NHH

INSTITUTT FOR FORETAKSØKONOMI

DEPARTMENT OF BUSINESS AND MANAGEMENT SCIENCE

FOR 22 2016

ISSN: 1500-4066

December 2016

Discussion paper

Non-parametric estimation of conditional densities: A new method

BY

Håkon Otneim AND **Dag Tjøstheim**

Norges
Handelshøyskole

NORWEGIAN SCHOOL OF ECONOMICS

Non-parametric estimation of conditional densities: A new method

Håkon Otneim [*] Dag Tjøstheim [†]

Abstract

Let $\boldsymbol{X} = (X_1, \ldots, X_p)$ be a stochastic vector having joint density function $f_{\boldsymbol{X}}(\boldsymbol{x})$ with partitions $\boldsymbol{X}_1 = (X_1, \ldots, X_k)$ and $\boldsymbol{X}_2 = (X_{k+1}, \ldots, X_p)$. A new method for estimating the conditional density function of $\boldsymbol{X}_1$ given $\boldsymbol{X}_2$ is presented. It is based on locally Gaussian approximations, but simplified in order to tackle the curse of dimensionality in multivariate applications, where both response and explanatory variables can be vectors. We compare our method to some available competitors, and the error of approximation is shown to be small in a series of examples using real and simulated data, and the estimator is shown to be particularly robust against noise caused by independent variables. We also present examples of practical applications of our conditional density estimator in the analysis of time series. Typical values for k in our examples are 1 and 2, and we include simulation experiments with values of p up to 6. Large sample theory is established under a strong mixing condition.

Keywords: Conditional density estimation, local likelihood, multivariate data, cross-validation.

1 Introduction

The need for expressing statistical inference in terms of conditional quantities is ubiquitous in most natural and social sciences. The obvious example is the estimation of the mean of some set of response variables conditioned on sets of explanatory variables taking specified values. Other common tasks are the forecasting of volatilities or quantiles of financial time series conditioned on past history. Problems of this kind often call for some sort of regression analysis, of which the literature provides an abundance of choices.

Conditional means, variances and quantiles are all properties of the conditional density, if it exists, as are all other probabilistic statements that we might ever want to make about the response variables given the explanatory variables. It is therefore clearly of interest to obtain good estimates of the entire conditional distribution in order to make use of all the evidence contained in the data, and to provide the user with a wide variety of options in analysing and visualising the relationships of the variables under study.

[*] Norwegian School of Economics, Department of Business and Management Science, Helleveien 30, 5045 Bergen, Norway.
E-mail: `hakon.otneim@nhh.no`

[†] University of Bergen, Department of Mathematics, P.B. 7803, 5020 Bergen, Norway

The classical method for non-parametric density estimation is the kernel estimator (Rosenblatt et al., 1956; Parzen, 1962), which in the decades following its introduction has been refined and developed in many directions. Especially the crucial choice of smoothing parameter, or bandwidth, has been addressed by several authors, including Silverman (1986), Sheather and Jones (1991) and Chacón and Duong (2010). The kernel estimator suffers greatly from the curse of dimensionality however, which quickly inhibits its use in multivariate problems. Several alternative methods of estimation has been proposed to improve performance if the subject of estimation is a joint multivariate density function, most recently the LGDE (locally Gaussian density estimator) by Otneim and Tjøstheim (2016), which the work in the present paper takes as its starting point. Very few methods exist for the non-parametric estimation of conditional densities though, especially if we do not wish to restrict ourselves to cases with one-dimensional response and/or explanatory variables. This lack of methodology is surprising, considering the aforementioned importance of estimating conditional densities; the practical use of which is of altogether greater interest than unconditional density estimates, as is illustrated by some of its possible applications in Section 5.

In this paper we present a new method for estimating conditional densities based on local Gaussian approximations. Let $\boldsymbol{X} = (X_1, \ldots, X_p)$ be a stochastic vector, and, assuming existence, denote by $f_{\boldsymbol{X}}(\cdot)$ its joint density function. Further, let $(\boldsymbol{X}_1; \boldsymbol{X}_2) = (X_1, \ldots, X_k; X_{k+1}, \ldots, X_p)$ be a partitioning of $\boldsymbol{X}$. Then the conditional density of $\boldsymbol{X}_1$ given $\boldsymbol{X}_2 = \boldsymbol{x}_2$ is defined by

$$f_{\boldsymbol{X}_1|\boldsymbol{X}_2}(\boldsymbol{x}_1|\boldsymbol{X}_2 = \boldsymbol{x}_2) = \frac{f_{\boldsymbol{X}}(\boldsymbol{x}_1, \boldsymbol{x}_2)}{f_{\boldsymbol{X}_2}(\boldsymbol{x}_2)}, \tag{1}$$

where $f_{\boldsymbol{X}_2}$ is the marginal density of $\boldsymbol{X}_2$.

The problem of estimating (1) is not trivial. We do not observe data directly from the density that we wish to estimate, so we need a different set of tools than those used in the unconditional case. A natural course of action is to follow Rosenblatt (1969) in obtaining good estimates of the numerator and denominator of (1) separately using the kernel estimator, and use the definition directly. Chen and Linton (2001) provide a discussion of choosing the bandwidths when using the kernel estimator to estimate the components, as do Bashtannyk and Hyndman (2001). Hall et al. (2004, chap. 5) give a unified approach to estimating conditional densities using the kernel estimator, which allows a mix of continuous and discrete variables, and automatically smooths out the irrelevant ones.

Unless one has a very good estimate of the marginal density, however, it is less than ideal to put a kernel estimate in the denominator of (1). This is remedied by Faugeras (2009), who writes the conditional density as a product of the marginal and copula density functions in the bivariate case,

$$f_{X_1|X_2}(x_1|X_2 = x_2) = f_{X_1}(x_1)c\left\{F_1(x_1), F_2(x_2)\right\}, \tag{2}$$

where f_{X_1} is the marginal density of X_1, F_1 and F_2 are the marginal distribution functions, c is the copula density of (X_1, X_2), and estimates those separately using the kernel estimator. The formula (2) can be generalized to the case of several covariates, but its practical use in higher dimensions is questionable because of boundary and dimensionality issues, unless one obtains better estimates of the multivariate copula density than provided by the kernel estimator, such as the local likelihood approach by Geenens et al. (2014).

Hyndman et al. (1996) starts to move away from the kernel estimator by adjusting the conditional mean to match a better performing regression technique, such as local polynomials, while Fan et al. (1996) estimate the conditional density directly using locally linear and locally quadratic fits, a method that Hyndman and Yao (2002) refine by constraining it to always be non-negative. The latter authors propose in the same paper a local likelihood approach which is based on some of the same machinery as we will employ in this paper, and Fan and Yim (2004) provide a cross-validation rule for bandwidth selection in the locally parametric models. These methods are to date implemented in the bivariate case only, however, where the response- and explanatory variables are both scalars.

Indeed, the main motivation behind our new method is to provide an estimator that can handle a greater number of variables without the requirement that either response or explanatory variables are scalar.

Holmes et al. (2012) develop a fast bandwidth selection algorithm, while correctly pointing out that bandwidth selection is a formidable computational and time-consuming task in non-parametric multivariate density estimation. We argue that the curse of dimensionality is an even bigger problem, because it will not be solved by clever algorithms, but is an inherent problem in all non-parametric analysis. We therefore base our method on the newly developed locally Gaussian density estimator (LGDE) (Otneim and Tjøstheim, 2016), which shows a promising robustness against dimensionality issues when estimating the multivariate unconditional density function. By exploiting locally the property of the Gaussian distribution that conditional densities are again Gaussian, we will see that conditional density estimates are readily available from the LGDE.

This paper is organized as follows: In Section 2 we give a short introduction to the LGDE method for multivariate *un*conditional density estimation, and in Section 3 we show that extracting conditional density estimates from the LGDE is straightforward and requires neither additional estimation steps, nor integration over the joint density estimate. In Section 4 we derive some large-sample properties for our estimator under a strong mixing condition, and proceed in Section 5 with a series of examples using real and simulated data, indicating the wide potential of conditional density estimation. Some concluding remarks and suggestions for further research follow in Section 6, and we include an appendix that contains the technical proofs.

2 A brief introduction to the LGDE

Because of its close relationship with our conditional density estimator, we include here a basic account of the LGDE. Suppose that we wish to estimate the full p-variate density $f_{\boldsymbol{X}}$ based on n independent observations $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$. Hjort and Jones (1996) provide a general setup for fitting a parametric family of densities $\psi(\cdot, \boldsymbol{\theta})$ *locally* to the unknown density by maximising the local log-likelihood function in each point $\boldsymbol{x}$;

$$\widehat{\boldsymbol{\theta}}(\boldsymbol{x}) = \arg\max_{\boldsymbol{\theta}} n^{-1} \sum_{i=1}^{n} K_{\boldsymbol{h}}(\boldsymbol{X}_i - \boldsymbol{x}) \log \psi(\boldsymbol{X}_i, \boldsymbol{\theta}) - \int K_{\boldsymbol{h}}(\boldsymbol{y} - \boldsymbol{x}) \psi(\boldsymbol{y}, \boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{y}, \tag{3}$$

so that the estimated density is given by $\widehat{f}_{\boldsymbol{X}}(\boldsymbol{x}) = \psi(\boldsymbol{x}, \widehat{\boldsymbol{\theta}}(\boldsymbol{x}))$. We use standard notation, letting $\boldsymbol{h}$ denote a diagonal matrix of bandwidths, $K(\cdot)$ a symmetric kernel function integrating to one, and $K_{\boldsymbol{h}}(\boldsymbol{x}) = |\boldsymbol{h}|^{-1} K(\boldsymbol{h}^{-1}\boldsymbol{x})$. Denote by ϕ and Φ the univariate standard normal density and distribution functions respectively,

$$\phi(z) = (2\pi)^{-1/2} \exp\left\{-z^2/2\right\}, \ \Phi(z) = \int_{-\infty}^{z} \phi(y) \, \mathrm{d}y.$$

According to Otneim and Tjøstheim (2016), we can write the p-variate density function $f_{\boldsymbol{X}}$ as

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = f_{\boldsymbol{Z}}\left(\Phi^{-1}\left(F_1(x_1)\right), \ldots, \Phi^{-1}\left(F_p(x_p)\right)\right) \prod_{i=1}^{p} \frac{f_i(x_i)}{\phi\left(\Phi^{-1}\left(F_i(x_i)\right)\right)} \tag{4}$$

where f_i and F_i, $i = 1, \ldots, p$, are the marginal densities and distribution functions of $f_{\boldsymbol{X}}$, and $f_{\boldsymbol{Z}}$ is the density function of a stochastic vector $\boldsymbol{Z} = (Z_1, \ldots, Z_p)$ with standard normal margins, and $Z_i = \Phi^{-1}\left(F_i(X_i)\right)$.

