Multivariate and conditional density estimation using local Gaussian approximations

Håkon Otneim

Dissertation for the degree of Philosophiae Doctor (PhD)

Department of Mathematics
University of Bergen

June 2016

Dissertation date: 23.09.2016

Preface

The present work has been carried out during my employment as PhD-student at the Department of Mathematics at the University of Bergen, Norway, lasting from August 2012 until November 2016. I spent the month of August 2015 on a research stay at Monash University, Melbourne, Australia, which was partially funded by the Meltzer Research Fund.

This thesis consists of two parts. In the first part, we will briefly introduce the problems of multivariate and conditional density estimation, motivate the need for new methods, and sketch the solutions that will be advocated in the second part.

The second part consists of three papers:

Paper 1 Håkon Otneim, Hans Arnfinn Karlsen, and Dag Tjøstheim. "Bias and bandwidth for local likelihood density estimation." *Statistics & Probability Letters 83.5 (2013): 1382-1387.*

Paper 2 Håkon Otneim and Dag Tjøstheim. "The locally Gaussian density estimator for multivariate data." Submitted for publication to *Statistics & Computing.*

Paper 3 Håkon Otneim and Dag Tjøstheim. "Non-parametric estimation of conditional densities: A new method."

Acknowledgements

First and foremost, I would like to thank my supervisor, Professor Dag Tjøstheim, for his great help, detailed feedback, contagious enthusiasm, and apparently endless patience during my work on this project. The fact that he gave the first lecture I ever attended at the University on calculating the mean and median of numbers almost ten years ago, and that we have spent the last few weeks fine tuning the asymptotic theory of our estimators, gives a telling picture of how much statistics he has taught me over the years. Our meetings have always been pleasant, and I have never been afraid to discuss freely in his presence, or even to ask stupid questions. I could not ask for a better mentor. Thank you.

I also want to thank my two co-supervisors, Associate Professors Hans Karlsen and Bård Støve. They have provided valuable feedback at key points during my time as PhD-student, and they have been great colleagues during the daily life at the department, and fun travel companions to various conferences throughout the world. I will also take this opportunity to thank all my colleagues in the statistics group at the Department of Mathematics, University of Bergen, past and present. You have created a fantastic and stimulating environment that, I can honestly say, has been absolutely crucial to my day-to-day well-being at work.

I am grateful to Professor Jiti Gao for hosting me at Monash University, Melbourne, Australia, as well as all the other people at the Department of Econometrics and Business Statistics at Monash who showed interest in my project and provided valuable input, especially Rob Hyndman and Anastasios Panagiotelis.

My trip to Monash was partially funded by the Meltzer Research Fund, for which I am very grateful.

The administrative staff at the Department of Mathematics are silent heroes in all our scientific efforts and progression, be it in providing equipment, handling the money, helping with forms and formalities, and, to be frank, tolerating crazy scientists on a daily basis. That is impressive!

A big thanks goes out to the open source software community, especially the master minds behind R, including all its maintainers and contributors, who slave along for no other compensation than the progress and democratization of statistics and data science.

Finally, I would like to thank my wonderful wife and companion in life, Karina, and our kids Kristian and Marie, who, without complaints, have endured my frustrations, workload and emotional ups and downs during the past four years.

Bergen, June 2016
Håkon Otneim

Abstracts

Paper 1 "Bias and bandwidth for local likelihood density estimation"

A local likelihood density estimator is shown to have asymptotic bias depending on the dimension of the local parameterization. Comparing with kernel estimation it is demonstrated using a variety of bandwidths that we may obtain as good, and potentially even better estimates using local likelihood. Boundary effects are also examined.

Paper 2 "The locally Gaussian density estimator for multivariate data"

It is well known that the Curse of Dimensionality causes the standard Kernel Density Estimator to break down quickly as the number of variables increases. In non-parametric regression, this effect is relieved in various ways, for example by assuming additivity or some other simplifying structure on the interaction between variables. This paper presents the Locally Gaussian Density Estimator (LGDE), which introduces a similar idea to the problem of density estimation.

The LGDE is a new method for the non-parametric estimation of multivariate probability density functions. It is based on preliminary transformations of the marginal observation vectors towards standard normality, and a simplified local likelihood fit of the resulting distribution with standard normal marginals. The LGDE is introduced, and asymptotic theory is derived. In particular, it is shown that the LGDE converges at a speed that does not depend on the dimension. Examples using real and simulated data confirm that the new estimator performs very well on finite sample sizes.

Paper 3 "Non-parametric estimation of conditional density functions: A new method"

Let $\mathbf{X} = (X_1, \ldots, X_p)$ be a stochastic vector having joint density function $f_{\mathbf{X}}(\mathbf{x})$ with partitions $\mathbf{X}_1 = (X_1, \ldots, X_k)$ and $\mathbf{X}_2 = (X_{k+1}, \ldots, X_p)$. A new method for estimating the conditional density function of $\mathbf{X}_1$ given $\mathbf{X}_2$ 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 estmator 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.

Contents

Preface i

Acknowledgements iii

Abstracts v

1 Introduction 1
- 1.1 The probability density function and its estimation 1
- 1.2 The curse of dimensionality 3
- 1.3 Is there a way around? 5
- 1.4 The conditional density function 7
- 1.5 Summary of papers 9
  - 1.5.1 Summary of Paper 1: "Bias and bandwidth for local likelihood density estimation" 9
  - 1.5.2 Summary of Paper 2: "The locally Gaussian density estimator for multivariate data" 10
  - 1.5.3 Summary of Paper 3: "Non-parametric estimation of conditional densities: A new method" 11

2 Computer code 13

3 Papers 19
- 3.1 Bias and bandwidth for local likelihood density estimation 21
- 3.2 The locally Gaussian density estimator for multivariate data 29
- 3.3 Non-parametric estimation of conditional density functions: A new method 63

Chapter 1

Introduction

1.1 The probability density function and its estimation

Let X be a stochastic variable, and denote by $F(x) = P(X \leq x)$ its probability distribution function. If $F(\cdot)$ is differentiable, the density function $f(\cdot)$ of $F(\cdot)$ is given by

$$f(x) = \frac{d}{dx}F(x). \tag{1.1}$$

Constructing estimates of the density function based on observed values of X is one of the fundamental tasks is statistics. Not only does a good density estimate provide an easily interpretable visualization of the behaviour of X — its realizations tend to fall in the higher density regions more often than the lower density regions — it is also an instrument which we may use to quantify further properties of X, such as moments, quantiles and probabilities:

$$\mathrm{E}(X) = \int x f(x)\,\mathrm{d}x, \qquad \mathrm{Var}(X) = \int \left(x - \mathrm{E}(x)\right)^2 f(x)\,\mathrm{d}x,$$

$$P(X \in A) = \int_A f(x)\,\mathrm{d}x, \qquad q_\alpha(X) = F^{-1}(\alpha), \text{ where } F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t.$$

Even if the theoretical density function does not exist in the strict mathematical sense (1.1), it is often useful to calculate a density estimate anyway, either for the purpose of data exploration, or as an intermediate step to more complex analyses. After all, real data is discrete by nature since in practice no continuous variable can be recorded with infinite accuracy. If it can, it is almost certainly not continuous.

The twentieth century saw the rise of three overall methods for estimating probability models in general, and probability density functions in particular. The first is the classical parametric approach, which is closely connected to the emergence of the maximum likelihood theory that was formalized by R.A. Fisher and others in the 1920s and 1930s (See Stigler (2007) for a comprehensive historical overview). It is generally easy to fit a parametric model once it has been specified, and the theoretical foundation upon which the classical parametric statistics rests is sound, solid and mature, and forms the backbone of most introductory courses in mathematical statistics to this day. The parameter estimates determine the full model estimate, and they can often be analysed and interpreted in their own right, as location-, scale- or rate-parameters in a density

function, or as a regression coefficient governing the influence of a specific explanatory variable in multiple linear regression.

It is not always straightforward to specify a parametric model, however, and even worse; the practitioner will sometimes impose a parametric structure for the data-set that is plain wrong. Whether the mistake stems from incompetence, old habits, or an error in judgement, important features of the data may be missed, interpretation of the parameter estimates will be questionable, and the ultimate decisions that are made based on the analysis, will at best be sub-optimal, and in some cases cause serious damage. The misuse of the parametric Gaussian copula has been blamed, rightly or not, to be the cause of the 2008 financial crisis (Jones, 2009).

We avoid the problems caused by misspecified models in the non-parametric paradigm. Instead of imposing a pre-specified structure, we estimate the probability model based solely on the data. The default method for non-parametric density estimation is the kernel estimator, which was introduced independently by Rosenblatt (1956) and Parzen (1962). Suppose that the random variables $X_1, \ldots, X_n$ are independent and identically distributed with density function $f(x)$. The kernel estimate of f based on observations $X_1, \ldots, X_n$ is given by

$$\widehat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right),$$

where $K(\cdot)$ is a symmetric probability density function, and h is the bandwidth, or window size, that determines the region of influence for each observation, the choice of which is crucial to the performance of the estimator. The standard text on the kernel density estimator is Silverman (1986), and the later years have seen advancements such as transformation techniques (Marron and Ruppert, 1994), adaptive bandwidths (Terrell and Scott, 1992), and higher- and infinite order kernels (Jones and Signorini, 1997; Politis and Romano, 1999), to name a few central references.

The kernel estimator will always converge towards the true density function under some regularity conditions, so one might ask why the maximum likelihood estimation of parameters is still in use today, considering the risk of misspecification. It all comes down to information. How much information does the data carry, and how much information do we have before the data is even collected? In the parametric case, we *assume* a certain structure, or family of models, based on experience perhaps, and the validity of this assumption can, and should, be formally tested. Given the parametric assumption, we can use all the information contained in the observations to estimate the relatively small amount of remaining unknowns; the parameters.

In the strictly nonparametric case, the information contained in the data must be used to estimate the *entire* model. Informally speaking, one might say that we have less information per unit unknown in the latter case. In statistical terms, this means that the variance of non-parametric methods is generally higher than for parametric methods. In other words, we must pay the price of variance in order to let the data speak for itself, and thereby reduce the misspecification.

The rapid emergence of powerful computing tools over the last few decades has fundamentally changed the field of statistics. This is of course also true in the special case of density estimation. The framework of semi-parametric statistical methods embraces the opportunities presented by the ability to calculate, evaluate, and optimize quickly

and efficiently, in ways that we could scarcely dream about just thirty years ago. The spirit of semi-parametric estimation is to avoid the "all-or-nothing"-situation with non- or fully parametric methods, but rather compromise, and strike a balance between their properties that is optimal in any given situation.

For example, we can use the logspline density estimator by Kooperberg and Stone (1991) to fit a cubic spline to the logarithm of the unknown density function, with the number and location of nodes being chosen automatically based on the data. If there are many observations, one may allow the number of nodes to grow large, increasing the flexibility of the method if there is enough information present to support it. We will use this method actively in this thesis as a tool on the way to a semi-parametric estimator for multivariate density functions. Other possibilities include the explicit mix of parametric and non-parametric estimates in the method by Hjort and Glad (1995), and the local polynomials by Fan and Gijbels (1996), which, although mainly a regression technique, can be formulated as a density estimator as well.

Finally, we mention the *local likelihood* by Hjort and Jones (1996) and Loader (1996), which works by fitting a parametric family locally to the unknown density. This concept will play a central role in this thesis, and will therefore be explained in detail later on.

1.2 The curse of dimensionality

This thesis concerns the estimation of multivariate probability density functions. Denote by $\boldsymbol{X} = (X_1, \ldots, X_p)$ a p-variate stochastic vector, and by $F(\boldsymbol{x}) = F(x_1, \ldots, x_p) = P(X_1 \leq x_1, \ldots, X_p \leq x_p)$ its distribution function. The multivariate density function, if it exists, is defined in the same manner as in (1.1):

$$f(\boldsymbol{x}) = \frac{\partial}{\partial x_1 \cdots \partial x_p} F(\boldsymbol{x}).$$

Most of the discussion in the preceding section holds true also in the multivariate case. Many parametric families of univariate density functions have multivariate generalizations, including of course the Gaussian distribution. Further, the famous Sklar's (1959) Theorem states that every continuous distribution function $F(x_1, \ldots, x_p)$, with $p \geq 2$, has a unique copula function C, such that

$$F(x_1, \ldots, x_p) = C(F_1(x_1), \ldots, F_p(x_p)),$$

where $F_i(\cdot)$, $i = 1, \ldots, p$ are the marginal distribution functions for the variables in $\boldsymbol{X}$. This means that we can estimate the marginal distributions of the variables separately from the dependence between them, which is governed by the copula function. There exists several parametric families of copulas, on which there is a rich literature. See for example Nelsen (2007) for an introduction to the topic.

The non-parametric kernel estimator also has a natural generalization to the multivariate case; let $K(\boldsymbol{x})$ be a p-variate density function that is radially symmetric about zero, and let $\boldsymbol{H}$ be a positive definite matrix of smoothing parameters. Using a random sample $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$, we estimate the unknown density $f(\boldsymbol{x})$ by

$$\widehat{f}(\boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\det(\boldsymbol{H})} K(\boldsymbol{H}^{-1}(\boldsymbol{x} - \boldsymbol{X}_i)),$$

which is usually simplified by restricting the smoothing matrix $\boldsymbol{H}$ to be diagonal.

So far, all seems well and good, but it turns out that both parametric and non-parametric density estimation methods have serious problems when implemented in the multivariate case. The reason is that we do not only have to construct an estimate that is accurate for the individual (marginal) behaviour of each variable, it must also capture the complete dependence structure *between* all the variables, which tends to grow much harder for each additional variable. Again, it all comes down to the amount of information required to construct an estimate with the desired accuracy, compared to the amount of information that is contained in the data. It is reasonable that we must require *more* data in the multivariate case, precisely to account for the extra information needed in order to estimate the dependence structure of the unknown distribution in addition to the marginal density functions. But how much more? Silverman (1986) gives an answer in the case of the non-parametric kernel estimator: According to one measure, we need 842 000 observations in order for the ten-dimensional kernel estimator to achieve the same accuracy as would only four observations in the one-dimensional case!

In the parametric framework, the problem is perhaps not so much the lack of data as the lack of parametric models to try. Whether we fit a full distribution, like the multivariate normal or the multivariate t-distribution, or just the copula-function, of which a moderate number of parametric families exists, we must essentially be able to summarize a possibly complex dependence structure between multiple stochastic variables by just a few parameters; and even if a given model does not strictly fail a goodness-of-fit test, the data may, perhaps more often than we care to admit, be far too sparse to make any sound inference about the validity of a given parametric family of densities.

The curse of dimensionality takes many forms, but in the specific problem of density estimation it means that the number of observations needed to keep the precision of a non-parametric density estimate grows sharply as the number of variables increases, or, equivalently, the precision of a non-parametric density estimate based on a fixed number of observations, decays sharply with dimension.

We can express the considerations above precisely in mathematical terms. Let $f(\cdot)$ be the density function of the d-dimensional stochastic vector $\boldsymbol{X}$, and denote by $\widehat{f}_P(\cdot)$ the parametric maximum likelihood estimate of f based on some parametric assumption, and by $\widehat{f}_{NP}(\cdot)$ a non-parametric kernel estimate. Both estimates are asymptotically normal under some regularity conditions, with

$$\sqrt{n}\left(\widehat{f}_P(\boldsymbol{x}) - f^*(\boldsymbol{x})\right) \xrightarrow{\mathcal{L}} N(0,\sigma_1^2), \qquad \text{and} \qquad \sqrt{nh^d}\left(\widehat{f}_{NP}(\boldsymbol{x}) - f(\boldsymbol{x})\right) \xrightarrow{\mathcal{L}} N(0,\sigma_2^2), \tag{1.2}$$

where f^* is the best approximant (in some sense) to f within the chosen parametric family, and $h \to 0$ is the smoothing parameter for the kernel estimator. The particular expressions of σ_1^2 and σ_2^2 depend on the situation at hand, but are not interesting in this context. We make two observations from these results. First, we see that the dimensionality of the problem has no effect on the *speed* of convergence of the maximum likelihood estimate. The asymptotic variance is not affected at all. The misspecification error, however, will not go away unless we happen to be working within the correct parametric family, which we have already argued becomes increasingly less likely in several dimensions. The non-parametric estimate, on the other hand, converges to the true density, but the rate of convergence slows down as $\sqrt{h^d}$, which means that adding just a few more variables to the problem, may increase the variance by several orders of

magnitude.

1.3 Is there a way around?

The expressions (1.2) demonstrate that choosing between a parametric and a non-parametric density estimate not only constitutes the choice between two extreme end-points on a spectrum, but also that the distance between the two possibilities becomes bigger in the multivariate case. We clearly need a semi-parametric compromise more than ever, but do existing methods automatically scale up to serve the purpose? Not quite, as we proceed to motivate briefly.

One of the main contributions of this thesis is to provide a version of the locally parametric estimator by Hjort and Jones (1996) that is especially designed to work in the multivariate case. Consider first the univariate problem, and assume that we wish to estimate the unknown density $f(x)$ based on the realized values of n independent and identically distributed variables $X_1, \ldots, X_n$, each having density function f. The Hjort and Jones (1996)-method requires us to choose a parametric family $\psi(\cdot; \boldsymbol{\theta})$ for $f(\cdot)$, but then provides a *locally parametric* estimate of f by fitting the parameter vector $\boldsymbol{\theta}$ locally. In each point x in the support of f, estimate the parameter vector by maximizing the *local log likelihood function*

$$L(\boldsymbol{\theta}; X_1, \ldots, X_n) = (nh)^{-1} \sum_{i=1}^{n} K(h^{-1}(X_i - x)) \log \psi(X_i; \boldsymbol{\theta}) - \int h^{-1} K(h^{-1}(x - y)) \psi(y; \boldsymbol{\theta}) \, \mathrm{d}y, \tag{1.3}$$

where K and h still denote the kernel function smoothing parameter correspondingly. This results in an estimated parameter vector $\widehat{\boldsymbol{\theta}}(\boldsymbol{x})$, which, when substituted back into the parametric family ψ, produces the local likelihood density estimate:

$$\widehat{f}_{LL}(x) = \psi(x, \widehat{\boldsymbol{\theta}}(\boldsymbol{x})).$$

In Figure 1.1 we see a simple illustration of the local likelihood procedure in a univariate example. In the left panel, we have plotted the Gamma(2,1)-density as a dashed line. Using the local likelihood function (1.3) to fit the Gaussian distribution locally based on 300 independent observations, results in two estimated parameters functions, $\widehat{\mu}(x)$ and $\widehat{\sigma}(x)$, that are plotted in the right hand panel. The solid line in the density plot displays the estimated density function, which is just the univariate Gaussian distribution, with these estimated parameter functions in place of the corresponding parameters, μ and σ.

