\gamma_0(w, y; A)' f(w, y)dy + \int \gamma_1(x, w; A)\gamma_1(x, w; A)' f(x, w)dx, \tag{6}$$

and component α_1 involving functions $\gamma_0, \gamma_1, \gamma_2$. Thus the component α_0 of the information operator is obtained from the differentiation of the part of the joint density $f(x, y; A)$ which depends on the value of parameter A at some point. It is called local component. The components of the density which depend on integrals of A contribute only to term α_1.

Assumption A.2 on the Hadamard derivative $D \log f(X, Y; A)$ and Assumption A.3 on the information operator I introduce constraints on the parameterization of the model. Indeed the family $f(x, y; A)$ can be parameterized in different ways. For instance, if A is differentiable, we can replace the initial functional parameter A by its derivative dA/dw, which provides the same information (up to a scalar parameter). However it is well-known that nonparametric estimators of A and dA/dw can have very different rates of convergence (see e.g. Silverman, 1978; Stone, 1983). The conditions of Assumptions A.2 - A.3 restrict the admissible choice of parameter A to ensure a one-dimensional nonparametric rate of convergence of the

---

[9]Available on the website.

[10]In this case it is seen that operator $Dlogf(X, Y; A)$ can be represented by a stochastic measure with a continuous and a degenerate component.

appropriate estimators.

iii) Invertibility of the information operator

In the rest of the paper we assume the following invertibility condition.

Assumption A.4: *The information operator I is invertible, with a continuous inverse I^{-1}.*

This assumption is the analogue of the usual invertibility condition of the information matrix, and is used to identify locally the functional parameter of interest [see Appendix 2.3, Lemma A.2 iv), for a detailed discussion of local identification]. Under Assumption A.3 and weak additional regularity conditions[11], the invertibility of I is implied by the following condition:

Sufficient invertibility condition: *the differential operator has a zero null space:*

$$\langle D \log f(X, Y; A_0), h \rangle = 0 \ P_{A_0}\text{-a.s.}, \ h \in L^2(\lambda) \implies h = 0.$$

This sufficient condition is easily checked in the examples.

3.2   Examples

The differential of the copula or of the conditional density, and the corresponding information operators are derived below for some constrained nonparametric families considered in Section 2.2 (see Appendix 1 for the derivations). For each example we select an appropriate functional parameter, in order to satisfy Assumptions A.2 and A.3. As seen from the case of Archimedean copula, the parameter choice is the difficult step when studying nonlinear dependence.

i) Truncated model

For the functional parameter $A = \log f^*$, the differential of $\log f(x \mid y; A)$, for $x \leq y$, is given by:

$$\langle D \log f(x \mid y; A), h \rangle = h(x) - \int f(z \mid y; A) h(z) dz = h(x) - E_A[h(X) \mid Y = y].$$

Let us now consider the conditional information operator $I_{X|Y}$. By definition we have:

$$
\begin{aligned}
\left(g, I_{X|Y} h \right)_{L^2(\lambda)} &= E_0 \left\{ (g(X) - E_0[g(X) \mid Y]) (h(X) - E_0[h(X) \mid Y]) \right\} \\
&= E_0 Cov_0 \left(g(X), h(X) \mid Y \right).
\end{aligned}
$$

---

[11] Available on the website.

It satisfies Assumption A.3 with:

$$\alpha_0\left(x;A\right)=f_X(x;A),\qquad \alpha_1(x,y;A)=-\int f\left(x\mid z;A\right)f\left(y\mid z;A\right)f_Y(z;A)dz.$$

ii) Dynamic model with proportional hazard

The differential of the copula density is given by (see Gagliardini and Gouriéroux, 2002b):

$$
\begin{aligned}
\langle D\log c\left(U_t,U_{t-1};a\right),h\rangle &= (1-a_{t-1}H_{0t})\left(h_{t-1}/a_{t-1}-E\left[h_{t-1}/a_{t-1}\mid U_t\right]\right)\\
&\quad -E\left\{(1-a_{t-1}H_{0t})\left(h_{t-1}/a_{t-1}-E\left[h_{t-1}/a_{t-1}\mid U_t\right]\right)\mid U_t\right\}\\
&= \gamma_0(U_t,U_{t-1})h\left(U_{t-1}\right)+\int \gamma_1\left(U_t,U_{t-1},w\right)h(w)dw,
\end{aligned}
$$

where $h_{t-1}=h(U_{t-1})$, $a_{t-1}=a(U_{t-1})$, $H_{0t}=H_0(U_t,a)$,

$$\gamma_0\left(u,v;a\right)=\frac{1-a(v)H_0(u;a)}{a(v)},$$

and γ_1 is given in Gagliardini and Gouriéroux (2002b). The copula information operator I_c satisfies Assumption A.3 with local component:

$$\alpha_0(w;a)=\frac{1}{a(w)^2}.$$

iii) Archimedean copula

Let us consider the Archimedean family of copula. The generator ϕ (or ϕ^{-1}) is a natural functional parameter for this family, but this parameter does not satisfy the differentiability condition given in Assumption A.2. The Lemma below[12] introduces an equivalent functional parameter in a one-to-one relationship with ϕ. Let us consider the transformed variables:

$$W=C(U,V),\qquad Z=V,$$

where U,V have joint c.d.f. C given in (2).

Lemma 1 *The joint p.d.f. of W and Z is given by:*

$$f(w,z)=\frac{f^*(w)}{\int_0^z f^*(v)dv}1_{w\leq z},\quad w,z\in(0,1),$$

---

[12]The proof is available on the website.

where the latent measure density f^* is:

$$f^*(w) = -\frac{\phi'' \left[\phi^{-1}(w)\right]}{\phi' \left[\phi^{-1}(w)\right]}, \quad w \in (0,1).$$ (7)

Moreover there is a one-to-one relationship between the c.d.f. F^* and the generator ϕ^{-1} since:

$$F^*(w) = -\phi' \left[\phi^{-1}(w)\right] \quad \Longleftrightarrow \quad \phi^{-1}(y) = \int_y^1 \frac{1}{F^*(w)} dw,$$

(for F^* satisfying the condition $\int_0^1 1/F^*(w) dw = \infty$).

The generator ϕ^{-1} and the measure density f^* are identifiable up to a multiplicative constant. This identification problem can be solved by imposing that f^* is a p.d.f. Then variables W and Z follow a truncation model [see example i)], with latent density f^* and $Z \sim U(0,1)$.

Now let us parameterize the copula density by means of function $a = f^*$. The expression of the copula density given in Section 2.2 iv) becomes:

$$c(u,v;a) = a\left[C(u,v;a)\right] \frac{F^*\left[C(u,v;a);a\right]}{F^*(u;a) F^*(v;a)},$$

where functional parameter a is a positive function defined on $[0,1]$ and such that:

$$\int_0^1 a(v) dv = 1.$$

The differential is:

$$\langle D \log c(u,v;a), h \rangle = \frac{h\left[C(u,v;a)\right]}{a\left[C(u,v;a)\right]} + \int_0^1 \gamma(u,v,w;a) h(w) dw,$$

where the expression of function γ is given in Appendix 1.1. The information operator satisfies Assumption A.3, with local component given by:

$$\alpha_0(w,a) = \frac{f_W(w;a)}{a(w)^2} = \frac{\phi^{-1}(w;a)}{a(w)},$$

where $f_W(.;a)$ is the p.d.f. of variable W, and α_1 is given by:

$$\alpha_1(x,y;a) = E_a\left\{\widetilde{\gamma}(W,Z,y) \mid W = x\right\} \phi^{-1}(x;a) + E_a\left\{\widetilde{\gamma}(W,Z,x) \mid W = y\right\} \phi^{-1}(y;a)$$
$$+ E_a\left\{\widetilde{\gamma}(W,Z,x)\widetilde{\gamma}(W,Z,y)\right\},$$

where function $\widetilde{\gamma}$ is such that $\widetilde{\gamma}\left(C(u,v),v,y\right) = \gamma(u,v,y)$ [13].

4 Minimum chi-square estimator

The aim of this section is to introduce a class of constrained nonparametric estimators with nice theoretical properties. We define the minimum chi-square estimators and study their consistency and asymptotic distribution. We first consider the i.i.d. framework, where observations (X_t, Y_t), t varying, correspond to either a cross-section or a i.i.d. time series and the focus is on the contemporaneous dependence between X_t and Y_t. Then the results are derived in the time series setting, where the focus is on serial dependence. The regularity assumptions and the proofs are gathered in Appendix 2.

4.1 Definition of the estimator

Let us consider the i.i.d. framework:

Assumption A.5.IID: *The variables (X_t, Y_t), t varying, are i.i.d., with distribution $f(x,y;A)$. The support of the p.d.f. is $[0,1]^2$.*

It is always possible to transform variables (X_t^*, Y_t^*) with values in $\mathbb{R}$ into variables with values in $[0,1]$ by applying the logit transformation. Therefore the assumption of compact support $[0,1]^2$ is not restrictive.

Let us introduce a kernel estimator of the unconstrained bivariate density function (Rosenblatt, 1956; Parzen, 1962):

$$\widehat{f}_T(x,y) = \frac{1}{Th_T^2} \sum_{t=1}^{T} K\left(\frac{x - X_t}{h_T}\right) K\left(\frac{y - Y_t}{h_T}\right), \tag{8}$$

where K is a kernel and h_T is a bandwidth[14]. Under standard regularity conditions (included in the set of assumptions in Appendix 2.1), the kernel density estimator is a consistent estimator of its mean and is asymptotically normal:

$$\sqrt{Th_T^2}\left[\widehat{f}_T(x,y) - E\widehat{f}_T(x,y)\right] \xrightarrow{d} N\left(0, \sigma^2(x,y;A_0)\right), \tag{9}$$

where $\sigma^2(x,y;A_0) = f(x,y;A_0)\left(\int K^2(w)dw\right)^2$. Moreover, we also get the consistency and asymptotic normality of integrals of f, that can be conditional and cross-moments, at rates depending on the number

---

[13]Another copula which defines a constrained nonparametric family is the extreme value copula (see e.g. Joe, 1997). The choice of the functional parameter and the corresponding differential and information operators are available on the website.

[14]The results can be easily generalized to the case where different bandwidths h_T and h_T' are introduced for processes X_t and Y_t, respectively.

of integrations (see Theorem 3 in Aït-Sahalia, 1993):

$$\sqrt{Th_T}\left[\int g(x)\widehat{f}_T(x,y)dx - E\int g(x)\widehat{f}_T(x,y)dx\right] \xrightarrow{d} N\left(0,\sigma^2\left(y,g;A_0\right)\right), \tag{10}$$

where $\sigma^2\left(y,g;A_0\right) = E_0\left[g(X_t)^2 \mid Y_t = y\right]f_Y(y)\int K^2(w)dw$, and:

$$\sqrt{T}\left[\int\int g(x,y)\widehat{f}_T(x,y)dxdy - E\int\int g(x,y)\widehat{f}_T(x,y)dxdy\right] \xrightarrow{d} N\left(0,\sigma^2\left(g;A_0\right)\right), \tag{11}$$

where $\sigma^2\left(g;A_0\right) = V_0\left[g\left(X_t,Y_t\right)\right]$ [15].

The unconstrained estimator of the bivariate density can be used to derive a minimum chi-square estimator of parameter A:

$$\widehat{A}_T = \arg\min_{A\in\Theta}Q_T(A) = \int_0^1\int_0^1\frac{\left[\widehat{f}_T(x,y) - f(x,y;A)\right]^2}{\widehat{f}_T(x,y)}\omega_T(x,y)dxdy, \tag{12}$$

where Θ is a subset of $\mathcal{A}$, ω_T is a smooth weighting function, converging pointwise to the constant function 1, when T tends to infinity. The constrained estimator of the bivariate density is deduced by:

$$\widehat{f}_T^0(x,y) = f(x,y;\widehat{A}_T). \tag{13}$$

The estimators $\widehat{A}_T$ and $\widehat{f}_T^0$ exist and are well defined (see Appendix 2.2).

The aim of this section is not to discuss the practical implementation of a nonparametric minimum chi-square estimator, but to prove the existence of nonparametrically efficient estimators. Nevertheless two approaches can be followed in practice to solve the optimization problem with respect to functional parameter.

i) The optimization can be performed over a finite dimensional sub-space of functions A by standard optimization software. When the dimension of the space tends to infinity sufficiently fast with T, the asymptotic properties of the estimator will stay the same.

ii) Alternatively the derivatives of the chi-square criterion are related to the information operator, and explicit expressions of the derivative are available for some examples. Then we can compute recursively the functional solution by a Newton-Raphson type algorithm, or apply a single step of the algorithm from a consistent, but inefficient functional estimator.

Finally, in some examples, such as the stochastic unit root model, the chi-square criterion can be con-

---

[15] The asymptotic bias in these nonparametric estimators will be carefully taken into account in Section 4.4 and in Appendix 2.

centrated with respect to some functional parameters, which diminishes the dimension of the optimization problem.

4.2 Consistency of the estimators

To prove the consistency of the minimum chi-square estimator $\widehat{A}_T$, it is shown in Appendix 2.3 that the optimization criterion Q_T converges to the chi-square proximity measure Q, defined by:

$$Q(A) = \int_0^1 \int_0^1 \frac{[f(x,y) - f(x,y;A)]^2}{f(x,y)} dxdy,$$

uniformly in $A \in \Theta$, and that Q is continuous.