We estimate $f_{\boldsymbol{Z}}$ by locally fitting the standardized normal distribution,

$$\psi(\boldsymbol{z}, \boldsymbol{\theta}) = \psi(\boldsymbol{z}, \boldsymbol{R}) = (2\pi)^{-p/2} |\boldsymbol{R}|^{-1/2} \exp\left\{-\frac{1}{2}\boldsymbol{z}^T \boldsymbol{R}^{-1} \boldsymbol{z}\right\}, \tag{5}$$

with $\boldsymbol{R} = \boldsymbol{R}(\boldsymbol{z}) = \{\rho_{ij}(\boldsymbol{z})\}$ denoting the correlation matrix, based on the marginally Gaussian pseudo-observations

$$\widehat{\boldsymbol{Z}}_j = \left(\Phi^{-1}(\widehat{F}_1(X_{j1})), \ldots, \Phi^{-1}(\widehat{F}_p(X_{jp}))\right)^T, \qquad j = 1, \ldots, n, \tag{6}$$

where $\widehat{F}_k(x_k), k = 1, \ldots, p$ are estimates of the marginal distribution functions, which, in our asymptotic results are assumed to be the empirical marginal distribution functions. There are several reasons for transforming all observation vectors to the standard Gaussian scale. First of all, it makes the choice of the Gaussian distribution as local parametric family in (5) very natural, where, in particular, we have fixed all means and standard deviations so they are equal to 0 and 1 respectively, reducing the number of parameters that we must estimate locally. Moreover, the normalisation (6) is a quick way to make the data more tidy, because the pseudo-observations will all be on the same scale, and there will not be any outliers which is otherwise known to create problems when using cross-validation to select bandwidths (Hall, 1987). In general, distributions become easier to estimate when they are closer to the Gaussian distribution, as shown and exploited by Wand et al. (1991) and Ruppert and Cline (1994).

In (5), each correlation $\rho_{ij}(\boldsymbol{z})$ depends on the coordinates of the entire $\boldsymbol{z}$-vector, making its estimation difficult because of the curse of dimensionality. In regression problems, this issue may be tackled by imposing an additive structure on the unknown regression function:

$$Y = f(X_1, \ldots, X_p) + \epsilon = f_1(X_1) + \cdots + f_p(X_p) + \epsilon,$$

possibly with higher order interactions if the data can support it. One motivation behind the LGDE is to introduce a similar idea to density estimation, and it is based on the fact that a *global* Gaussian fit is produced by calculating the correlation coefficients between each pair of variables by using only the corresponding observation vectors. In order to circumvent the curse of dimensionality, Otneim and Tjøstheim (2016) carry this procedure over to the local case by restricting $\rho_{ij}(\boldsymbol{z})$ so that it is only allowed to depend on its own variables; i.e. $\rho_{ij}(\boldsymbol{z}) = \rho_{ij}(z_i, z_j)$. The corresponding estimate $\widehat{\rho}(z_i, z_j)$ is computed from the corresponding simplified pairwise local log likelihood so that we can take

$$\widehat{\rho}_{ij}(z_1, \ldots, z_p) = \widehat{\rho}_{ij}(z_i, z_j). \tag{7}$$

This technique effectively reduces the estimation of $f_{\boldsymbol{X}}$ to a series of bivariate local problems, which is reflected in the rate of convergence in the following asymptotic result, that holds under some standard regularity conditions (Otneim and Tjøstheim, 2016) and proven for sets of iid observations:

$$\sqrt{nh_n^2}\left(\widehat{f}_{\boldsymbol{X}}(\boldsymbol{x}) - f_0(\boldsymbol{x})\right) \xrightarrow{\mathcal{L}} N\left(0, \sigma^2_{f_{\boldsymbol{X}}}\right), \tag{8}$$

where, in general, $f_0(\boldsymbol{x}) \neq f_{\boldsymbol{X}}(\boldsymbol{x})$ is the population density towards which the LGDE converges. Here, $f_0(\boldsymbol{x})$ is the simplified density obtained from (4) and (5) by replacing $f_{\boldsymbol{Z}}(\boldsymbol{z})$ with $\Psi(\boldsymbol{z}, \boldsymbol{R}_0)$, where $\boldsymbol{R}_0 = \{\rho_{0,ij}(z_i, z_j)\}$ and $\rho_{0,ij}$ is the true local Gaussian correlation between Z_i and Z_j, as will be defined in Section 4.

Otneim and Tjøstheim (2016) propose two methods for bandwidth selection. Cross-validation is used to determine the bandwidths that minimise the estimated Kullback-Leibler distance between the density estimate and the true density. They also employ the k-nearest neighbor technique in order to obtain adaptive bandwidths, but simulation results suggest that, of the two, the global bandwidth selector performs better. Indeed, as already mentioned, Hall (1987) shows that the performance of cross-validation bandwidth selection depends on the tails of the underlying distribution not being thicker than the tails of the kernel function. By transforming the data to marginal standard normality, and using the Gaussian kernel function, it follows that the cross-validation procedure is well suited for selecting the LGDE bandwidths.

3 Estimating the conditional density

Conditional density estimates are in principle available from any non-parametric estimate of the unconditional density of all variables. Let us return to the problem in Section 1, and suppose that we obtain an estimate $\tilde{f}_{\boldsymbol{X}}$ of $f_{\boldsymbol{X}}$ in the process of estimating the left hand side of (1). The corresponding marginal density $\tilde{f}_{\boldsymbol{X}_2}$ that ideally we should put in the denominator of (1) is given by

$$\tilde{f}_{\boldsymbol{X}_2} = \int \tilde{f}_{\boldsymbol{X}} \, \mathrm{d}\boldsymbol{x}_1,$$

but one must usually turn to numerical methods in order to obtain this integral, which can be a costly affair in terms of computing power, especially when there are many variables over which to integrate. Thus, estimating the marginal density directly from the data is often quicker, but introduces a new source of uncertainty that, again, will be difficult to handle in case of several explanatory variables.

We proceed to show that this problem is completely circumvented if we use the LGDE strategy for estimation. As is well known for a multivariate Gaussian distribution, every conditional density that can be formed by partitioning the Gaussian vector and computing the fraction (1), is again Gaussian, and where the (conditional) mean and (conditional) covariance matrix in that Gaussian can be easily computed; see e.g. Johnson and Wichern (2007, Chap. 4). This is of course also the case for the fraction of Gaussians that are local approximations, and we can obtain estimates by using these formulas. In more detail, starting from the p-variate density in (4),

$$\begin{aligned}f_{\boldsymbol{X}_1|\boldsymbol{X}_2}(\boldsymbol{x}_1|\boldsymbol{X}_2=\boldsymbol{x}_2) &= \frac{f_{\boldsymbol{X}}(\boldsymbol{x})}{f_{\boldsymbol{X}_2}(\boldsymbol{x}_2)}\\ &= \frac{f_{\boldsymbol{Z}}(z_1,\dots,z_p)}{f_{\boldsymbol{Z}_2}(z_{k+1},\dots,z_p)}\prod_{i=1}^{k}\frac{f_i(x_i)}{\phi\,(z_i)},\end{aligned}$$

where $f_{\boldsymbol{Z}}/f_{\boldsymbol{Z}_2}$ can be seen locally as a fraction of a p-variate and a $p-k$-variate Gaussian function, each with all expectations equal to zero, and with correlation matrices $\boldsymbol{R}(\boldsymbol{z})$ and $\boldsymbol{R}_{22}(\boldsymbol{z})$ respectively. The latter notation is natural because of the pairwise analysis, so that $\boldsymbol{R}_{22}(\boldsymbol{z})$ is *exactly equal* to the lower right block of $\boldsymbol{R}(\boldsymbol{z})$. Thus, in every grid point $\boldsymbol{z}$, $f_{\boldsymbol{Z}_2}$ is exactly the marginal density of the $p-k$ last variables of $f_{\boldsymbol{Z}}$, and we can use the basic result for the multivariate normal distribution mentioned above to rewrite the fraction. Partition $\boldsymbol{R}(\boldsymbol{z})$ into four blocks, of which the lower right block is $\boldsymbol{R}_{22}(\boldsymbol{z})$:

$$\boldsymbol{R}(\boldsymbol{z}) = \begin{pmatrix}\boldsymbol{R}_{11} & \boldsymbol{R}_{12}\\ \boldsymbol{R}_{21} & \boldsymbol{R}_{22}\end{pmatrix}$$

Then

$$f_{\boldsymbol{Z}}/f_{\boldsymbol{Z}_2} = \Psi^*(z_1,\dots,z_k;\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^*), \tag{9}$$

where $\Psi^*(\cdot)$ is the general k-variate Gaussian density with expectation vector and covariance matrix given by

$$\boldsymbol{\mu}^* = \boldsymbol{R}_{12}\boldsymbol{R}_{22}^{-1}\boldsymbol{z}_2, \tag{10}$$
$$\boldsymbol{\Sigma}^* = \boldsymbol{R}_{11} - \boldsymbol{R}_{12}\boldsymbol{R}_{22}^{-1}\boldsymbol{R}_{21}, \tag{11}$$

where $\boldsymbol{z_2} = (z_{k+1},\dots,z_p)$. Note that we may use correlation- and covariance matrices interchangeably, because all standard deviations are equal to one in $f_{\boldsymbol{Z}}$ and $f_{\boldsymbol{Z}_2}$.

We can now obtain an estimate of $f_{\boldsymbol{X}_1|\boldsymbol{X}_2=\boldsymbol{x}_2}$ by plugging in local likelihood estimates of $\boldsymbol{R}(\boldsymbol{z}) = \{\rho_{ij}(z_i,z_j)\}$, resulting in

$$\widehat{f}_{\boldsymbol{X}_1|\boldsymbol{X}_2}(\boldsymbol{x}_1|\boldsymbol{X}_2=\boldsymbol{x}_2) = \Psi^*\left(\widehat{\boldsymbol{z}};\widehat{\boldsymbol{\mu}^*}(\widehat{\boldsymbol{z}}),\widehat{\boldsymbol{\Sigma}^*}(\widehat{\boldsymbol{z}})\right)\prod_{i=1}^{k}\frac{\widehat{f}_i(x_i)}{\phi\,(\widehat{z}_i)}, \tag{12}$$

where $\widehat{\boldsymbol{\mu}^*}(\widehat{\boldsymbol{z}})$ and $\widehat{\boldsymbol{\Sigma}^*}(\widehat{\boldsymbol{z}})$ are obtained by substituting local correlation estimates into equations (10) and (11), and where we write $\widehat{z}_i = \Phi^{-1}(\widehat{F}_i(x_i))$. Moreover, the second factor in (12) requires estimates $\widehat{f}_i(x_i)$ of the marginal densities $f_i(x_i)$, $i=1,\dots,k$. As we will see in the next section, this can be any smooth estimate, and will not affect the asymptotic results as long as they converge faster than $\sqrt{nh^2}$. The current implementation of the LGDE uses the logspline estimator by Stone et al. (1997) for this purpose. It is interesting to note that the computation resulting in (9), (10) and (11) can be done directly on estimated quantities using results on fractions of exponential functions.

We modify the LGDE algorithm in Otneim and Tjøstheim (2016) according to the discussion above, and estimate conditional densities by following these steps:

1. Transform each marginal observation vector to pseudo-standard normality using (6).

2. Estimate the local correlation matrix of the transformed data by fitting the Gaussian family (5) using the local likelihood function in (3) and the simplification (7). In practice, this amounts to fitting the bivariate version of (5) to each pair of approximately marginally standard normal variables $(\widehat{Z}_i, \widehat{Z}_j)$, and let $\widehat{\boldsymbol{R}}(\boldsymbol{z}) = \{\widehat{\rho}(z_i, z_j)\}_{i,j=1,\ldots,p}$.

3. Calculate the local mean and covariance matrix of $\widehat{f}_{\boldsymbol{Z}}/\widehat{f}_{\boldsymbol{Z}_2}$ using the formulas (10) and (11), so that the conditional density estimate becomes as given in (12)

4. Normalize the density estimate so that it integrates to one.

Again, we point out that our simplification of the dependence structure (7) in general will result in an estimate of an approximation $f_0(\cdot)$ of the true density $f(\cdot)$. We proceed in the next section to discuss the nature of the simplification, to discuss regularity conditions, and to explore the large sample properties of our method.

4 Regularity conditions and asymptotic theory

The following theorems on consistency relative to f_0 and asymptotic normality state analogous results to those found in Otneim and Tjøstheim (2016), but they are proven under a new set of regularity conditions that allow for dependence between the observations $X_1, \ldots, X_n$.