The setup by Hjort and Jones (1996) provides a good compromise between fully parametric, and fully non-parametric methods. The smoothing parameter h controls not only the smoothness of the density estimate, but also the degree to which we trust the local parametric family to be close to the truth. If the bandwidth grows to infinity, the local likelihood function (1.3) will produce constant parameter estimates, which corresponds to a standard, global, maximum likelihood estimate of the density. On the other hand, if we let the bandwidth go to zero at a certain rate as the sample size increases, we have a non-parametric density estimator with theoretical properties

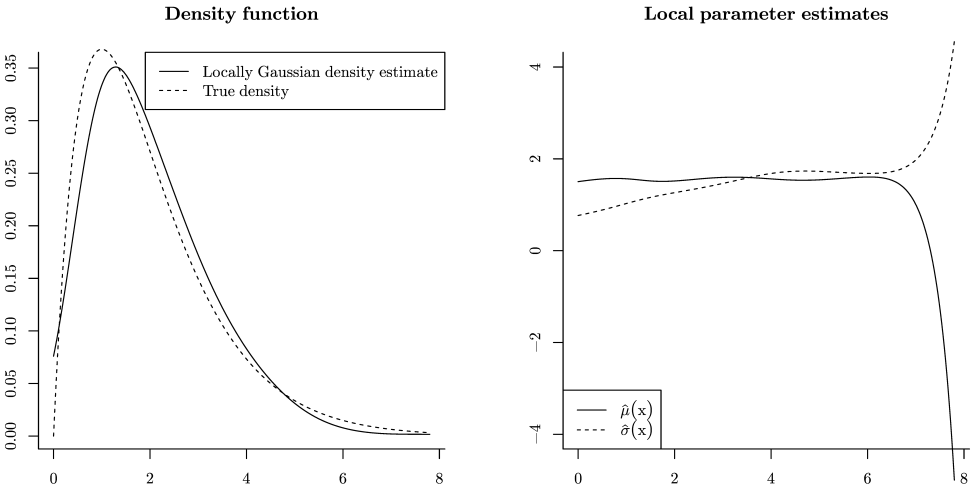


Figure 1.1: A univariate locally Gaussian density estimate of the Gamma(2,1)-density, based on 300 observations

comparable to the kernel estimator (Hjort and Jones, 1996). In practical situations, with a fixed and finite sample size, we will usually find ourselves somewhere in between these two extremes, and seek a bandwidth that is optimal in that particular case.

Consider again the multivariate case where $\boldsymbol{X}$ is a p-variate stochastic vector with density function $f(\boldsymbol{x})$. The general local likelihood approach is not well suited to tackle the curse of dimensionality on its own, however, and the reason is simple: Instead of estimating the unknown, p-variate density function directly, we must rather estimate several completely unknown, p-variate, parameter functions. If we, for example, were to fit the three-variate Gaussian distribution to some data using this method directly, we would have to produce nine estimated p-variate functions, one for each parameter in the parametric family. It seems as if we have achieved very little, except for a heavy computational burden.

The core contribution of this thesis is a simplified version of the Hjort and Jones (1996)-strategy that is especially designed to perform well in the multivariate case, have simple theoretical properties, and be easily implemented. We introduce the *Locally Gaussian Density Estimator* (LGDE), and the general idea is to proceed according to the following algorithm:

1. Transform the observations to approximate marginal standard normality using the logspline density estimator by Stone et al. (1997).

2. To *each pair* (Z_i, Z_j) of transformed variables, fit the standardized bivariate Gaussian distribution

$$\psi(z_i, z_j; \rho_{ij}) = \frac{1}{2\pi\sqrt{1-\rho_{ij}^2}} \exp\left(-\frac{1}{2(1-\rho_{ij}^2)}[z_i^2 - 2\rho_{ij} z_i z_j + z_j^2]\right) \tag{1.4}$$

   locally, using the local likelihood function (1.3).

3. Collect all the pairwise estimated local correlations in one $p \times p$ local correlation matrix $\widehat{\boldsymbol{R}}(\boldsymbol{z}) = \{\widehat{\rho}_{ij}(z_i, z_j)\}$, which is then used in the standardized p-variate Gaussian density function in order to produce a density estimate on the marginally Gaussian scale:

$$\widehat{f}(\boldsymbol{z}) = \frac{1}{(2\pi)^{p/2}|\widehat{\boldsymbol{R}}(\boldsymbol{z})|^{1/2}} \exp\left(-\frac{1}{2}\boldsymbol{z}^T \widehat{\boldsymbol{R}}(\boldsymbol{z})^{-1}\boldsymbol{z}\right). \tag{1.5}$$

4. Transform back to the original scale.

5. If necessary, normalize the density estimate so that it integrates to one.

The details of these points will of course be presented in due time, when the necessary notation has been established. A central result deserves reproduction in this preliminary discussion, however: Denote by $\widehat{f}_{LGDE}(\boldsymbol{x})$ the resulting density estimate from the procedure described above. We will see that, under some regularity conditions,

$$\sqrt{nh^2}\left(\widehat{f}_{LGDE}(\boldsymbol{x}) - f_0(\boldsymbol{x})\right) \xrightarrow{\mathcal{L}} N(0, \sigma_3^2),$$

which, we will argue, represents a true middle ground between the convergence results in (1.2). First of all, our new estimator does not converge to the true and unknown density function f, but rather to its best approximant (in some sense) within the class of densities that admits the pairwise dependence structure as described in the algorithm above. It turns out that this class of functions is very flexible, and provides a good approximation in many cases. On the other hand, the convergence rate is $\sqrt{nh^2}$, no matter what the dimensionality of the problem is, which is not surprising at all, precisely because of the assumed pairwise dependence structure. Simply stated: we are able to reduce the variance significantly by imposing a restriction on the dependence structure of the stochastic vector, but pay by introducing some misspecification error, but, as we will see, that is a very reasonable price in many cases.

1.4 The conditional density function

Obtaining the estimate of a high-dimensional joint density function, however fundamental it may be, is of somewhat limited practical use. The *conditional* density function, on the other hand, is extremely useful in the formulation of a large, and diverse, set of statistical methods, including regression analysis, dependence modelling, time series, and the construction of Bayesian networks. Non-parametric estimates of the conditional density can be useful in all stages of the analysis.

Partition the stochastic vector $\boldsymbol{X} = (X_1, \ldots, X_p)$ into two sub-vectors $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, such that $\boldsymbol{X}_1 = (X_1, \ldots, X_k)$ and $\boldsymbol{X}_2 = (X_{k+1}, \ldots, X_p)$. The Conditional density of $\boldsymbol{X}_1|\boldsymbol{X}_2 = \boldsymbol{x}_2$ is given by

$$f_{\boldsymbol{X}_1|\boldsymbol{X}_2=\boldsymbol{x}_2}(\boldsymbol{x}_1|\boldsymbol{X}_2 = \boldsymbol{x}_2) = \frac{f_{\boldsymbol{X}}(\boldsymbol{x}_1, \boldsymbol{x}_2)}{f_{\boldsymbol{X}_2}(\boldsymbol{x}_2)} \tag{1.6}$$

We will drop the subscripts on the density functions for easier notation.

Rosenblatt (1969) made the first systematic attempt at constructing a non-parametric estimate of (1.6) by simply estimating the numerator using the kernel estimator, and

putting the marginal density of this estimate in the denominator. If we use the Gaussian kernel function, this simplifies to

$$f_{\boldsymbol{X}_1|\boldsymbol{X}_2=\boldsymbol{x}_2}(\boldsymbol{x}_1|\boldsymbol{X}_2=\boldsymbol{x}_2) = \frac{\widehat{f}_{\boldsymbol{H}}(\boldsymbol{x}_1,\boldsymbol{x}_2)}{\int \widehat{f}_{\boldsymbol{H}}(\boldsymbol{x}_1,\boldsymbol{x}_2)\,\mathrm{d}\boldsymbol{x}_1} = \frac{\widehat{f}_{\boldsymbol{H}}(\boldsymbol{x}_1,\boldsymbol{x}_2)}{\widehat{f}_{\boldsymbol{H}^*}(\boldsymbol{x}_2)}, \tag{1.7}$$

where $\boldsymbol{H}^*$ is the lower right block of $\boldsymbol{H}^*$ corresponding to the variables in $\boldsymbol{X}_2$. It is not necessarily true, however, that the bandwidths that are optimal for estimating the joint density $f(\boldsymbol{x})$, are optimal for estimating a functional of f, such as its derivatives, quantiles, or, in this case, the conditional density of a subset of the variables. This has triggered some attempts to improve the basic estimator (1.7), which include a collection of bandwidth selectors by Bashtannyk and Hyndman (2001), a fast bandwidth selection algorithm by Holmes et al. (2012), and the work by Li and Racine (2007), who provide the practitioner with a conditional kernel density estimator that works in the general multivariate situation, with a mix of continuous and discrete variables. The latter work is also implemented in the R programming language (R Core Team, 2015) through the `np`-package (Hayfield and Racine, 2008), which makes it very appealing to the general practitioner. Faugeras (2009) acknowledges the problematic aspect of putting a possibly low-valued and high-variance kernel density estimate in the denominator of (1.7), and reformulates the problem using the copula density, with promising results.

Hyndman et al. (1996) starts a line of non-parametric conditional density estimators that move away from the kernel estimator, and towards the semi-parametric framework. The authors adjust the kernel estimator, so that it has mean equal to some better performing non-parametric regression of the conditional mean than a standard kernel smoother. Fan et al. (1996) formulate the problem as a locally linear or locally quadratic least squares fit, while Hyndman and Yao (2002) restrict this method so that it is always non-negative, and also introduce locally linear or locally quadratic models that are fitted using local likelihood. This work has been implemented in R through the `hdrcde`-package (Hyndman et al., 2013). Finally, Fan and Yim (2004) provide a new method for selecting bandwidths for the local polynomial fits above, using cross-validation.

The second main contribution of this thesis is to develop a new conditional density estimator, and it turns out the locally Gaussian multivariate density estimator that was sketched in the preceding section serves as a natural starting point for this purpose. It is well known that a subset of the variables in a multivariate normal distribution is again jointly normally distributed, and further, that the conditional density constructed from response- and explanatory variables that are all jointly normal, is also normally distributed. Specifically, if $\boldsymbol{Z} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{Z}$ is partitioned into the subsets $(\boldsymbol{Z}_1, \boldsymbol{Z}_2)$, and the mean vector and covariance matrix is partitioned correspondingly;

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix},$$

then $\boldsymbol{Z}_1|\boldsymbol{Z}_2 = \boldsymbol{z}_2$ is jointly normal with expectation vector $\boldsymbol{\mu}^*$ and covariance matrix $\boldsymbol{\Sigma}^*$, where (Johnson and Wichern, 2007, Chap. 4)

$$\boldsymbol{\mu}^* = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2), \tag{1.8}$$

$$\boldsymbol{\Sigma}^* = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}. \tag{1.9}$$

This result follows from the fact that the variables in $\boldsymbol{Z}_2$ are jointly normal, and direct manipulation of the fraction (1.6), and points to a natural extension of the LGDE, so that it can also be used for the estimation of conditional density functions.

Consider separate locally Gaussian estimates of the numerator and denominator of (1.6), produced using the algorithm in the preceding section. Both estimates will be on the form (1.5), up to the final back-transformation to the original scale. The key observation to make here is that, in each point $\boldsymbol{z}$, the local correlation matrix of the marginal density estimate in the denominator, $\widehat{\boldsymbol{R}}_{22}(\boldsymbol{z})$ say, is exactly the lower right block of the local correlation matrix $\widehat{\boldsymbol{R}}(\boldsymbol{z})$ in the numerator, which is a consequence of our pairwise estimation procedure. It follows that we can use the results (1.8) and (1.9) directly to express the fraction of locally Gaussian density estimates as *one* locally Gaussian density estimate, with local correlation and local covariance matrix given by

$$\begin{aligned}\widehat{\boldsymbol{\mu}}^*(\boldsymbol{z}) &= \widehat{\boldsymbol{R}}_{12}(\boldsymbol{z})\widehat{\boldsymbol{R}}_{22}^{-1}(\boldsymbol{z})\boldsymbol{z}_2,\\ \widehat{\boldsymbol{\Sigma}}^*(\boldsymbol{z}) &= \widehat{\boldsymbol{R}}_{11}(\boldsymbol{z}) - \widehat{\boldsymbol{R}}_{12}(\boldsymbol{z})\widehat{\boldsymbol{R}}_{22}^{-1}(\boldsymbol{z})\widehat{\boldsymbol{R}}_{21}(\boldsymbol{z}),\end{aligned}$$

where the indices mean the block-wise decomposition of $\widehat{\boldsymbol{R}}$, following the same pattern as in (1.8) and (1.9), and the expectations and variances are 0 and 1 respectively, as in the pairwise normal distributions (1.4).

With minimal effort, we can therefore present a semi-parametric conditional density estimator that allows several response variables as well as several explanatory variables. Its finite sample performance turns out to be very good, and, in most cases examined by us, superior to the few competitors that have publicly available computer implementations. The practical contribution of this work is the code for the implementation of the new estimator, and the theoretical contribution consists of asymptotic results that are parallel to those derived for the *un*conditional version of the LGDE, but proven under a strong mixing condition, allowing dependent observations.

1.5 Summary of papers

1.5.1 Summary of Paper 1: "Bias and bandwidth for local likelihood density estimation"

Otneim, Håkon, Hans Arnfinn Karlsen, and Dag Tjøstheim. "Bias and bandwidth for local likelihood density estimation." Statistics & Probability Letters *83.5 (2013): 1382-1387.*

The article *Locally parametric nonparametric density estimation* by Hjort and Jones (1996) is the main reference for this thesis. In it, the authors provide a complete framework for using local likelihood to produce non-parametric, but locally parametric, probability density estimates. Their work appear in the same issue of the *Annals of Statistics* as *Local likelihood density estimation* by Loader (1996), who tackle the same problem, using the same tool, but Loader (1996) is more in line with the contemporary regression literature, in modelling the log-density function as a local polynomial. Hjort and Jones (1996), on the other hand, provide the framework for fitting a parametric family of densities locally.

In Paper 1 *Bias and bandwidth for local likelihood density estimation*, we lay some ground work. By fitting the normal distribution locally using the machinery by Hjort and Jones (1996), we are able to demonstrate that it is a very appealing estimator for many univariate density functions, compared to the traditional kernel estimator. In particular, the locally Gaussian estimator appears to be much more robust against oversmoothing than the kernel estimator.

We also investigate some theoretical issues of the locally parametric estimator. It has been known for a while (cf. Tjøstheim and Hufthammer (2013)), that special care must be taken when deriving the asymptotic properties of a locally parametric density estimator with more than one parameter. In the two-parameter case, for example the frequently used univariate normal distribution, it turns out that the covariance matrix of the local parameters converges as $(nh^3)^{-1}$, instead of the usual non-parametric rate of $(nh)^{-1}$, where h is the smoothing parameter. We show in Paper 1 that the density estimate has an additional bias term of the same order.

Further, we investigate the issue of estimating density functions with bounded support, which is known to cause bias problems when using the kernel estimator. Hjort and Jones (1996) show that, if the chosen local parametric family has the same support as the unknown density function, there will be no boundary issues. We show in Paper 1 that if that is not the case (for example if we fit the normal distribution locally to a density with bounded support), we can use the bias corrections that Jones (1993) propose for the kernel estimator, for local likelihood estimates also.

1.5.2 Summary of Paper 2: "The locally Gaussian density estimator for multivariate data"

In this paper, we give a detailed account of the locally Gaussian density estimator, that was sketched in Section 1.3. It is a well known problem that the non-parametric kernel estimator, even though easily defined and calculated in higher dimensions, does not work very well in the multivariate case. The explanation is simple: The estimation of a density $f(\boldsymbol{x})$ requires the number of observations to grow enormously fast as the number of variables increases in order to keep a fair amount of accuracy, and that is usually not practically possible.

Turning directly to local likelihood will not help us either, because that would entail the estimation of just another multivariate function, the parameter $\theta(\boldsymbol{x})$, which, to complicate matters even more, may itself be a vector of several components. We try to simplify by describing the p-variate density function with a set of bivariate parameter functions, by means of a simplified version of a local multivariate Gaussian fit, as described in Section 1.3.

Simplification of reality is absolutely necessary when estimating a multivariate density function non-parametrically. We show through simulations that our estimator can be trusted to provide very good results in a variety of situations, and still be robust, so that it does not perform particularly worse than its competitors in difficult situations.

We prove large-sample results for the LGDE by mostly turning to existing theory, but intermixed with some recent theory of copula estimation.

1.5.3 Summary of Paper 3: "Non-parametric estimation of conditional densities: A new method"

As indicated in Section 1.4, we can use the local correlations that we estimate along the way in order to produce the LGDE, to rather estimate a conditional density function. This, we believe, can be of great practical use, because the estimator performs well, and can be applied to a wide variety of problems, as we set out to demonstrate in Paper 3: *Non-parametric estimation of conditional densities: A new method*.

The disposition of Paper 3 is of course very similar to that of Paper 2, with key parts being the asymptotic theory and the presentation of practical examples. While much of the theoretical considerations can be transferred directly from the unconditional to the conditional case, we prove the large-sample properties of the conditional density estimator under a new set of conditions. An important difference is that we replace the requirement of independent samples with a strong mixing condition.

Chapter 2

Computer code

All computer programs used in this thesis have been written in the R-programming language (R Core Team, 2015). Being a self-taught programmer, I do not claim that the code fulfills good programming standards to any particular high degree, but the functions nevertheless do their job well enough.

I have compiled the key elements of my code into an R-package, that will enable the reader to quickly put the methods described in the following pages into practice. The architecture of the larger simulation experiments is not suitable for general publication, however, but will of course be made available upon request to ensure easy reproducibility. The code can be installed into R by issuing the following commands (requires the excellent `devtools`-package (Wickham and Chang, 2015)):

```
library(devtools)
install_github("hotneim/lgde")
```

There are two main functions in this package:

`multiLocal()`	Estimates the multivariate density function of a data matrix using the method described in Paper 2. Each row represents on observation, and each column represents one variable. Optional arguments are `bandwidths`, which, if not supplied, will use cross-validation to calculate the smoothing parameters, and `grid`, which specifies the points where the estimate should be evaluated. The `grid`-argument must have the same number of columns as the data-matrix.
`condLocal()`	Estimates the conditional density using the method described in Paper 3. Takes data of the same format as for the `multiLocal`-function, but one must also supply a vector of conditions on the explanatory variables. The function always assumes that the response variables come first, so if `data` is an $n \times 4$-matrix with columns `X1`, `X2`, `X3` and `X4`, supplying `cond = c(1, 2)` means that we estimate the joint conditional density of $(X_1, X_2 \vert X_3 = 1, X_4 = 2)$. The `grid` must have the same number of columns as there are response variables, and the `bandwidths`-argument has the same meaning as above.

Although mainly invoked under the hood by the functions above, the following routines may come in handy as well:

`transLocal()`	Transforms a multivariate data set to approximate standard normal marginals, by estimating the marginals using the `logspline`-package (Kooperberg, 2016).
`HLocal()`	Calculates the cross-validation bandwidths as described in Paper 2. Assumes standard normal marginals, so for a general `data`-matrix, use `HLocal(transLocal(data)$transformed.data)`. This is by far the most time consuming element of the estimation process.
`pluginLocal()`	A quick and dirty plug-in approximation to the cross-validation bandwidth selection routine. Takes two arguments: `n` is the number of observations and `nvar` is the number of variables. The function returns a bandwidth object that can be used directly in the `multiLocal`- and `condLocal`-functions, which contains $1.75 \times n^{-1/6}$ in all of its elements.