Proposition 2 : *Under the regularity assumptions of Appendix 2 including the bandwidth condition $h_T = c_T T^{-\alpha}$, $\lim_{T \to \infty} c_T = c > 0$, with $0 < \alpha < 1/d$, where d is the dimension of the observable variables (that is $d = 2$ in the present framework), the chi-square estimator $\widehat{A}_T$ is consistent in norm:*

$$\left\| \widehat{A}_T - A_0 \right\|_{L^2(\lambda)} \xrightarrow{p} 0.$$

Proof. *See Appendix 2.3.* ∎

Let us now consider the constrained density estimator $\widehat{f}_T^0$, and its consistency in L^1-norm, where L^p is the space of p-integrable functions on $[0,1]^2$. The convergence of $\widehat{A}_T$ to A_0 and the continuity of the chi-square measure Q imply the convergence of $Q(\widehat{A}_T)$ to $Q(A_0) = 0$. By using the Cauchy-Schwarz inequality:

$$\left\| f(.,.;\widehat{A}_T) - f(.,.) \right\|_{L^1} \leq \left\| \frac{f(.,.;\widehat{A}_T) - f(.,.)}{\sqrt{f(.,.)}} \right\|_{L^2} \left\| \sqrt{f(.,.)} \right\|_{L^2} = Q(\widehat{A}_T)^{1/2},$$

we deduce the following proposition.

Proposition 3 : *Under the assumptions of Proposition 2, the constrained density estimator $\widehat{f}_T^0$ is consistent in L^1 norm:* $\left\| \widehat{f}_T^0 - f \right\|_{L^1} \xrightarrow{p} 0.$

4.3 Asymptotic expansion of the minimum chi-square estimator

Let us now consider the asymptotic expansion of the minimum chi-square estimator which relates the difference between the estimator and the true value with the efficient functional score. The expansion and the asymptotic normality require a stronger condition on the bandwidth.

18

Proposition 4 : *Assume the regularity conditions of Appendix 2 and let bandwidth h_T be such that:*

$$h_T = c_T T^{-\alpha}, \ lim_{T \to \infty} c_T = c > 0, \ with \ \frac{1}{4m}\left(1 + \frac{2m-1}{4m^2+2m+1}\right) < \alpha < \frac{1}{2d}\left(1 - \frac{1}{2}\frac{2m-1}{4m^2+2m+1}\right),$$

$$(14)$$

where m is the degree of differentiability of the density f and d the dimension of the observed variables. Then minimum chi-square estimator $\widehat{A}_T$ is such that:

$$I\left(\widehat{A}_T - A_0\right) = \psi_T + r_T,$$

$$(15)$$

where $\psi_T \in L^2(\lambda)$ is defined by:

$$(\psi_T, h)_{L^2(\lambda)} = \int\int \delta\widehat{f}_T(x,y)\,\omega_T(x,y)\,\langle D\log f(x,y;A_0), h\rangle\,dxdy, \ h \in L^2(\lambda),$$

$\delta\widehat{f}_T = \widehat{f}_T - f$, *and r_T is a residual term which can be neglected asymptotically, both in norm and pointwise:*
i) $\|r_T\|_{L^2(\lambda)} = o_p(1/\sqrt{T})$ and
ii) $r_T(w) = o_p(1/\sqrt{Th_T})$ λ-a.s. in $w \in (0,1)$.

Proof. See Appendix 2.4 i)-iv). ∎

In the bivariate case $d = 2$, the bandwidth condition can be satisfied whenever the density is twice differentiable ($m \geq 2$). The expansion of Proposition 4 is the analogue of the usual asymptotic expansion of the maximum likelihood estimator in a parametric model with likelihood function $L_T(\theta)$:

$$I\left(\widehat{\theta}_T - \theta_0\right) \simeq \frac{\partial L_T}{\partial \theta}(\theta_0).$$

Moreover in this parametric framework it is known that the ML estimator is asymptotically equivalent to a GMM estimator based on the moment restrictions $E_{\theta_0}\left[\frac{\partial L_T}{\partial \theta}(\theta)\right] = 0$, which justifies the terminology efficient score for $\frac{\partial L_T}{\partial \theta}(\theta)$. Since:

$$\psi_T \simeq \int\int \widehat{f}_T(x,y)\,D\log f(x,y;A_0)dxdy \simeq \frac{1}{T}\sum_{t=1}^T D\log f(X_t, Y_t; A_0),$$

function ψ_T has the interpretation of the efficient score[16].

---

[16]Note that $\int\int f(x,y)D\log f(x,y;A_0)dxdy = 0$, which follows from the unit mass restriction: $\int\int f(x,y;A)dxdy = 1$, for any $A \in \mathcal{A}$.

When the differential operator admits the decomposition (5) into a singular and an integral component, the efficient score ψ_T is given by:

$$
\begin{aligned}
\psi_T(w) \;=\; & \int \delta\widehat{f}_T(w,y)\omega_T(w,y)\gamma_0(w,y)\,dy + \int \delta\widehat{f}_T(x,w)\omega_T(x,w)\gamma_1(x,w)dx \\
& + \int\int \delta\widehat{f}_T(x,y)\omega_T(x,y)\gamma_2(x,y,w)dxdy.
\end{aligned}
\tag{16}
$$

Moreover when the information operator satisfies Assumption A.3, the first order condition is equivalent to:

$$
\alpha_0(w)\delta\widehat{A}_T(w) + \int \alpha_1(w,v)\delta\widehat{A}_T(v)dv \simeq \psi_T(w),
\tag{17}
$$

where $\delta\widehat{A}_T = \widehat{A}_T - A_0$.

The proposition below gives the norm and pointwise asymptotic expansion of the minimum chi-square estimator [see Appendix 2.4 v)].

Corollary 5 *Under the Assumptions of Proposition 4, the minimum chi-square estimator is such that:*

i) $\widehat{A}_T - A_0 = I^{-1}\psi_T + \widetilde{r}_T, \quad where\; \|\widetilde{r}_T\|_{L^2(\lambda)} = o_p(1/\sqrt{T});$

ii) $\widehat{A}_T(w) - A_0(w) = \alpha_0(w)^{-1}\left(\psi_T(w) - E\psi_T(w)\right) + I^{-1}E\psi_T(w) + o_p(1/\sqrt{Th_T}),\; \lambda\text{-a.s. in } w \in (0,1).$

In particular a bias term appears in the asymptotic expansion of the minimum chi-square estimator $\widehat{A}_T$; it is induced by the expectation of the efficient score $E\psi_T$, which vanishes only asymptotically.

Finally the expansion of the constrained estimator of the density is deduced by a δ-method (see website).

Proposition 6 : *Under the Assumptions of Proposition 4, the constrained density estimator is such that:*

$$
\widehat{f}_T^0(x,y) - f(x,y) = \left\langle Df(x,y;A_0), \delta\widehat{A}_T \right\rangle + o_p(1/\sqrt{Th_T}), \quad \lambda_2\text{-a.s. in } (x,y).
$$

4.4 Asymptotic distribution of the minimum chi-square estimator

The asymptotic distribution of the minimum chi-square estimator $\widehat{A}_T$ is derived from the asymptotic expansion given in Corollary 5. We distinguish the pointwise estimation of A, that is estimation of $A(w)$ for any w, and the estimation of linear forms[17] of A, such as $\int_0^1 g(w)' A(w)dw$, for which different orders are expected $1/\sqrt{Th_T}$ and $1/\sqrt{T}$, respectively.

---

[17]In the mathematical literature, $\int_0^1 g(w)' A(w)dw$ is usually called a functional of A. However, we prefer the name linear form in order to avoid confusion with the word functional used to characterize parameter A. Moreover, $\int_0^1 g(w)' A(w)dw$ is the counterpart in our setting of the linear forms of a finite-dimensional parameter usually considered in econometrics.

i) Pointwise estimation

Let us assume that the differential operator admits the decomposition (5) into singular and integral components. We deduce from (16), (10), and (11) that $\sqrt{Th_T}\left(\psi_T - E\psi_T\right)(w)$ is pointwise asymptotically normal (see Appendix 2.5).

Lemma 7 : *Under the regularity assumptions in Appendix 2 and when the differential admits the decomposition (5):*

$$\sqrt{Th_T}\left(\psi_T - E\psi_T\right)(w) \xrightarrow{d} N\left[0, \left(\int K^2(x)dx\right)\alpha_0(w)\right], \quad \lambda\text{-}a.s. \ in \ w.$$

The pointwise asymptotic distribution of $\widehat{A}_T$ follows from Corollary 5 ii).

Proposition 8 : *Under the Assumptions of Proposition 4 the estimator $\widehat{A}_T$ is λ-a.s. pointwise asymptotically normal:*

$$\sqrt{Th_T}\left(\widehat{A}_T(w) - A_0(w) - I^{-1}E\psi_T(w)\right) \xrightarrow{d} N\left(0, \left(\int K^2(x)dx\right)\alpha_0(w)^{-1}\right),$$

λ-*a.s. in* $w \in (0,1)$.

Let us briefly discuss the expressions of the asymptotic variance and of the asymptotic bias. The asymptotic variance is given by the inverse of the local component of the information operator $\alpha_0(w)^{-1}$ (up to a multiplicative constant). To provide insight into this result we remark that, since linear forms of A converge at a parametric rate $1/\sqrt{T}$ [see subsection ii) below], for the computation of the asymptotic variance of the pointwise estimator we can neglect any dependence of the density $f(x,y;A)$ on linear forms of A. Then the relevant component of the information operator is the local component α_0, as illustrated in Section 3.1 ii), and the asymptotic variance of the estimator is essentially its inverse. The asymptotic bias is such that (see website):

$$I^{-1}E\psi_T(w) = \frac{h_T^m}{m!}\left(\int K(u)u^m du\right)I^{-1}b(w) + o\left(h_T^m\right), \tag{18}$$

where function $b \in L^2(\lambda)$ is given by:

$$b(w) = \int \Delta^m f(w,y)\gamma_0(w,y)\,dy + \int \Delta^m f(x,w)\gamma_1(x,w)dx + \int\int \Delta^m f(x,y)\gamma_2(x,y,w)\,dxdy,$$

with $\Delta^m f = \partial^m f/\partial x^m + \partial^m f/\partial y^m$. In particular, the asymptotic variance and the asymptotic bias have the standard orders for one-dimensional nonparametric estimation, that is $1/Th_T$ and h_T^m, respectively. By choosing the bandwidth h_T such that:

$$h_T = O\left(T^{-\frac{1}{2m+1}}\right), \tag{19}$$

the minimum chi-square estimator admits the optimal one-dimensional rate of convergence $T^{-\frac{m}{2m+1}}$ (Stone, 1983). Under bandwidth condition (14) the optimal choice (19) is admissible whenever $\frac{1}{2m+1} < \frac{1}{2d}\left(1 - \frac{1}{2}\frac{2m-1}{4m^2+2m+1}\right)$. For instance, $m \geq 2$ is sufficient in the bivariate case $d = 2$. Finally, the minimum chi-square estimator is pointwise asymptotically unbiased if $\sqrt{Th_T}h_T^m = o(1)$, that is if $h_T = o\left(T^{-\frac{1}{2m+1}}\right)$.

From Proposition 6 we deduce the asymptotic distribution of the constrained density estimator.

Corollary 9 : *Under the Assumptions of Proposition 6, the constrained density estimator*
$\sqrt{Th_T}\left(\widehat{f}_T^0(x,y) - f(x,y) - \left\langle D\log f(x,y;A_0), I^{-1}E\psi_T\right\rangle\right)$ *is asymptotically normal, with asymptotic variance:*

$$\left(\int K^2(x)dx\right)f(x,y)^2\left[\gamma_0(x,y)'\alpha_0(x)^{-1}\gamma_0(x,y) + \gamma_1(x,y)'\alpha_0(y)^{-1}\gamma_1(x,y)\right].$$

Moreover the asymptotic bias is such that:

$$\left\langle D\log f(x,y;A_0), I^{-1}E\psi_T\right\rangle = \frac{h_T^m}{m!}\left(\int K(u)u^m du\right)\left\langle D\log f(x,y;A_0), I^{-1}b\right\rangle + o(h_T^m).$$

In particular the constrained estimator has a one-dimensional nonparametric convergence rate, and, if $h_T = o\left(T^{-\frac{1}{2(m+1)}}\right)$:

$$\sqrt{Th_T^2}\left[\widehat{f}_T(x,y) - \widehat{f}_T^0(x,y)\right] \simeq \sqrt{Th_T^2}\left[\widehat{f}_T(x,y) - f(x,y)\right] \xrightarrow{d} N\left[0, f(x,y)\left(\int K^2(w)dw\right)^2\right].$$

The discrepancy $\sqrt{Th_T^2}\left[\widehat{f}_T(x,y) - \widehat{f}_T^0(x,y)\right]$, x,y varying, between the unconstrained and the constrained estimators can be used as a basis for a (pointwise) misspecification test.

ii) Estimation of linear forms

Let us now consider the estimation of a linear form $G = \int g(v)'A_0(v)dv$, with $g \in L^2(\lambda)$. We deduce from Corollary 5 i):

$$\sqrt{T}\left(\widehat{G}_T - G\right) = \sqrt{T}\left(g, \delta\widehat{A}_T\right)_{L^2(\lambda)} \simeq \sqrt{T}\left(g, I^{-1}\psi_T\right)_{L^2(\lambda)} = \sqrt{T}\left(I^{-1}g, \psi_T\right)_{L^2(\lambda)},$$

since I^{-1} is self-adjoint on $L^2(\lambda)$. The following Lemma provides the asymptotic distribution of $\sqrt{T}(g, \psi_T - E\psi_T)_{L^2(\lambda)}$, for $g \in L^2(\lambda)$.