The simplification (7) means that we estimate the local correlations pairwise, which also means that it suffices to derive most of the asymptotic theory in the bivariate case. Consider, for the time being, a pair (Z_i, Z_j) of marginally standard normal random variables. Denote by $\rho_0(z_i, z_j) = \rho_0(\boldsymbol{z})$ the local Gaussian correlation between them, as will be defined below, and by $\widehat{\rho}(\boldsymbol{z})$ its estimate, calculated using the bandwidths $\boldsymbol{h} = (h_i, h_j)$ according to the algorithm in Section 3. Denote further by $L_n(\rho(\boldsymbol{z}), \boldsymbol{z})$ the local log-likelihood function in (3) with the bivariate version of (5) as parametric family $\psi(\cdot, \rho)$. For a fixed $\boldsymbol{h} > 0$ (where all statements about the vector $\boldsymbol{h}$ in this section are element-wise), denote by $\rho_{\boldsymbol{h}}$ the local correlation that satisfies

$$\frac{\partial L_n(\boldsymbol{\rho}; \boldsymbol{z})}{\partial \rho} \rightarrow \int K_h(\boldsymbol{y} - \boldsymbol{z}) u(\boldsymbol{y}, \rho_{\boldsymbol{h}}) \left\{ f_{ij}(\boldsymbol{y}) - \psi(\boldsymbol{y}, \rho_{\boldsymbol{h}}) \right\} \, \mathrm{d}\boldsymbol{y} = 0 \tag{13}$$

as $n \rightarrow \infty$, where $u(\cdot, \rho) = \partial \log \psi(\cdot, \rho)/\partial \rho$, and f_{ij} is the joint density of (Z_i, Z_j). We assume hereafter that $\rho_{\boldsymbol{h}}$ exists and is unique for any $\boldsymbol{h} > 0$ (see also Hjort and Jones (1996) and discussion in Otneim and Tjøstheim (2016)). By letting $\boldsymbol{h} = \boldsymbol{h}_n \rightarrow 0$, at an appropriate rate (see Assumption C), the local correlation in the expression above, as mentioned in the previous section, satisfies

$$\psi(\boldsymbol{z}, \rho_0(\boldsymbol{z})) = f_{ij}(\boldsymbol{z}), \tag{14}$$

and we require the population value $\rho_0(\boldsymbol{z})$ to satisfy (14), cf. Hjort and Jones (1996) and Tjøstheim and Hufthammer (2013). Assuming (14) is not enough to ensure uniqueness of ρ_0 just by itself, though, even in our restricted case with f_{ij} having standard normal margins, and the expectations and standard deviations of $\psi(\cdot, \rho)$ being equal to zero and one respectively. Consider for example the case where f_{ij} is the bivariate Gaussian distribution with correlation coefficient $\rho^* \neq 0$. It is obvious that $\rho_0(\boldsymbol{z}) = \rho^*$ is the population parameter, but in the point $\boldsymbol{z} = \boldsymbol{0}$, we see that $\rho_0 = -\rho^*$ also satisfies (14).

In this and more general situations, such problems are avoided by approximating with a Gaussian in successively smaller neighborhoods. We must therefore make the following assumption that guarantees a well defined population parameter at the point $\boldsymbol{z}$:

Assumption A. For any sequence $\boldsymbol{h}_n$ tending to zero as $n \to \infty$ there exists for the bivariate marginally standard Gaussian vector (Z_i, Z_j) a unique $\rho_{\boldsymbol{h}_n}(\boldsymbol{z})$ that satisfies (13), and there exists a $\rho_0(\boldsymbol{z})$ such that $\rho_{\boldsymbol{h}_n} \to \rho_0(\boldsymbol{z})$.

See Tjøstheim and Hufthammer (2013) for a discussion of Assumption A, and see Berentsen et al. (2016) for a discussion of an alternative neighborhood-free approach to defining the population parameter by means of matching the partial derivatives of the locally Gaussian approximation with the true underlying density function. Assumption A essentially ensures that we estimate the joint densities of each pair of transformed variables consistently, but the joint density $f_0(\boldsymbol{z}) = \Psi(\boldsymbol{z}, \boldsymbol{R}_0)$, where $\boldsymbol{R}_0 = \{\rho_{0,ij}(z_i, z_j)\}_{i<j}$, and $\Psi(\cdot, \boldsymbol{R})$ is the standardized multivariate Gaussian density function with correlation matrix $\boldsymbol{R}$, is not necessarily equal to the true density of the standardized variables, which we for simplicity denote by $f(\boldsymbol{z})$. For this to be true, $f(\boldsymbol{z})$ must be on the form

$$f(\boldsymbol{z}) = \Psi(\boldsymbol{z}, \boldsymbol{R}_0), \tag{15}$$

and this is a restriction of a general density because the entire dependence structure must be contained in the pairwise correlation functions $\rho_{0,ij}(z_i, z_j)$, which is true for distributions with the Gaussian copula (for which the correlation functions are constant in *all* directions), or a stepwise Gaussian distribution as described by Tjøstheim and Hufthammer (2013), but it is difficult (but not paramount for our estimation procedure) to find more analytic examples.

The class of density functions satisfying (15), $H(f_0)$ say, is much richer than the Gaussian case, however, and our performance in estimating a given unconditional density $f(\cdot)$ is clearly sensitive to the distance from $f(\cdot)$ to its best approximant in $H(f_0)$.

Imposing a sparsity requirement like (7) can be viewed in one of two ways. First, as a modeling assumption that can be formally tested, and then discarded if the test should fail. On the other hand, it can be viewed as a simplification of reality that arises due to computational necessity, much like additivity in non-parametric regression as explained in Section 2. We focus on the latter interpretation, and so the method must therefore be judged first and foremost by its performance in practical situations, like those being presented in Section 5. We also refer to Otneim and Tjøstheim (2016) for comprehensive simulations and discussions.

Next, we introduce time series dependence. A strictly stationary series of stochastic variables $\{X_n\}, n = 1, 2, \ldots$ is said to be α-mixing if $\alpha(m) \to 0$, where

$$\alpha(m) = \sup_{A \in \mathcal{F}_{-\infty}^{0}, B \in \mathcal{F}_{m}^{\infty}} |P(A)P(B) - P(AB)|, \tag{16}$$

and where $\mathcal{F}_i^j$ is the σ-algebra generated by $\{X_m, i \leq m \leq j\}$ (Fan and Yao, 2003, p. 68). We require the mixing coefficients (16) of our observations to tend to zero at an appropriate rate, which means that we can turn to standard theorems in order to establish the asymptotic properties of our estimator.

Assumption B. For each pair (i, j), $1 \leq i \leq p$, $1 \leq j \leq p$, $i \neq j$, $\{(Z_i, Z_j)\}_n$ is α-mixing with the mixing coefficients satisfying $\sum_{m \geq 1} m^{\lambda} \alpha(m)^{1-2/\delta} < \infty$ for some $\lambda > 1 - 2/\delta$ and $\delta > 2$.

The next assumption links allowable bandwidth rates with the mixing rate:

Assumption C. $n \to \infty$, and each of the bandwidths h tend to zero such that $nh^{\frac{\lambda+2-2/\delta}{\lambda+2/\delta}} = O(n^{\epsilon_0})$ for some constant $\epsilon_0 > 0$.

In the current context $\{(Z_i, Z_j)\}_n$ is a bivariate process with standard normal margins. In the statement of Theorem 3, Assumption B means that the general p-variate observations $\{\boldsymbol{X}_n\}$ are α-mixing with the specified convergence rate for the mixing coefficients. This distinction has no practical importance when transforming back and forth between these two scales, because the mixing properties of a process are conserved under any measurable transformation (Fan and Yao, 2003, p. 69).

We need a compact parameter space and some regularity conditions on the kernel function in order to prove consistency and asymptotic normality for the local correlations:

Assumption D. The parameter space Θ for ρ is a compact subset of $(-1, 1)$.

Assumption E. The kernel function satisfies $\sup_{\boldsymbol{z}} |K(\boldsymbol{z})| < \infty$, $\int |K(\boldsymbol{y})| \,\mathrm{d}\boldsymbol{y} < \infty$, $\partial/\partial z_i K(\boldsymbol{z}) < \infty$ and $\lim_{z_i \to \infty} |z_i K(z_i)| = 0$ for $i = 1, 2$.

Theorem 1. *Let $\{(Z_i, Z_j)\}_n$ be identically distributed bivariate stochastic vectors with standard normal margins. Denote by $\rho_0(\boldsymbol{z})$ the local Gaussian correlation between Z_i and Z_j, and by $\widehat{\rho}_n(\boldsymbol{z})$ its local likelihood estimate. Then, under assumptions A-E, $\widehat{\rho}_n(\boldsymbol{z}) \xrightarrow{P} \rho_0(\boldsymbol{z})$ as $n \to \infty$.*

Proof. See Appendix A.1. □

Fan and Yao (2003, pp. 76-77) provide a general central limit theorem for nonparametric regression. It is applicable to the local correlations, with obvious adaptations in order to achieve consistent notation. Assume now that $\{\boldsymbol{Z}_n\}$ is a sequence of p-variate observations having standard normal margins, and denote by $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_{p(p-1)/2})$ the vector of local correlations, which has one component for each pair of variables. The local correlations are estimated one by one using the scheme described above, and denote by $\widehat{\boldsymbol{\rho}}$ the estimate of $\boldsymbol{\rho}$. Further, as all bandwidths are assumed to tend to zero at the same rate, statements like h^2 are taken to mean the product of any two bandwidths h_i and h_j.

The local correlation estimates are then jointly asymptotically normal:

Theorem 2. *Under assumptions A-E,*

$$\sqrt{nh_n^2}\,(\widehat{\boldsymbol{\rho}}_n - \boldsymbol{\rho}_0) \xrightarrow{\mathcal{L}} N(\boldsymbol{0}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\Sigma}$ is a diagonal matrix with components

$$\boldsymbol{\Sigma}^{(k,k)} = \frac{f_k(\boldsymbol{z}_k) \int K^2(\boldsymbol{y}_k)\, d\boldsymbol{y}_k}{u^2(\boldsymbol{z}_k, \rho_{0,k}(\boldsymbol{z}_k))\psi^2(\boldsymbol{z}_k, \rho_{0,k}(\boldsymbol{z}_k))},$$

where $k = 1, \ldots, p(p-1)/2$ runs over all pairs of variables, f_k is the corresponding bivariate marginal density of the pair $\boldsymbol{Z}_k$, $\psi(\cdot)$ is defined in (5) and $u(\cdot)$ is defined in the paragraph following equation (13).

When comparing with the corresponding result in Otneim and Tjøstheim (2016), we see that the mixing has no effect on the asymptotic covariance matrix compared with the iid case. See Appendix A.2 for proof.

The preceding theorems lead up to the following asymptotic result for the locally Gaussian conditional density estimates, which is analogous to the corresponding result in Otneim and Tjøstheim (2016) in the unconditional case. Denote by $f_0(\boldsymbol{x}_1|\boldsymbol{X}_2 = \boldsymbol{x}_2)$ the locally Gaussian conditional density function of $\boldsymbol{X}_1|\boldsymbol{X}_2 = \boldsymbol{x}_2$ (where $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$ does not necessarily have standard normal marginals), which is obtained by replacing $f_{\boldsymbol{Z}}/f_{\boldsymbol{Z}_2}$ with $\Psi^*(\boldsymbol{z}; \boldsymbol{\mu}_0^*, \boldsymbol{\Sigma}_0^*)$ in equation (12). The parameters $\boldsymbol{\mu}_0^*$ and $\boldsymbol{\Sigma}_0^*$ are again obtained from formulas (10) and (11) using the population values of the local correlations as defined in Assumption A.