The following R-packages have been used, either in the functions mentioned above, or along the way in simulation experiments, or the production of graphics and tables:

Package name	Author(s) (Year)	Purpose
`copula`	Hofert et al. (2015)	Simulation from and evaluation of copula models
`doMC`	Revolution Analytics (2014)	Parallel processing
`extrafont`	Chang (2014)	Fonts in graphics
`devtools`	Wickham and Chang (2015)	Make an R-package
`fastICA`	Marchini et al. (2013)	Projection pursuit
`gdata`	Warnes et al. (2015)	The `upperTriangle`-function
`ks`	Duong (2015)	Multivariate kernel density estimation
`MBESS`	Kelley and Lai (2012)	the `cor2cov`-function
`mvtnorm`	Genz et al. (2016)	Implementation of the multivariate normal distribution
`np`	Hayfield and Racine (2008)	Non-parametric conditional density estimation
`Rlab`	Boos and Nychka (2012)	Generate random Bernoulli variables
`sn`	Azzalini (2015)	The skew-normal and skew-t-distributions
`stringi`	Gagolewski and Tartanus (2015)	Handle strings
`stringr`	Wickham (2015)	Handle strings
`SuppDists`	Wheeler (2016)	Generate inverse Gaussian variables

Bibliography

Adelchi Azzalini. *The R package* `sn`*: The skew-normal and skew-t distributions (version 1.2-4)*. Università di Padova, Italia, 2015. URL `http://azzalini.stat.unipd.it/SN`.

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

Dennis D. Boos and Douglas Nychka. *Rlab: Functions and Datasets Required for ST370 class*, 2012. URL `http://CRAN.R-project.org/package=Rlab`. R package version 2.15.1.

Winston Chang. *extrafont: Tools for using fonts*, 2014. URL `http://CRAN.R-project.org/package=extrafont`. R package version 0.17.

Tarn Duong. *ks: Kernel Smoothing*, 2015. URL `http://CRAN.R-project.org/package=ks`. R package version 1.9.4.

Jianqing Fan and Irene Gijbels. *Local polynomial modelling and its applications: monographs on statistics and applied probability 66*, volume 66. CRC Press, 1996.

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.

Marek Gagolewski and Bartek Tartanus. *R package stringi: Character string processing facilities*, 2015. URL `http://stringi.rexamine.com/`.

Alan Genz, Frank Bretz, Tetsuhisa Miwa, Xuefei Mi, Friedrich Leisch, Fabian Scheipl, and Torsten Hothorn. *mvtnorm: Multivariate Normal and t Distributions*, 2016. URL `http://CRAN.R-project.org/package=mvtnorm`. R package version 1.0-5.

Tristen Hayfield and Jeffrey S Racine. Nonparametric econometrics: The np package. *Journal of Statistical Software*, 27(5), 2008. URL `http://www.jstatsoft.org/v27/i05/`.

Nils Lid Hjort and Ingrid K Glad. Nonparametric density estimation with a parametric start. *The Annals of Statistics*, pages 882–904, 1995.

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

Marius Hofert, Ivan Kojadinovic, Martin Maechler, and Jun Yan. *copula: Multivariate Dependence with Copulas*, 2015. URL `http://CRAN.R-project.org/package=copula`. R package version 0.99-13.

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.

Rob J Hyndman, Jochen Einbeck, and Matt Wand. *hdrcde: Highest density regions and conditional density estimation*, 2013. URL `http://CRAN.R-project.org/package=hdrcde`. R package version 3.1.

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

MC Jones. Simple boundary correction for kernel density estimation. *Statistics and Computing*, 3(3):135–146, 1993.

MC Jones and DF Signorini. A comparison of higher-order bias kernel density estimators. *Journal of the American Statistical Association*, 92(439):1063–1073, 1997.

Sam Jones. The formula that felled wall street. *The Financial Times*, 22(04), 2009.

Ken Kelley and Keke Lai. *MBESS: MBESS*, 2012. URL `http://CRAN.R-project.org/package=MBESS`. R package version 3.3.3.

Charles Kooperberg. *logspline: Logspline Density Estimation Routines*, 2016. URL `http://CRAN.R-project.org/package=logspline`. R package version 2.1.9.

Charles Kooperberg and Charles J Stone. A study of logspline density estimation. *Computational Statistics & Data Analysis*, 12(3):327–347, 1991.

Qi Li and Jeffrey S Racine. *Nonparametric econometrics: theory and practice*. Princeton University Press, 2007.

Clive R Loader. Local likelihood density estimation. *The Annals of Statistics*, 24(4): 1602–1618, 1996.

JL Marchini, C Heaton, and BD Ripley. *fastICA: FastICA Algorithms to perform ICA and Projection Pursuit*, 2013. URL `http://CRAN.R-project.org/package=fastICA`. R package version 1.2-0.

James Stephen Marron and David Ruppert. Transformations to reduce boundary bias in kernel density estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 653–671, 1994.

Roger B Nelsen. *An introduction to copulas*. Springer Science & Business Media, 2007.

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

Dimitris N Politis and Joseph P Romano. Multivariate density estimation with general flat-top kernels of infinite order. *Journal of Multivariate Analysis*, 68(1):1–25, 1999.

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/`.

Revolution Analytics. *doMC: Foreach parallel adaptor for the multicore package*, 2014. URL `http://CRAN.R-project.org/package=doMC`. R package version 1.3.3.

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

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

Bernard W Silverman. *Density estimation for statistics and data analysis*, volume 26. CRC press, 1986.

Abe Sklar. *Fonctions de répartition à n dimensions et leurs marges*. Université Paris 8, 1959.

Stephen Stigler. The epic story of maximum likelihood. *Statistical Science*, 22(4):598–620, 2007.

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.

George R Terrell and David W Scott. Variable kernel density estimation. *The Annals of Statistics*, pages 1236–1265, 1992.

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

Gregory R. Warnes, Ben Bolker, Gregor Gorjanc, Gabor Grothendieck, Ales Korosec, Thomas Lumley, Don MacQueen, Arni Magnusson, Jim Rogers, and others. *gdata: Various R Programming Tools for Data Manipulation*, 2015. URL `http://CRAN.R-project.org/package=gdata`. R package version 2.17.0.

Bob Wheeler. *SuppDists: Supplementary Distributions*, 2016. URL `http://CRAN.R-project.org/package=SuppDists`. R package version 1.1-9.2.

Hadley Wickham. *stringr: Simple, Consistent Wrappers for Common String Operations*, 2015. URL `http://CRAN.R-project.org/package=stringr`. R package version 1.0.0.

Hadley Wickham and Winston Chang. *devtools: Tools to Make Developing R Packages Easier*, 2015. URL `http://CRAN.R-project.org/package=devtools`. R package version 1.8.0.

Chapter 3

Papers

I

Paper I

3.1 Bias and bandwidth for local likelihood density estimation

Håkon Otneim, Hans Arnfinn Karlsen, and Dag Tjøstheim

Statistics & Probability Letters , **83**, 1382-1387 (2013)

Statistics and Probability Letters 83 (2013) 1382–1387

Contents lists available at SciVerse ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Bias and bandwidth for local likelihood density estimation

Håkon Otneim *, Hans Arnfinn Karlsen, Dag Tjøstheim

University of Bergen, Department of Mathematics, P.O. Box 7800, N-5020 Bergen, Norway

ARTICLE INFO

Article history:
Received 15 November 2012
Received in revised form 4 February 2013
Accepted 5 February 2013
Available online 13 February 2013

Keywords:
Local likelihood
Density estimation
Bandwidth selection

ABSTRACT

A local likelihood density estimator is shown to have asymptotic bias depending on the dimension of the local parameterization. Comparing with kernel estimation it is demonstrated using a variety of bandwidths that we may obtain as good and potentially even better estimates using local likelihood. Boundary effects are also examined.

1. Introduction

Estimation of probability density functions is not a matter of black or white in choosing between a parametric and a non-parametric estimator. Considerable work has been done recently on estimators which can be thought of as compromises between the two mindsets. Such estimators are particularly useful when global parametric fits fail to capture important structures, but a non-parametric estimate would still neglect a priori information.

Tibshirani and Hastie (1987) introduced the term *local likelihood* to describe their method of estimating a regression line locally using only the observations within a certain window. The book-length treatment by Fan and Gijbels (1996) generalized local lines to local polynomials and obtained important advantages. See also Loader (1996) for a direct link between local polynomials and local likelihood. Here we will focus on the method by Hjort and Jones (1996) which produces *density* estimates by fitting a parametric density family $\phi(x, \theta)$ locally, that is to let the parameter θ depend on x. For a random sample $x_1, x_2, \ldots, x_n$ from the unknown density $f(x)$, the local maximum likelihood estimator $\widehat{\theta}(x)$ solves the local likelihood equations (L_n denotes the local log-likelihood)

$$U_j = \frac{\partial L_n}{\partial \theta_j} = n^{-1} \sum_{i=1}^{n} K_h(x - x_i) u_j(x_i, \theta(x)) - \int K_h(x - y) u_j(y, \theta(x)) \phi(y, \theta(x))\, \mathrm{d}y = 0, \tag{1}$$

for each x, where $j = 1, \ldots, p$ runs over the components of θ. Further, we denote by $u_j(x, \theta) = \partial/\partial\theta_j \log \phi(x, \theta)$ the local score functions, and the kernel K is a symmetric density function with $K_h(t) = h^{-1}K(t/h)$, h being the bandwidth. The density estimate is given as $\widehat{f}(x) = \phi(x, \widehat{\theta}(x))$. The bandwidth h tunes the level of parameterization in the method. As $h \to \infty$, the local likelihood equations (1) turn global, but if we employ small bandwidths, properties of the local likelihood density estimates are largely separated from those of the parametric family so the method becomes essentially non-parametric.

We make two contributions in this paper. The first is a theoretical one. Building on results by Hufthammer and Tjøstheim (2008) we demonstrate in Section 2 that the leading terms of the bias of $\widehat{f}(x)$ depend on the dimension of θ. Coming to the second contribution, several papers on the theory of local likelihood estimation exist, see e.g. Park et al. (2002), Eguchi and

* Corresponding author. Tel.: +47 55582838.
E-mail address: hakon.otneim@math.uib.no (H. Otneim).

0167-7152/$ – see front matter
doi:10.1016/j.spl.2013.02.003

Copas (1998) and Hall and Tao (2002). To the best of our knowledge, however, very little has been done to examine the finite sample properties of local likelihood estimates and to compare them with kernel estimates. The second contribution seeks to rectify this. Our numerical experiments suggest that local likelihood density estimates perform well, and potentially better than corresponding kernel estimates. In particular, this is exemplified by using a local Gaussian family for a variety of bandwidths in Section 3. Finally, we demonstrate in Section 4 that boundary effects as a function of the bandwidth can be dealt with using familiar methods from non-parametric theory. We believe these results to be novel as well.

2. Asymptotic bias

For iid observations, the law of large numbers implies that the likelihood equations (1) converge in probability to

$$\int K_h(x-y)u_j(y,\theta)\{f(y)-\phi(y,\theta)\}\,\mathrm{d}y=0,\quad j=1,\ldots,p, \tag{2}$$

as $n\to\infty$ and the bandwidth h is held fixed. Following Hjort and Jones (1996) the population parameter $\theta_0(x)=\theta_{0,h}(x)$ is defined as the unique solution to (2) and again following Hjort and Jones (1996) we assume its existence throughout this paper. Let $\phi_0(x)=\phi(x,\theta_0)$. Hjort and Jones (1996) investigate the asymptotic bias of the local likelihood estimate by assessing the size of $\mathrm{E}\widehat{f}(x)-f(x)=\mathrm{E}(\widehat{f}(x)-\phi_0(x))+(\phi_0(x)-f(x))$ which in the one-parameter case is of order $O((nh)^{-1}+h^2)$. They go on to conjecture that the h^2-part is the first step in a pattern known to exist for local polynomials: h^2-convergence for one and two parameters, h^4-convergence for three and four parameters and so on. We will see below that the analogy of the $(nh)^{-1}$-part also depends on the dimension p of θ, and that it seems to *increase* with larger p as $h\to 0$. We use the fact established by Hufthammer and Tjøstheim (2008) (see also Tjøstheim and Hufthammer, 2013, for higher dimensional θ) that the covariance matrix of $\widehat{\theta}$, say $\Sigma_{\widehat{\theta}}$, is of order $O((nh^3)^{-1})$ in the two-parameter case. As a direct consequence we will see below that $\mathrm{E}(\widehat{f}(x)-\phi_0(x))$ is of order $O((nh^3)^{-1})$ as well. This leads to our main result $\mathrm{E}\widehat{f}(x)-f(x))=O((nh^3)^{-1}+h^2)$ stated with regularity conditions at the end of this section.

Recall that $U_i(\theta)=\partial/\partial\theta_i L_n(\theta,X)$, and put $V_{ij}=\partial/\partial\theta_j U_i(\theta_0)$ and $W_{ijk}=\partial/\partial\theta_k V_{ij}(\theta_0)$. Collect these quantities in the 2×2 matrices $I=\mathrm{E}\{-V_{i,j}\}_{i,j=1,2}$ and $J_i=\mathrm{E}\{W_{ijk}\}_{j,k=1,2}$ for $i=1,2$, assuming they exist. Finally, write the estimate and population parameter in terms of their components as $\widehat{\theta}=(\widehat{\theta}_1,\widehat{\theta}_2)^T$ and $\theta_0=(\theta_{0,1},\theta_{0,2})^T$ and the difference between them as $Z_i=\widehat{\theta}_i-\theta_{0,i}$.

We approximate Eq. (1) by a second order Taylor polynomial about θ_0 and take its expectation, noting that $\mathrm{E}U_i(\theta_0)=0$ by (2). Following the lines of Cox and Snell (1968) for ordinary maximum likelihood estimates,

$$0=U_i(\widehat{\theta})\sim\sum_{j=1}^{2}\left[\mathrm{E}Z_j\mathrm{E}V_{ij}+\mathrm{Cov}(Z_j,V_{ij})\right]+\frac{1}{2}\sum_{j=1}^{2}\sum_{k=1}^{2}\left[\mathrm{E}(Z_jZ_k)\mathrm{E}W_{ijk}+\mathrm{Cov}(Z_jZ_k,W_{ijk})\right], \tag{3}$$

for $i=1,2$. Here we neglect the remainder; it being evaluated in Eq. (5) below. Taking both components into account, we can write Eq. (3) more compactly using matrix notation,

$$I\,\mathrm{E}(Z)\sim\left\{\sum_{j=1}^{2}\mathrm{Cov}(Z_j,V_{\ell j})+\frac{1}{2}\mathrm{Tr}(\Sigma_{\widehat{\theta}}J_\ell)+\frac{1}{2}\sum_{j,k=1}^{2}\mathrm{Cov}(Z_jZ_k,W_{\ell jk})\right\}_{\ell=1,2}, \tag{4}$$

where $\mathrm{Tr}(\cdot)$ denotes the trace of a matrix and where $Z=(Z_1,Z_2)^T$. Direct calculations and applications of the Schwarz inequality show that the first and third term on the right hand side of (4) are not larger than $(nh)^{-1}$ asymptotically. Further, Hufthammer and Tjøstheim (2008) show that the matrix I is $O(h^2)$, and inspection of the components in J_1 and J_2 using the same methodology reveals that these two matrices are of the same order as I.

It follows immediately that $\mathrm{E}\widehat{\theta}-\theta_0$ has the same asymptotic order as the covariance matrix of $\widehat{\theta}$, which we have already noted is $O((nh^3)^{-1})$.

We need to check that the remainder in the Taylor expansion above is dominated by its preceding terms. We do this by assuming that $W_{i,j,k}(\theta)$ is continuously differentiable and that $\widehat{\theta}$ converges almost surely to the population parameter θ_0, conditions for which are provided by Theorem 3 in Tjøstheim and Hufthammer (2013) in a more general bivariate time series setting. The Taylor expansion (3) can be written exactly as

$$0=U_i(\theta_0)+\sum_{j=1}^{2}(\widehat{\theta}_j-\theta_{0,j})V_{i,j}(\theta_0)+S_i(\theta_0)+(S_i(\xi)-S(\theta_0)), \tag{5}$$

where $\xi=(\xi_1,\xi_2)^T$ is a random quantity determined by the mean value theorem satisfying

$$\|\widehat{\theta}-\xi\|\le\|\theta_0-\widehat{\theta}\|$$

with probability one, and where $S(\xi)=(S_1(\xi),S_2(\xi))^T$ is given by

$$S_i(\xi)=\frac{1}{2}\sum_{j=1}^{2}\sum_{k=1}^{2}(\widehat{\theta}_j-\theta_{0,j})(\widehat{\theta}_k-\theta_{0,k})W_{i,j,k}(\xi).$$

Since W is assumed to be continuously differentiable, $|S_i(\xi) - S_i(\widehat{\theta})| \leq C\|\theta_0 - \xi\|^3 \leq C\|\theta_0 - \widehat{\theta}\|^3$ with probability one for some constant C. The remainder is therefore asymptotically negligible. Our argument is completed by applying the delta method to see that $\mathrm{E}(\widehat{f}(x) - \phi_0(x)) \sim \mathrm{E}(\widehat{\theta}(x) - \theta_0(x)) = O((nh^3)^{-1})$.

To sum up, we have the following main result: If $\widehat{f}(x)$ is the density estimate resulting from a local likelihood fit using a two-parameter family, and if

(1) there exists a unique solution θ_0 to Eq. (2),
(2) the local likelihood function is four times continuously differentiable with respect to the parameter, which holds trivially for the Gaussian family, and
(3) the parameter estimate $\widehat{\theta}$ converges almost surely to the population parameter θ_0; conditions for this being given by Theorem 3 in Tjøstheim and Hufthammer (2013),

then

$$\mathrm{E}\widehat{f}(x) - f(x) = O((nh^3)^{-1} + h^2). \tag{6}$$

Hufthammer and Tjøstheim (2008) show using the delta method that local likelihood density estimates regain the usual asymptotic variance of order $(nh)^{-1}$ in spite of the variance of the parameter estimates being of larger order. This happens due to some cancellations that do not occur for the bias. The variance of order $O((nh)^{-1})$ and squared bias of order $(h^2 + (nh^3)^{-1})^2$ is therefore balanced asymptotically by choosing the bandwidth to be proportional to $n^{-1/5}$, which is standard procedure in kernel estimation (see e.g. Silverman (1986)). This parallel will be exploited in Section 3.2 on bandwidth selection.

It is worth noting that, Tjøstheim and Hufthammer (2013) go even further and show that the variance of the local parameters is $O((nh^6)^{-1})$ when using the bivariate Gaussian family with its five parameters, and we conjecture that there is a corresponding bias term of the same order.