Lemma 10 *Under the regularity Assumptions in Appendix 2:*

$$\sqrt{T}(g, \psi_T - E\psi_T)_{L^2(\lambda)} \xrightarrow{d} N\left[0, (g, Ig)_{L^2(\lambda)}\right], \quad for\ g \in L^2(\lambda).$$

Proof. *We have:*

$$
\sqrt{T}\,(g,\psi_T - E\psi_T)_{L^2(\lambda)} \;=\; \sqrt{T}\int\int \left(\widehat{f}_T(x,y) - E\widehat{f}_T(x,y)\right)\omega_T(x,y)\,\langle D\log f(x,y;A_0),g\rangle\,dxdy
$$

$$
\simeq\; \sqrt{T}\int\int \left(\widehat{f}_T(x,y) - E\widehat{f}_T(x,y)\right)\langle D\log f(x,y;A_0),g\rangle\,dxdy.
$$

By using (11) in Section 4.1, the latter expression is asymptotically normal. Its variance is given by:

$$
\sigma^2(g) = V_0\left[\langle D\log f(X_t,Y_t;A),g\rangle\right] = E_0\left[\langle D\log f(X_t,Y_t;A),g\rangle^2\right] = (g,Ig)_{L^2(\lambda)}\,.
$$

∎

The asymptotic distribution of a linear form follows.

Proposition 11 *Under the Assumptions in Proposition 4, the estimator $\widehat{G}_T = \int g(v)'\widehat{A}_T(v)dv$ of a linear form of A is asymptotically normal, with parametric rate of convergence:*

$$
\sqrt{T}\left(\widehat{G}_T - G - \left(g,I^{-1}E\psi_T\right)_{L^2(\lambda)}\right) \xrightarrow{d} N\left(0,\left(g,I^{-1}g\right)_{L^2(\lambda)}\right),\;\; for\;\; g\in L^2(\lambda)\,.
$$

From (18) the bias term is such that:

$$
\left(g,I^{-1}E\psi_T\right)_{L^2(\lambda)} = \frac{h_T^m}{m!}\left(\int K(u)u^m du\right)\left(g,I^{-1}b\right)_{L^2(\lambda)} + o\left(h_T^m\right),
$$

and is negligible whenever $h_T = o\left(T^{-\frac{1}{2m}}\right)$.

4.5   Time series framework

The previous results are easily extended to the time series framework by introducing some mixing condition and considering the transition density, which includes all the relevant information for a Markov process.

Assumption A.5.TS: *Process X_t, t varying, is a strictly stationary, homogeneous Markov process, with transition density $f(x_t \mid x_{t-1};A)$, and β-mixing coefficients such that: $\beta_k = O\left(\rho^k\right)$, $\rho < 1$. The support of the stationary p.d.f. is $[0,1]$.*

Moreover the minimum chi-square estimator is now defined by minimizing a chi-square divergence between the conditional distribution in the family and its unconstrained kernel estimator:

$$
\widehat{A}_T = \arg\min_{A\in\Theta} Q_T(A) = \int_0^1\int_0^1 \frac{\left[\widehat{f}_T(x|y) - f(x|y;A)\right]^2}{\widehat{f}_T(x|y)}\omega_T(x,y)\widehat{f}_{Y,T}(y)dxdy. \tag{20}
$$

We also need some regularity assumptions, valid for the time series framework. They are deduced from the set in Appendix 2.1 by considering the conditional distribution $f(x|y; A)$, instead of the joint one, the conditional differential operator $D \log f(X|Y; A)$ and the conditional information operator $I_{X|Y}$. They are referred to as Assumptions TS in Appendix 2.1.

The derivation of the asymptotic properties of the minimum chi-square estimator in the time series framework is completely analogous to Section 4.2-4.4 [18].

Proposition 12 : *Under regularity Assumptions TS in Appendix 2 the minimum chi-square estimator $\widehat{A}_T$ is consistent.*

The asymptotic expansion of the chi-square estimator in the time series framework is given by:

$$I_{X|Y} \left(\widehat{A}_T - A_0 \right) \simeq \widetilde{\psi}_T,$$

where function $\widetilde{\psi}_T \in L^2 (\lambda)$ is defined by:

$$\left(\widetilde{\psi}_T, h \right)_{L^2(\lambda)} = E_0 \left[\frac{\delta \widehat{f}_T (X \mid Y)}{f (X \mid Y)} \omega_T(X, Y) \left\langle D \log f(X \mid Y; A_0), h \right\rangle \right], \ h \in L^2 (\lambda),$$

and the omitted residual term is asymptotically negligible both pointwise and in norm. In particular, when the conditional information operator $I_{X|Y}$ admits a representation as in Assumption A.3 with $\widetilde{\alpha}_0$, $\widetilde{\alpha}_1$, say, the asymptotic expansion becomes:

$$\widetilde{\alpha}_0(w)\delta \widehat{A}_T(w) + \int \widetilde{\alpha}_1 (w, v) \, \delta \widehat{A}_T(v) dv \simeq \widetilde{\psi}_T(w).$$

The asymptotic distribution of $\widehat{A}_T$ is immediately deduced from that of $\widetilde{\psi}_T$:

$$\sqrt{Th_T} \left(\widetilde{\psi}_T(w) - E\widetilde{\psi}_T(w) \right) \xrightarrow{d} N \left[0, \left(\int K^2(x)dx \right) \widetilde{\alpha}_0(w) \right], \ \lambda\text{-a.s. in } w,$$

$$\sqrt{T} \left(g, \widetilde{\psi}_T - E\widetilde{\psi}_T \right)_{L^2(\lambda)} \xrightarrow{d} N \left[0, \left(g, I_{X|Y}g \right)_{L^2(\lambda)} \right], \ \text{for } g \text{ in } L^2 (\lambda).$$

Note that the asymptotic variance $\left(g, I_{X|Y}g \right)_{L^2(\lambda)} = V_0 \left[\left\langle D \log f(X_t \mid X_{t-1}; A_0), g \right\rangle \right]$ includes no cross-term, since $\left\langle D \log f(X_t \mid X_{t-1}; A_0), g \right\rangle$ is a martingale difference sequence.

We deduce:

---

[18] It will not be repeated in the Appendix and is available upon request.

Proposition 13 : *Under regularity Assumptions TS in Appendix 2 we have:*

$$\sqrt{Th_T}\left(\widehat{A}_T\left(v\right) - A_0(v) - I_{X|Y}^{-1} E\widetilde{\psi}_T(v)\right) \xrightarrow{d} N\left(0, \left(\int K^2(x)dx\right)\widetilde{\alpha}_0\left(v\right)^{-1}\right), \quad \lambda\text{-}a.s \ in \ v,$$

and:

$$\sqrt{T}\left(g, \widehat{A}_T - A_0 - I_{X|Y}^{-1} E\widetilde{\psi}_T\right)_{L^2(\lambda)} \xrightarrow{d} N\left[0, \left(g, I_{X|Y}^{-1} g\right)_{L^2(\lambda)}\right], \quad for \ g \ in \ L^2\left(\lambda\right).$$

5   Nonparametric efficiency for linear forms

The aim of this section is to show that a minimum chi-square estimator is nonparametrically efficient for linear forms[19]. We first review the approach to derive the nonparametric efficiency bound.

5.1   Nonparametric efficiency bound for linear forms

The nonparametric "efficiency bound" for function A is defined from the parametric efficiency bound. The idea is to consider linear forms of A, such as $\int A(v)' g(v)dv$, for $g \in L^2(\lambda)$, which is a scalar parameter that can be consistently estimated at rate $1/\sqrt{T}$, and to construct the semi-parametric bound $B_A(g)$, say, for this scalar parameter. A consistent estimator $\widehat{A}_T$ is said to be nonparametrically efficient for linear forms if the asymptotic variance of $\int \widehat{A}_T(v)' g(v)dv$ is $B_A(g)$, for any $g \in L^2(\lambda)$, that is if $\widehat{A}_T$ provides efficient estimators for all linear forms. This concept of nonparametric efficiency differs from minimax efficiency, for which the estimator is nonparametrically efficient when its Mean Squared Error is pointwise minimal in a suitable class of estimators (see e.g. Donoho and Liu, 1991; Fan, 1993, and references therein). The choice between these different definitions of efficiency is not clear in practice, especially since the estimated functional parameter $\widehat{A}_T$ can be used for several purposes. As an illustration we can consider the canonical decomposition described in Section 2.2 v). We are interested in both canonical correlation (which is a scalar function of A) and canonical directions (which are related to pointwise estimation of A). In such a case it seems natural to impose at any "efficient" pointwise estimator to be at least efficient for linear forms.

The semi-parametric bound $B_A(g)$ can be derived using Stein's heuristic (see Stein, 1956; Severini and Tripathi, 2001). The approach consists in two steps:

i) First introduce a one dimensional parametric model $A\left(.; \theta\right)$, and derive the Cramer-Rao lower bound $B_A(g, \theta)$ for $\int_0^1 A(v; \theta)' g(v)dv$ in this model.

---

[19]For expository purpose we will consider linear forms of A. The results extend immediately to nonlinear functions of A whose derivative is bounded in $L^2\left(\lambda\right).$

ii) Then the nonparametric efficiency bound is defined by:

$$B_A(g) = \max B_A(g, \theta), \; g \text{ varying},$$

where the maximization is performed on all possible parametric specifications $A(., \theta)$.

Since a parameter is defined up to an invertible transformation, for any parametric specification we can select the parameter θ such that:

$$\int A(v; \theta)' g(v) dv = \theta.$$

In a neighbourhood of θ_0, this condition is equivalent to:

$$\int g(v)' \frac{\partial A}{\partial \theta}(v; \theta_0) \, dv = 1.$$

Then we compute:

$$\begin{aligned} B_A(g) &= \max B_A(g, \theta), & (21) \\ \text{s.t.} \quad : \quad & \int g(v)' \frac{\partial A}{\partial \theta}(v; \theta_0) \, dv = 1, \end{aligned}$$

where the maximization is performed over all parameterizations $A(., \theta)$. The nonparametric efficiency bound is the mapping $g \rightarrow B_A(g)$.

Proposition 14 : *i) In the i.i.d. framework the nonparametric efficiency bound is given by:*

$$B_A(g) = (g, I^{-1}g)_{L^2(\lambda)}.$$

ii) In the time series framework the nonparametric efficiency bound is given by:

$$B_A(g) = (g, I_{X|Y}^{-1} g)_{L^2(\lambda)}.$$

Proof. *See Appendix 3.* ∎

5.2 Nonparametric efficiency of the minimum chi-square estimator

From Propositions 11 and 13, if the bandwidth h_T is such that $h_T = o\left(T^{-\frac{1}{2m}}\right)$, we immediately deduce that the estimator $\widehat{G}_T = \int g(v)' \widehat{A}_T(v) dv$ is asymptotically unbiased and reaches the nonparametric efficiency bound for linear forms.

Corollary 15 : *The minimum chi-square estimator $\widehat{A}_T$ is nonparametrically efficient for linear forms.*

The efficiency property of the minimum chi-square estimator is important in practice. Indeed a number of inefficient nonparametric estimation methods have been introduced for some specific copulas (see e.g. Genest and Rivest, 1993, for Archimedean copulas, Abdous et al., 2000, and references therein for extreme value copulas). Similarly the usual estimator of the transformation in transformed regression model, based on the ratio of partial derivatives of the conditional distribution due to the nonparametric identification constraint suggested by Ridder (Ridder, 1990), is consistent, asymptotically normal (Horowitz, 1996; Gorgens and Horowitz, 1999), but in general inefficient. However these inefficient nonparametric estimators can be used as a first step of a Newton-Raphson type algorithm to compute the efficient chi-square estimator.

6 Constrained estimation. Identifying restrictions

Up to now we have assumed that the functional parameter A is free to vary over an open sphere in $L^2(\lambda)$. However this condition is not met in some examples described in Section 2.2. We consider therefore in this section the case of a constrained functional parameter. From the examples, two sources of constraints can be distinguished. First, when one component of A is a marginal distribution, f_Y say, this component satisfies the unit mass restriction $\int f_Y(y)dy = 1$. Second, some parameters may be not identified unless additional restrictions are imposed. This is the case for the copula parameter a in the proportional hazard and in Archimedean copulas [examples iii) and iv)], since a and ka, where k is a positive constant, define the same copula. A possible identifying restriction is: $\int a(v)dv = 1$.