Following the algorithm in Section 3, we must estimate the local Gaussian correlation for pairs of variables $\widehat{\boldsymbol{Z}}_n = \{(\widehat{Z}_i, \widehat{Z}_j)\}_n$ as defined in equation (6), that are not exactly marginally standard normal, because the distribution functions $F_i(\cdot)$, $i = 1, \ldots, p$ must be estimated from the data. In the same way as for the iid case in Otneim and Tjøstheim (2016), we need some extra assumptions on the pairwise copulas between the components in $\boldsymbol{X}$ to ensure that using the empirical distribution distribution functions instead of the true distributions will not affect the asymptotic distribution of the LGDE conditional density estimate. The following assumptions are taken directly from Geenens et al. (2014), who derive the asymptotic properties of a local likelihood copula density estimator in the bivariate case, that is also based on transformations to marginal standard normality.

Assumption F. The marginal distribution functions $F_1, \ldots, F_p$ are strictly increasing on their support.

Assumption G. Each pairwise copula C_{ij} of (X_i, X_j) is such that $(\partial C_{ij}/\partial u)(u, v)$ and $(\partial^2 C_{ij}/\partial u^2)(u, v)$ exist and are continuous on $\{(u, v) : u \in (0, 1), v \in [0, 1]\}$, and $(\partial C_{ij}/\partial v)(u, v)$ and $(\partial^2 C_{ij}/\partial v^2)(u, v)$ exist and are continuous on $\{(u, v) : u \in [0, 1], v \in (0, 1)\}$. In addition, there are constants K_i and K_j such that

$$\left|\frac{\partial^2 C_{ij}}{\partial u^2}(u, v)\right| \leq \frac{K_i}{u(1-u)} \qquad \text{for } (u, v) \in (0, 1) \times [0, 1],$$
$$\left|\frac{\partial^2 C_{ij}}{\partial v^2}(u, v)\right| \leq \frac{K_j}{v(1-v)} \qquad \text{for } (u, v) \in [0, 1] \times (0, 1).$$

Assumption H. Each density $c_{i,j}$ of $C_{i,j}$ exists, is positive, and admits continuous partial derivatives to the fourth order on the interior of the unit square. In addition, there is a constant K_{00} such that

$$c(u, v) \leq K_{00} \min\left(\frac{1}{u(1-u)}, \frac{1}{v(1-v)}\right) \text{ for all } (u, v) \in (0, 1)^2.$$

These smoothness assumptions are quite weak, as can be seen from the discussion in Geenens et al. (2014). Finally, we need to assume that the final back-transformation of the density estimate converge faster than the nonparametric rate of $\sqrt{nh^2}$:

Assumption I. The estimates of the marginal densities and quantile functions that are used for the back-transformations in (12), are asymptotically normal with convergence rates faster than $\sqrt{nh^2}$.

As we use the logspline-estimator (Stone et al., 1997) for the back-transformations in all our examples, we discuss its large sample properties in light of assumption I in Appendix B. Another possible candidate is the basic univariate kernel estimator, which, under some regularity conditions, converges as $\sqrt{nh}$.

Theorem 3. *Let $\{\boldsymbol{X}_n\}$ be a strictly stationary process with density function $f_{\boldsymbol{X}}(\boldsymbol{x})$. Partition $\boldsymbol{X}$ into $\boldsymbol{X}_1 = (X_1, \ldots, X_k)$ and $\boldsymbol{X}_2 = (X_{k+1}, \ldots, X_p)$, and let $\widehat{f}_{\boldsymbol{X}_1|\boldsymbol{X}_2}(\boldsymbol{x}_1|\boldsymbol{X}_2 = \boldsymbol{x_2})$ be the estimate of the conditional density $f_{\boldsymbol{X}_1|\boldsymbol{X}_2}$ that is obtained using the procedure in Section 3. Then, under assumptions A-I,*

$$\sqrt{nh_n^2}\left(\widehat{f}_{\boldsymbol{X}_1|\boldsymbol{X}_2}(\boldsymbol{x}_1|\boldsymbol{X}_2 = \boldsymbol{x_2}) - f_0(\boldsymbol{x}_1|\boldsymbol{X}_2 = \boldsymbol{x_2})\right)$$
$$\xrightarrow{\mathcal{L}} N\big(0, \psi^*(\boldsymbol{z}; \boldsymbol{\mu}_0^*, \boldsymbol{\Sigma}_0^*)^2 g(\boldsymbol{x})^2 \boldsymbol{u}^T(\boldsymbol{z}; \boldsymbol{\mu}_0^*, \boldsymbol{\Sigma}_0^*)\, \boldsymbol{\Sigma}\, \boldsymbol{u}(\boldsymbol{z}; \boldsymbol{\mu}_0^*, \boldsymbol{\Sigma}_0^*)\big),$$

where

$$g(\boldsymbol{x}) = \prod_{i=1}^{k} f_i(x_i)/\phi(z_i),$$
$$\boldsymbol{z} = \{z_i\}_{i=1,\ldots,p} = \{\Phi^{-1}(F_i(x_i))\}_{i=1,\ldots,p},$$

and $\boldsymbol{u}(\boldsymbol{z}) = \nabla \log \psi^*(\boldsymbol{z}, \boldsymbol{\mu}_0^*, \boldsymbol{\Sigma}_0^*)$, where the gradient is taken with respect to the vector of local correlations.

See Appendix A.3 for a proof.

5 Examples

The asymptotic results of the preceding section will not give us the complete picture on how the LGDE estimator of conditional densities behaves in practice for a finite sample. We must also take into account that the simplification (7) of the dependence structure could introduce an approximation error in practical applications, the size of which depends on the problem at hand. We proceed to apply our new estimator to a series of problems using real and simulated data, and compare it with existing methods.

It is customary in the copula literature to generate pseudo-observations by means of the marginal empirical distribution functions, and this is why we can prove Theorem 3 by mostly referring to existing results. The back-transformation (12) must be smooth and invertible, making a standard marginal kernel estimate a natural choice. Extensive testing, however, has revealed that we obtain better finite sample performance if we use the logspline method by Stone et al. (1997) for marginal density and distribution estimates, not only in the back-transformation (12), but also in generating the marginally Gaussian pseudo-observations (6). The following examples, as well as the computer code that accompany this article as supplementary material, therefore use the logspline estimator for both of these purposes. We argue in Appendix B that the asymptotic properties of the logspline estimator do not change when applied to α-mixing data compared to independent data.

5.1 Conditional density estimation

5.1.1 Simulated data with relevant variables

In this section, we wish to investigate the sensitivity of various methods with respect to the number of explanatory variables in the problem, and begin by presenting some

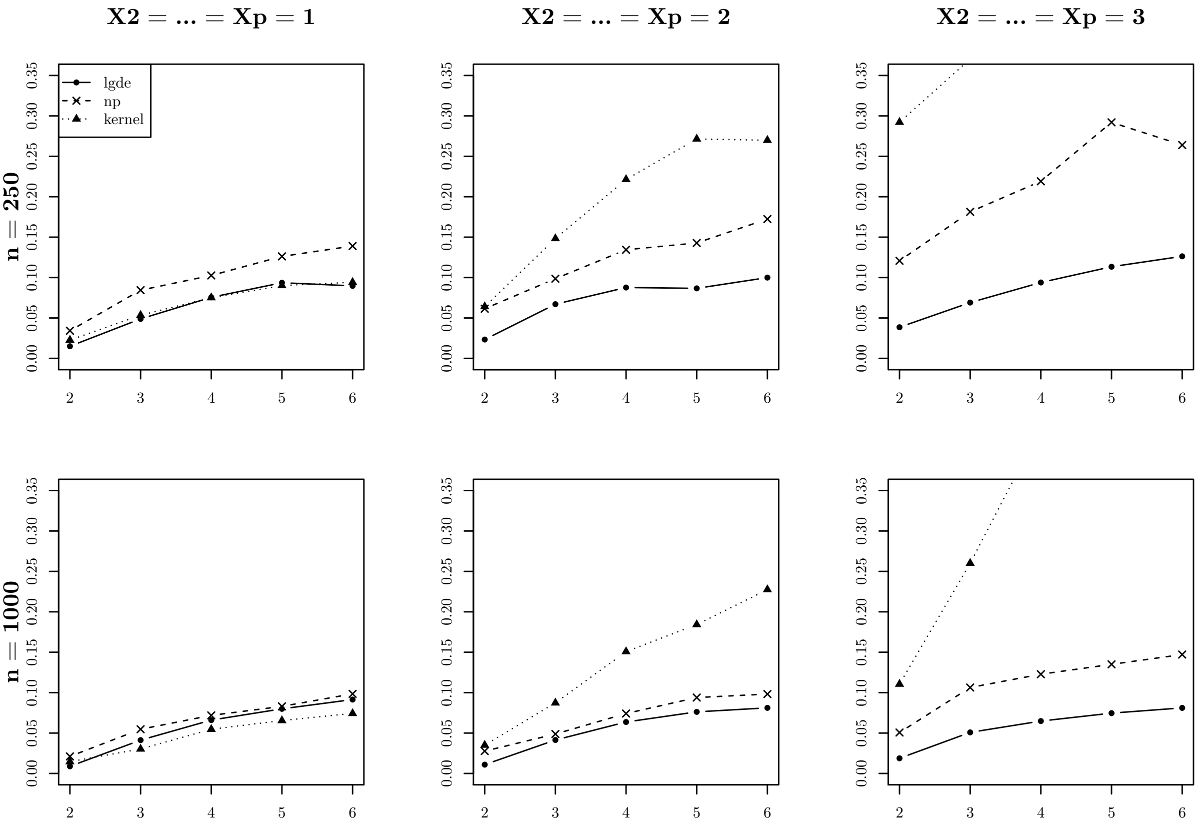

Figure 1: The integrated squared error of conditional density estimates of $f_{X_1|X_2,\dots,X_p}$ as a function of p, generated from a density with exponential margins and a Joe copula with Kendall's Tau equal to 0.6.

simulation experiments in which we generate data from test distributions, measure the integrated squared error (ISE) of our conditional density estimate, and compare it with the two natural competitors which are readily available for implementation: the naïve approach, where the numerator and denominator of (1) are estimated separately using the multivariate kernel estimator with the plug-in bandwidth selector of Wand and Jones (1994), and the specialized kernel method by Hall et al. (2004), which we denote by the name of the software package written in the R programming language (R Core Team, 2015) from which it can be calculated: "NP" (Hayfield et al., 2008).

The first test distribution has standard exponentially distributed margins, and the dependence structure is defined by the Joe copula (see e.g. Nelsen (2013, p. 116, distribution 6)) with parameter $\theta = 3.83$, which corresponds to a Kendall's Tau of 0.6 between all pairs of variables. For each dimension p, ranging from 2 to 6, we generate $2^7 = 128$ data sets, and estimate the conditional density of $X_1|X_2 = \cdots = X_p = c$, with c being equal to 1,2 and 3 in this example. We calculate the ISE of the density estimates numerically over 2000 equally spaced grid points, and graph the mean of the estimated errors as a function of the dimension for two different sample sizes ($n = 250$ and $n = 1000$), see Figure 1.