3. The practical implementation

3.1. Choosing a parametric family

All the theoretical derivations performed by Hjort and Jones (1996) show that the parametric family $\phi(x, \theta)$ should be as close to the true density as possible in order to maximize performance. From a practical point of view, this means that we should consider the nature of our data and make sure that our candidate family is actually able to reach the unknown f with a proper set of parameters, which is the essence in assuming the existence of θ_0 as the unique solution of Eq. (2).

There are, however, situations in which the parameter estimates themselves are more interesting than the density estimates. Consider for example the idea in Tjøstheim and Hufthammer (2013) of estimating the local dependence between stochastic variables as the local correlation resulting from a local bivariate Gaussian fit. This leads to the question to which degree the Gaussian family could serve as a general family for local likelihood density estimates. It is indeed a flexible family that can approximate many types of curves locally. We have carried out a number of experiments, some of which are described below, in which the Gaussian family with good precision estimates non-Gaussian densities. A somewhat different approach is explored by Loader (1996), who fits local polynomials using the same likelihood equations as we do.

We estimate four different distributions by fitting a Gaussian family and using a Gaussian kernel as well, and compare the results with the traditional kernel estimator. They are a bimodal normal distribution with parameters $(\mu_1, \mu_2, \sigma_1, \sigma_2, p) = (3, 5, 1, 0.5, 0.65)$, a t-distribution with one degree of freedom, a gamma distribution with parameters $(\alpha, \beta) = (2, 1)$ and a normal inverse Gaussian distribution with probability density function

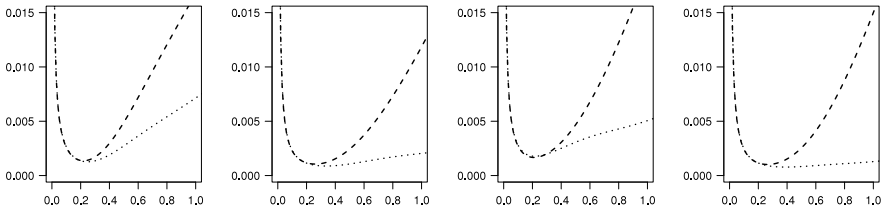
$$f(x) = \sqrt{\chi}(\psi + \gamma^2) \exp\left(\sqrt{\chi\psi}\right) \exp((x - \mu)\gamma) \frac{K_{-1}\left(\sqrt{(\chi + (x - \mu)^2)(\psi + \gamma^2)}\right)}{\pi\sqrt{(\chi + (x - \mu)^2)(\psi + \gamma^2)}}$$

with parameters $(\chi, \psi, \mu, \gamma) = (2, 1, 0, 1)$. In order to make a best case comparison, we calculate the best possible bandwidth in each case in terms of minimum mean integrated squared error (MISE).

The results from our simulations are summarized in Fig. 1 and in Table 1. The mean squared errors have been calculated using 200 data sets at each bandwidth. We make two suggestions based on our simulations. First, the Gaussian family is capable of producing just as good, and in some cases better, density estimates than the non-parametric kernel estimator for a wide variety of probability density functions. Second, the local likelihood density estimator is much less prone to oversmoothing than the kernel estimator. This behavior is not unexpected considering what happens to the two types of estimates as $h \to \infty$. It is interesting to note, however, that the two errors seem to separate in size quickly after the optimal bandwidth, so choosing a too large bandwidth will yield a smaller error when employing the local likelihood methodology.

3.2. Choosing the bandwidth

Based on our remarks concluding Section 2 on the parallel asymptotics between local likelihood and kernel estimation, we find it natural to base a data driven bandwidth selector on existing theory. When using the kernel estimator, we minimize



(a) The bimodal Gaussian distribution. (b) The Student's t-distribution. (c) The gamma distribution. (d) The normal inverse Gaussian distribution.

Fig. 1. Mean integrated squared error for four different distributions fitted locally with the Gaussian parametric family (dotted line) and the kernel estimator (dashed line) as a function of the bandwidth.

Table 1
Performance of local likelihood and kernel density estimators based on 200 data sets, each consisting of 1000 observations. We denote by h_{OPT} the bandwidth that minimizes MISE using the two methods (calculated through simulations), while h^*_{PM} and h_{SJ} denote the bandwidths calculated by our modified Park/Marron approach for local likelihood and the Sheather/Jones selector for the kernel estimator, respectively. MISE is the observed mean integrated squared error at the optimal bandwidths, and we integrate over a region so large that increasing it even further gives the same result. Note that, h_{SJ} does not seem to work for the t-distribution in the kernel case. This is probably because the squared density integral which appears in the derivation of h_{SJ} does not exist for this distribution.

Distribution	Local likelihood			Kernel estimator		
	h_{OPT}	h^*_{PM}	MISE	h_{OPT}	h_{SJ}	MISE
Bimodal normal distribution	0.24	0.32	0.00132	0.22	0.23	0.00138
t-distribution	0.35	0.29	0.00089	0.27	0.02	0.00107
Gamma distribution	0.22	0.27	0.00172	0.21	0.22	0.00163
Normal inverse Gaussian dist.	0.41	0.31	0.00083	0.27	0.27	0.00098

the asymptotic mean integrated squared error (AMISE) by utilizing the bandwidth $h_{\text{AMISE}} = \{R(K)/\sigma_K^4 R(f'')\}^{1/5} n^{-1/5}$, where $R(g) = \int g^2(x)\,dx$ and $\sigma_g^2 = \int x^2 g(x)\,dx$. Unfortunately, the optimal bandwidth depends on the unknown f through $R(f'')$, so Park and Marron (1990) suggest to use the bandwidth that solves the equation

$$h = \frac{R(K)}{R(\tilde{f}''_{\alpha(h)})\sigma_K^4} n^{-1/5}, \tag{7}$$

where $\alpha(h)$ is a known function of the bandwidth h which is optimized for estimating $R(f'')$, and $\tilde{f}_{\alpha(h)}(x)$ is the kernel estimate of f obtained using this bandwidth. As a crude modification to the local likelihood case, we propose to substitute $\tilde{f}''$ in (7) by the corresponding local likelihood estimate $\widehat{f}''$. According to Hjort and Jones (1996), $R(f'')$ should actually be replaced by $R(f'' - \phi_0'')$, but it is unclear how ϕ_0'' can be estimated. They indicate that this could be done using the same bandwidth h as in the estimation of f, but that gives inadequate results in our examples.

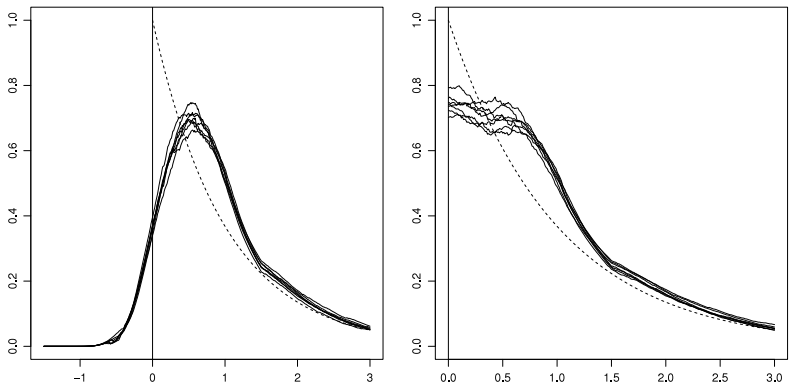
Sheather and Jones (1991) improve Park and Marron's method in terms of convergence rates, but it would make little sense to translate this improvement to our rather unpolished modification. We note, however, that in our examples, the performance of local likelihood density estimates using our modified Park/Marron bandwidth selector performs at least as well as, and in some cases better than kernel estimates with Sheather/Jones-bandwidths (not given in Table 1).

4. Estimating densities with bounded support

A well known problem with the kernel estimator is the increased bias that arises when estimating a density with bounded support. We assume throughout this section, without loss of generality, that our unknown density f is positive on $[0, \infty)$ only. We have in fact included one such density, the gamma distribution, in our comparison in Fig. 1 and Table 1 without boundary correction, but this density tends to zero as $x \to 0$ which makes it easier to handle. In a more general case, if $\tilde{f}(x)$ denotes the kernel estimate, Marron and Ruppert (1994) show that, for $x < h$,

$$\mathrm{E}\tilde{f}(x) = a_0(p)f(x) - a_1(p)hf'(x) + \frac{a_2(p)h^2}{2}f''(x) + O(h^2), \tag{8}$$

where $x = ph$, $p \in [0, 1]$, the kernel K has support $[-1, 1]$ (we use the uniform kernel $K(x) = \frac{1}{2}\mathbf{1}_{[-1,1]}$ in the following simulations), and $a_i(p) = \int_{-1}^{p} u^i K(u)\,du$. We regain the usual expansion $\mathrm{E}\tilde{f}(x) \sim f(x) + (h^2/2)f''(x)$ for $x > h$. Hjort and Jones (1996) show that no additional bias occurs at boundaries in local likelihood estimates if the parametric family has the same support as the unknown density. See Fig. 3(b) for an example where the parametric family respects the boundary of the unknown density. We have argued, however, that the Gaussian family produces good estimates in a variety of settings,



(a) No reflection. (b) With reflection method.

Fig. 2. Seven local likelihood density estimates of the exponential distribution $f(x) = \exp(-x)$ without any adjustment (a) and using the reflection method (b). Each data set consist of 1000 observations, and a bandwidth of 1.5 has been used.

but since the normal distribution has unbounded support, similar effects as (8) arise. Write Eq. (2) as

$$0 = \int_0^{x+h} K_h^*(x-y)u_0(y)f(y)\,\mathrm{d}y - \int_{x-h}^{x+h} K_h(x-y)u_0(y)\phi_0(y)\,\mathrm{d}y, \tag{9}$$

where the notation K_h^* indicates that we may actually use a modification of the kernel K_h in the local likelihood equations. We proceed in the exact same manner as Marron and Ruppert (1994), and, starting with $K_h^* = K_h$, arrive at

$$0 = a_0(p)u_0(x)f(x) - ha_1(p)[u_0(x)f(x)]' + \frac{h^2a_2(p)}{2}[u_0(x)f(x)]'' - u_0(x)\phi_0(x) + \frac{h^2}{2}[u_0(x)\phi_0(x)]'', \tag{10}$$

which results in

$$\mathrm{E}\widehat{f}(x) = a_0(p)f(x) - ha_1(p)[u_0(x)f(x)]'/u_0(x) + O(h^2 + (nh^3)^{-1}) \tag{11}$$

in the two-parameter case, where $a_i(p) = \int_{-1}^{p} u^i K(u)\,\mathrm{d}u$.

Jones (1993) presents several methods for dealing with boundary bias for the kernel estimator, and we will briefly look at two of these below, to see that they apply just as well for local likelihood estimates.

4.1. The reflection method: h-convergence

A very simple method for ensuring consistency (but with bias of order h) is to reflect our data set about zero which amounts to putting $\tilde{f}_R(x) = \tilde{f}(x) + \tilde{f}(-x)$. We see that this works for local likelihood as well by expanding (9), but this time with $-x$ in place of x:

$$\begin{aligned} 0 = {} & a_0(-p)u_0(x)f(x) - h\{2pa_0(-p) + a_1(-p)\}[u_0(x)f(x)]' \\ & + h^2\left\{2p^2a_0(-p) + 2pa_1(-p) + \frac{a_2(-p)}{2}\right\}[f(x)u_0(x)]'' - u_0(-x)\phi_0(-x) - \frac{h^2}{2}[u_0(-x)\phi_0(-x)]'' + o(h^2), \end{aligned} \tag{12}$$

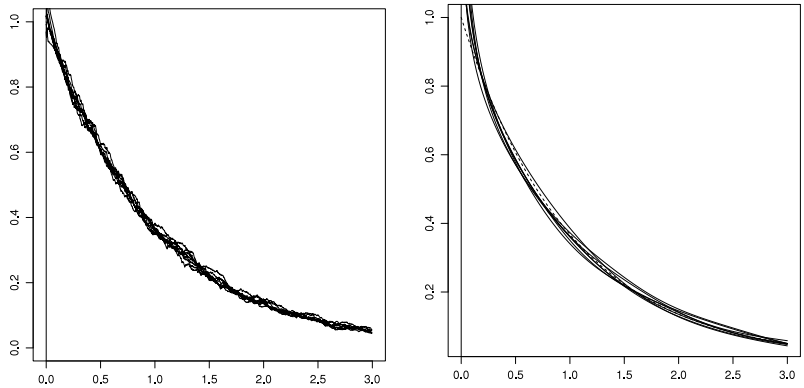
where we have used the same calculations as Marron and Ruppert (1994) again. By adding (10) and (12) together and using that $a_0(-p) + a_0(p) = 1$ and $a_1(-p) = a_1(p)$, it follows that

$$\mathrm{E}\left(\widehat{f}(x) + \widehat{f}(-x)\right) = f(x) - 2h\{pa_0(-p) + a_1(-p)\} + O(h^2 + (nh^3)^{-1})$$

in the two-parameter case. The reflection method thus works in the same way for local likelihood density estimates as it does for the kernel method. See Fig. 2 for a simple illustration using the exponential distribution.

4.2. Kernel modification: h^2-convergence

Jones (1993) proposes to modify the kernel function K to a function K_h^* in order to obtain $a_0^*(p) = 1$ and $a_1^*(p) = 0$, where $a_i^*(p)$ is defined in the same way as for K_h, and thereby regaining the usual $O(h^2)$ convergence. From (10) we see that, this



(a) Kernel corrected estimates. (b) Gamma parametric family.

Fig. 3. Seven realizations using (A) the kernel correction ($n = 10\,000$, $h = 0.1$) and (B) using the Gamma parametric family ($n = 1000$, $h = 0.5$). The true density is the exponential density $f(x) = \exp(-x)$.

will work also in the local likelihood case if we in Eq. (9) use the boundary kernel given by

$$K^*(x) = \frac{(a_2(p) - a_1(p)x)K(x)}{a_0(p)a_2(p) - a_1^2(p)},$$

for which the desired properties are easy to verify.

The downside of such a kernel modification is that the parametric family could in some cases, local Gaussian being one of them, exhibit large departures from the true density near the boundary, and in practice we need smaller bandwidths, and hence more data to make the method essentially non-parametric in this area. See Fig. 3(a) for an example with 10 000 observations, in which the bandwidth cannot be much larger than 0.1.

Acknowledgment

The authors are most grateful to the anonymous referee who provided several valuable comments and suggestions.

References

Cox, D., Snell, E., 1968. A general definition of residuals. Journal of the Royal Statistical Society. Series B 248–275.
Eguchi, S., Copas, J., 1998. A class of local likelihood methods and near-parametric asymptotics. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 60 (4), 709–724.
Fan, J., Gijbels, I., 1996. Local Polynomial Modelling and Its Applications. Vol. 66. Chapman & Hall/CRC.
Hall, P., Tao, T., 2002. Relative efficiencies of kernel and local likelihood density estimators. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64 (3), 537–547.
Hjort, N., Jones, M., 1996. Locally parametric nonparametric density estimation. The Annals of Statistics 1619–1647.
Hufthammer, K., Tjøstheim, D., 2008. Local gaussian likelihood and local gaussian correlation. Doctoral Thesis, Chapter 3, University of Bergen.
Jones, M., 1993. Simple boundary correction for kernel density estimation. Statistics and Computing 3 (3), 135–146.
Loader, C., 1996. Local likelihood density estimation. The Annals of Statistics 24 (4), 1602–1618.
Marron, J., Ruppert, D., 1994. Transformations to reduce boundary bias in kernel density estimation. Journal of the Royal Statistical Society. Series B 653–671.
Park, B., Kim, W., Jones, M., 2002. On local likelihood density estimation. Annals of Statistics 1480–1495.
Park, B., Marron, J., 1990. Comparison of data-driven bandwidth selectors. Journal of the American Statistical Association 85 (409), 66–72.
Sheather, S., Jones, M., 1991. A reliable data-based bandwidth selection method for kernel density estimation. Journal of the Royal Statistical Society. Series B 683–690.
Silverman, B., 1986. Density Estimation for Statistics and Data Analysis. Vol. 26. Chapman & Hall/CRC.
Tibshirani, R., Hastie, T., 1987. Local likelihood estimation. Journal of the American Statistical Association 82 (398), 559–567.
Tjøstheim, D., Hufthammer, K., 2013. Local Gaussian correlation: a new measure of dependence. Journal of Econometrics 172, 33–48.

Paper II

3.2 The locally Gaussian density estimator for multivariate data

Håkon Otneim and Dag Tjøstheim

Submitted for publication to *Statistics & Computing*.

II

The Locally Gaussian Density Estimator for Multivariate Data

Håkon Otneim Dag Tjøstheim

Abstract

It is well known that the Curse of Dimensionality causes the standard Kernel Density Estimator to break down quickly as the number of variables increases. In non-parametric regression, this effect is relieved in various ways, for example by assuming additivity or some other simplifying structure on the interaction between variables. This paper presents the Locally Gaussian Density Estimator (LGDE), which introduces a similar idea to the problem of density estimation.

The LGDE is a new method for the non-parametric estimation of multivariate probability density functions. It is based on preliminary transformations of the marginal observation vectors towards standard normality, and a simplified local likelihood fit of the resulting distribution with standard normal marginals. The LGDE is introduced, and asymptotic theory is derived. In particular, it is shown that the LGDE converges at a speed that does not depend on the dimension. Examples using real and simulated data confirm that the new estimator performs very well on finite sample sizes.

1 Introduction

The *Curse of Dimensionality* precludes the use of many common statistical methods in higher dimensions. The problem is a consequence of the geometry of Euclidean spaces, and will not be solved when the next generation of computing power arrives; it will potentially get worse, as the amount and complexity of data increase. There exist techniques for multivariate data analysis that relieve the effects of the Curse of Dimensionality in various ways. This is especially true for non-parametric regression analysis, but to a much smaller extent in density estimation. In this paper, we present a new estimator for probability density functions that is especially designed to be flexible, yet robust, when fitted to increasingly higher dimensional data (dimensions 2-10 in this paper) of unknown parametric origin.

Suppose that the observations $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are independent and identically distributed with an unknown density function, $f(\boldsymbol{x})$, that we wish to estimate. Classical statistics provides two fundamentally different approaches to the problem. If we know the functional form of the unknown density up to a set of parameters, they can be estimated efficiently by maximum likelihood. If a parametric assumption cannot be supported by the data, or prior knowledge, a non-parametric method such as the kernel estimator is the natural alternative. It is well known, however, that the kernel estimator breaks down quickly as the dimension of our data increases. Silverman (1986) shows that we need close to a million ten-dimensional observations in order to produce a kernel density estimate with the same accuracy as would only four observations in one dimension.

Techniques for dimensionality reduction exist, including the widely used Principal Component Analysis. The reduced observation vector may still have too many dimensions to produce a fully non-parametric estimate, though, possibly forcing the experimenter to choose a parametric model far from the true distribution. In many cases, that means fitting the multivariate normal distribution, because the parameter estimates are quick to calculate and easy to interpret.