6.1 Restricted information operator

Let us assume that functional parameter A satisfies n linear constraints:

$$\int A(v)' g_i(v)dv = (A, g_i)_{L^2(\lambda)} = k_i, \ \ i = 1, ..., n,$$

where $g_i \in L^2(\lambda)$, $k_i \in \mathbb{R}$, $i = 1, ..., n$. Let us denote by $\widetilde{\mathcal{A}} \subset \mathcal{A}$ the subset of functional parameters satisfying these restrictions. The tangent space H of $\widetilde{\mathcal{A}}$ at $A_0 \in \widetilde{\mathcal{A}}$ does not depend on A_0, has a finite codimension, and it is given by:

$$H = \{h \in L^2(\lambda) : (h, g_i)_{L^2(\lambda)} = 0, \ \ i = 1, ..., n\}.$$

The differential operator $D \log f(X, Y; A_0)$ can be restricted to the linear space $H \subset L^2(\lambda)$, and under Assumption A.2 it follows that $D \log f(X, Y; A_0) : H \to L^2(P_{A_0})$ is a bounded operator. The information

operator I_H is the bounded linear operator from H in itself defined by:

$$(g, I_H h)_{L^2(\lambda)} = E_0 \left[\langle D \log f(X, Y; A_0), g \rangle \langle D \log f(X, Y; A_0), h \rangle \right], \ \ h, g \in H.$$

Under Assumption A.3 the restricted information operator I_H admits a decomposition into singular and integral components:

$$I_H h(w) = \alpha_{0,H}(w) h(w) + \int \alpha_{1,H}(w, v) h(v) dv, \ \ h \in H, \tag{22}$$

with $\alpha_{0,H} = \alpha_0$. Under this decomposition it is possible to derive sufficient conditions for the boundedness and invertibility of I_H, see website. In particular, under weak regularity assumptions, a sufficient condition for the invertibility of operator I_H is:

the differential operator has a zero null space on H:

$$\langle D \log f(X, Y; A_0), h \rangle = 0 \ P_{A_0}\text{-a.s.}, \ h \in H \implies h = 0. \tag{23}$$

To illustrate these results, let us consider the example of the proportional hazard copula [see example iii) in Section 2.2 and Section 3.2 ii)]. The functional parameter a of the copula is subject to the identifying constraint: $\int_0^1 a(v) dv = 1$. The corresponding tangent space H is given by:

$$H = \left\{ h \in L^2(\lambda) : \int_0^1 h(v) dv = 0 \right\}.$$

Let us show that the copula information operator I_H^c is invertible. Indeed let us consider a function $h \in H$ such that:

$$\langle D \log c(U_t, U_{t-1}; a_0), h \rangle = 0, \ P_{A_0}\text{-a.s.}$$

Then by using the differential of the proportional hazard copula [see section 3.2 ii)], we deduce that:

$$(1 - a_{0t-1} H_{0t})(h_{t-1}/a_{0t-1} - E[h_{t-1}/a_{0t-1} \mid U_t])$$
$$= [1 - a_0(U_{t-1}) H_0(U_t)] \{ h(U_{t-1})/a_0(U_{t-1}) - E[h(U_{t-1})/a_0(U_{t-1}) \mid U_t] \}$$
$$\text{is a function of } U_t \text{ only.}$$

This implies that h/a_0 is a constant. Since $\int_0^1 h(v) dv = 0$, it follows that $h = 0$. Thus sufficient condition (23) is satisfied and I_H^c is invertible. On the contrary, the copula information operator is not invertible when it is defined on the entire space $L^2(\lambda)$, since the differential $D \log c(U_t, U_{t-1}; a_0)$ has a non zero null space,

consisting in functions ka_0, where k is a constant.

6.2 The minimum chi-square estimator

Let $\widetilde{\Theta}$ be a subset of $\widetilde{\mathcal{A}}$. The minimum chi-square estimator is obtained by minimizing the chi-square divergence under the constraints:

$$\widehat{A}_T = \arg\min_{A \in \widetilde{\Theta}} Q_T(A) = \int_0^1 \int_0^1 \frac{\left[\widehat{f}_T(x,y) - f(x,y;A)\right]^2}{\widehat{f}_T(x,y)} \omega_T(x,y) dx dy. \tag{24}$$

The consistency of the constrained estimator is proved as in Section 4.2. Let us now focus on the asymptotic expansion. Under the regularity assumptions in Appendix 2, it can be verified that the first order condition is equivalent to:

$$I_H \delta \widehat{A}_T \simeq P_H \psi_T,$$

where P_H denotes the orthogonal projector on H [20]. The asymptotic expansion of the minimum chi-square estimator $\widehat{A}_T$ is derived as in Section 4.3 by replacing the information operator I with the restricted one, I_H, and the efficient score ψ_T with the projection $P_H \psi_T$:

$$\widehat{A}_T - A_0 \simeq I_H^{-1} P_H \psi_T.$$

Let us derive the asymptotic distribution, and neglect the bias terms for expository purposes. Without loss of generality, let functions $g_i \in L^2(\lambda)$, $i = 1, ..., n$, be orthonormal. Since H has finite codimension, we have $P_H \psi_T(w) = \psi_T(w) - \sum_{i=1}^n (g_i, \psi_T)_{L^2(\lambda)} g_i(w) = \psi_T(w) + o_p(1/\sqrt{T})$ [see Lemma 10], and we get from Lemma 7:

$$\sqrt{Th_T} P_H \psi_T(w) \xrightarrow{d} N\left[0, \left(\int K^2(x)dx\right)\alpha_0(w)\right], \quad \lambda\text{-a.s. in } w.$$

Moreover, from Lemma 10:

$$\sqrt{T}(g, P_H \psi_T)_{L^2(\lambda)} = \sqrt{T}(P_H g, \psi_T)_{L^2(\lambda)} \xrightarrow{d} N\left[0, (P_H g, I_H P_H g)_{L^2(\lambda)}\right], \quad g \in L^2(\lambda).$$

We deduce:

---

[20]The residual term of this asymptotic expansion has the same order as in Proposition 4 and can be neglected for the derivation of the asymptotic distribution of $\widehat{A}_T$.

Proposition 16 : *Under the regularity assumptions in Appendix 2 including bandwidth condition (14):*

$$\sqrt{T} \left(g, \widehat{A}_T - A_0 \right)_{L^2(\lambda)} \xrightarrow{d} N \left[0, \left(g, P_H I_H^{-1} P_H g \right)_{L^2(\lambda)} \right], \ g \in L^2(\lambda).$$

When in addition the differential operator admits a decomposition (5):

$$\sqrt{Th_T} \left(\widehat{A}_T(v) - A_0(v) \right) \xrightarrow{d} N \left(0, \left(\int K^2(x) dx \right) \alpha_{0,H}(v)^{-1} \right),$$

λ-*a.s in* v.

6.3 The nonparametric efficiency bound for linear forms

The following proposition reports the efficiency bound $B_A(g)$ for linear forms $(g, A)_{L^2(\lambda)}$, $g \in L^2(\lambda)$, under the constraint $A \in \widetilde{\mathcal{A}}$.

Proposition 17 : *The nonparametric efficiency bound is given by:*

$$B_A(g) = (g, P_H I_H^{-1} P_H g)_{L^2(\lambda)}, \ \ g \in L^2(\lambda).$$

The constrained minimum chi-square estimator is therefore nonparametrically efficient for linear forms.

7 Concluding remarks

The analysis of nonlinear dependence is crucial for financial applications and requires an appropriate specification of the joint density for often a rather large dimension. To avoid the curse of dimensionality and to select models with structural interpretations the density cannot be let unconstrained. At the opposite a standard parametric specification is generally too restrictive to get the expected fit. In this paper we have considered the intermediate case in which the conditional distribution or the copula depends on one-dimensional functional parameters. The functional parameter is defined up to a one to one transformation. We have explained what representation of the functional parameter has to be selected to get results on the information operator, efficiency bound, and efficient estimators similar to the standard results of the pure parametric framework. The approach has been illustrated by discussing different families of constrained nonparametric densities, useful for financial and duration analysis.

Appendix 1

Differential and information operators for the Archimedean copula

A.1.1 Differential of the copula

The log-copula density is given by:

$$\log c(u, v; a) = \log a\left[C(u, v; a)\right] + \log F^*\left[C(u, v; a); a\right] - \log F^*(u; a) - \log F^*(v; a),$$

where $a = f^*$. Let us derive the differential with respect to a. We get:

$$\langle D\log c(u, v; a), h\rangle = \frac{h\left[C(u, v; a)\right]}{a\left[C(u, v; a)\right]} + \langle D\log F^*\left[C(u, v; a); a\right], h\rangle - \langle D\log F^*(u; a), h\rangle - \langle D\log F^*(v; a), h\rangle$$

$$+ \left(\frac{d\log a}{dw}\left[C(u, v; a)\right] + \frac{a\left[C(u, v; a)\right]}{F^*\left[C(u, v; a); a\right]}\right)\langle DC(u, v; a), h\rangle. \tag{a.1}$$

Let us now derive the differentials of $C(u, v; a)$ and $F^*(u, v; a)$ w.r.t. a.

i) Differential of $C(u, v; a)$. We get:

$$
\begin{aligned}
\langle DC(u, v; a), h\rangle &= \left\langle D\phi\left[\phi^{-1}(u; a) + \phi^{-1}(v; a); a\right], h\right\rangle \\
&\quad + \phi'\left[\phi^{-1}(u; a) + \phi^{-1}(v; a); a\right]\left\{\left\langle D\phi^{-1}(u; a), h\right\rangle + \left\langle D\phi^{-1}(v; a), h\right\rangle\right\} \\
&= \left\langle D\phi\left(\phi^{-1}\left[C(u, v; a); a\right]; a\right), h\right\rangle \\
&\quad + \phi'\left(\phi^{-1}\left[C(u, v; a); a\right]; a\right)\left\{\left\langle D\phi^{-1}(u; a), h\right\rangle + \left\langle D\phi^{-1}(v; a), h\right\rangle\right\}.
\end{aligned}
$$

By the implicit function theorem we have:

$$\left\langle D\phi\left[\phi^{-1}(y; a); a\right], h\right\rangle = -\phi'\left[\phi^{-1}(y; a); a\right]\left\langle D\phi^{-1}(y; a), h\right\rangle,$$

and thus we get:

$$\langle DC(u, v; a), h\rangle = F^*\left[C(u, v; a); a\right]\left\{\left\langle D\phi^{-1}\left[C(u, v; a); a\right], h\right\rangle - \left\langle D\phi^{-1}(u; a), h\right\rangle - \left\langle D\phi^{-1}(v; a), h\right\rangle\right\}. \tag{a.2}$$

ii) Differential of $F^*(y; a)$. We have:

$$\left\langle D\log F^*(y; a), h\right\rangle = \frac{1}{F^*(y; a)}\int_0^y h(v)dv = E_a\left[h(W)/a(W) \mid Z = y\right]. \tag{a.3}$$

31

By inserting (a.2) and (a.3) in (a.1) we get:

$$\langle D \log c(u,v;a), h \rangle = \frac{h\left[C(u,v;a)\right]}{a\left[C(u,v;a)\right]} + E_a\left[h(W)/a(W) \mid Z = C(u,v;a)\right]$$

$$- E_a\left[h(W)/a(W) \mid Z = u\right] - E_a\left[h(W)/a(W) \mid Z = v\right]$$

$$+ \left\{ a\left[C(u,v;a)\right] + \frac{d \log a}{dw}\left[C(u,v;a)\right] F^*\left[C(u,v;a); a\right] \right\}$$

$$\cdot \left\{ \left\langle D\phi^{-1}\left[C(u,v;a); a\right], h \right\rangle - \left\langle D\phi^{-1}(u;a), h \right\rangle - \left\langle D\phi^{-1}(v;a), h \right\rangle \right\}. \tag{a.4}$$

Let us finally compute the differential of $\phi^{-1}(y;a)$ with respect to a.

iii) Differential of $\phi^{-1}(y;a)$. We have:

$$\phi^{-1}(y;a) = \int_y^1 \frac{dw}{\int_0^w a(v)dv}.$$

Let us consider the first order expansion:

$$\phi^{-1}(y; a+h) = \int_y^1 \frac{dw}{\int_0^w a(v)\,dv + \int_0^w h(v)dv} \simeq \int_y^1 \frac{1}{\int_0^w a(v)\,dv}\left[1 - \frac{\int_0^w h(v)dv}{\int_0^w a(v)\,dv}\right] dw$$

$$\simeq \phi^{-1}(y;a) - \int_y^1 \frac{\int_0^w h(v)dv}{F^*(w;a)^2} dw.$$

Thus:

$$\langle D\phi^{-1}(y;a), h \rangle = -\int_y^1 \frac{1}{F^*(w;a)^2}\left(\int_0^w h(v)dv\right) dw$$

$$= \left(\int_w^1 \frac{dv}{F^*(v;a)^2}\right)\left(\int_0^w h(v)dv\right) \Big|_y^1 - \int_y^1 \left(\int_w^1 \frac{dv}{F^*(v;a)^2}\right) h(w)dw$$

$$= -\left(\int_y^1 \frac{dv}{F^*(v;a)^2}\right)\int_0^y h(v)dv - \int_y^1 \left(\int_w^1 \frac{dv}{F^*(v;a)^2}\right) h(w)dw.$$

$$= k(y;a)\int_0^y h(v)dv + \int_y^1 k(w;a)h(w)dw, \tag{a.5}$$

where $k(y;a) = -\int_y^1 \left(1/F^*(v)^2\right) dv$.