The basic kernel estimator performs well in the center of the distribution, especially in the example with sample size 1000. When we condition on values that are farther out in tail, however, it quickly deteriorates as the dimension increases. This behavior is of course expected because of the curse of dimensionality. The NP-estimator is clearly a major improvement to naïve kernel estimation of conditional densities, but in this

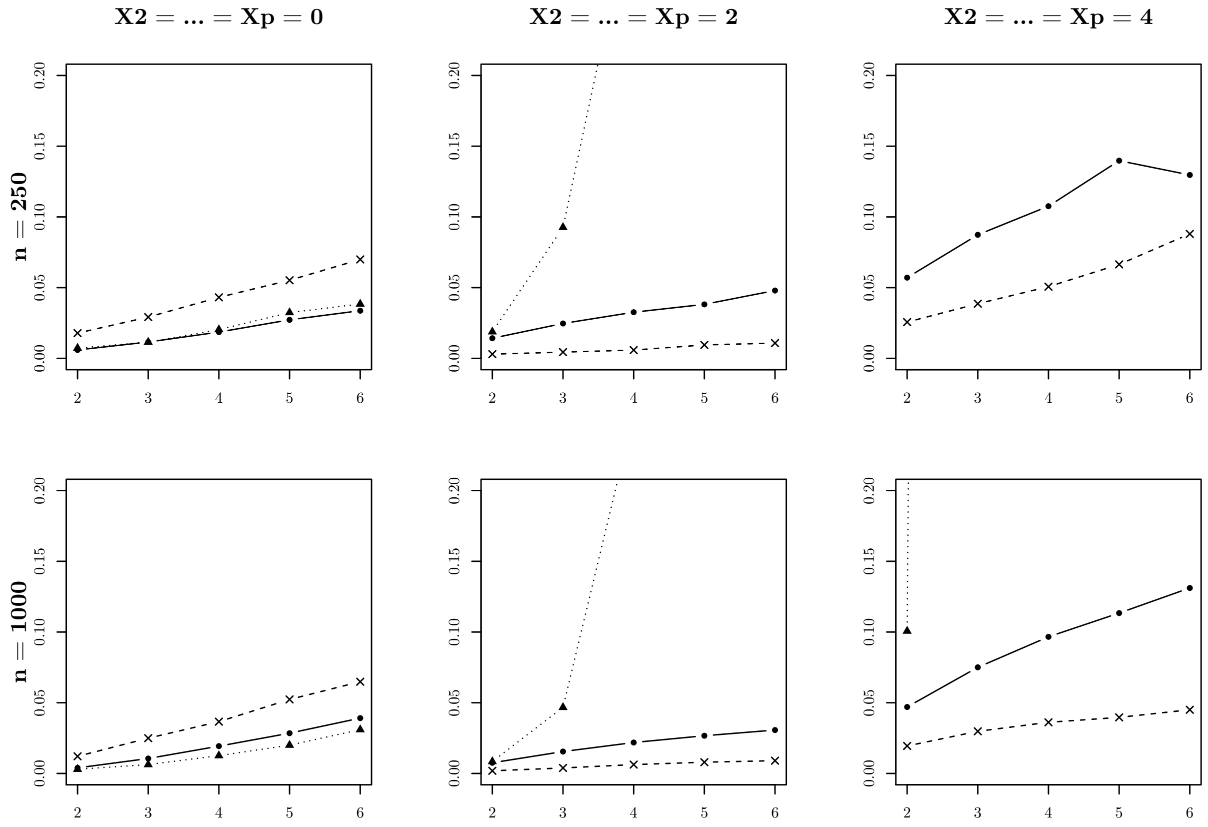

Figure 2: The integrated squared error of conditional density estimates of $f_{X_1|X_2,\dots,X_p}$ as a function of p, generated from the multivariate t-distribution with 4 degrees of freedom.

example we see that the LGDE approach is the overall best performer. It matches the purely non-parametric methods in lower-dimensional cases, but also boasts a greater robustness against increasing dimensionality than its competitors. The tail behavior of the LGDE is much better than the other two methods. It is governed by a Gaussian distribution, which again is determined locally by the behavior of $f_{X_1|X_2,\dots,X_p}$ in the tail.

5.1.2 Simulated data from a heavy-tailed distribution

Otneim and Tjøstheim (2016) show that the unconditional version of the LGDE does not work very well when fitted to the heavy-tailed $t(4)$-distribution. The reason for this is not entirely clear, but one explanation is that the cross-validated bandwidths are too small. The conditional version of the LGDE also starts to struggle when presented with data from this distribution, as can be seen in Figure 2. It is expected that using the t-distribution in the same pairwise and local manner as we use the Gaussian distribution here, will improve this fit, and we discuss this more closely in Section 6. The conditional density estimator by Hall et al. (2004) is the best alternative in this case if the explanatory variables are not in the center of the distribution.

5.1.3 Simulated data with irrelevant variables

One challenge in estimating conditional densities is to discover, and take account of, independence between variables. We have not addressed this problem explicitly in the derivation of our estimator, contrary to the NP-estimator by Hall et al. (2004), which smooths irrelevant variables away automatically. In our next example, however, most

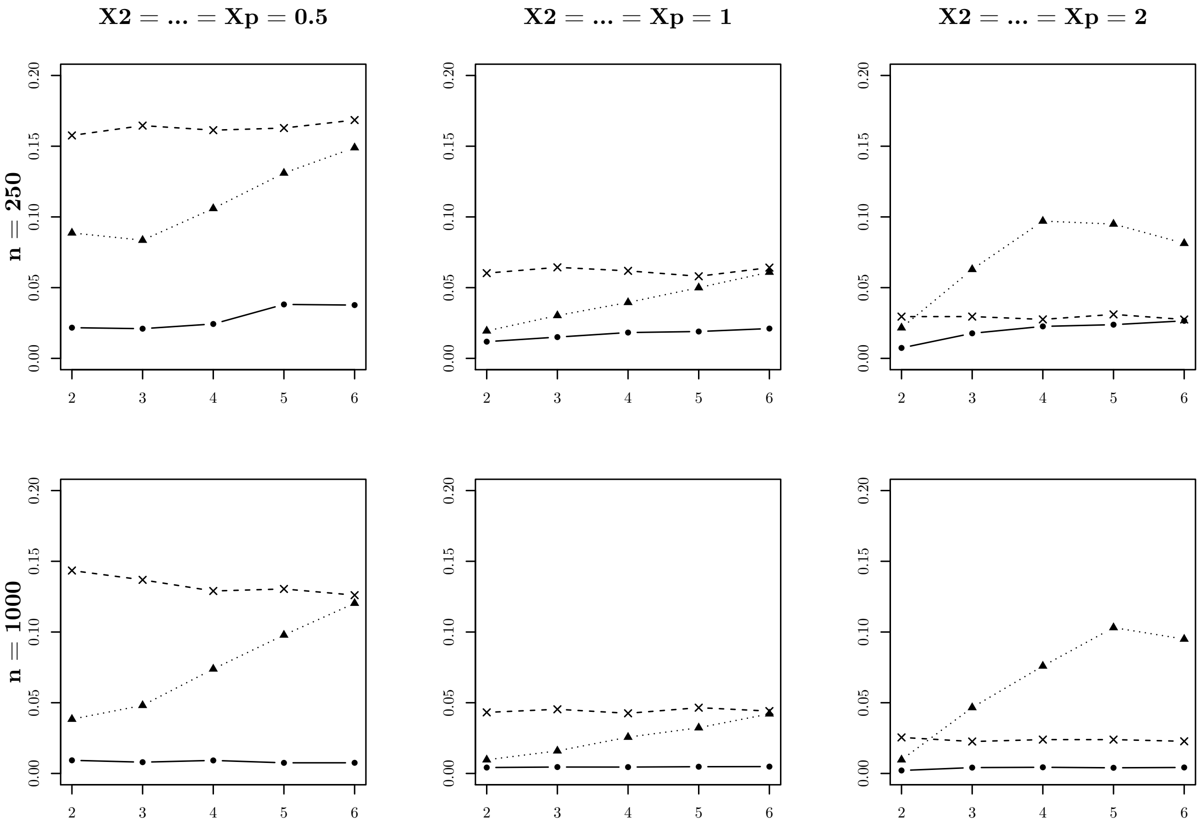

Figure 3: The integrated squared error of conditional density estimates of $f_{X_1|X_2,\ldots,X_p}$ as a function of p, generated from a density in which the first two variables are marginally log-normal with a $t(10)$-copula, and the rest of the variables are multivariate $t(5)$-distributed, independently from (X_1, X_2).

of the explanatory variables are independent from the response variable, but they are mutually dependent themselves. In the two-dimensional case with $\boldsymbol{X} = (X_1, X_2)$, we generate data from a bivariate distribution with log-normal margins that has been assembled using the t-copula with 10 degrees of freedom. For all dimensions greater than two, the remaining variables $X_3, \ldots, X_p$ are drawn from a multivariate t-distribution with 5 degrees of freedom, but independent from (X_1, X_2).

It turns out that our approach handles this case very well, see Figure 3. None of the methods have errors that grow sharply with the dimension, which indicate that they more or less ignore the extra noise that the extra dimensions contains. The LGDE-method is clearly the best, however, according to this particular choice of error measure. The explanation for this is the equivalence between independence and the local correlation being equal to zero between marginally Gaussian variables, which in turn means that, by construction, variables that are independent from the response variable will have very little influence in the final conditional density estimate.

5.1.4 Real data with irrelevant variables

We can explore this property using a real data set as well. Consider a subset of the data set which is also analyzed in Otneim and Tjøstheim (2016) comprising daily log-returns on the S&P 500 stock index observed on 1443 days from January 3rd 2005 until July 14th, 2010. In this example we will use only the first 500 observations, so the financial

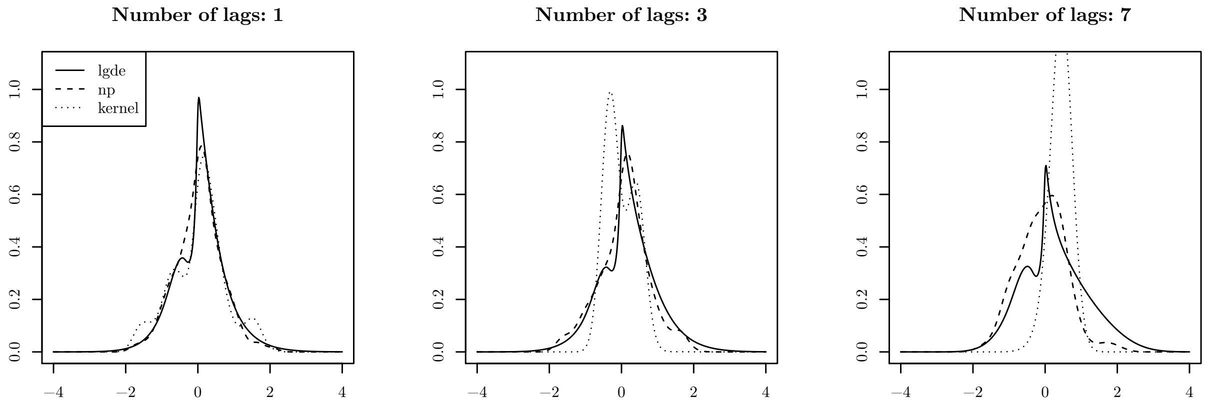

Figure 4: Estimate of the conditional density of the US log-returns conditioned on $X_{t-1} = \cdots = X_{t-k} = -1$ with $k = 1, 3, 7$ respectively.

crisis of 2008 is not included in this particular analysis.

We know that there is very little extra information given the first lag in this time series, thus estimating the marginal density of these log returns by conditioning on more and more lags will not introduce more information, but rather noise, that should ideally be ignored by the estimation routine.

Figure 4 displays the marginal density estimates of the data, calculated using the three competing methods and conditioned on the preceding 1, 3 and 7 days' values respectively being equal to −1. All methods perform similarly in the first case in which we condition on only one variable. In the second panel we condition on three lags, which amounts to a four dimensional problem in terms of density estimation, and the naïve kernel estimator, not surprisingly, struggles in this case. The other two methods, however, the NP and the LGDE, remain largely unchanged, which indicates that they, for the most part, ignore the additional two variables of data. When conditioning on 7 lags, the kernel estimator should not be trusted. The NP-estimator also appears to loose some characteristics, like the sharpness of its peak and the fatness of its right tail. The LGDE, on the other hand, seems to be the better performer in this case. Although the estimate is slightly deformed compared to the other two figures, its main characteristics are conserved. The tails in particular shows great robustness compared to the other two methods, and we believe that this behavior to a large part explains its good performance in simulation experiments, and we will also exploit this feature in Section 5.3.