There is a middle road, however, which can be labelled broadly as semi-parametric density estimation. Methods include the local likelihood estimators by Hjort and Jones (1996) and Loader (1996) and the combination of non-parametric and parametric estimates provided by Hjort and Glad (1995). A semi-parametric model can be considered as a trade-off between non-parametric flexibility and parametric performance, making them very attractive in practical use, exemplified by the recent work by Geenens (2014), who shows that the local likelihood variety by Loader (1996) combined with a pre-transformation of the data solves the long standing problem of estimating densities restricted to the unit interval.

Geenens et al. (2014) extend this methodology to the bivariate case, and provide a non-parametric estimator of the copula density by first transforming the data to approximate standard normality, upon estimating the transformed density using local likelihood. Although we are mainly interested in estimates of the density function on the original scale in the present work, we will see that our approach is an attempt to extend the Geenens et al. (2014)-methodology to the multivariate case, which becomes clear when we show that their theoretical contributions are directly applicable to our new method.

In this paper, we use the local likelihood function that was proposed by Hjort and Jones (1996) to fit a parametric distribution *locally* to an unknown multivariate density. Just as in Geenens et al. (2014), we pre-transform the observations so that they have approximate standard normal margins. We then fit the multivariate normal distribution locally by carrying a simplified estimation procedure over from the global case. Asymptotic properties of the estimator are presented, and we show through simulations and a real data example that the estimator works very well for a large class of non-Gaussian data.

Our main motivation for transforming the data to standard normal marginals will become clear in our formal presentation of the *Local Gaussian Density Estimator* (LGDE) in Section 2, but it carries several advantages with it besides. The transformed multivariate density to be estimated has unbounded support, it has short tails, and all its variables are on the same scale. Furthermore, several authors have noted that densities become easier to estimate when they are transformed towards normality, see e.g. Wand et al. (1991) and Ruppert and Cline (1994).

In Section 3, we present the asymptotic theory of our estimator, including a discussion on the existence of a least false density function within the restrictions that we impose, and towards which our estimate converges. In particular, our density estimate converges at a rate that does not depend on the dimension. Practical issues, like bandwidth selection and choice of the kernel function, are examined in Section 4. Sections 5 and 6 concern the application of the LGDE on simulated and real data, respectively. The treatment is brought to an end in Section 7, where we make some concluding remarks, and discuss various aspects of our approach.

2 Description of the estimator

2.1 Motivation

A large number of dimensions does not necessarily mean trouble when we face the problem of density estimation. If we know that the observations are Gaussian, the means, variances and covariances are estimated based on the first and second empirical moments only. Dependence between a large set of stochastic variables is a complicated matter, though, multivariate normality being a rather restrictive property. In general, dependence between two variables must be modelled taking all other variables into account, resulting in a daunting task of estimation if we have no prior assumptions on the general structure. The same problem arises in the regression setting with a large set of explanatory variables. Estimating the nonlinear regression $Y = f(X_1, \ldots, X_p) + \epsilon$ using observations $(Y_i, X_{1i}, \ldots, X_{pi})$, $i = 1, \ldots, n$ is more or less impossible for a moderately large p because of the Curse of Dimensionality. A common simplification is the additive model, in which we assume $f(X_1, \ldots, X_p) = f_1(X_1) + \cdots + f_p(X_p)$, disregarding any interactions between the variables. The simplification may well be restrictive, but it is computationally *possible* and may be our best guess in many situations. Producing reliable estimates of the complete dependence structure without some sort of restriction is simply not an option. The LGDE has a similar flavour.

Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be a random sample from the p-variate distribution with density function $f(\boldsymbol{x})$. The observations, as well as the variable $\boldsymbol{x} = (x_1, \ldots, x_p)^T$, are column vectors of length p, so that $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})^T$. Denote by $F(\boldsymbol{x})$ the cumulative distribution function (cdf) corresponding to f, and further, let $f_i(x_i)$ and $F_i(x_i)$ denote the marginal densities and cdfs respectively for $i = 1, \ldots, p$. The univariate standard normal density and distribution function are identified by ϕ and Φ:

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

We transform each observation vector to standard normality using the marginal cdfs (assuming these known at the present stage) and the Gaussian quantile function, so that observation number j becomes

$$\boldsymbol{Z}_j = \left(\Phi^{-1}\left(F_1(X_{j1})\right), \ldots, \Phi^{-1}\left(F_p(X_{jp})\right)\right)^T.$$

The marginal distributions of the transformed data are now standard normal, and the joint density function, $f_{\boldsymbol{Z}}(\boldsymbol{z})$ say, is given by

$$f_{\boldsymbol{Z}}(\boldsymbol{z}) = f\left(F_1^{-1}\left(\Phi(z_1)\right), \ldots, F_p^{-1}\left(\Phi(z_p)\right)\right) \times \prod_{i=1}^{p} q_i(\Phi(z_i))\phi(z_i),$$

where $q_i(z_i) = \mathrm{d}/\mathrm{d}z\ F_i^{-1}(z_i)$, $i = 1, \ldots, p$, are the marginal quantile density functions. By a change of variables, we express the original density in terms of $f_{\boldsymbol{Z}}$ and the marginal distributions as

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

The decomposition of the density in (1) is parallel to what we find in the copula framework of analysis. Sklar's (1959) theorem states that any multivariate cdf can be

expressed by a unique copula function of its marginals, enabling us to model dependence between variables separately from the their individual marginal distributions. The copula function is simply a cdf with standard uniform margins; the transformed density $f_{\boldsymbol{Z}}$ in (1) has standard normal margins, but contains complete information on the dependence between the variables constituting our data, and its estimation is the main contribution of this paper.

In fact, we believe that analyzing the Gaussian observations $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$, which in practice must be estimated from data making them into Gaussian pseudo-observations, instead of uniform ones, is advantageous in many situations, especially in the non-parametric paradigm, because distributions of real data are usually closer to being Gaussian than uniform, with less tail distortions in the former case. This is illustrated by Berentsen et al. (2014) in identification of copula structures, and discussed in detail by Mikosch (2006).

2.2 Estimation of the marginals

The usual way of producing uniform pseudo-observations from the copula of $f(\cdot)$ is to transform each marginal with the empirical distribution functions, which we denote by $\tilde{F}_{k,n}(\cdot)$, $k = 1, \ldots, p$, so that the pseudo-observations on the standard normal scale is given by

$$\widehat{\boldsymbol{Z}}_j = \left(\Phi^{-1}\left(\tilde{F}_{1,n}(X_{j1})\right), \ldots, \Phi^{-1}\left(\tilde{F}_{p,n}(X_{jp})\right)\right)^T. \tag{2}$$

It is well known from the estimation of copulas that using pseudo-observations in general affect the copula estimate (Genest and Segers, 2010), because the marginal empirical distribution functions have the same $\sqrt{n}$-convergence rate as the final, empirical or parametric, copula estimate. In our case, however, we estimate the transformed density $f_{\boldsymbol{Z}}$ semi-parametrically, which results in a slower convergence of the order $\sqrt{nh^2}$, where $h \to 0$ is the smoothing parameter. It is natural then, that using pseudo-observations instead of genuine observations from $f_{\boldsymbol{Z}}$ will not affect the asymptotic distribution of the density estimate. It turns out that we can use the theory presented by Geenens et al. (2014) directly, to show that this is, in fact, the case.

Contrary to Geenens et al. (2014), we are interested in density estimates on the original scale, that is, we estimate the density function $f(\boldsymbol{x})$ of $\boldsymbol{X}$, and not the copula density $c(u_1, \ldots, u_p)$ associated with f. This is not a dramatic change, however, because the relation between the two quantities are determined by the marginal distributions only, with

$$f(\boldsymbol{x}) = c(F_1(x_1), \ldots, F_p(x_p)) \prod_{i=1}^{p} f_i(x_i). \tag{3}$$

The marginal quantile- and density functions must again be estimated from the data, but the empirical distribution function is not suitable to use in the back-transformation, as it is neither invertible nor differentiable. Any other suitable estimates for these quantities will work, however, and will not cause any trouble in the asymptotic results as long as they converge faster than the $\sqrt{nh^2}$-rate that we will see holds for the multivariate density estimate.

A natural method for producing estimates of the quantile- and density function is the one-dimensional kernel estimator. We achieve much better results in our finite-sample simulation experiments if we rather employ the logspline method by Stone et al.

(1997) for this purpose, and we prove that the logspline marginal estimates achieve an appropriate convergence rate, under a regularity condition, in Theorem 5 in Section 3.

2.3 Estimation of the joint dependence function

Let $\psi(\cdot;\boldsymbol{\theta})$ be a parametric family of p-variate density functions. Hjort and Jones (1996) estimate the unknown density f using the sample $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n$ by fitting ψ locally. The parameter estimate $\widehat{\boldsymbol{\theta}}_n = \widehat{\boldsymbol{\theta}}_n(x)$ maximises the *local log-likelihood function*

$$L_n(\boldsymbol{\theta},\boldsymbol{x}) = n^{-1}\sum_{i=1}^{n} K_{\boldsymbol{h}}(\boldsymbol{X}_i-\boldsymbol{x})\log\psi\left(\boldsymbol{X}_i,\boldsymbol{\theta}\right) - \int K_{\boldsymbol{h}}(\boldsymbol{y}-\boldsymbol{x})\psi(\boldsymbol{y},\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{y}, \tag{4}$$

where $K(\cdot)$ is a kernel function that integrates to one and is symmetric about the origin, $\boldsymbol{h}$ is a positive definite matrix of bandwidths, and $K_{\boldsymbol{h}}(\boldsymbol{x}) = |\boldsymbol{h}|^{-1}K(\boldsymbol{h}^{-1}\boldsymbol{x})$. For small bandwidths, the local estimate $\widehat{f}(\boldsymbol{x}) = \psi(\boldsymbol{x},\widehat{\boldsymbol{\theta}}_n(\boldsymbol{x}))$ is close to $f(\boldsymbol{x})$ in the limit as $n\to\infty$, because, if the bandwidth matrix $\boldsymbol{h}$ is held fixed and $u_j(\cdot,\boldsymbol{\theta}) = \partial/\partial\theta_j \log\psi(\cdot,\boldsymbol{\theta})$ denotes the score functions, we have

$$0 = \frac{\partial L_n(\widehat{\boldsymbol{\theta}}_n,x)}{\partial\theta_j} \xrightarrow{P} \int K_{\boldsymbol{h}}(\boldsymbol{y}-\boldsymbol{x})u_j(\boldsymbol{y},\boldsymbol{\theta}_{\boldsymbol{h},K}(\boldsymbol{y}))\left\{f(\boldsymbol{y})-\psi(\boldsymbol{y},\boldsymbol{\theta}_{\boldsymbol{h},K}(\boldsymbol{y}))\right\}\,\mathrm{d}\boldsymbol{y}$$

for some value of $\boldsymbol{\theta}_{\boldsymbol{h},K}(x)$ towards which $\widehat{\boldsymbol{\theta}}_n(\boldsymbol{x})$ converges in probability. For finite sample sizes, however, the Curse of Dimensionality comes into play as the dimension of $\boldsymbol{x}$ increases, making the local estimates difficult to obtain at every point in the sample space. One solution might be to increase the bandwidths so that the estimation becomes almost parametric. We propose a different path around the *Curse*, directly exploiting the decomposition (1). The first step is to choose a standardised multivariate normal distribution as parametric family in (4) for modelling $f_{\boldsymbol{Z}}$ in (1) locally:

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

where $\boldsymbol{R}$ denotes the correlation matrix. We refer to its local likelihood estimate $\widehat{\boldsymbol{R}} = \widehat{\boldsymbol{R}}(\boldsymbol{z})$ as the *local Gaussian correlation*, see Tjøstheim and Hufthammer (2013) for theory and applications in the bivariate case, including a time series setting. Using a univariate local fit, the local Gaussian expectations and variances in (5) are constant and equal to zero and one respectively, reflecting our knowledge that the margins of the unknown density function $f_{\boldsymbol{Z}}$ are standard normal.

Fitting the Gaussian distribution according to the scheme described above results in a local correlation matrix at each point. Specifically, the estimated local correlations are written $\widehat{\rho}_{ij} = \widehat{\rho}_{ij}(z_1,\ldots,z_p)$, $i,j=1,\ldots,p$, indicating that each parameter depends on all variables. The dependence between variables is captured in the variation of the parameter estimates in the p-dimensional Euclidean space, and its estimate maximises the local likelihood function (4). As mentioned before, however, the quality of the estimate deteriorates quickly with the dimension.

If the data were jointly normally distributed, however, there would be no dimensionality problem, since the entire distribution would be characterised by the global correlation

coefficients between pairs of variables, and their empirical counterparts are easily computed from the data. A local Gaussian fit would then coincide with a global fit, and result in estimates of the form $\widehat{\rho}_{ij} = \widehat{\rho}_{ij}(Z_i, Z_j)$, where the arguments indicate which of the transformed observation vectors have been used to obtain the estimate. This points to a natural simplification that we may use in order to estimate the density $f_{\boldsymbol{Z}}$, analogous to the additive regression model. Allow the local correlations to depend on their own variables only:

$$\widehat{\rho}_{ij}(z_1, \ldots, z_p) = \widehat{\rho}_{ij}(z_i, z_j). \tag{6}$$

The resulting estimation is carried out in four steps:

1. Estimate the marginal distributions using the logspline method, and transform each observation vector to pseudo-standard normality as described in Section 2.1.

2. Estimate the joint density of the transformed data using the Hjort and Jones (1996) local likelihood function (4), the standardised normal parametric family (5) and the simplification (6). In practice, this means fitting the bivariate version of (5) to each pair of the transformed variables (Z_i, Z_j). Put the estimated local correlations into the estimated local correlation matrix: $\widehat{\boldsymbol{R}}(\boldsymbol{z}) = \{\widehat{\rho}_{ij}(z_i, z_j)\}_{i,j=1,\ldots,p}$.

3. Let $\widehat{f}_{\boldsymbol{Z}}(z) = \psi\left(\boldsymbol{z}, \widehat{\boldsymbol{R}}(\boldsymbol{z})\right)$ and obtain the final estimate of $f(\boldsymbol{x})$ by replacing $f_{\boldsymbol{Z}}$ with $\widehat{f}_{\boldsymbol{Z}}$, and the marginal distribution and density functions with their estimates in (1):

$$\widehat{f}(\boldsymbol{x}) = \widehat{f}_{\boldsymbol{Z}}\left(\Phi^{-1}\left(\widehat{F}_1(x_1)\right), \ldots, \Phi^{-1}\left(\widehat{F}_p(x_p)\right)\right) \times \prod_{i=1}^{p} \frac{\widehat{f}_i(x_i)}{\phi\left(\Phi^{-1}\left(\widehat{F}_i(x_i)\right)\right)}. \tag{7}$$

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

The existence of population values corresponding to the estimated local correlations is discussed in the following section. It is clear that the simplification (6) represents an approximation for most multivariate distributions. The authors are aware of no other distributions than those possessing the Gaussian copula, or step functions thereof as in Tjøstheim and Hufthammer (2013), for which (6) is an exact property of the true local correlations. In that case, the local correlations are constant, or stepwise constant, in *all* its variables. As is the case for the additive regression model, exploring how restrictive (6) is in practice, can primarily be done by trying it out on large classes of simulation models and empirical data sets. A start is made in Sections 5 and 6 with a fairly varied set of examples.

It is not difficult to find examples where (6) is not satisfied. There is an analogous, but nevertheless different, formulation of (6) that may be easier to compare to the additive regression assumption. The characteristic function of a Gaussian distribution is (in the standardised case) given by $\exp(-\boldsymbol{x}^T\boldsymbol{R}\boldsymbol{x}/2)$ where again R is the correlation matrix. The analog to assumption (6) would then be that $\boldsymbol{x}^T\boldsymbol{R}(\boldsymbol{x})\boldsymbol{x} = \sum_{i,j} x_i x_j \rho_{ij}(x_i, x_j)$, whereas in the regression case $\mathrm{E}(Y|X = x) = f(\boldsymbol{x}) = \sum_i f_i(x_i)$. The regression case can be successively generalised to higher order interaction (see e.g. Sperlich et al. (2002)), the first step being $f(x) = \sum_{i,j} f_{ij}(x_i, x_j)$, but the effect of the curse of dimensionality is quickly felt for higher order interactions.

The pairwise additivity in (6) is the natural assumption to make in the local Gaussian case, because the multivariate Gaussian itself is based on pairwise covariances, but unlike the additive regression, there is a distributional extension that can be made. The Gaussian distribution is a member of the much larger family of elliptical distributions. This family can again be characterised by pairwise covariances. The standard normal marginals can be replaced by 'standard' univariate elliptical distributions, and the interaction may be described by local multivariate elliptical distributions with pairs of local covariances because the characteristic function is on the form $g(\boldsymbol{x}^T\boldsymbol{R}\boldsymbol{x})$. Clearly, a separate investigation is required to examine this closer.

In a sense, our approach is comparable to the popular vine-copulas (Bedford and Cooke, 2002) within the parametric framework, that are popular for approximating dependence between several variables with pairwise copulas. This entails of course that the type of parametric model has to be chosen. Recent work by Nagler and Czado (2015) gives a promising method for using pair-copulas for non-parametric multivariate density estimation.

3 Asymptotic theory

Let us establish some notation, and then formulate results regarding the asymptotic behaviour of the LGDE. We will proceed by first proving some convergence results on the local Gaussian correlation for marginally standard normal variables, and then state asymptotic normality for the multivariate density estimate.

Product kernels will be used in theory, as well as in practice, so the matrix of bandwidths, $\boldsymbol{h}$, is diagonal, and $\boldsymbol{h} \to 0$ means that each element of $\boldsymbol{h}$ tends to zero.

For each pair of variables, we maximise the local log-likelihood function $L_n(\rho_{ij}, z_i, z_j)$ in order to obtain the estimated local correlation for that pair. Indeed, the simplification (6) means that we can develop most of the asymptotic theory by looking only at the bivariate case. Keep therefore the pair of indices (i, j) fixed for the time being, so that $\boldsymbol{z} = (z_i, z_j)^T$, $\boldsymbol{h} = \operatorname{diag}(h_i, h_j)$, and, for simplicity, write $\rho_{ij}(z_i, z_j) = \rho(\boldsymbol{z})$. The standardised Gaussian family $\psi(\boldsymbol{z}, \rho)$ will represent the bivariate version of (5) in the following. It will be seen in Lemma 1 below that we may ignore the fact that in practice Z_i has to be estimated by $\widehat{Z}_i = \Phi^{-1}(\tilde{F}_i(X_i))$, because we can translate to our use the results of Geenens et al. (2014), who show that, under some smoothing conditions, using pseudo-observations do not affect the asymptotic properties of their copula density estimator.