By inserting (a.5) in (a.4), we get the differential of the copula density, which is of the form:

$$\langle D \log c(u,v;a), h \rangle = \frac{h\left[C(u,v;a)\right]}{a\left[C(u,v;a)\right]} + \int_0^1 \gamma(u,v,w;a)\, h(w)dw, \quad \text{say.}$$

A.1.2 The information operator

Let us now compute the information operator I_c of the copula. We get:

$$
\begin{aligned}
E_0 \left[\langle D \log c(U, V; a_0), g \rangle \langle D \log c(U, V; a_0), h \rangle \right] &= E_0 \left\{ g\left[C_0(U,V)\right] h\left[C_0(U,V)\right] / a_0 \left[C_0(U,V)\right]^2 \right\} \\
&+ \int E_0 \left\{ g\left[C_0(U,V)\right] \gamma\left(U,V,y\right) / a_0 \left[C_0(U,V)\right] \right\} h(y) dy \\
&+ \int E_0 \left\{ \gamma\left(U,V,y\right) h\left[C_0(U,V)\right] / a_0 \left[C_0(U,V)\right] \right\} g(y) dy \\
&+ \int \int E_0 \left\{ \gamma\left(U,V,x\right) \gamma\left(U,V,y\right) \right\} g(x) h(y) dx dy.
\end{aligned}
$$

Let us consider the four terms of the decomposition separately. The first one is:

$$
E_0 \left\{ g\left[C_0(U,V)\right] h\left[C_0(U,V)\right] / a_0 \left[C_0(U,V)\right]^2 \right\} = \int g(w) h(w) \frac{f_W(w; a_0)}{a_0(w)^2} dw,
$$

where $f_W(.; a_0)$ is the density of W. The second term is:

$$
\begin{aligned}
&\int E_0 \left\{ g\left[C_0(U,V)\right] \gamma\left(U,V,y\right) / a_0 \left[C_0(U,V)\right] \right\} h(y) dy \\
&= \int E_0 \left\{ g(W) \widetilde{\gamma}\left(W,Z,y\right) / a_0(W) \right\} h(y) dy, \quad \text{say}, \\
&= \int \int g(w) \frac{E_0 \left\{ \widetilde{\gamma}\left(W,Z,y\right) \mid W = w \right\} f_W\left(w, a_0\right)}{a_0(w)} h(y) dw dy,
\end{aligned}
$$

where function $\widetilde{\gamma}$ is such that $\widetilde{\gamma}\left(C(u,v), v, y\right) = \gamma(u,v,y)$. Similarly we get the third and fourth terms:

$$
\int \int g(y) \frac{E_0 \left\{ \widetilde{\gamma}\left(W,Z,y\right) \mid W = w \right\} f_W\left(w, a_0\right)}{a_0(w)} h(w) dy dw,
$$

and :

$$
\int \int g(x) E_0 \left\{ \widetilde{\gamma}\left(W,Z,x\right) \widetilde{\gamma}\left(W,Z,y\right) \right\} h(y) dx dy, \quad \text{respectively}.
$$

Thus I_c satisfies Assumption A.3, with local component $\alpha_0(w; a) = f_W(w; a) / a_0(w)^2$, and:

$$
\begin{aligned}
\alpha_1(x, y; a) &= E_a \left\{ \widetilde{\gamma}\left(W,Z,y\right) \mid W = x \right\} f_W\left(x, a\right) / a(x) \\
&+ E_a \left\{ \widetilde{\gamma}\left(W,Z,x\right) \mid W = y \right\} f_W\left(y, a\right) / a(y) + E_a \left\{ \widetilde{\gamma}\left(W,Z,x\right) \widetilde{\gamma}\left(W,Z,y\right) \right\}.
\end{aligned}
$$

Finally the density of W is given by (see website):

$$
f_W(w) = 1 + \frac{1}{\phi'\left[\phi^{-1}(w)\right]} F^*(w) + \phi^{-1}(w) f^*(w) = \phi^{-1}(w) f^*(w).
$$

Asymptotic properties of minimum chi-square estimator

The asymptotic properties of the minimum chi-square estimator are derived in this Appendix for $d = 2$ observed variables and one-dimensional functional parameter. However the results are easily extended to any number d of observed variables with still one-dimensional functional parameter. This multivariate framework is considered only when we discuss the regularity conditions on the bandwidth and its optimal choice.

A.2.1 Set of regularity conditions

Let us first describe the set of additional regularity conditions which are used for asymptotic analysis.

Assumption A.5: (X_t, Y_t), t varying, is a strictly stationary process, with β-mixing coefficients $\beta(k)$ such that: $\beta(k) = O(\rho^k)$, $\rho < 1$.

Assumption A.6: The stationary density f of (X_t, Y_t) has compact support $[0,1]^d$, vanishes at its boundary, and is of class $C^m\left([0,1]^d\right)$.

Assumption A.7: There exist $C > 0$, $\gamma > 0$, and an increasing sequence of sets $\Omega_T \subset (0,1)^d$, $T \in \mathbb{N}$, such that:

$$\inf_{(x,y) \in \Omega_T} f(x,y) > C(\log T)^{-\gamma}, \quad \text{for any } T.$$

Assumption A.8: The conditional density $f_h(z \mid w)$ of (X_t, Y_t) given $(X_{t-h}, Y_{t-h}) = w$ is such that:

$$\sup_{h \in \mathbb{N}} \sup_{z,w \in [0,1]^d : f(w) > 0} f_h(z \mid w) < +\infty.$$

Assumption A.9: The kernel K is of class C^m, with derivatives in $L^2(\mathbb{R})$, and is Lipschitz. Moreover the kernel K is of order $m \geq 2$, that is:

$$\int u^s K(u) du = 0, \quad s = 1, ..., m-1, \quad \text{and} \quad \int |u|^m K(u) du < +\infty.$$

Assumption A.10: The bandwidth h_T is such that:

$$h_T = c_T T^{-\alpha}, \ lim_{T \to \infty} c_T = c > 0, \quad with \quad \frac{1}{4m}\left(1 + \frac{2m-1}{4m^2 + 2m + 1}\right) < \alpha < \frac{1}{2d}\left(1 - \frac{1}{2}\frac{2m-1}{4m^2 + 2m + 1}\right).$$

Assumption A.11: There exists compact sets $\widetilde{\Omega}_T, \Omega_T$ such that $\widetilde{\Omega}_T \subset \Omega_T \subset [0,1]^d$, weighting function ω_T has support in Ω_T, is smaller than 1 with restriction $\omega_T|_{\widetilde{\Omega}_T} = 1$, $T \in \mathbb{N}$, and $\lambda_d\left(\widetilde{\Omega}_T\right) \to 1$, as $T \to \infty$,

where λ_d is the Lebesgue measure.

Assumption A.12: *Set Ω_T in Assumption A.11 satisfies Assumption A.7.*

Assumption A.13: *For any $A, A_0 \in \Theta$: $f(X,Y;A)/f(X,Y) \in L^2(P_{A_0})$. Moreover the first-order expansion of the density is such that:*

$$f(x,y;A+h) = f(x,y;A) + \langle Df(x,y;A), h \rangle + R(x,y;A,h), \quad A, A+h \in \Theta,$$

where:

i) $Df(X,Y;A)/f(X,Y)$ is a bounded operator from $L^2(\lambda)$ in $L^2(P_{A_0}), \forall A, A_0 \in \Theta$;

ii) the residual term satisfies: $\|R(X,Y;A,h)/f(X,Y)\|_{L^2(P_{A_0})} = O\left(\|h\|_{L^2(\lambda)}^2\right), \quad A, A+h \in \Theta.$

iii) Moreover: $R(x,y;A_0,h) = O\left(\left(\|h(x)\| + \|h(y)\| + \|h\|_{L^2(\lambda)}\right)^2\right), \quad \lambda_2\text{-a.s. in } (x,y) \in (0,1)^2.$

Assumption A.14: *There exist $p > 1$ such that:*

$$\sup_{A \in \Theta} \left\| \frac{f(.,.;A)^2}{f(.,.)} \right\|_{L^p} < \infty.$$

Assumption A.15: *The set Θ is bounded and closed with respect to the norm $\|.\|_{L^2(\lambda)}$.*

Assumption A.16: *The set $\{f(.,.;A), A \in \Theta\}$ is bounded and closed in $L^2(\mu)$ for any measure μ on $(0,1)^d$ with compact support and continuous density w.r.t λ_d.*

Assumption A.17: *The information operator I is such that:*

$$\inf_{h:\|h\|_{L^2(\lambda)}=1} (h, Ih)_{L^2(\lambda)} > 0.$$

Assumption A.18: *With probability approaching 1:*

$$\text{for any } g \in L^2(\lambda) : \widehat{A}_T + tg \in \Theta \text{ for } t \text{ in a neighbourhood of } 0.$$

Assumption A.19: *The operator $\frac{Df(X,Y;A)}{f(X,Y)}$ is Lipschitz with respect to A at A_0 :*

$$\left\| \frac{Df(X,Y;A_0+h)}{f(X,Y)} - \frac{Df(X,Y;A_0)}{f(X,Y)} \right\|_L \leq C \|h\|_{L^2(\lambda)},$$

for some constant C, where $\|.\|_L$ denotes the L^2-norm on the space of bounded linear operators from $L^2(\lambda)$ into $L^2(P_{A_0})$.

Assumption A.20: *There exists $p > 1$ such that:*

$$\|\langle D \log f(.,.; A_0), g \rangle \langle D \log f(.,.; A_0), h \rangle f(.,.)\|_{L^p} = O\left(\|g\|_{L^2(\lambda)} \|h\|_{L^2(\lambda)}\right).$$

Assumption A.21: *There exists $\beta_2 > q/4$ such that:*

$$\lambda_2(\widetilde{\Omega}_T^c) = O\left(T^{-\beta_2}\right),$$

where p is the value given in Assumption A.20, $1/p + 1/q = 1$, and $\widetilde{\Omega}_T$ is defined in Assumption A.11.

Assumptions A.5-A.10 are standard conditions on the distribution and serial dependence of process (X_t, Y_t), on the kernel K and on bandwidth h_T. In particular, when $d = 2$, it is easily checked that Assumption A.10 can be satisfied whenever $m \geq 2$. Assumptions A.5-A.10 are used to prove standard results on the convergence of kernel density estimator $\widehat{f}_T$, which will be used later on to derive the asymptotic properties of the minimum chi-square estimator. In particular we have Lemma A.1 below, which follows from Theorem 2.2 in Bosq (1998) [see the website for additional results on kernel estimators used in the proofs].

Lemma A.1 : *Under Assumptions A.5, A.6, A.7 and A.9:*

i) If the bandwidth h_T is such that $h_T = c_T T^{-\alpha}$, $\lim_{T \to \infty} c_T = c > 0$, with $0 < \alpha < 1/d$, then:

$$\tau_{T,1} := \sup_{(x,y) \in \Omega_T} \left| \frac{\widehat{f}_T(x,y) - f(x,y)}{\widehat{f}_T(x,y)} \right| = o_p(1).$$

ii) If the bandwidth h_T satisfies Assumption A.10, then there exists $\beta_1 > \frac{1}{4}\left(1 + \frac{1}{2}\frac{2m-1}{4m^2+2m+1}\right)$ such that:

$$\tau_{T,1} = o_p\left(T^{-\beta_1}\right).$$

Assumptions A.11 and A.12 explain how the sequence of weighting functions ω_T with compact support Ω_T converges to the constant function 1 on $(0,1)^d$. Assumption A.13 introduces integrability conditions for the family $f(x,y; A)$ and for its Hadamard derivative with respect to A. It is used to define the chi-square criterion and to prove its continuity. In particular for $A = A_0$ Assumption A.13 i) reduces to Assumption A.2 i). Assumption A.14 implies that the densities in the family have similar patterns at the boundary. Assumption A.15 describes the set of admissible values of functional parameter A. Assumption A.16 is useful to prove the existence of the minimum chi-square estimator $\widehat{A}_T$, by using the fact that the criterion defining $\widehat{A}_T$ is a distance in an Hilbert space. Assumption A.17 is useful for local identification of the functional parameter, and implies in particular that the information operator I is invertible, that is

Assumption A.4 (see Yosida, 1995, Theorem 2, p.320). When the information operator I satisfies Assumption A.3, sufficient conditions for Assumption A.17 can be easily derived. Assumption A.18 is used to derive first order expansions. When the parameter set Θ has a non-empty interior (w.r.t. $\|.\|_{L^2(\lambda)}$) containing the true functional parameter A_0, Assumption A.18 is satisfied whenever $\widehat{A}_T$ is consistent. Finally, the last three Assumptions A.19-A.21 are technical conditions required to bound the residual terms in the asymptotic expansions.

In the time series framework of Section 4.5, this set of assumptions is replaced by a similar one, in which $f(x,y)$, $Df(X,Y;A_0)$ and I are replaced by $f(x|y)$, $Df(X|Y;A_0)$ and $I_{X|Y}$, respectively, in Assumptions A.13, A.14, A.16, A.17, A.19 and A.20. These new assumptions are refereed to as TS.

A.2.2 Existence of the minimum chi-square estimator

The chi-square criterion Q_T is the $L^2(\mu_T)$-distance between the constrained nonparametric family of densities $f(.,.;A)$, $A \in \Theta$, and the unconstrained kernel estimator $\widehat{f}_T$, where measure μ_T has density $\omega_T/\widehat{f}_T$. Under Assumption A.16 the projection of $\widehat{f}_T$ on the set $\{f(.,.;A), A \in \Theta\}$ with respect to $L^2(\mu_T)$ is well-defined. We deduce that a solution $\widehat{A}_T$ exists.

A.2.3 Consistency of the minimum chi-square estimator: proof of Proposition 2

As usual the proof is based on the analysis of the asymptotic criterion:

$$Q_\infty(A) = Q(A) = \int \int \frac{[f(x,y) - f(x,y;A)]^2}{f(x,y)} dx dy, \quad A \in \Theta,$$

We have the following Lemma A.2. In particular, the consistency of the minimum chi-square estimator is a direct consequence of Lemma A.2 i), iv) and v).

Lemma A.2: *i) Under Assumption A.13 the chi-square criterion Q is continuous.*

ii) Under Assumptions A.13 and A.15: $Q(A_0 + h) = (h, Ih)_{L^2(\lambda)} + O\left(\|h\|^3_{L^2(\lambda)}\right)$, for $A_0 + h \in \Theta$.

iii) Under Assumptions A.13 and A.15: $\sup_{A \in \Theta} Q(A) < \infty$.

iv) Under Assumptions A.13, A.15 and A.17 parameter A_0 is locally identified, that is:

$$\forall \varepsilon > 0 : \inf_{A \in \Theta \cap B_\varepsilon(A_0)^c} Q(A) > Q(A_0),$$

where $B_\varepsilon(A_0)$ denotes a ball of radius ε around A_0, w.r.t. the norm $\|.\|_{L^2(\lambda)}$.

v) Under Assumptions A.1, A.2, A.5-A.9, A.11-A.15 and bandwidth condition $h_T = c_T T^{-\alpha}$,

$lim_{T \to \infty} c_T = c > 0$, with $0 < \alpha < 1/d$, the criterion Q_T converges in probability to Q uniformly in $A \in \Theta$.