5.1.5 Melbourne temperature data: comparison with local polynomials

The local polynomial conditional density estimators of Hyndman et al. (1996) and Hyndman and Yao (2002) is in its current implementation restricted to the case where the explanatory and response variables are both scalar, and is therefore not included in the simulation experiments of the preceding subsection. We will, however, compare these estimators to our approach using the Melbourne temperature data that is presented by Hyndman et al. (1996). The data consists of daily recordings of the maximum air temperature in Melbourne, Australia from 1981 until 1990. It is known that a low maximum temperature one day most often results in a similar temperature the next day. Local meteorological conditions, however, have the effect that a high maximum temperature is often followed by either a large, or a much smaller observation, making the corresponding conditional density bimodal. The Hyndman et al. (1996)-estimator, which in this exam-

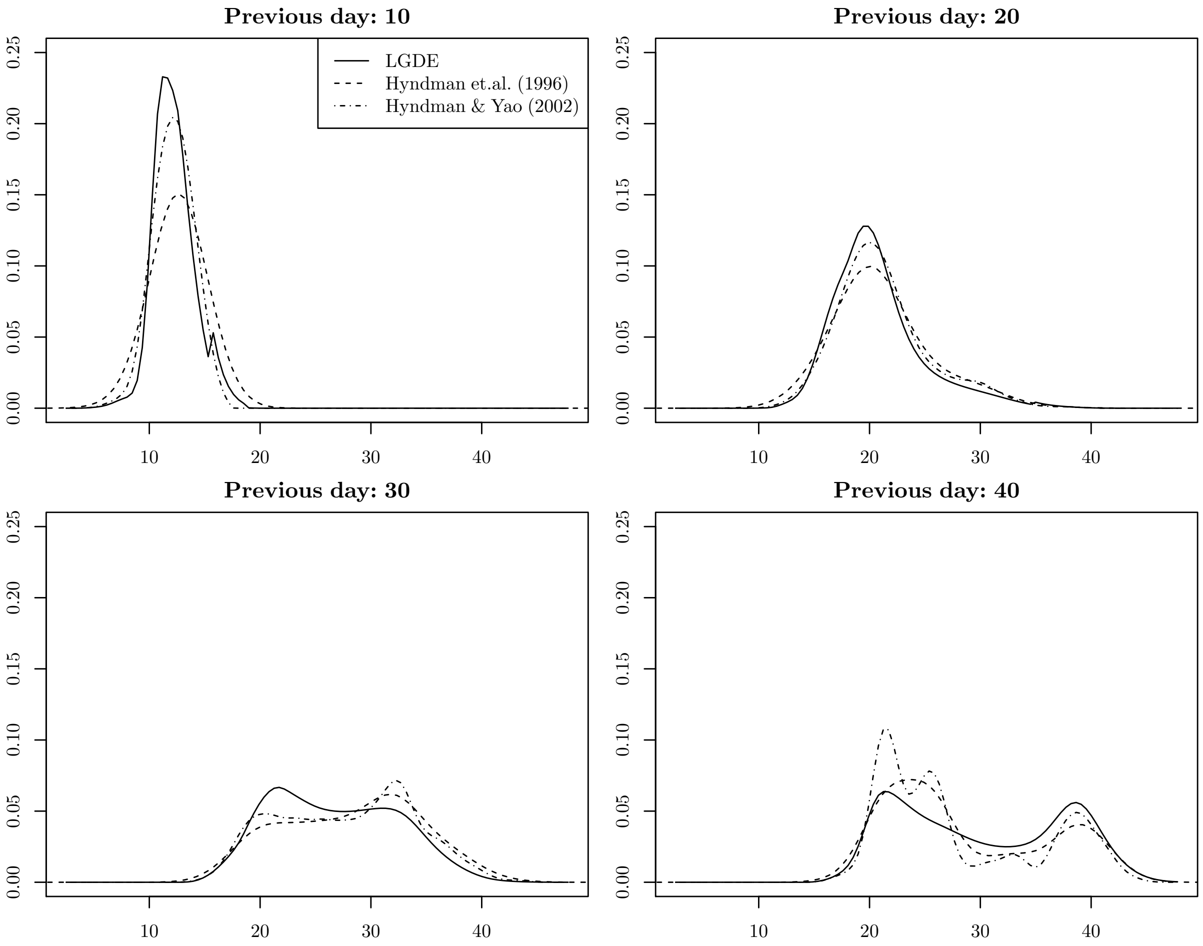

Figure 5: Australian temperature data, with estimated conditional density of the maximum daily air temperature, given a preceding recording of 10, 20, 30 and 40 degrees Celsius respectively.

ple is a local polynomial of order zero, recovers this phenomenon nicely, and although our locally Gaussian estimator is not identical, it gives a similar picture, see Figure 5. The Hyndman and Yao (2002)-estimator is a locally quadratic polynomial, and mostly agrees with the other methods, but seems to be slightly overfitting the density in the lower right panel.

It is interesting to note that the bimodality of the LGDE-estimator is mirrored compared with the local polynomials in the lower left panel.

5.2 Partial correlation and covariance

The partial autocorrelation function for a stationary time series at lag k is the correlation between X_t and X_{t-k}, given the values of the intervening lags (Brockwell and Davis, 2013, p. 98). The concept of partial correlation is very important, especially in the analysis of conditional dependencies in Bayesian networks. Partial local correlation is a natural extension of local correlation in light of the new theory allowing for dependent observations. Consider for example the nonlinear AR(1) model

$$X_t = 0.8X_{t-1} + 0.5\sqrt{|X_{t-1}|} + Z_t,$$

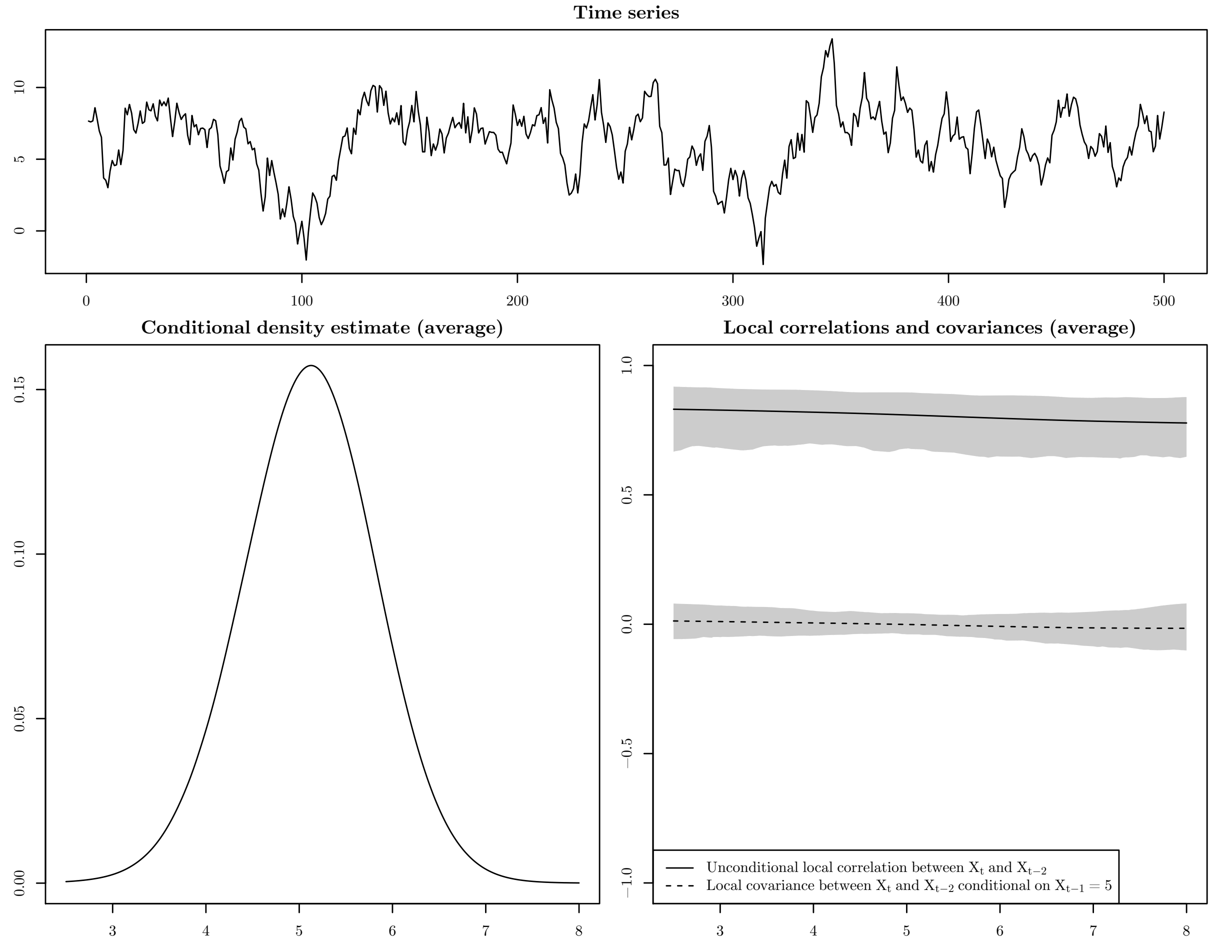

Figure 6: The top panel displays a simulated time series. The lower left panel displays the average of the estimated conditional densities of $X_t|X_{t-2} = 5$, and the lower right panel shows the unconditional diagonal local correlation between X_t and X_{t-2}, as well as the same quantity when conditioned on the intermediate value X_{t-1}, with 95% empirical confidence intervals.

where the Z_ts are independent standard normal innovations. One realization of length 500 is plotted in the upper panel of Figure 6. There is strong serial dependence in this model. Indeed, if we estimate the joint density of the lagged values X_t and X_{t-2} using the LGDE methodology, the estimated local correlation is close to 1. This can be seen in the lower right panel of Figure 6, in which the local correlation for 300 realizations has been averaged and plotted as a solid line along the diagonal $x_t = x_{t-2}$, along with the empirical 95% confidence interval. We do know from the Markov property of $\{X_t\}$, however, that X_t is independent of X_{t-2} given X_{t-1}, and this is clearly reflected in the estimated local covariance between the two variables for the joint *conditional* density of $(X_t, X_{t-2})|X_{t-1} = x_{t-1}$ (where $x_{t-1} = 5$ in this particular case), that has been plotted as a dashed line. We use the term local covariance here, instead of local correlation, because the diagonal elements in $\mathbf{\Sigma}$ as defined by (11) are no longer 1. As seen in the lower right panel of Figure 6, the local covariance practically vanishes when the intermediate variable is conditioned upon.

The average of the estimated conditional densities in question has been plotted along its diagonal in the lower left panel of Figure 6.

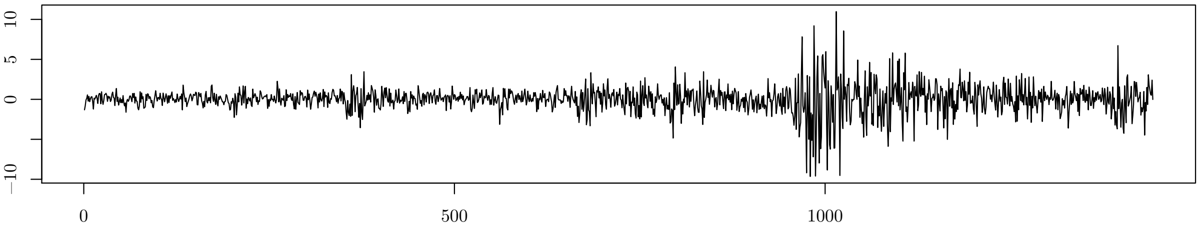

Figure 7: Value of the portfolio over a period of 1442 days.