Denote by I the limiting integral in (5). As the sample size increases to infinity, the local score function $\partial L_n(\boldsymbol{z}, \rho)/\partial\rho$ satisfies the equation

$$I = \int K_{\boldsymbol{h}}(\boldsymbol{y} - \boldsymbol{z}) u(\boldsymbol{y}, \rho) \big\{ f_{ij}(\boldsymbol{y}) - \psi(\boldsymbol{y}, \rho) \big\} \, \mathrm{d}\boldsymbol{y} = 0, \tag{8}$$

where $f_{ij}(\boldsymbol{z})$ is the joint density of (Z_i, Z_j), and the expression for $u(\cdot) = \partial\psi(\cdot)/\partial\rho$ of course is known in our case, and has been written out explicitly in Appendix A.2. Thus, as mentioned before, the estimate $\widehat{\rho}_n(\boldsymbol{z})$ aims at the solution of (8), which we denote by $\rho_{\boldsymbol{h},K}(\boldsymbol{z})$. There are two problems in perceiving $\rho_{\boldsymbol{h},K}(\boldsymbol{z})$ as the 'true' local correlation function, however. First, it is hard to do any general analysis on existence and uniqueness based on the integral in (8), considering that $\rho = \rho(\boldsymbol{z})$ is an unknown function of $\boldsymbol{z}$. Second, $\rho_{\boldsymbol{h},K}(\boldsymbol{z})$ depends on the bandwidths as well as the kernel function $K(\cdot)$, while the true local correlation function for a given pair of variables should be a property of their unknown bivariate density f_{ij} only.

By letting the bandwidth tend to zero as the sample size increases, we solve the second problem and make the first easier. To see this, we reproduce the Taylor expansions of (8) in powers of $\boldsymbol{h}$ as provided by Hjort and Jones (1996). Let the index $\boldsymbol{h}, K$ to functions $\psi(\boldsymbol{z})$ and $u(\boldsymbol{z})$ mean that we insert the parameter value $\rho_{\boldsymbol{h},K}$. It follows that,

$$u_{\boldsymbol{h},K}(\boldsymbol{z})\big\{f_{ij}(\boldsymbol{z}) - \psi_{\boldsymbol{h},K}(\boldsymbol{z})\big\} = \frac{1}{2}\sum_{k=i,j}\sigma_{K_k}^2 h_k^2\big\{u_{\boldsymbol{h},K}(\psi_{\boldsymbol{h},K} - f_{ij})\big\}''(\boldsymbol{z}) + O((h_1^2+h_2^2)^2), \quad (9)$$

where $\sigma_{K_i}^2 = \int y^2 K_i(y)\,\mathrm{d}y$, and the cross-term is zero because of the symmetry of K. The differentiation on the right hand side is taken with respect to z_k. There is only one such equation for each local correlation, and it follows readily that the limit $\rho_0(\boldsymbol{z}) = \lim_{\boldsymbol{h}\to 0}\rho_{\boldsymbol{h},K}(\boldsymbol{z})$ must satisfy $\psi(\boldsymbol{z},\rho) = f_{ij}(\boldsymbol{z})$. This is not enough to ensure the uniqueness of ρ_0, though. It is essential that $\rho_0(\boldsymbol{z})$ is the result of a limiting process as $\boldsymbol{h}\to 0$ in (8). Said in another way, this means that the local fit is done in a neighbourhood of $\boldsymbol{z}$ that shrinks to zero with $\boldsymbol{h}$. Such a process eliminates fits of Gaussians that just pass through the point $\boldsymbol{z}$.

For a fixed $\boldsymbol{h}$, the $\rho_{\boldsymbol{h}}$ can be obtained by minimizing the penalty function

$$q_{\boldsymbol{h},K} = \int K_{\boldsymbol{h}}(\boldsymbol{y}-\boldsymbol{z})\big\{\psi(\boldsymbol{y},\rho) - \log\psi(y,\rho)f_{ij}(\boldsymbol{y})\big\}\,\mathrm{d}\boldsymbol{y}.$$

As seen in Hjort and Jones (1996), this can be interpreted as a locally weighted Kullback-Leibler distance from $f(\cdot)$ to $\psi(\cdot,\rho(\cdot))$.

Let $\boldsymbol{h}_n$ be a sequence of bandwidths tending to zero as $n\to\infty$. If $\{\rho_{\boldsymbol{h}_n,K}(\boldsymbol{z})\}$ converges towards the value $\rho_0(\boldsymbol{z})$, we take this to be the population parameter. This essentially requires then (cf. Hjort and Jones (1996) and Tjøstheim and Hufthammer (2013)), that there is a unique maximum of the local likelihood function once $\boldsymbol{h}$ is small enough, and we include this as an assumption in the following theorem. This is akin to the assumption of a unique maximum in global maximum likelihood estimation. The continuity of ψ as a function of ρ ensures that the population parameter as defined above automatically satisfies $\psi(\boldsymbol{z},\rho_0) = f_{ij}(\boldsymbol{z})$ (Even if a unique maximum should not exist, our approach could still, as a purely data algorithmic tool, produce a good approximation to the theoretical density $f(\boldsymbol{x})$).

The following theorems provide conditions for the consistency and asymptotic normality of the local correlation estimate $\widehat{\rho}_n(\boldsymbol{z})$, provided that the marginals of the observations are standard normally distributed.

Theorem 1. *Let $\{\boldsymbol{Z}_n\}$ be a sequence of bivariate iid random variables with with standard normal marginals. Assume that*

(i) for a sequence $\boldsymbol{h}_n$, $n = 1,2,\ldots$, converging to zero as n tends to infinity, there exists a unique minimiser ρ_0 of $q(\rho)$ such that $\rho_{h_n,K}(\boldsymbol{z})\to\rho_0(\boldsymbol{z})$,

(ii) the parameter space Θ for ρ is a compact subset of $(-1,1)$.

Then, for each $\boldsymbol{z}$ at which ρ_0 exists, $\widehat{\rho}_n(\boldsymbol{z}) \xrightarrow{P} \rho_0(\boldsymbol{z})$ as $n\to\infty$.

See Appendix A.1 for a proof of this result. The local correlation estimate is asymptotically normal according to the following theorem:

Theorem 2. *Denote by $f_{ij}(\boldsymbol{z})$ the joint density function of $\{\boldsymbol{Z}_n\}$. Assume that the conditions of Theorem 1 are satisfied, and further that*

(iii) the sequence of bandwidths, $\boldsymbol{h}_n$, satisfies $\boldsymbol{h}_n \to 0$, $\lim_n nh_{in}h_{jn} = \infty$, and

(iv) the kernel function satisfies $\sup_{\boldsymbol{z}} |K(\boldsymbol{z})| < \infty$, $\int |K(\boldsymbol{y})| d\boldsymbol{y} < \infty$, $\partial/\partial z_k K(\boldsymbol{z}) < \infty$ and $\lim_{z_k \to \infty} |z_k K(z_k)| = 0$ for $k = 1, 2$;

Then

$$\sqrt{nh_{in}h_{jn}}\,(\widehat{\rho}_n - \rho_0) \xrightarrow{\mathcal{L}} N(0, M/J^2),$$

where

$$M = f_{ij}(\boldsymbol{z}) \left(\int K^2(\boldsymbol{y})\, d\boldsymbol{y} \right), \quad J = u(\boldsymbol{z}, \rho_0(\boldsymbol{z}))\psi(\boldsymbol{z}, \rho_0(\boldsymbol{z})),$$

The preceding result is contained in the following, and more general, Theorem 3, regarding joint asymptotic normality of the local correlations $\{\widehat{\rho}_{ij,n}\}_{i<j}$. Assume now that the observations $\{\boldsymbol{Z}_n\}$ are p-variate with standard normal marginals, and that we calculate one local correlation for each pair of variables. There are $d = p(p-1)/2$ pairs, and denote by $\boldsymbol{\rho} = \{\rho_k\}_{k=1,\ldots,d}$ the vector of local correlations, and by $\widehat{\boldsymbol{\rho}}_n$ its estimate. In order to stress that $\boldsymbol{\rho}$ is a vector, and not a matrix, we use the single index k to identify the individual components. The matrix of bandwidths is defined as before by $\boldsymbol{h} = \text{diag}(h_1, \ldots, h_p)$, but the symbol h^2 now means the product of any two bandwidths which we do not need to specify in the asymptotic analysis, because we assume that they all tend to zero at the same rate.

Theorem 3. *Let $\{\boldsymbol{Z}_n\}$ be a sequence of p-variate iid marginally standard normal random variables. Enumerate each pair of variables by $k = 1, \ldots, d$, and for each of the pairs, calculate the local Gaussian correlation. Assume that the conditions (i) - (iv) of Theorems 1 and 2 are satisfied.*

The local Gaussian correlations are jointly asymptotically normal, with

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

where $\boldsymbol{\Sigma}$ is the diagonal matrix in which element (k,k) is the corresponding asymptotic variance M/J^2 that was defined in Theorem 2:

$$\boldsymbol{\Sigma}^{(k,k)} = \frac{f_k(\boldsymbol{z}_k) \int K^2(\boldsymbol{y}_k)\, d\boldsymbol{y}_k}{u^2(\boldsymbol{z}_k, \rho_{0,k}(\boldsymbol{z}_k))\psi^2(\boldsymbol{z}_k, \rho_{0,k}(\boldsymbol{z}_k))}. \tag{10}$$

See Appendix A.2 for proof.

The following lemma ensures that the asymptotic theory does not change, even though we in practice use marginally Gaussian pseudo-observations in the estimation of the local correlations:

Lemma 1. *Assume that conditions (i)-(iv) of Theorems 2 and 3 are satisfied, and assume further that*

(v) the marginal distribution functions $F_1, \ldots, F_p$ are strictly increasing on their support,

(vi) 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

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

and,

(vii) 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

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

Then, Theorems 2 and 3 hold when the marginally Gaussian variables $\boldsymbol{Z}_n$ are replaced with the pseudo-observations $\widehat{\boldsymbol{Z}}_n$ as defined by (2).

Assumptions (v)-(vii) are reproductions of the corresponding assumptions in Geenens et al. (2014), who discuss their reasonableness, and prove a very similar result. See Appendix A.3 for a proof. We are now ready to state our main asymptotic result regarding the convergence of the density estimate:

Theorem 4. *Assume that we fit the LGDE to a sequence of p-variate iid random variables $\{\boldsymbol{X}_n\}$ with density function $f(\boldsymbol{x})$. Assume that each pair of the transformed observation vectors $\{\boldsymbol{Z}_n\}$ satisfies conditions (i) - (vii) of Theorems 1 to 3 and Lemma 1. Assume further that*

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

Let $f_0(\boldsymbol{x})$ be the LGDE density function, which is obtained by replacing $f_Z(\cdot)$ with $\psi(\cdot, \boldsymbol{R}_0)$ in (1). Then, in all $\boldsymbol{x}$ for which $F_i(x_i) \in (0,1)$, $i = 1,\ldots,p$, with $\widehat{f}(\boldsymbol{x})$ estimated by the LGDE,

$$\sqrt{nh_n^2}\left(\widehat{f}(\boldsymbol{x}) - f_0(\boldsymbol{x})\right) \xrightarrow{\mathcal{L}} N(0, \psi(\boldsymbol{z}, R_0(\boldsymbol{z}))^2 g(\boldsymbol{x})^2 \boldsymbol{u}^T(\boldsymbol{z}, \boldsymbol{R}_0(\boldsymbol{z}))\boldsymbol{\Sigma}\, \boldsymbol{u}(\boldsymbol{z}, \boldsymbol{R}_0(\boldsymbol{z}))), \tag{11}$$

where

$$g(\boldsymbol{x}) = \prod f_i(x_i)/\phi(\Phi^{-1}(F_i(x_i))),$$

and

$$\boldsymbol{z} = \left\{\Phi^{-1}(F_i(x_i))\right\}_{i=1\ldots,p}.$$

We refer to Appendix A.4 for a proof.

The trade-off we make by choosing the LGDE for non-parametric density estimation is now evident. Not surprisingly, the approximation (6) keeps the convergence rate of the density estimate at $\sqrt{nh^2}$ regardless of the dimension, compared to the multivariate kernel estimator that converges as $\sqrt{nh^p}$. The price paid is that the LGDE converges to an approximation $f_0(\boldsymbol{x})$ of the unknown density $f(\boldsymbol{x})$, rather than $f(\boldsymbol{x})$ itself. Simulations provided in Section 5, however, indicate that the trade-off is very favourable for a large class of distributions.

Assumption (viii) of Theorem 4 is important to ensure that the estimated back-transformation of the density estimate does not influence the limiting distribution (11). We conclude this section by presenting a result that justifies our use of the logspline estimator by Stone et al. (1997) for this purpose.

Stone (1990) derives the asymptotic properties of the logspline estimator for density, distribution, and quantile functions. The results in that article are proven under the assumption that the unknown density has compact support, however, which is too restrictive for our purpose, but on the other hand, fairly simple to relax by using a truncation argument. Indeed, the following result replaces the compactness assumption with a condition on the marginal density functions limiting their tail thickness.

Theorem 5. *Denote by $\widehat{f}_i(\cdot)$ and $\widehat{F}_i(\cdot)$ the logspline estimates of the marginal density and distribution functions respectively. Assume that $f_i(\cdot)$ is twice continuously differentiable, and that there exist constants $M > 0$, $\epsilon \in (0, 1/2)$, $\gamma > 2\epsilon/(1-2\epsilon)$, and $x_0 > 0$ such that $f_i(x) \leq M|x|^{-(5/2+\gamma)}$ for all $|x| > x_0$, $i = 1, \ldots, p$. Then*

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

and

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

where the asymptotic variances σ_1^2 and σ_2^2 are specified by Stone (1990).

We refer to Appendix A.5 for a proof of this result and a discussion of the conditions. If we need γ to approach zero, we see from this result that the convergence rate of the logspline density estimate approaches $n^{-1/4}$. In that case, it follows immediately that we must also require the bandwidths to converge to zero fast enough so that $n^{1/2}h^2 \to 0$ in assumption (iii), in order for $n^{1/4}$ to dominate nh^2 in the limit.

4 Bandwidth selection

The general local likelihood density estimator by Hjort and Jones (1996) requires three distinct modelling choices to be made by the practitioner. She must pick (i) a parametric family $\psi(\cdot, \boldsymbol{\theta})$ for local approximation, (ii) a kernel function $K(\cdot)$, and (iii) a smoothing matrix $\boldsymbol{h}$.

We have already settled the first point. Transforming the marginals to standard normality leaves the standardised multivariate normal family (5) as the logical choice for the parametric family, with the additional restriction (6) to open up for high-dimensional applications. Points (ii) and (iii) are traditional non-parametric problems, but we argue that they have natural solutions when using the LGDE as well.

We use the bivariate Gaussian product kernel function $K(\boldsymbol{z}) = (2\pi)^{-1}\exp\{-\boldsymbol{z}^T\boldsymbol{z}/2\}$ for two reasons. First, K and ψ both being Gaussian functions means that the integral in the likelihood function (4) has a closed form expression, which greatly simplifies its numerical optimisation; Second, we will see below that the Gaussian kernel works very well in conjunction with our bandwidth selector. Previous developments in this paper imply that it is enough to look at the bivariate case.

There is a subtle difference between smoothing local likelihood- and kernel density estimates. As the bandwidth goes to infinity, the kernel estimate loses all structure and approaches zero at every point. The local likelihood estimate, on the other hand, is smoothed towards a global maximum likelihood fit by the parametric family. One can thus interpret bandwidth selection in the latter case as determining to which degree one believes the parametric family to be the true underlying distribution of the data.

In most practical situations, however, we need a data-driven bandwidth selection routine, and to this end, we adapt to our needs general schemes for model selection that already exist. The principle of cross-validation has been applied in many statistical methods. Stone (1974) provides a thorough treatment on the topic, Stone (1984) treats bandwidth selection for kernel density estimates by cross validation, and Berentsen and Tjøstheim (2014) use cross validation to select bandwidths for bivariate local likelihood density estimates. The latter authors note, however, that the procedure is sensitive to outliers, so raw data must be screened in advance. Hall (1987) investigates this phenomenon and shows that the kernel function and the true density must have approximately the same tail thickness for cross-validation to work properly. This is the second reason why the Gaussian kernel is such a natural choice for the LGDE; the density and the kernel both having Gaussian tails means that no screening of the data is needed.

The Kullback-Leibler divergence between the true density and its estimate is defined by

$$\begin{aligned}\mathrm{KL}\left(f, \widehat{f}\right) &= \int f(\boldsymbol{z}) \log\left\{f(\boldsymbol{z})/\widehat{f}(\boldsymbol{z})\right\}\,\mathrm{d}\boldsymbol{z} \\ &= \int f(\boldsymbol{z})\log f(\boldsymbol{z})\,\mathrm{d}\boldsymbol{z} - \int f(\boldsymbol{z})\log\widehat{f}(\boldsymbol{z})\,\mathrm{d}\boldsymbol{z},\end{aligned}$$

where the last term depends on the bandwidth. It can be estimated by cross-validation, and so for each pair of variables, we choose the bandwidth $\boldsymbol{h} = (h_1, h_2)$ that maximises

$$CV(\boldsymbol{h}) = n^{-1}\sum_{i=1}^{n}\log\widehat{f}_{\boldsymbol{h}}^{(-i)}(\boldsymbol{Z}_i),$$

where $\widehat{f}_{\boldsymbol{h}}^{(-i)}(\cdot)$ is the bivariate local Gaussian density estimate calculated using the bandwidth $\boldsymbol{h}$, and without the observation with index i.

We also obtain adaptive bandwidths using the k-nearest-neighbour strategy, for which the bandwidth used in a particular point $\boldsymbol{z}$ is taken to be the Euclidean distance to the kth nearest observation measured from $\boldsymbol{z}$. That way, we allow more details to appear in areas with much data while keeping a fairly large bandwidth in the tails of the distribution. We choose k using cross validation as above, as the maximiser of

$$CV(k) = n^{-1}\sum_{i=1}^{n}\log\widehat{f}_{k}^{(-i)}(\boldsymbol{Z}_i),$$

where $\widehat{f}_{k}^{(-i)}(\cdot)$ in the same way as above denotes the cross-validated density estimate that is calculated using as bandwidth the distance to the kth nearest neighbour of $\boldsymbol{Z}_i$.

1.	χ^2_3 marginals, Gaussian copula with all parameters equal to 0.5
2.	$t(10)$ marginals, Clayton copula with parameter 0.9
3.	Log-normal marginals, t-copula with 10 degrees of freedom
4.	Uniform marginals with observations taken directly from the Clayton copula with parameter 0.9
5.	Mixture of two Gaussians centred at $(0, \ldots)^T$ and $(4, \ldots)^T$ respectively
6.	Multivariate $t(4)$ distribution

Table 1: Test distributions

To avoid overfitting, we must keep k from becoming too small. In practice, and in the subsequent simulation experiments, we do this by requiring k to be at least 20, which for moderate sample sizes seems to be a reasonable number.

5 Simulated data

We have developed a routine in the R programming language (R Core Team, 2015) for the practical implementation of our estimator, which accompany this paper as supplementary material. In it, the logspline estimator has been used not only in the final back transformation of the density estimate, but also for estimating the marginal cumulative distribution functions that is used to produce the pseudo-observations (2). Strictly speaking, following Lemma 1 and its proof, the marginal empirical distribution functions should be used for this purpose, but the logspline estimator, in our experience thus far, has better finite sample properties. We believe that the asymptotic properties are the same, since different convergence speeds, which is the essence of the argument proving Lemma 1, still hold, as shown in Theorem 5.