Proof of Lemma A.2: The proofs of i)-iii) are simple and given on the website. Let us first focus on the proof of iv). From Lemma A.2 ii), for any h such that $A_0 + h \in \Theta \backslash B_\varepsilon(A_0)$ we get:

$$
\begin{aligned}
Q(A_0 + h) &\geq (h, Ih)_{L^2(\lambda)} \left[1 - C \|h\|_{L^2(\lambda)} \frac{\|h\|_{L^2(\lambda)}^2}{(h, Ih)_{L^2(\lambda)}} \right], \text{ for some constant } C > 0, \\
&\geq (h, Ih)_{L^2(\lambda)} \left[1 - C \sup_{h \in (\Theta - A_0)} \|h\|_{L^2(\lambda)} \left(\inf_{h : \|h\|_{L^2(\lambda)} = 1} (h, Ih)_{L^2(\lambda)} \right)^{-1} \right].
\end{aligned}
$$

From Assumption A.17 we have $k = \inf_{h : \|h\|_{L^2(\lambda)} = 1} (h, Ih)_{L^2(\lambda)} > 0$. Moreover without loss of generality we can assume that $\sup_{h : A_0 + h \in \Theta} \|h\|_{L^2(\lambda)} < k/2C$. Then for any h such that $A_0 + h \in \Theta \backslash B_\varepsilon(A_0)$, we get:

$$
Q(A_0 + h) \geq \frac{1}{2} (h, Ih)_{L^2(\lambda)}.
$$

Thus:

$$
\inf_{A \in \Theta \backslash B_\varepsilon(A_0)} Q(A) \geq \inf_{h \in (\Theta - A_0) \backslash B_\varepsilon(0)} Q(A_0 + h) \geq \frac{\varepsilon^2}{2} \inf_{h \in (\Theta - A_0) \backslash B_\varepsilon(0)} \frac{(h, Ih)_{L^2(\lambda)}}{\|h\|_{L^2(\lambda)}^2} \geq \frac{k}{2} \varepsilon^2 > 0 = Q(A_0), \quad \text{(a.6)}
$$

and iv) is proved.

Let us finally consider v). By developing the criterion functions we get:

$$
\begin{aligned}
Q_T(A) - Q(A) &= \left(\int \int \widehat{f}_T(x, y) \omega_T(x, y) dx dy - 1 \right) - 2 \left(\int \int f(x, y; A) \omega_T(x, y) dx dy - 1 \right) \\
&\quad + \int_0^1 \int_0^1 f(x, y; A)^2 \left(\frac{1}{\widehat{f}_T(x, y)} - \frac{1}{f(x, y)} \right) \omega_T(x, y) dx dy \\
&\quad + \int_0^1 \int_0^1 \frac{f(x, y; A)^2}{f(x, y)} (\omega_T(x, y) - 1) \, dx dy \\
&\equiv S_{1,T} + S_{2,T} + S_{3,T} + S_{4,T}, \text{ say.}
\end{aligned}
$$

Let us now check that each term converges in probability to 0, uniformly in $A \in \Theta$. We have:

$$
\begin{aligned}
|S_{1,T}| &= \left| \int \int \widehat{f}_T(x, y) (\omega_T(x, y) - 1) \, dx dy \right| \leq \int \int \left| \widehat{f}_T(x, y) \right| |\omega_T(x, y) - 1| \, dx dy \\
&\leq \int \int \left| \widehat{f}_T(x, y) \right| \mathbb{I}_{\widetilde{\Omega}_T^c}(x, y) dx dy \\
&\leq \int \int \left| \widehat{f}_T(x, y) - f(x, y) \right| dx dy + \int \int f(x, y) \mathbb{I}_{\widetilde{\Omega}_T^c}(x, y) dx dy \\
&\leq \left(\int \int \left[\widehat{f}_T(x, y) - f(x, y) \right]^2 dx dy \right)^{1/2} + P_{A_0} \left[(X_t, Y_t) \in \widetilde{\Omega}_T^c \right] = o_p(1),
\end{aligned}
$$

uniformly in $A \in \Theta$, from Theorem 4.1 in Gouriéroux and Tenreiro (2001) and Assumption A.11. The proof is similar for $S_{2,T}$:

$$
\begin{aligned}
|S_{2,T}| & \leq 2 \int \int f(x,y;A) \mathbb{I}_{\widetilde{\Omega}_T^c}(x,y) dx dy \leq 2 \left(\int \int \frac{f(x,y;A)^2}{f(x,y)} dx dy \right)^{\frac{1}{2}} \left(\int \int \mathbb{I}_{\widetilde{\Omega}_T^c}(x,y) f(x,y) dx dy \right)^{\frac{1}{2}} \\
& \leq 2 \left(\sup_{A \in \Theta} Q(A) + 1 \right)^{\frac{1}{2}} P_{A_0} \left[(X_t, Y_t) \in \widetilde{\Omega}_T^c \right]^{1/2} = o_p(1)
\end{aligned}
$$

in probability uniformly in $A \in \Theta$ due to Assumption A.11 and Lemma A.2 iii).

Let us now consider $S_{3,T}$:

$$
\begin{aligned}
|S_{3,T}| & \leq \int_0^1 \int_0^1 \frac{f(x,y;A)^2}{f(x,y)} \left| \frac{\widehat{f}_T(x,y) - f(x,y)}{\widehat{f}_T(x,y)} \right| \omega_T(x,y) dx dy \\
& \leq \sup_{(x,y) \in \Omega_T} \left| \frac{\widehat{f}_T(x,y) - f(x,y)}{\widehat{f}_T(x,y)} \right| \cdot \int_0^1 \int_0^1 \frac{f(x,y;A)^2}{f(x,y)} dx dy \\
& \leq \left(\sup_{A \in \Theta} Q(A) + 1 \right) \sup_{(x,y) \in \Omega_T} \left| \frac{\widehat{f}_T(x,y) - f(x,y)}{\widehat{f}_T(x,y)} \right| = o_p(1),
\end{aligned}
$$

in probability uniformly in $A \in \Theta$ due to Lemma A.1 i) and Lemma A.2 iii).

Finally, the last term $S_{4,T}$ is such that:

$$
\begin{aligned}
|S_{4,T}| & \leq \int \int \frac{f(x,y;A)^2}{f(x,y)} |\omega_T(x,y) - 1| dx dy \leq \int \int \frac{f(x;y;A)^2}{f(x,y)} \mathbb{I}_{\widetilde{\Omega}_T^c}(x,y) dx dy \\
& \leq \left\| \frac{f(.,.;A)^2}{f(.,.)} \right\|_{L^p} \left\| \mathbb{I}_{\widetilde{\Omega}_T^c} \right\|_{L^q}, \quad \text{where } \frac{1}{p} + \frac{1}{q} = 1, \\
& \leq \sup_{A \in \Theta} \left\| \frac{f(.,.;A)^2}{f(.,.)} \right\|_{L^p} \cdot \lambda_2 \left(\widetilde{\Omega}_T^c \right)^{1/q} = o_p(1), \text{ uniformly in } A \in \Theta,
\end{aligned}
$$

due to Assumptions A.11 and A.14.   Q.E.D.

A.2.4 Asymptotic expansion of the minimum chi-square estimator

i) Expansion of the first order condition

From Assumption A.18, with probability approaching 1, $\widehat{A}_T$ satisfies the set of first order conditions:

$$
\int \int \frac{\widehat{f}_T(x,y) - f(x,y;\widehat{A}_T)}{\widehat{f}_T(x,y)} \left\langle Df \left(x,y;\widehat{A}_T \right), g \right\rangle \omega_T(x,y) dx dy = 0, \ \forall g \in L^2(\lambda).
$$

Let us denote $\delta\widehat{A}_T = \widehat{A}_T - A_0$. We expand the functions involved in the first order condition. We can write:

$$
\begin{aligned}
f\left(x, y; \widehat{A}_T\right) &= f(x,y) + \left\langle Df\left(x,y; A_0\right), \delta\widehat{A}_T \right\rangle + R\left(x,y; \delta\widehat{A}_T\right), \\
\left\langle Df\left(x,y; \widehat{A}_T\right), g \right\rangle &= \left\langle Df\left(x,y; A_0\right), g \right\rangle + \widetilde{R}\left(x,y; \delta\widehat{A}_T, g\right).
\end{aligned}
$$

where R and $\widetilde{R}$ are residual terms. By writing:

$$
\frac{1}{\widehat{f}_T(x,y)} = \frac{1}{f(x,y)}\left(1 - \frac{\delta\widehat{f}_T(x,y)}{\widehat{f}_T(x,y)}\right),
$$

and using the definitions of efficient score ψ_T and information operator I, the first order condition can be rewritten as:

$$
\left(g, \psi_T - I\delta\widehat{A}_T\right)_{L^2(\lambda)} + R\left(\delta\widehat{A}_T, g\right) = 0, \ \forall g \in L^2(\lambda), \tag{a.7}
$$

where the residual term $R\left(\delta\widehat{A}_T, g\right)$ is:

$$
\begin{aligned}
&R\left(\delta\widehat{A}_T, g\right) \\
&= -\int\int \delta\widehat{f}_T(x,y) \left\langle D\log f\left(x,y; A_0\right), g\right\rangle \frac{\delta\widehat{f}_T(x,y)}{\widehat{f}_T(x,y)} \omega_T(x,y)dxdy \\
&\quad -\int\int \left\langle D\log f\left(x,y; A_0\right), \delta\widehat{A}_T\right\rangle \left\langle D\log f\left(x,y; A_0\right), g\right\rangle f(x,y) \left[\left(1 - \frac{\delta\widehat{f}_T(x,y)}{\widehat{f}_T(x,y)}\right)\omega_T(x,y) - 1\right] dxdy \\
&\quad -\int\int R\left(x,y; \delta\widehat{A}_T\right)\left\langle D\log f\left(x,y; A_0\right), g\right\rangle \left(1 - \frac{\delta\widehat{f}_T(x,y)}{\widehat{f}_T(x,y)}\right)\omega_T(x,y)dxdy \\
&\quad +\int\int \frac{\delta\widehat{f}_T(x,y)}{f(x,y)}\widetilde{R}\left(x,y; \delta\widehat{A}_T, g\right)\left(1 - \frac{\delta\widehat{f}_T(x,y)}{\widehat{f}_T(x,y)}\right)\omega_T(x,y)dxdy \\
&\quad -\int\int \frac{\left\langle Df\left(x,y; A_0\right), \delta\widehat{A}_T\right\rangle}{f(x,y)}\widetilde{R}\left(x,y; \delta\widehat{A}_T, g\right)\left(1 - \frac{\delta\widehat{f}_T(x,y)}{\widehat{f}_T(x,y)}\right)\omega_T(x,y)dxdy \\
&\quad -\int\int \frac{R\left(x,y; \delta\widehat{A}_T\right)}{f(x,y)}\widetilde{R}\left(x,y; \delta\widehat{A}_T, g\right)\left(1 - \frac{\delta\widehat{f}_T(x,y)}{\widehat{f}_T(x,y)}\right)\omega_T(x,y)dxdy \\
&\equiv R_1\left(\delta\widehat{A}_T, g\right) + R_2\left(\delta\widehat{A}_T, g\right) + R_3\left(\delta\widehat{A}_T, g\right) + R_4\left(\delta\widehat{A}_T, g\right) + R_5\left(\delta\widehat{A}_T, g\right) + R_6\left(\delta\widehat{A}_T, g\right).
\end{aligned}
$$

ii) A bound for the residual term

The following Lemma provides a bound for the residual term $R\left(\delta\widehat{A}_T, g\right)$ [see the website for the proof].

Lemma A.3: *Under Assumptions A.1-A.2, A.5-A.13, A.19 and A.20 the residual term $R\left(\delta\widehat{A}_T, g\right)$ in (a.7)*

is such that:

$$R\left(\delta\widehat{A}_T, g\right) = \|g\|_{L^2(\lambda)} \; O_p\left[\tau_{T,1}^2 + (\tau_{T,1} + \tau_{T,2})\left\|\delta\widehat{A}_T\right\|_{L^2(\lambda)} + \left\|\delta\widehat{A}_T\right\|_{L^2(\lambda)}^2\right],$$

where

$$\tau_{T,1} := \sup_{(x,y)\in\Omega_T}\left|\frac{\delta\widehat{f}_T(x,y)}{\widehat{f}_T(x,y)}\right|, \quad \tau_{T,2} := \lambda_2\left(\widetilde{\Omega}_T^c\right)^{1/q},$$

q is defined as in Assumption A.21, and the O_p term is uniform w.r.t. g.

iii) The residual term is negligible in norm: proof of Proposition 4 i)

Let us now show that the residual term is negligible in norm with respect to the other terms in the expansion (a.7) of the first order condition.