Table 1: Proportion of observations exceeding the estimated VaR

	Level		
Method	0.005	0.01	0.05
LGDE	0.014	0.017	0.072
np	0.084	0.097	0.161
Kernel	0.117	0.134	0.187
Gaussian	0.045	0.064	0.125

5.3 Forecasting the value-at-risk of a portfolio

There is a vast literature available on portfolio optimization theory. A vital element when selecting the optimal distribution of wealth over a set of assets is the estimation of risk, of which the Value-at-Risk (VaR) is a common measure. The VaR of a portfolio at level α is simply the upper $(1-\alpha)$-quantile of the loss-distribution of the portfolio, which usually needs to be estimated from past data.

We look at the S&P 500 data from Section 5.1.4, as well as the corresponding log-returns on the British FTSE 100 index and the Norwegian OBX, and consider the observations on all 1443 days. In this toy example, we will show that our conditional density estimator may well be used as an instrument in estimating the VaR.

We wish to estimate the daily VaR of a portfolio consisting of each of these indices, equally weighted, conditioned on the observed log-returns on preceding days. The log-returns of this portfolio is plotted in Figure 7. Denote by $(X_1, \ldots, X_4)$ the four-dimensional vector that we observe each day, in which X_1 is the value of the portfolio that day, and $X_2, \ldots, X_4$ are the values of its individual components on the preceding day. On each day we estimate the conditional density of $X_1 | X_2 = x_2, \ldots, X_4 = x_4$ and calculate the α-level VaR by numerical integration. We do the same by using the non-parametric kernel estimator by Hall et al. (2004), naive kernel estimator, as well as by assuming the data to be jointly Gaussian and calculating the quantile from a fully parametric fit. We start our analysis on day number 500, and for computational feasibility, we calculate the bandwidths for all methods on the first day of analysis only, and keep them constant throughout the period.

Table 1 displays the result of our analysis. For each method we count the proportion of observations that exceed the estimated VaR on the corresponding day. We see that all methods under-estimate the risk, but the LGDE-approach is clearly the better performer, which we believe is due to its tendency to allow fat tails in the density estimates, see e.g. Figure 4, even though it has a *local* Gaussian tail.

A thorough treatment of this topic would include pre-filtering of the data using for example a GARCH-type model as found in Palaro and Hotta (2006), as well as implementation of the LGDE in optimization over the portfolio weights, but that is beyond the scope of this paper.

6 Conclusion and further work

Constructing non-parametric estimates of conditional density functions is a fundamental problem in statistics, but it is difficult, because many of the existing methods rely either on the traditional kernel density estimator, or on separate estimates of the numerator and denominator in the definition of the conditional density, or, most often, both. This could work in lower dimensional problems, especially if we keep ourselves away from the tails of the distribution in question.

We have shown, however, that by using the LGDE methodology, both of these problems tend to disappear. The simplified locally Gaussian estimates cope far better in higher dimensions than the kernel estimator, and it provides an explicit expression of the conditional density estimates, without the need for separate estimates of the numerator and denominator. The result is a general conditional density estimator for continuous data that is robust against dimensionality issues, modeling error, as well as noise induced by irrelevant variables.

These properties have been demonstrated through examples and asymptotic derivations. A more comprehensive theoretical analysis of the LGDE-framework and its possible generalizations remains to be developed, and will be the subject of later studies. For example, the degree to which a general multivariate density function can be characterized by pairwise locally Gaussian correlations, or the distance between $f(\boldsymbol{x})$ and $f_0(\boldsymbol{x})$ in keeping with the notation from Section 4, is a challenge, cf. Otneim and Tjøstheim (2016). Further, if the LGDE-approach can be labeled as a two-fold approximation compared to the fully non-parametric, or p-fold, estimation procedure in which we omit the simplification (7), it might be worthwhile to develop a general procedure allowing for a k-fold model, in which each local correlation depends on k variables, with k increasing, and these variables being selected based on data analogously to variable selection methods in regression. In theory, this can be generalized even further by replacing the normal distribution as a building block, with another member of the family of elliptical distributions that also organizes its parameters in a covariance-like matrix structure. Deriving conditional densities from such a general model requires more work, but should in principle be possible.

A Proofs

A.1 Proof of Theorem 1

Except from a slight modification that accounts for the replacement of independence with α-mixing, the proof of Theorem 1 is identical to the corresponding proof in Otneim and Tjøstheim (2016), which again is based on the global maximum likelihood case covered by Severini (2000). For each location $\boldsymbol{z}$, that we for simplicity suppress from notation, denote by $Q_{\boldsymbol{h}_n,K}(\rho)$ the expectation of the local likelihood function $L_n(\rho, \boldsymbol{Z})$. Consistency follows from uniform convergence in probability of $L_n(\rho, \boldsymbol{Z})$ towards $Q_{\boldsymbol{h}_n,K}$ (ρ), conditions for which are provided in Corollary 2.2 by Newey (1991).

The result requires compact support of the parameter space, equicontinuity and Lipschitz continuity of the family of functions $\{Q_{\boldsymbol{h}_n,K}(\rho)\}$, as well as pointwise convergence of the local likelihood functions. Compactness is covered by Assumption D, and the demonstration of equi- and Lipschitz continuity in Otneim and Tjøstheim (2016) does not rely on the independent data assumption. Pointwise convergence follows from a standard non-parametric law of large numbers in the independent case. Our assumption B about α-mixing data, however, ensures that pointwise convergence still holds, see for example Theorem 1 by Irle (1997), conditions for which are straightforward to verify in our local likelihood setting.

The rest of the proof is identical to the corresponding argument by (Severini, 2000, pp. 105-107).

A.2 Proof of Theorem 2

Consider first the bivariate case, in which there is only one local correlation to estimate. The first part of the proof goes through exactly as in the iid-case of Otneim and Tjøstheim (2016). We follow the argument for global maximum likelihood estimators as presented in Theorem 7.63 by Schervish (1995). The statement of Theorem 2 follows provided that

$$Y_n(\boldsymbol{z}) = \sum_{i=1}^{n} K\left(|\boldsymbol{h}_n|^{-1}(\boldsymbol{Z}_i - \boldsymbol{z})\right) u(\boldsymbol{Z}_i, \rho_0) = \sum_{i=1}^{n} V_{ni}, \tag{17}$$

is asymptotically normal, and this follows from a standard Taylor expansion. In the iid-case, the limiting distribution of (17) is derived using the same technique as when demonstrating asymptotic normality for the standard kernel estimator, for example as in the proof of Theorem 1A by Parzen (1962). We establish asymptotic normality of (17) in case of α-mixing data, however, by going through the steps used in proving Theorem 2.22 in Fan and Yao (2003). Let $W_i = h^{-1}V_{ni}$, then

$$\begin{aligned}\frac{1}{nh^2}\mathrm{Var}(Y_n(\boldsymbol{z})) &= \frac{1}{nh^2}\left\{\sum_{i=1}^{n}\mathrm{Var}(V_{ni}) + 2\sum\sum_{1\leq i<j\leq n}\mathrm{Cov}(V_{ni}, V_{nj})\right\} \\ &= \mathrm{Var}(W_1) + 2\sum_{j=1}^{n}(1 - j/n)\mathrm{Cov}(W_1, W_{j+1}),\end{aligned}$$

where

$$\begin{aligned}\mathrm{Var}(W_1) &= \mathrm{E}(W_1^2) - (\mathrm{E}(W_1))^2 \\ &= \int h^{-2}u^2(\boldsymbol{z}, \rho_0)K^2(h^{-1}(\boldsymbol{y} - \boldsymbol{z}))f(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y} + O(h^2) \\ &= \int u^2(\boldsymbol{z} + h\boldsymbol{v})K^2(\boldsymbol{v})f(\boldsymbol{z} + h\boldsymbol{v})\,\mathrm{d}\boldsymbol{v} + O(h^2) \\ &\to u^2(\boldsymbol{z}, \rho_0)f(\boldsymbol{z})\int K^2(\boldsymbol{v})\,\mathrm{d}\boldsymbol{v} \overset{\mathrm{def}}{=} M(\boldsymbol{z}) \text{ as } \boldsymbol{h} \to 0,\end{aligned}$$

and

$$|\mathrm{Cov}(W_1, W_{j+1})| = |\mathrm{E}(W_1W_{j+1}) - \mathrm{E}(W_1)\mathrm{E}(W_{j+1})| = O(h^2),$$

using the same argument once again. Therefore,

$$\left|\sum_{j=1}^{m_n}\mathrm{Cov}(W_1, W_{j+1})\right| = O(m_nh^2).$$

Fan and Yao (2003) require that

$$\mathrm{E}(u(\boldsymbol{Z}_n, \rho_0(\boldsymbol{z}))^\delta) < \infty \tag{18}$$

for some $\delta > 2$, but this is of course true for our transformed data, because it is marginally normal. In proposition 2.5(i) by Fan and Yao (2003) we can therefore use $p = q = \delta > 2$ in order to obtain, for some constant C,

$$|\mathrm{Cov}(W_|, W_{j+1})| \leq C\alpha(j)^{1-2/\delta}h^{4/\delta-2}.$$

Let $m_n = (h_n^2|\log h_n^2|)^{-1}$. Then $m_n \to \infty$, $m_n h^2 \to 0$, and

$$\sum_{j=m_n+1}^{n-1} |\mathrm{Cov}(W_1, W_{j+1})| \leq C\frac{h^{4/\delta-2}}{m_n^\lambda} \sum_{j=m_n+1}^{n} j^\lambda \alpha(j)^{1-2/\delta} \to 0,$$

which follows from assumption B. Thus,

$$\sum_{j=1}^{n-1} \mathrm{Cov}(W_1, W_{j+1}) \to 0,$$

and it follows that

$$\frac{1}{nh^2}\mathrm{Var}(Y_n(\boldsymbol{z})) = M(\boldsymbol{z})(1 + o(1)).$$

The proof now continues exactly as in Fan and Yao (2003) using the "big block small block" technique, but with the obvious replacement of h with h^2 to accommodate the bivariate case.