The large sample properties in Section 3 show clearly that we trade asymptotic unbiasedness for faster convergence if we choose the LGDE instead of the kernel density estimator for multivariate data. We proceed to investigate the practical consequences of doing so in a series of controlled experiments using simulated data.

There are many ways to evaluate the performance of a density estimator. When introducing a new estimator, we seek a presentation that emphasises the advantageous aspects, as well as the fallacies one may encounter in practical applications. We believe that the LGDE enjoys two particularly beneficial properties that we wish to confirm:

- It approximates the unknown density by simplifying the dependence structure in a way that is exact for distributions having the Gaussian copula ($\boldsymbol{R}(\boldsymbol{z}) = \boldsymbol{R}$). Therefore, the LGDE should work particularly well for distributions for which the joint structure is not too far from normal. This is confirmed in our simulations, but it also works well for many non-Gaussian joint structures.

- In the tail of the distribution, where there is little or no data, the LGDE does what is perhaps most natural. It fits a Gaussian tail, based on the general direction towards the main body of the data. The influence from the data will not change much from point to point in the tail, nor will the local parameter estimates. The kernel estimator, on the other hand, assigns density estimates in the tail by adding up values far out in the tail of the kernels, which may well be zeroes if the kernel is compactly supported. This effect becomes increasingly troublesome as the number

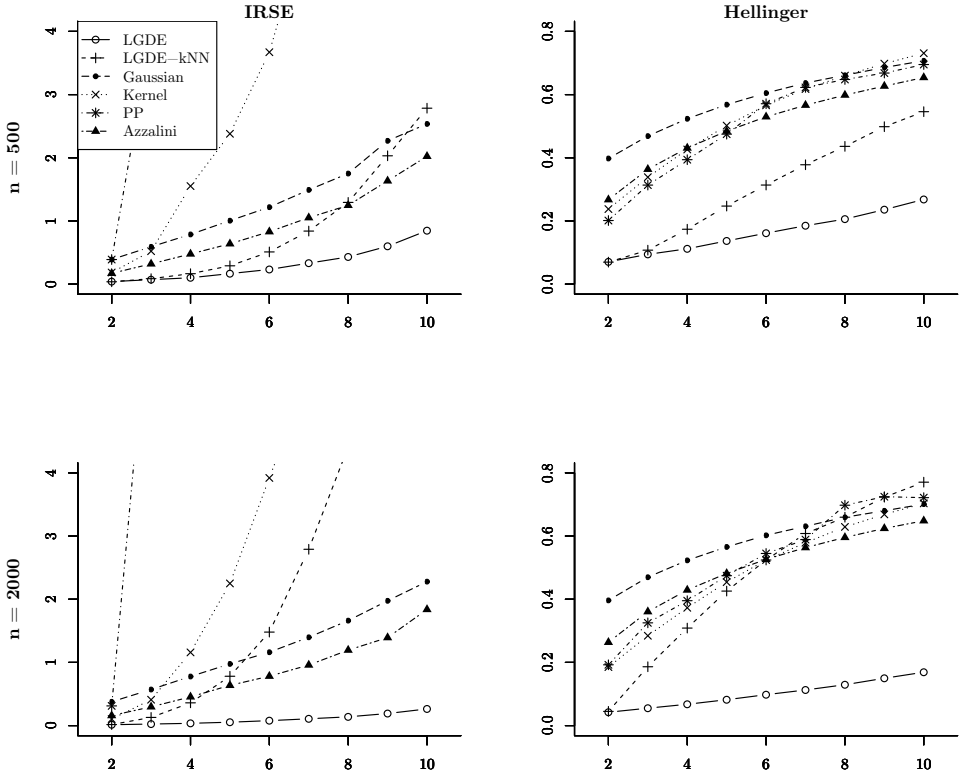


Figure 1: χ^2-distributed marginals with 3 degrees of freedom, Gaussian-copula with all correlations equal to 0.5.

of variables increases, and is indeed a demonstration of the Curse of Dimensionality (see e.g. Hastie et al. (2009), section 2.5).

We calculate density estimates for data from a selection of distributions (listed in Table 1) that can be generalised to higher dimensions in a natural way. These include various copula models, a multivariate t-distribution as well as a mixture of two Gaussians. We use the integrated *relative* squared error (IRSE) as a measure of discrepancy between the estimate and the true density because it is more natural to compare across dimensions than the more common ISE. Further, the relative error emphasises the performance in the tails. We also report the Hellinger distance (H, see Van der Vaart (2000), p. 211) from the density estimate to the true density, so that

$$IRSE(\widehat{f}) = \int \frac{\left(\widehat{f}(\boldsymbol{x}) - f(\boldsymbol{x})\right)^2}{f(\boldsymbol{x})}\,\mathrm{d}\boldsymbol{x},$$
$$H^2(\widehat{f}) = 1 - \int \sqrt{f(x)\widehat{f}(x)}\,\mathrm{d}x.$$

For each distribution listed in Table 1, we generate data sets comprising $n = 500$ and $n = 2000$ observations and estimate their density using the LGDE with the two bandwidth selection algorithms of Section 4 at $m = 4000$ grid points, $\{\boldsymbol{y}_j, j = 1, \ldots, m\}$,

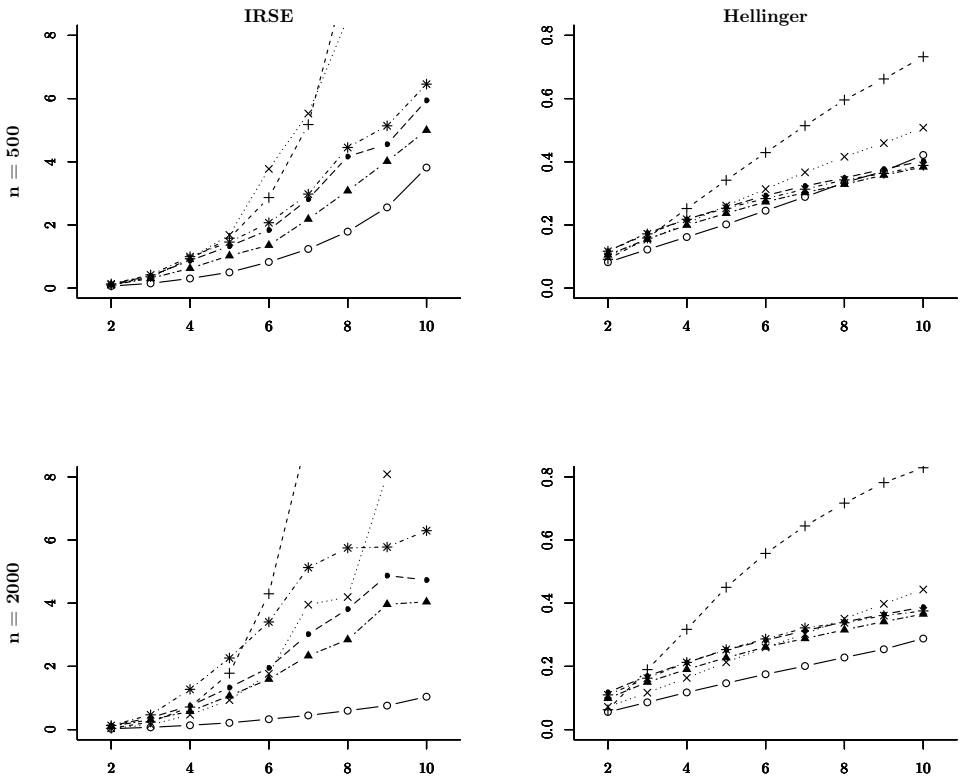


Figure 2: t-distributed marginals with 10 degrees of freedom, Clayton-copula with parameter equal to 0.9.

that we generate from the same distribution, but independently from the data. We repeat the procedure $2^7 = 128$ times, and report the median of the estimated IRSE and Hellinger error, which we obtain by Monte Carlo integration;

$$IRSE(\widehat{f}) \approx m^{-1} \sum_{j=1}^{m} \frac{\left(\widehat{f}(\boldsymbol{y}_j) - f(\boldsymbol{y}_j)\right)^2}{f(\boldsymbol{y}_j)^2},$$

$$H(\widehat{f}) \approx \sqrt{1 - m^{-1} \sum_{j=1}^{m} \sqrt{\widehat{f}(\boldsymbol{y}_j)/f(\boldsymbol{y}_j)}}.$$

We do the same for the kernel estimator, using a multivariate generalization of the cross-validation algorithm by Bowman (1984) (plug-in bandwidths give similar results, but are not included in the figures), the flexible, but parametric, skewed t-distribution by Azzalini (2005), as well as the Projection Pursuit algorithm (PP) by Friedman et al. (1984). PP estimates the univariate densities of a small number of highly non-Gaussian linear projections of the data, and uses these to build a multivariate density estimate. The latter is included for completeness and reference only, and we do point out that PP cannot be expected to fare well in our simulation study. First, Friedman et al. (1984)

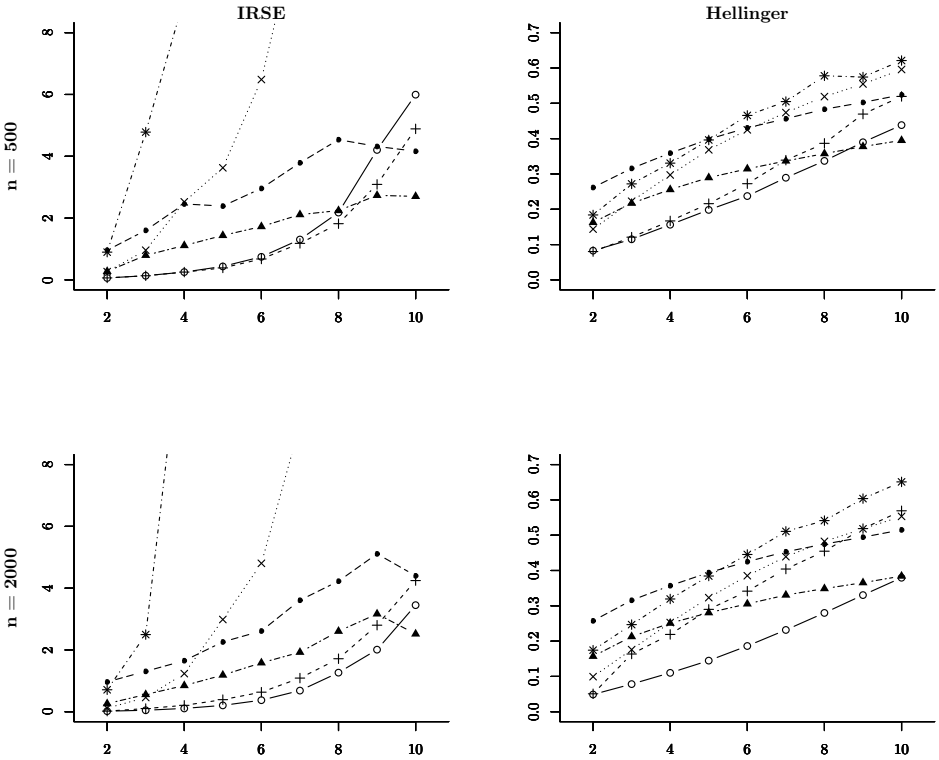


Figure 3: Log-normally-distributed marginals with $\mu = 0$ and $\sigma = 0.4$, t-copula with all correlations equal to 0.7 and 10 degrees of freedom.

state clearly that PP is inaccurate in the tails, which will be greatly emphasised by the IRSE. Second, PP seems to be very good at recovering sharp structures in high dimensional data, but all but one of our test distributions are unimodal, and do not have dramatic features. We also note here that the authors are not aware of publicly available software that chooses the optimal number of projections for the PP. In these experiments, we therefore choose the number of projections that actually minimises the error. At last, we compute the error of the *global* Gaussian fit and compare it to the *local* Gaussian fit, in order to quantify the severity of parametric miss-specification side by side with the Curse of Dimensionality.

Figures 1-6 display the results from our simulations. Each figure represents one distribution. The upper panels report results for the sample size $n = 500$, and the lower panels show results for $n = 2000$. The panels on the left hand side report IRSE, while the right hand panels display the Hellinger error. The horizontal axis represents the number of variables.

Let us briefly comment on the individual figures.

Fig. 1. The marginals are χ^2-distributed with 3 degrees of freedom, and the dependence is governed by the Gaussian copula. In this situation the simplification (6) is theoretically true, so the LGDE naturally outperforms all its competitors.

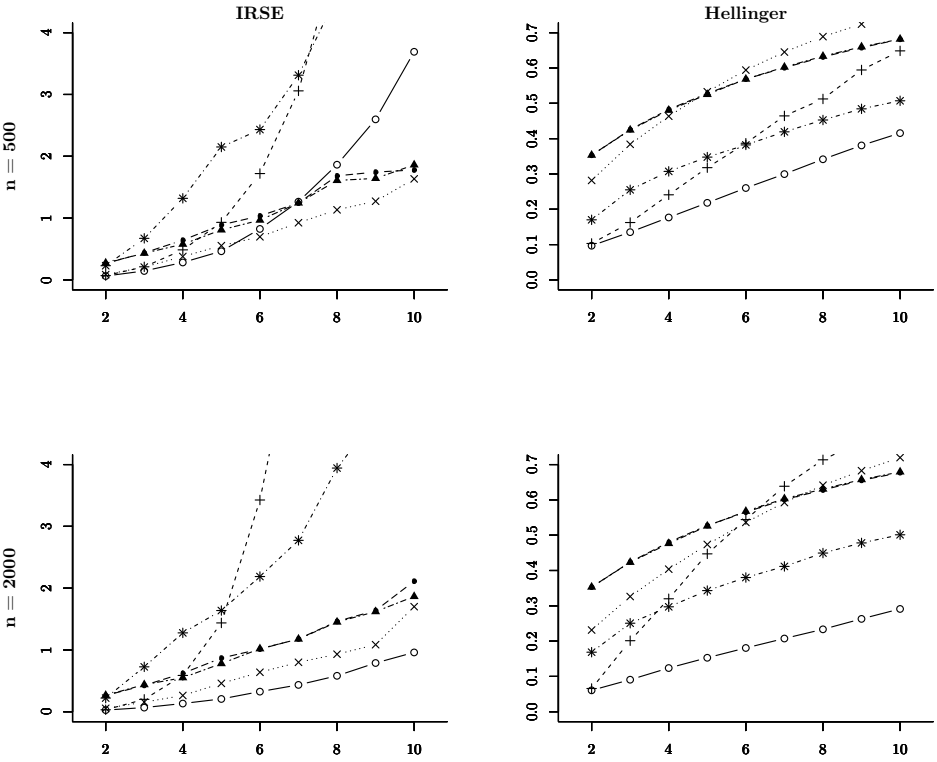


Figure 4: Uniform marginals with observations drawn directly from the Clayton copula in Figure 2.

Fig. 2. The marginals are t-distributed with 10 degrees of freedom, but due to the Clayton copula, the distribution is asymmetric even after the initial transformation. The LGDE with a global choice of bandwidths is clearly the best estimator if evaluated using IRSE or the Hellinger distance. Note that the parametric skewed t-distribution beats all other nonparametric competitors.

Fig. 3. We introduce asymmetrical marginals and use a t-copula with 10 degrees of freedom. The LGDE is the overall best performer.

Fig. 4. In this case, we generate observations directly from the Clayton copula, meaning that the marginals are uniformly distributed, and nonparametric methods can be expected to exhibit boundary issues. We see clearly that the LGDE with the global bandwidth selector is the best alternative here.

Fig. 5. Mixtures of distributions are not easy to recover under the restriction (6), but the LGDE performs reasonably well in this case. The PP has been shown to estimate the main body of mixture distributions very well (Hwang et al., 1994). This is the only example for which the kNN bandwidth selector performs acceptably.

Fig. 6. The LGDE does not seem to cope very well with the $t(4)$-distribution in higher dimensions. When weighing up the tail error, we see that fitting the Gaussian

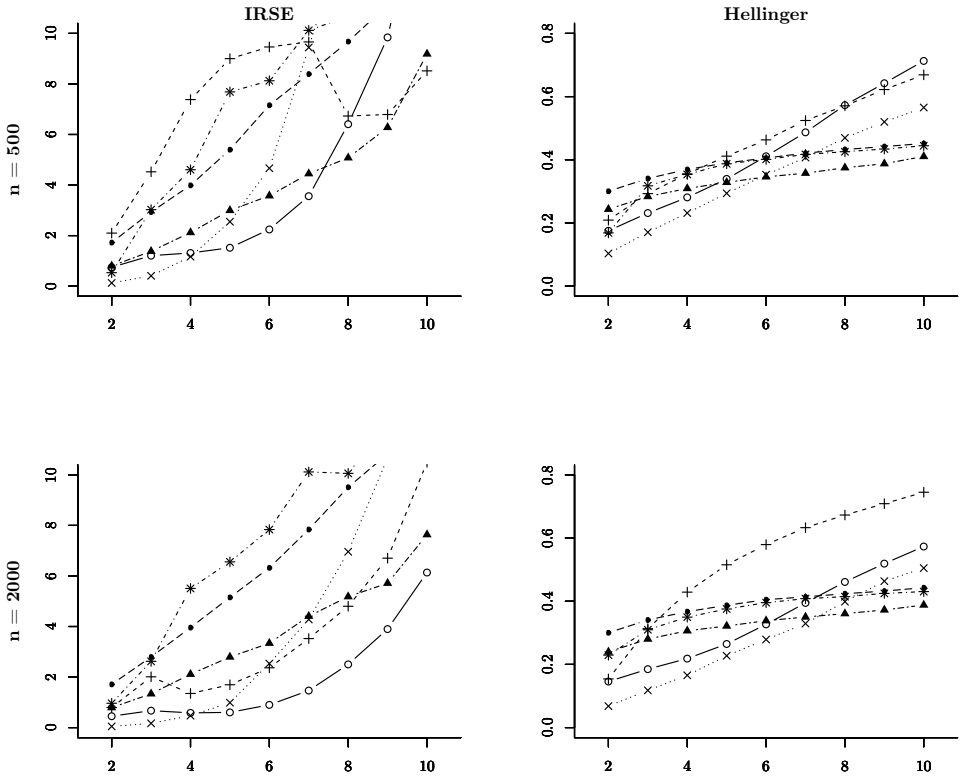


Figure 5: Mixture of two Gaussians; $0.7N(\mu = 0, \sigma = 1, \rho = 0.5) + 0.3N(\mu = 4, \sigma = 1, \rho = 0.1)$

distribution globally is actually better than a local Gaussian fit, suggesting that the cross-validation bandwidth is too small in this case. The skewed t-distribution is naturally the best estimator here, because it contains the true distribution as a special case. A local t-distribution estimator as discussed at the end of Section 2.3 would be expected to do better than the LGDE in this situation.