Lemma A.4: *Under Assumptions A.1-A.2, A.4-A.13, and A.19-A.21:*

i) $\left\|\delta\widehat{A}_T\right\|_{L^2(\lambda)} = O_p\left(1/\sqrt{T} + h_T^m\right).$

ii) $\left(g, I\delta\widehat{A}_T\right)_{L^2(\lambda)} = (g, \psi_T)_{L^2(\lambda)} + o_p\left(1/\sqrt{T}\right),$ *uniformly in* $g \in L^2(\lambda).$

iii) $\left\|I\delta\widehat{A}_T - \psi_T\right\|_{L^2(\lambda)} = o_p\left(1/\sqrt{T}\right).$

Proof of Lemma A.4: i) To get a bound for $\left\|\delta\widehat{A}_T\right\|_{L^2(\lambda)}$, we first consider the expansion of $\left(g, \delta\widehat{A}_T\right)_{L^2(\lambda)}$ for any g. Since $\tau_{T,1} = o_p(T^{-1/4})$, $\tau_{T,2} = o(T^{-1/4})$ [see Lemma A.1 ii) and Assumption A.21], from Lemma A.3 we get:

$$R\left(\delta\widehat{A}_T, g\right) = \|g\|_{L^2(\lambda)}\left[o_p\left(1/\sqrt{T}\right) + o_p\left(T^{-1/4}\left\|\delta\widehat{A}_T\right\|_{L^2(\lambda)}\right) + O_p\left(\left\|\delta\widehat{A}_T\right\|_{L^2(\lambda)}^2\right)\right].$$

Then the first order condition is such that:

$$\left(g, I\delta\widehat{A}_T\right)_{L^2} = (g, \psi_T)_{L^2(\lambda)} + \|g\|_{L^2(\lambda)}\left[o_p\left(T^{-1/2}\right) + o_p\left(T^{-1/4}\left\|\delta\widehat{A}_T\right\|_{L^2(\lambda)}\right) + O_p\left(\left\|\delta\widehat{A}_T\right\|_{L^2(\lambda)}^2\right)\right], \quad \text{(a.8)}$$

uniformly in $g \in L^2(\lambda)$. Since I^{-1} is bounded (Assumption A.4), we get:

$$\left(g, \delta\widehat{A}_T\right)_{L^2(\lambda)} = (g, I^{-1}\psi_T)_{L^2(\lambda)} + \|g\|_{L^2(\lambda)}\left[o_p\left(1/\sqrt{T}\right) + o_p\left(T^{-1/4}\left\|\delta\widehat{A}_T\right\|_{L^2(\lambda)}\right) + O_p\left(\left\|\delta\widehat{A}_T\right\|_{L^2(\lambda)}^2\right)\right].$$

Let us now deduce a bound for $\left\| \delta \widehat{A}_T \right\|_{L^2(\lambda)}$. Since $(g, \psi_T - E\psi_T)_{L^2(\lambda)} = O_p(1/\sqrt{T})$ [see Lemma 10 in the text], $(g, E\psi_T)_{L^2(\lambda)} = O\left(\|g\|_{L^2(\lambda)} h_T^m \right)$ [see website] and I^{-1} is bounded (Assumption A.4) we get:

$$\left(g, I^{-1} \psi_T \right)_{L^2(\lambda)} = O_p\left(\|g\|_{L^2(\lambda)} \left(1/\sqrt{T} + h_T^m \right) \right).$$

Thus:

$$\left(g, \delta\widehat{A}_T \right)_{L^2(\lambda)} = \|g\|_{L^2(\lambda)} \left[O_p\left(1/\sqrt{T} + h_T^m \right) + o_p\left(T^{-1/4} \left\| \delta\widehat{A}_T \right\|_{L^2(\lambda)} \right) + O_p\left(\left\| \delta\widehat{A}_T \right\|_{L^2(\lambda)}^2 \right) \right], \quad g \in L^2(\lambda).$$

We get:

$$
\begin{aligned}
\left\| \delta\widehat{A}_T \right\|_{L^2(\lambda)} &= \sup_{g \in L^2(\lambda): \|g\|_{L^2(\lambda)} \leq 1} \left(g, \delta\widehat{A}_T \right)_{L^2(\lambda)} \\
&= O_p\left(1/\sqrt{T} + h_T^m \right) + o_p\left(T^{-1/4} \left\| \delta\widehat{A}_T \right\|_{L^2(\lambda)} \right) + O_p\left(\left\| \delta\widehat{A}_T \right\|_{L^2(\lambda)}^2 \right),
\end{aligned}
$$

that is $\left\| \delta\widehat{A}_T \right\|_{L^2(\lambda)} = O_p\left(1/\sqrt{T} + h_T^m \right)$.

ii) Since $h_T^{2m} = o\left(T^{-1/2} \right)$ by the bandwidth condition A.10, we deduce ii) directly from (a.8). Note that the o_p term is uniform w.r.t. g.

iii) Finally we have:

$\left\| I\delta\widehat{A}_T - \psi_T \right\|_{L^2(\lambda)} = \sup_{g \in L^2(\lambda): \|g\|_{L^2(\lambda)} \leq 1} \left(g, I\delta\widehat{A}_T - \psi_T \right)_{L^2(\lambda)} = o_p\left(1/\sqrt{T} \right).$ Q.E.D.

iv) The residual term is negligible pointwise: proof of Proposition 4 ii)

Let us now focus on pointwise expansions, which provide the pointwise rate of convergence of the residual term in the expansion of the first order condition. Intuitively, pointwise expansions are derived from the first order condition corresponding to a variation g of the functional parameter A which involves only its value at a point $x_0 \in [0,1]$. Such a variation will be approached by localization. More precisely we consider variations which are more and more concentrated around x_0 as $T \to \infty$, at a higher speed than the kernel localization. For expository purpose let us consider the case where A has a single component.

Let $\varphi \in C_0^\infty$ be a symmetric kernel with compact support, and let $\widetilde{h}_T$ be a bandwidth converging to 0. For any $x_0 \in [0,1]$, let us define the function:

$$g_{T,x_0}(x) = \frac{1}{\sqrt{\widetilde{h}_T}} \varphi\left(\frac{x - x_0}{\widetilde{h}_T} \right), \quad x \in [0,1].$$

Then:

$$\|g_{T,x_0}\|^2_{L^2(\lambda)} = \int \frac{1}{\widetilde{h}_T} \varphi \left(\frac{x-x_0}{\widetilde{h}_T} \right)^2 dx = \int \varphi(u)^2 \, du,$$

is constant in T. Moreover for any $h \in L^2(\lambda)$:

$$
\begin{aligned}
(g_{T,x_0}, h)_{L^2(\lambda)} &= \int \frac{1}{\sqrt{\widetilde{h}_T}} \varphi \left(\frac{x-x_0}{\widetilde{h}_T} \right) h(x) dx = \sqrt{\widetilde{h}_T} \int \varphi(u) h(x_0 + \widetilde{h}_T u) du \\
&= \sqrt{\widetilde{h}_T} h(x_0) + \sqrt{\widetilde{h}_T} \int \varphi(u) \left[h(x_0 + \widetilde{h}_T u) - h(x_0) \right] du.
\end{aligned}
$$

We will now derive a result similar to Lemma A.4 ii), in which the function $g = g_{T,x_0}$ depends on T.

Lemma A.5: *Let $g_T \in L^2(\lambda)$ for any T, such that $\|g_T\|_{L^2(\lambda)} \le const.$ Under the Assumptions of Lemma A.4:*

$$\sqrt{T} \left(g_T, I\delta\widehat{A}_T \right)_{L^2(\lambda)} = \sqrt{T} \, (g_T, \psi_T)_{L^2(\lambda)} + O_p \left(T^{-\beta^*} \right),$$

for some $\beta^* > \frac{1}{4} \frac{2m-1}{4m^2+2m+1}$.

Proof of Lemma A.5: Since the first order condition holds for any given T:

$$\left(g_T, I\delta\widehat{A}_T \right)_{L^2(\lambda)} = (g_T, \psi_T)_{L^2(\lambda)} + R \left(\delta\widehat{A}_T, g_T \right).$$

From Lemma A.1 ii), A.3, A.4 i) and Assumptions A.10, A.21 we get:

$$
\begin{aligned}
R \left(\delta\widehat{A}_T, g_T \right) &= \|g_T\|_{L^2(\lambda)} O_p \left[T^{-2\beta_1} + \left(T^{-\beta_1} + T^{-\beta_2/q} \right) \left(T^{-1/2} + h_T^m \right) + \left(T^{-1/2} + h_T^m \right)^2 \right] \\
&= O_p(T^{-\beta^*-1/2}),
\end{aligned}
$$

for $\beta^* = \min \left\{ 2 \left(\beta_1 - \frac{1}{4} \right), \beta_1, \frac{\beta_2}{q}, \alpha m - \frac{1}{4}, \frac{1}{2}, 2 \left(\alpha m - \frac{1}{4} \right) \right\} > \frac{1}{4} \frac{2m-1}{4m^2+2m+1}.$ Q.E.D.

Let us now apply Lemma A.5 to sequence $g_T = g_{T,x_0}$, where the bandwidth for localization $\widetilde{h}_T$ is such that[21]:

$$\widetilde{h}_T = o(h_T), \quad h_T = o(\widetilde{h}_T T^{2\beta^*}), \quad \sqrt{T\widetilde{h}_T}\widetilde{h}_T^m = o(1).$$

We get:

$$\sqrt{Th_T/\widetilde{h}_T} \left(g_{T,x_0}, I\delta\widehat{A}_T \right)_{L^2(\lambda)} = \sqrt{Th_T/\widetilde{h}_T} \, (g_{T,x_0}, \psi_T)_{L^2(\lambda)} + O_p \left(T^{-\beta^*} \sqrt{h_T/\widetilde{h}_T} \right). \tag{a.9}$$

---

[21] For instance $\widetilde{h}_T = \widetilde{c}_T T^{-\delta}$, $lim_{T\to\infty} \widetilde{c}_T = \widetilde{c} > 0$, with $\max \left\{ \alpha, \frac{1-\alpha}{2m} \right\} < \delta < \alpha + 2\beta^*$. This is possible under Assumption A.10.

Let us consider the RHS of (a.9). We get:

$$\sqrt{Th_T/\widetilde{h}_T}\,(g_{T,x_0},\psi_T)_{L^2(\lambda)} + O_p\left(T^{-\beta^*}\sqrt{h_T/\widetilde{h}_T}\right)$$

$$=\sqrt{Th_T}\psi_T(x_0) + \sqrt{Th_T}\int\varphi(u)\left[\psi_T\left(x_0+\widetilde{h}_Tu\right)-\psi_T(x_0)\right]du + o_p(1). \tag{a.10}$$

Let us now show that the second term is negligible. We have:

$$\sqrt{Th_T}\int\varphi(u)\left[\psi_T\left(x_0+\widetilde{h}_Tu\right)-\psi_T(x_0)\right]du$$

$$=\quad \sqrt{Th_T}\int\varphi(u)\left[(\psi_T-E\psi_T)\left(x_0+\widetilde{h}_Tu\right)-(\psi_T-E\psi_T)(x_0)\right]du$$

$$+\sqrt{Th_T}\int\varphi(u)\left[E\psi_T\left(x_0+\widetilde{h}_Tu\right)-E\psi_T(x_0)\right]du$$

$$=\quad O_p\left(\sqrt{Th_T}\frac{\widetilde{h}_T^2}{2}\frac{d^2}{dx^2}(\psi_T-E\psi_T)(x_0)\int u^2\varphi(u)\,du\right) + O(\sqrt{Th_T}\widetilde{h}_T^m),$$

if the kernel φ is of order $m \geq 2$. Since $\psi_T(x_0)$ involves partial moments of kernel estimators of a density [see (16)], we have $(\psi_T-E\psi_T)(x_0) = O_p\left[(Th_T)^{-1/2}\right]$ (see Lemma 7 in the text). Since each differentiation diminishes the rate of convergence of a kernel estimator by the factor h_T (see Theorem 3 in Aït-Sahalia, 1993), we deduce $\frac{d^2}{dx^2}(\psi_T-E\psi_T)(x_0) = O_p\left[(Th_T)^{-1/2}h_T^{-2}\right]$. Finally, since $\widetilde{h}_T = o(h_T)$ and $\sqrt{Th_T}\widetilde{h}_T^m = o(1)$, we conclude:

$$\sqrt{Th_T}\int\varphi(u)\left[\psi_T\left(x_0+\widetilde{h}_Tu\right)-\psi_T(x_0)\right]du = o_p(1), \tag{a.11}$$

and from (a.9), (a.10) it follows:

$$\sqrt{Th_T/\widetilde{h}_T}\left(g_{T,x_0}, I\delta\widehat{A}_T\right)_{L^2(\lambda)} = \sqrt{Th_T}\psi_T(x_0) + o_p(1). \tag{a.12}$$

Let us now consider the LHS of (a.12). We get:

$$\sqrt{Th_T/\widetilde{h}_T}\left(g_{T,x_0}, I\delta\widehat{A}_T\right)_{L^2(\lambda)} = \sqrt{Th_T}I\delta\widehat{A}_T(x_0) + \sqrt{Th_T}\int\varphi(u)\left[I\delta\widehat{A}_T\left(x_0+\widetilde{h}_Tu\right)-I\delta\widehat{A}_T(x_0)\right]du.$$

Thus, from (a.12) we get:

$$\sqrt{Th_T}I\delta\widehat{A}_T(x_0) = -\sqrt{Th_T}\int\varphi(u)\left[\left(I\delta\widehat{A}_T\right)\left(x_0+\widetilde{h}_Tu\right)-\left(I\delta\widehat{A}_T\right)(x_0)\right]du + \sqrt{Th_T}\psi_T(x_0) + o_p(1),$$

λ-a.s. in $x_0 \in [0,1]$. This is an integral equation for $\sqrt{Th_T}I\delta\widehat{A}_T$ with a unique solution in $L^2(\lambda)$ [up to order $o_p(1)$ pointwise]. By substitution and using (a.11), the solution is of the form $\sqrt{Th_T}I\delta\widehat{A}_T = \sqrt{Th_T}\psi_T$

$+o_p(1)$, pointwise, which gives Proposition 4 ii).

v) Pointwise expansion of the minimum chi-square estimator: proof of Corollary 5

Part i) follows from Lemma A.4 ii) and boundedness of I^{-1} (Assumption A.4). Let us now focus on part ii).