We expand the argument to the multivariate case using the Cramèr-Wold device. Let $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_d)^T$ be the vector of local correlations, where $d = p(p-1)/2$, write $\boldsymbol{u}(\boldsymbol{z}, \boldsymbol{\rho}_0) = (u_1(\boldsymbol{z}, \boldsymbol{\rho}_0), \ldots, u_d(\boldsymbol{z}, \boldsymbol{\rho}_0))$ and let $\boldsymbol{S}_n(\boldsymbol{z}) = \{S_{ni}(\boldsymbol{z})\}_{i=1}^d$, where

$$S_{ni} = \sum_{n=1}^{n} u_k(\boldsymbol{Z}_t, \boldsymbol{\rho}_0)K(|\boldsymbol{h}|^{-1}(\boldsymbol{Z}_t - \boldsymbol{z})).$$

We must show that

$$\sum_k a_k S_{nk} \xrightarrow{\mathcal{L}} \sum_k a_k Z_k^*, \tag{19}$$

where $\boldsymbol{a} = (a_1, \ldots, a_d)^T$ is an arbitrary vector of constants, and $\boldsymbol{Z}^* = (Z_1^*, \ldots, Z_k^*)$ is a jointly normally distributed random vector. Because of Slutsky's Theorem, it suffices to show that the left hand side of (19) is asymptotically normal. This follows from observing that it is on the same form as the original sequence comprising S_n, with

$$\sum_k a_k S_{nk} = \sum_n u^*(\boldsymbol{Z}_n, \boldsymbol{\rho}_0)K(|\boldsymbol{h}|^{-1}(\boldsymbol{Z}_n - \boldsymbol{z})),$$

where $u^*(\boldsymbol{Z}_n, \boldsymbol{\rho}_0) = \sum_k a_k u_k(\boldsymbol{Z}_n, \boldsymbol{\rho}_0)$. It is well known that any measurable mapping of a mixing sequence of random variables inherit the mixing properties of the original series, so condition B is therefore satisfied by the linear combination. The new sequence of observations satisfies (18) because it follows from Jensen's inequality that for $\delta > 2$,

$$\begin{aligned}\left[\frac{u^*(\boldsymbol{Z}_t, \boldsymbol{\rho}_0)}{\sum_k a_k}\right]^\delta &= \left[\frac{\sum_k a_k u_k(\boldsymbol{Z}_t, \boldsymbol{\rho}_0)}{\sum_k a_k}\right]^\delta \\ &\leq \frac{\sum_k a_k [u_k(\boldsymbol{Z}_t, \boldsymbol{\rho}_0)]^\delta}{\sum_k a_k},\end{aligned}$$

so that

$$\mathrm{E}[u^*(\boldsymbol{Z}_t, \boldsymbol{\rho}_0)]^\delta \leq \sum_k a_k \mathrm{E}[u_k(\boldsymbol{Z}_t, \boldsymbol{\rho}_0)]^\delta \left[\sum_k a_k\right]^{\delta-1} < \infty.$$

The off-diagonal elements in the asymptotic covariance matrix are zero using the same arguments as in Otneim and Tjøstheim (2016).

A.3 Proof of Theorem 3

The key to proving 3 is to show that the asymptotic distribution of (17) remains unchanged when the marginally standard normal stochastic vectors $\boldsymbol{Z}_n$ are replaced with the pseudo-observations

$$\widehat{\boldsymbol{Z}}_n = \left(\Phi^{-1}(\widehat{F}_1(X_{j1})), \ldots, \Phi^{-1}(\widehat{F}_p(X_{jp}))\right)^T,$$

where $\widehat{F}_i(\cdot)$, $i = 1, \ldots, p$ are the marginal empirical distribution functions. This is shown in the independent case under assumptions F-G in Otneim and Tjøstheim (2016), by providing a slight modification to Proposition 3.1 by Geenens et al. (2014). The essence in that proof is the convergence of the empirical copula process, which remain unchanged if we replace the assumption of independent observations with α-mixing, according to Bücher and Volgushev (2013).

The multivariate delta method states that if $\sqrt{nh^2}(\theta_n - \theta) \overset{\mathcal{L}}{\to} N(0, A)$ and $q : R^n \to R$ has continuous first partial derivatives, then $\sqrt{nh^2}(q(\theta_n) - q(\theta)) \overset{\mathcal{L}}{\to} N(0, \nabla q(\theta)^T A \nabla q(\theta))$ (Schervish, 1995, p. 403)). In our case, $q(\boldsymbol{\rho}) = \Psi(\boldsymbol{z}, \boldsymbol{R})g(\boldsymbol{x})$, and

$$\nabla q(\boldsymbol{\rho}) = \Psi(\boldsymbol{z}, \boldsymbol{R})g(\boldsymbol{x})\boldsymbol{u}(\boldsymbol{z}, \boldsymbol{R}),$$

from which the result follows immediately.

B Large sample properties of the logspline estimator

The current implementation of our method in the R programming language (R Core Team, 2015) uses the logspline method by Stone et al. (1997) for marginal density estimation. The asymptotic theory for the logspline estimator is derived by Stone (1990), but restricted to density functions with compact support. Otneim and Tjøstheim (2016) relax this requirement using a truncation argument, so that the requirement of compact support can be replaced by an assumption on the tails of the unknown density not being too heavy.

In particular, Stone (1990) denotes by $\epsilon \in (0, 1/2)$ a tuning parameter that determines the asymptotic rate at which new nodes are added to the logspline procedure. If ϵ is close to zero, new nodes are added quickly to the procedure, and as $\epsilon \to 1/2$, new nodes are added very slowly. Stone (1990) then provides the following asymptotic results (again, under the assumption that the true density $f(\boldsymbol{x})$ has compact support):

$$\sqrt{n^{0.5+\epsilon}} \left(\widehat{f}_i(x) - f(x)\right) \overset{\mathcal{L}}{\to} N(0, \sigma_1^2),$$

and

$$\sqrt{n^{0.5}} \left(\widehat{F}_i(x) - F(x)\right) \overset{\mathcal{L}}{\to} N(0, \sigma_2^2).$$

Otneim and Tjøstheim (2016) show that these results hold if there exist constants $M > 0$, $\gamma > 2\epsilon/(1-2\epsilon)$, and $x_0 > 0$ such that $f(x) \leq M|x|^{-(5/2+\gamma)}$ for all $|x| > x_0$, so the 'worst case scenario' with respect to assumption I when using the logspline estimator for the final back-transformation, is ϵ being close to zero. In that case, we must require the bandwidths to tend to zero fast enough so that $n^{1/2}h^2 \to 0$, but on the other hand, that will allow γ to approach zero, and thus the tail-thickness of the density to approach that of $|x|^{-5/2}$.

What remains here is to show that these results hold also in the case where the observations are α-mixing. This is easily done by replacing the use of the iid central limit theorem (clt) in the proof of Theorem 3 in Stone (1990), with a corresponding clt that holds under our mixing condition. For example, Theorem A by Peligrad (1992) proves the clt under α-mixing provided that the mixing coefficients satisfy $\sum_{n=1}^{\infty} \alpha(n)^{1-2/\delta} < \infty$. This condition follows from our assumption B.

C Supplementary material

The file `code.zip`, that accompanies this article, contains the data sets that has been used, as well as routines for implementing the conditional density estimator in the R programming language (R Core Team, 2015).

References

David M Bashtannyk and Rob J Hyndman. Bandwidth selection for kernel conditional density estimation. *Computational Statistics & Data Analysis*, 36(3):279–298, 2001.

Geir Drage Berentsen, Ricardo Cao, Mario Francisco-Fernández, and Dag Tjøstheim. Some properties of local gaussian correlation and other nonlinear dependence measures. *Journal of Time Series Analysis*, 2016.

Peter J Brockwell and Richard A Davis. *Time series: theory and methods*. Springer Science & Business Media, 2013.

Axel Bücher and Stanislav Volgushev. Empirical and sequential empirical copula processes under serial dependence. *Journal of Multivariate Analysis*, 119:61–70, 2013.

José E Chacón and T Duong. Multivariate plug-in bandwidth selection with unconstrained pilot bandwidth matrices. *Test*, 19(2):375–398, 2010.

Xiaohong Chen and Oliver B Linton. The estimation of conditional densities. *LSE STICERD Research Paper No. EM415*, 2001.

Jianqing Fan and Qiwei Yao. *Nonlinear time series: nonparametric and parametric methods*. Springer Science & Business Media, 2003.

Jianqing Fan and Tsz Ho Yim. A crossvalidation method for estimating conditional densities. *Biometrika*, 91(4):819–834, 2004.

Jianqing Fan, Qiwei Yao, and Howell Tong. Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika*, 83(1):189–206, 1996.

Olivier P Faugeras. A quantile-copula approach to conditional density estimation. *Journal of Multivariate Analysis*, 100(9):2083–2099, 2009.

Gery Geenens, Arthur Charpentier, and Davy Paindaveine. Probit transformation for nonparametric kernel estimation of the copula density. *arXiv preprint arXiv:1404.4414*, 2014.

Peter Hall. On Kullback-Leibler loss and density estimation. *The Annals of Statistics*, 15(4):1491–1519, 1987.

Peter Hall, Jeffrey Scott Racine, and Qi Li. Cross-validation and the estimation of conditional probability densities. *Journal of the American Statistical Association*, 99 (468):1015–1026, 2004.

Tristen Hayfield, Jeffrey S Racine, et al. Nonparametric econometrics: The np package. *Journal of statistical software*, 27(5):1–32, 2008.

Nils Lid Hjort and MC Jones. Locally parametric nonparametric density estimation. *The Annals of Statistics*, pages 1619–1647, 1996.

Michael P Holmes, Alexander G Gray, and Charles Lee Isbell. Fast nonparametric conditional density estimation. *arXiv preprint arXiv:1206.5278*, 2012.

Rob J Hyndman and Qiwei Yao. Nonparametric estimation and symmetry tests for conditional density functions. *Journal of nonparametric statistics*, 14(3):259–278, 2002.

Rob J Hyndman, David M Bashtannyk, and Gary K Grunwald. Estimating and visualizing conditional densities. *Journal of Computational and Graphical Statistics*, 5(4): 315–336, 1996.

A Irle. On consistency in nonparametric estimation under mixing conditions. *Journal of multivariate analysis*, 60(1):123–147, 1997.

Richard A. Johnson and Dean W. Wichern. *Applied Multivariate Statistical Analysis, Sixth Edition*. Pearson Education Iternational, 2007.

Roger B Nelsen. *An introduction to copulas*, volume 139. Springer Science & Business Media, 2013.

Whitney K Newey. Uniform convergence in probability and stochastic equicontinuity. *Econometrica*, 59(4):1161–1167, 1991.

Håkon Otneim and Dag Tjøstheim. The locally gaussian density estimator for multivariate data. *Statistics and Computing*, pages 1–22, 2016. ISSN 1573-1375. doi: 10.1007/s11222-016-9706-6. URL `http://dx.doi.org/10.1007/s11222-016-9706-6`.

Helder P Palaro and Luiz Koodi Hotta. Using conditional copula to estimate value at risk. *Journal of Data Science*, 4:93–115, 2006.

Emanuel Parzen. On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3):1065–1076, 1962.

Magda Peligrad. On the central limit theorem for weakly dependent sequences with a decomposed strong mixing coefficient. *Stochastic processes and their applications*, 42 (2):181–193, 1992.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. URL `https://www.R-project.org/`.

Murray Rosenblatt. Conditional probability density and regression estimators. *Multivariate analysis II*, 25:31, 1969.

Murray Rosenblatt et al. Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 27(3):832–837, 1956.

David Ruppert and Daren BH Cline. Bias reduction in kernel density estimation by smoothed empirical transformations. *The Annals of Statistics*, 22(1):185–210, 1994.

Mark J Schervish. *Theory of statistics*. Springer Science & Business Media, 1995.

Thomas A. Severini. *Likelihood Methods in Statistics*. Oxford science publications. Oxford University Press, 2000. ISBN 9780198506508.

Simon J Sheather and Michael C Jones. A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 683–690, 1991.

Bernard W Silverman. Density estimation for statistics and data analysis. *Monographs on Statistics and Applied Probability*, 26, 1986.

Charles J Stone. Large-sample inference for log-spline models. *The Annals of Statistics*, pages 717–741, 1990.

Charles J Stone, Mark H Hansen, Charles Kooperberg, Young K Truong, et al. Polynomial splines and their tensor products in extended linear modeling: 1994 Wald Memorial Lecture. *The Annals of Statistics*, 25(4):1371–1470, 1997.

Dag Tjøstheim and Karl Ove Hufthammer. Local gaussian correlation: a new measure of dependence. *Journal of Econometrics*, 172(1):33–48, 2013.

Matt P. Wand, James Stephen Marron, and David Ruppert. Transformations in density estimation. *Journal of the American Statistical Association*, 86(414):343–353, 1991.

MP Wand and MC Jones. Multivariate plug-in bandwidth selection. *Computational Statistics*, 9(2):97–116, 1994.