Since I^{-1} is bounded, the same argument as in section A.2.4 iv) above can be used to bound pointwise the

residual in the asymptotic expansion of the minimum chi-square estimator $\delta \widehat{A}_T \simeq I^{-1} \psi_T$. Thus we have:

$$\sqrt{Th_T} \delta \widehat{A}_T(w) = \sqrt{Th_T} I^{-1} \psi_T(w) + o_p(1), \quad \lambda\text{-a.s. in } w \in [0,1].$$

Let us now carefully separate the bias term. We get:

$$
\begin{aligned}
\sqrt{Th_T} \delta \widehat{A}_T(w) &= I^{-1} \left[\sqrt{Th_T} \left(\psi_T(w) - E_0 \psi_T(w) \right) \right] + \sqrt{Th_T} I^{-1} E_0 \psi_T(w) + o_p(1) \\
&= \sqrt{Th_T} \alpha_0(w)^{-1} \left(\psi_T(w) - E_0 \psi_T(w) \right) + \sqrt{Th_T} I^{-1} E_0 \psi_T(w) + o_p(1),
\end{aligned}
$$

since the contribution of the integral component of I to $I^{-1} \left(\psi_T - E_0 \psi_T \right)$ is of order $O_p \left(1/\sqrt{T} \right)$ [see Lemma

10]. Therefore Corollary 5 is proved.

A.2.5 Asymptotic distribution of ψ_T: proof of Lemma 7

Under the regularity conditions in Appendix 2.1 we have:

$$
\begin{aligned}
\sqrt{Th_T} \left(\psi_T - E \psi_T \right)(w) &= \sqrt{Th_T} \int \left(\widehat{f}_T(w,y) - E\widehat{f}_T(w,y) \right) \gamma_0(w,y)\, dy \\
&\quad + \sqrt{Th_T} \int \left(\widehat{f}_T(x,w) - E\widehat{f}_T(x,w) \right) \gamma_1(x,w) dx + o_p(1).
\end{aligned}
$$

Theorem 3 in Ait-Sahalia (1993) applies and it follows:

$$\sqrt{Th_T} \left(\psi_T - E \psi_T \right)(w) \xrightarrow{d} N(0, \sigma^2(w)),$$

where the asymptotic variance is given by:

$$
\begin{aligned}
\sigma^2(w) &= \left(\int K(z)^2 dz \right) \left(E \left[\gamma_0(X_t, Y_t) \gamma_0(X_t, Y_t)' \mid X_t = w \right] f_X(w) \right. \\
&\quad \left. + E \left[\gamma_1(X_t, Y_t) \gamma_1(X_t, Y_t)' \mid Y_t = w \right] f_Y(w) \right) \\
&= \left(\int K(z)^2 dz \right) \alpha_0(w), \quad \text{from equation (6).}
\end{aligned}
$$

Appendix 3

Nonparametric efficiency bound

In this Appendix we derive the nonparametric efficiency bound in the i.i.d. framework [proof of Proposition 14 i)]; the proof in the time series framework is similar and is available on the website. Let us introduce a one dimensional parametric model $A(.,\theta)$ and derive its Cramer-Rao bound. The score is given by:

$$\frac{\partial \log f}{\partial \theta}(x,y;A(\theta_0)) = \left\langle D \log f(x,y;A_0), \frac{dA}{d\theta}(\theta_0) \right\rangle.$$

The Fisher information is:

$$E_0\left[\left(\frac{\partial \log f}{\partial \theta}(X_t,Y_t;A(\theta_0))\right)^2\right] = E_0\left[\left\langle D \log f(X,Y;A_0), \frac{dA}{d\theta}(\theta_0) \right\rangle^2\right] = \left(\frac{dA}{d\theta}(\theta_0), I\frac{dA}{d\theta}(\theta_0)\right)_{L^2(\lambda)}.$$

Thus the Cramer-Rao bound is given by:

$$B_A(g,\theta) = \left(\frac{dA}{d\theta}(\theta_0), I\frac{dA}{d\theta}(\theta_0)\right)_{L^2(\lambda)}^{-1}.$$

The parametric specification can be chosen such that $\int g(v)' A(v,\theta)\,dv = \theta$, which is equivalent (in a neighborhood of θ_0) to the constraint:

$$\int g(v)'\frac{dA}{d\theta}(v,\theta_0)dv = \left(g, \frac{dA}{d\theta}(\theta_0)\right)_{L^2(\lambda)} = 1. \tag{a.13}$$

Thus both the Cramer-Rao bound and the constraint (a.13) depend on the parameterization only by means of the function $\delta(.) = dA/d\theta(.,\theta_0)$. Therefore problem (21) in the text is equivalent to:

$$\min_{\delta \in L^2(\lambda)} (\delta, I\delta)_{L^2(\lambda)},$$
$$\text{s.t} \quad : \quad (g,\delta)_{L^2(\lambda)} = 1.$$

By Cauchy-Schwarz inequality we have:

$$1 = (g,\delta)_{L^2(\lambda)}^2 = \left(I^{-1/2}g, I^{1/2}\delta\right)_{L^2(\lambda)}^2 \leq \left(I^{-1}g,g\right)_{L^2(\lambda)} (\delta, I\delta)_{L^2(\lambda)}.$$

Therefore $(\delta, I\delta)_{L^2(\lambda)} \geq \left(I^{-1}g,g\right)_{L^2(\lambda)}^{-1}$ and the bound is reached for $\delta^* = I^{-1}g \in L^2(\lambda)$. Thus we deduce:

$$B_A(g) = \left(g, I^{-1}g\right)_{L^2(\lambda)}.$$

Acknowledgements

References

Abdous, B., Ghoudi, K., and A. Khoudraji, 2000, Nonparametric estimation of the limit dependence function of multivariate extremes, Extremes 2, 245-268.

Aït-Sahalia, Y., 1993, The delta and bootstrap methods for nonparametric kernel functionals, MIT Discussion Paper.

Bec, F., Ben Salem, M., and M. Carrasco, 2004, Tests for unit-root versus threshold specifications with an application to the PPP, Journal of Business and Economic Statistics 22, 382-395.

Begun, J. M., Hall, W. J., Huang, W.-M., and J. A. Wellner, 1983, Information and asymptotic efficiency in parametric-nonparametric models, Annals of Statistics 11, 432-452.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and J. A. Wellner, 1993, Efficient and adaptive estimation in semiparametric models (Johns Hopkings University Press, Baltimore).

Bosq, D., 1998, Nonparametric statistics for stochastic processes (Springer, New-York).

Chen, X., and Y. Fan, 2003, Semiparametric estimation of copula-based time series models, Working Paper.

Deheuvels, P., 1981, A nonparametric test for independence, Publications de l'Institut de Statistique de l'Université de Paris 26, 29-50.

Donoho, D., and R. Liu, 1991, Geometrizing rate of convergence II, Annals of Statistics 19, 633-667.

Duffie, D., and K. Singleton, 1999, Simulating correlated defaults, Stanford University, Working Paper.

Dunford, N., and J. Schwartz, 1968, Linear operators: Part I (Wiley, New York).

Embrechts, P., Höing, A., and A. Juri, 2003, Using copulae to bound the value-at-risk for functions of dependent risks, Finance and Stochastics 7, 145-167.

Engle, R., and J. Russell, 1998, The autoregressive conditional duration model, Econometrica 66, 1127-1162.

Fan, J., 1993, Local linear regression smoothers and their minimax efficiency, Annals of Statistics 21, 196-216.

Fermanian, J. D., and O. Scaillet, 2003, Nonparametric estimation of copulas for time series, Journal of Risk 5, 25-54.

Gagliardini, P., and C. Gouriéroux, 2002a, Constrained nonparametric dependence with applications to finance and insurance, CREST Working Paper.

Gagliardini, P., and C. Gouriéroux, 2002b, Duration time series models with proportional hazard, CREST Working Paper.

Gagliardini, P., and C. Gouriéroux, 2005, Stochastic migration models with application to corporate risk, forthcoming in Journal of Financial Econometrics.

Genest, C., and R. J. Mc Kay, 1986, Copules archimediennes et familles de lois bidimensionelles dont les marges sont données, Canadian Journal of Statistics 14, 145-159.

Genest, C., and L.-P. Rivest, 1993, Statistical inference procedures for bivariate Archimedean copulas, Journal of the American Statistical Association 88, 1034-1043.

Ghysels, E., Gouriéroux, C., and J. Jasiak, 2004, Stochastic volatility duration models, Journal of Econometrics 119, 413-433.

Gill, R. D., and A. W. van der Vaart, 1993, Non- and semi-parametric maximum likelihood estimators and the von Mises method-II, Scandinavian Journal of Statistics 20, 271-288.

Gorgens, T., and J. Horowitz, 1999, Semi-parametric estimation of a censored regression model with an unknown transformation of the dependent variable, Journal of Econometrics 90, 155-191.

Gouriéroux, C., and J. Jasiak, 2001, State-space models with finite dimensional dependence, Journal of Time Series Analysis 22, 665-678.

Gouriéroux, C., and J. Jasiak, 2002, Nonlinear autocorrelograms: an application to intertrade durations, Journal of Time Series Analysis 33, 127-154.

Gouriéroux, C., and A. Monfort, 2002a, Equidependence in qualitative and duration models with application to credit risk, CREST Working Paper.

Gouriéroux, C., and A. Monfort, 2002b, Term structure in duration models, CREST Working Paper.

Gouriéroux, C., and C. Y. Robert, 2005, Stochastic unit root models, CREST Working Paper.

Gouriéroux, C., and C. Tenreiro, 2001, Local power properties of kernel based goodness of fit tests, Journal of Multivariate Analysis 78, 161-190.

Han, A., 1987a, Nonparametric analysis of a generalized regression model, Journal of Econometrics 35, 305-316.

Han, A. 1987b, A nonparametric analysis of transformations, Journal of Econometrics 35, 191-209.

Härdle, W., and A. Tsybakov, 1997, Local polynomial estimation of the volatility function, Journal of Econometrics 81, 223-242.

Hastie, T. J., and R. J. Tibshirani, 1990, Generalized additive models (Chapman and Hall, London).

Holly, A., 1995, A random linear functional approach to efficiency bounds, Journal of Econometrics 65, 235-261.

Horowitz, J., 1996, Semiparametric estimation of a regression model with an unknown transformation of the dependent variable, Econometrica 64, 103-137.

Ichimura, M., 1993, Semi-parametric least-squares and weighted SLS estimation of single index models, Journal of Econometrics 58, 71-120.

Joe, H., 1997, Multivariate models and dependence concepts (Chapman and Hall, London).

Koshevnik, Y., and B. Levit, 1976, On a nonparametric analogue of the information matrix, Theory of Probability and its Applications 21, 738-753.

Lancaster, H., 1968, The structure of bivariate distributions, Annals of Mathematical Statistics 29, 719-736.

Li, D. X., 2000, On default correlation: a copula function approach, Journal of Fixed Income 9, 43-54.

Nelsen, R., 1999, An introduction to copulas (Springer, New York).

Oakes, D., 1989, Bivariate survival models induced by frailties, Journal of the American Statistical Association 84, 487-493.

Parzen, E., 1962, On estimation of a probability density function and the mode, Annals of Mathematical Statistics 33, 1065-1076.

Powell, J., Stock, J., and T. Stoker, 1989, Semi-parametric estimation of index coefficients, Econometrica 57, 474-523.

Rahbek, A., and N. Shephard, 2002, Inference and ergodicity in the autoregressive conditional root model, University of Copenhagen and Nuffield College, Working Paper.

Ridder, G., 1990, The nonparametric identification of generalized accelerated failure time models, Review of Economic Studies 57, 167-182.

Rosenblatt, M., 1956, Remarks on some nonparametric estimates of a density function, Annals of Mathematical Statistics 27, 832-835.

Rudin, W., 1973, Functional analysis (McGraw-Hill, New York).

Schönbucher, P. J., and D. Schubert, 2001, Copula-dependent default risk in intensity models, University of Bonn, Working Paper.

Scott, D., 1992, Multivariate density estimation. Theory, practice and visualization (Wiley, New York).

Severini, T. A., and G. Tripathi, 2001, A simplified approach to computing efficiency bounds in semiparametric models, Journal of Econometrics 102, 23-66.

Silverman, B., 1978, Weak and strong consistency of the kernel estimate of a density and its derivatives, Annals of Statistics 6, 177-184.

Silverman, B., 1986, Density estimation for statistics and data analysis (Chapman and Hall, London).

Sklar, A., 1959, Fonctions de répartition à n dimensions et leurs marges, Publications de l'Institut de Statistique de l'Université de Paris 8, 229-231.

Stein, C., 1956, Efficient nonparametric testing and estimation, in: J. Neyman, ed., Proceedings of the third Berkeley symposium on mathematical statistics and probability, Vol. 1 (University of California Press, Berkeley) 187-196.

Stone, C., 1983, Optimal uniform rates of convergence of nonparametric estimators of a density function or its derivatives, in: M. H. Rizvi, J. S. Rustagi and D. Siegmund, eds., Recent advances in statistics, essays in honor of Herman Chernoff's sixtieth birthday (Academic Press, New-York).

Vaart, A. W. van der, and J. A. Wellner, 1996, Weak convergence and empirical processes (Springer, New York).

Van den Berg, G., 2001, Duration models: specification, identification, and multiple durations, in: J. J. Heckman and E. Leamer, eds., Handbook of econometrics, Vol. 5 (North-Holland, Amsterdam).

Yosida, K., 1995, Functional analysis (Springer, New York).