6 Real data

Let us apply the LGDE to a real data set. We have observed 1443 daily log-returns from January 3rd, 2005, until July 14th, 2010 for five stock indices: the S&P 500 in the US, the FTSE 100 in the UK, the German DAX 30, the French CAC 40 and, finally, the Norwegian OBX. These data have been analysed by Støve et al. (2014) using local Gaussian correlation in describing financial contagion. We estimate their joint density using the LGDE and the kernel estimator. The true density function is of course not available for a formal comparison, but a simple visual check will indicate some appealing properties of our approach. We point out that this data set does not satisfy the assumption of independence in our asymptotic results.

Figure 7a displays a cross section of the density estimate along a series of 2000 equally spaced points going diagonally through the data. The kernel estimator has been

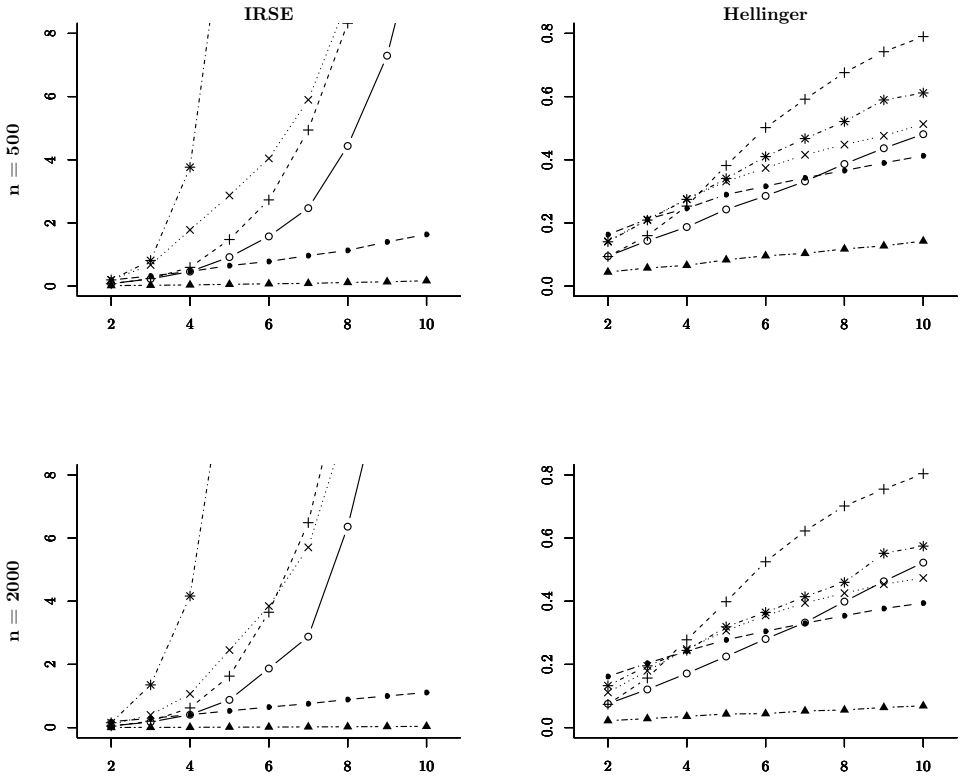
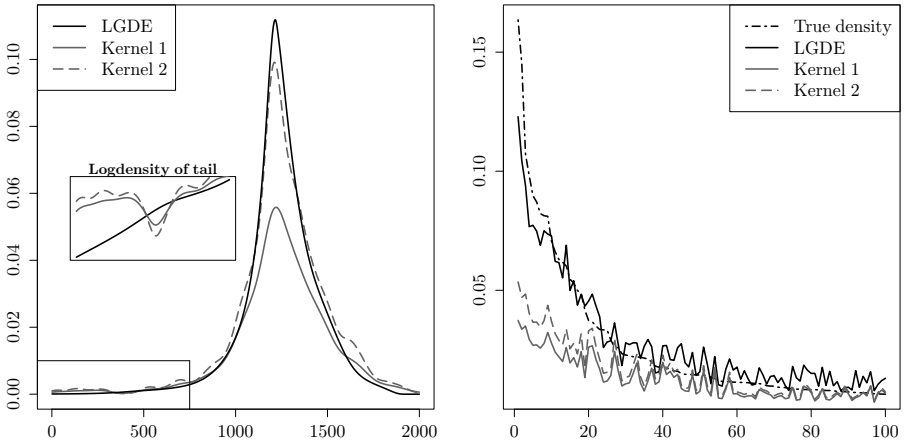


Figure 6: Multivariate t-distribution with 4 degrees of freedom.

calculated using the Wand and Jones (1994) plug-in bandwidth selection routine (Kernel 1), and we include a less smooth version, for which these bandwidths have been reduced by a factor of 1.5 (Kernel 2). The LGDE has a sharp peak, and it decays smoothly towards zero. The kernel estimator, on the other hand, does not seem to pick up the peak equally well unless we reduce the bandwidths. In that case, however, its tails are very wiggly and unstable, which is indicated in the smaller plot, where the logarithm of the left tails of the density estimates are displayed. We have also included some contour plots in Figure 7, and the same picture appears. The top two plots are projections of the LGDE estimate on the US-UK-axes and the France-Norway-axes. The other variables are held constant equal to zero. The second and third row show the kernel estimates with plug-in bandwidths and plug-in bandwidths divided by 1.5 respectively. Again, we do not know the true density, so it is hard to compare the quality of the estimates, but the LGDEs are altogether more pleasing; their tails are smooth, and, if inspected carefully, their main bodies display structure that are not visible in the kernel estimates. The Pearson correlation coefficients for the two pairs in question, are 0.52 and 0.82 respectively.

Another way to evaluate the performance is to apply the LGDE on new observations generated from a parametric model that has been fitted to the original data. If our choice of model is not too far off, the fitted parametric density presumably shares key characteristics with the true density. Pair Copula Constructions (PCC) are very flexible



(a) Density estimate of real data. The horizontal axis follows the index of the grid points. The box show the logarithms of the left tails.

(b) Density estimates for data generated from the fitted PCC-model. The variability does *not* indicate variability of the estimated densities.

in modelling high dimensional dependence, see Aas et al. (2009) for details. In particular, we fit a so-called R-vine to the original log-returns and generate new samples of the same size. We cannot plot five-dimensional densities in their entirety, but Figure 7b indicates our results. In order to create a suitable grid, we generate 1000 observations from the 5-variate Gaussian distribution with the same mean and covariance matrix as the original data, and evaluate the density function of the fitted PCC-model at these points; they are then sorted by decreasing density value. Finally, we pick the first 100 points, and plot their density value sequentially, as can be seen in the plot under the heading 'True density'. We then generate 1000 data sets from the PCC-model, and estimate the density at these points by the LGDE and the Kernel estimator, using the same bandwidths as in the left hand figure. The results at each grid point are averaged and plotted along with the true density in Figure 7b. Although the LGDE does not seem to coincide with the "true" density perfectly, it is evident that the LGDE does a much better job than the kernel estimator. We see clearly in this plot a substantial improvement over the kernel estimator in the center of the distribution. The variability in the curves does not imply non-smoothness of the estimated density surfaces, as subsequent grid points may be far from each other in space.

7 Discussion

Building on existing methods, we offer a new way to tackle the fundamental problem of non-parametric density estimation in higher dimensions. Instead of converging painfully slow to the correct answer, as the traditional kernel estimator does, or quickly to something potentially very wrong due to a parametric assumption, the LGDE converges much faster to something potentially much less wrong. We observe this phenomenon in our subsequent analysis of simulated and real data.

The LGDE is not perfect; it cannot be, since the Curse of Dimensionality forces

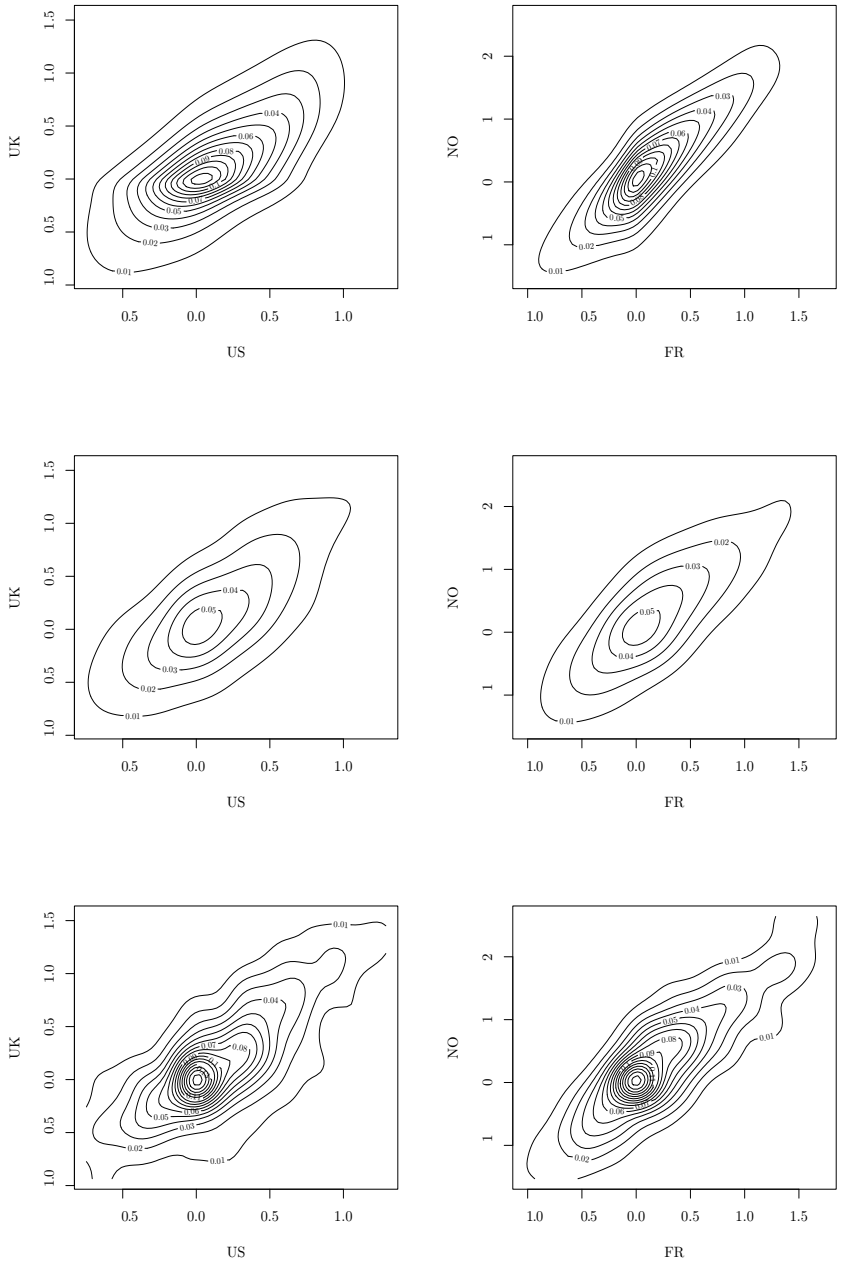


Figure 7: Projections of density estimates of the log-return data. The first two variables in the left hand plots, the last two variables in the right hand plots. The other variables are constant and equal to zero. The first row are LGDE estimates, the second row shows kernel estimates using a plug-in bandwidth selector. The third row of plots displays kernel estimates also, but all bandwidths have been reduced by a factor of 1.5.

us to compromise, and there might be other ways to do just that. It is tempting to search for further analogies to the nonparametric regression setting, in which one can include higher order interaction terms in order to improve the fit. There is, however, no obvious way to do the same thing for the LGDE, as we depend crucially on the pairwise structure of the covariance matrix of the Gaussian distribution. This, on the other hand, points to another possible extension as indicated at the end of Section 2.3. The general elliptical distribution has the characteristic function $\exp(i\boldsymbol{\mu}^T\boldsymbol{t})\Psi(\boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t})$. Its density function is symmetric about $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$ is a symmetric, positive definite matrix that is proportional to the covariance matrix if it exists. For the Gaussian distribution we have that $\Psi(x) = \exp(-x/2)$, but our trick does not depend on any particular choice of Ψ, but rather the covariance structure of its argument. The elliptical distributions could therefore, in principle, replace the Gaussian in the LGDE, and we conjecture that the results for heavy-tailed distributions will be improved if, for example, the t-distribution with a fixed degree of freedom that is estimated from the data, is implemented using our method.

The practitioner must keep in mind two issues when applying the LGDE to a multivariate data set. First, the density estimate does not integrate to one by definition, and so it must be normalised in order to be a proper density function. This is true also for the Projection Pursuit estimates, and can be accomplished by straightforward Monte Carlo integration. Second, we compute the elements of the local correlation matrix individually, so it is not positive definite by definition. In practise, however, the authors have not experienced serious problems as a result. When the number of dimensions is less than 10 or so, the local correlation matrix is positive definite at most grid points, except, perhaps, some in the far tail of the distribution. If the number of variables increases towards 20, one might experience negative definite correlation matrices at central points. One can then increase the bandwidths slightly until the problem goes away. In one example, the authors fitted the LGDE to a 19-variate data set, and obtained positive definite local correlation matrices at almost all grid points by multiplying the cross-validation bandwidths by 1.5.

By its very construction, the LGDE works best when estimating densities that share key characteristics with the Gaussian, such as unimodality and simple dependence structures. Our simplification (6) does not necessarily provide an optimal description of the Gaussian mixture that is the subject of estimation in Figure 5. Although not obvious from our particular choices of discrepancy measures, this is a typical case for which the Projection Pursuit algorithm will give informative results (see e.g. Hwang et al. (1994)). There is a potential for synergy between the LGDE and PP here. One can use PP as a first exploratory step to reveal multimodality in the distribution. If the least Gaussian projection of the data is unimodal, we will apply the LGDE with some confidence; if not, we could estimate the location and the weight of a mixture using PP, and estimate the individual components by the LGDE. Another example that is far from the Gaussian, but for which the LGDE works well, is that of Figure 4.

Finally, we acknowledge that the kernel density estimator has been around for a long time, and that many improvements have been made upon it such as variable bandwidths and higher order kernels. Politis and Romano (1999) use infinite order kernels in order to lessen the impact of the curse of dimensionality asymptotically. We have compared the LGDE to the basic kernel estimator, however, because most such improvements can be applied directly to the local likelihood case as well. For instance, Otneim et al. (2013) show that the bias corrections by Jones (1993) for densities with bounded support carry

over to the Hjort and Jones (1996) local likelihood unaltered. The implementations of such improvements for the LGDE may be the topic of later studies.

A Proofs

A.1 Proof of Theorem 1

The method of proof is the same as that of Severini (2000, pp. 105-107) for ordinary maximum likelihood estimates. The proof requires the additional assumption of uniform convergence in probability of the local likelihood function:

$$\sup_{\rho\in\Theta} |L_n(\rho, \boldsymbol{Z}) - q_{\boldsymbol{h}_n,K}(\rho)| \xrightarrow{P} 0 \quad \text{as} \quad n \to \infty. \tag{12}$$

The bivariate version of (5) satisfies condition (12) provided that condition (ii) is fulfilled. To see this, consider $\psi(\cdot, \rho)$ as a function of the parameter; it is bounded and differentiable to any order on the compact set Θ, and so is its logarithm. Thus $g(\rho) = \log \psi(\cdot, \rho) f(\cdot) - \psi(\cdot; \rho)$ is uniformly continuous there, so for every $\epsilon > 0$ there exists a $\delta > 0$ such that if $|\rho_1 - \rho_2| < \delta$ then $|g(\rho_1) - g(\rho_2)| < \epsilon$. Multiplying with a kernel and integrating over a different variable conserves this property, because if $|\rho_1 - \rho_2| < \delta$, then

$$\begin{aligned}\left| \int K_{\boldsymbol{h}_n}(\boldsymbol{y} - \boldsymbol{z}) g(\rho_1) \, \mathrm{d}\boldsymbol{y} - \int K_{\boldsymbol{h}_n}(\boldsymbol{y} - \boldsymbol{z}) g(\rho_2) \, \mathrm{d}\boldsymbol{y} \right| &\leq \int K_{\boldsymbol{h}_n}(\boldsymbol{y} - \boldsymbol{z}) |g(\rho_1) - g(\rho_2)| \, \mathrm{d}\boldsymbol{y} \\ &< \epsilon \int K_{\boldsymbol{h}_n}(\boldsymbol{y} - \boldsymbol{z}) \, \mathrm{d}\boldsymbol{y} = \epsilon.\end{aligned}$$

The ϵ and δ do not depend on $\boldsymbol{h}$ nor n, so $\{q_{\boldsymbol{h}_n,K}(\rho)\}$ form an equicontinuous family of functions. Further, and again exploiting the smoothness of $\psi(\cdot, \rho)$ on a compact set Θ, the local likelihood functions are Lipschitz continuous there by the mean value theorem. The conditions in Corollary 2.2 by Newey (1991) are thus satisfied, and condition (12) follows thereof. It follows from the uniform convergence that

$$\sup_{\rho\in\Theta} L_n(\rho, \boldsymbol{Z}) = L_n(\widehat{\rho}, \boldsymbol{Z}) \xrightarrow{P} \sup_{\rho\in\Theta} q(\rho) = q(\rho_0).$$

The rest of the argument follows exactly that of Severini (2000) pp. 105-107 for ordinary maximum likelihood estimates.

A.2 Proof of Theorems 2 and 3

We establish joint asymptotic normality of the local correlation vector by first following the standard argument for ordinary maximum likelihood estimates in the bivariate, and thus one-parameter case, and then apply a central limit argument, which amounts to a proof of Theorem 2. Then we make use of the Cramèr-Wold device to include the multi-parameter case. In the end, we show that the off-diagonal elements in the covariance matrix vanish asymptotically. In the bivariate case, we must verify the following conditions in order to use Theorem 7.63 in Schervish (1995) and Theorem 1A of Parzen (1962):

(I) The parametric family $\psi(\boldsymbol{z}, \rho)$ is continuously differentiable with respect to ρ;

(II) $\int |u(\boldsymbol{z}, \rho_0) f(\boldsymbol{z})| < \infty$;

(III) There exists a function $T_r(\boldsymbol{z}, \rho)$ such that for each $\rho_0 \in \text{int}(\Theta)$ and,

$$\sup_{|\rho - \rho_0| \leq r} \left| \partial^2 L_n(\rho_0, \boldsymbol{z})/\partial\rho^2 - \partial^2 L_n(\rho, \boldsymbol{z})/\partial\rho^2 \right| \leq T_r(\boldsymbol{z}, \rho_0)$$

with $\lim_{r \to 0} \text{E} T_r(\boldsymbol{Z}, \rho_0) = 0$ (stochastic equicontinuity).

The parametric family is Gaussian, so condition (I) is obviously true. The local score function $u(\boldsymbol{z}, \rho) = \partial \log \psi(\boldsymbol{z}, \rho)/\partial\rho$ in the bivariate Gaussian case is given by

$$u(z_1, z_2, \rho) = \frac{\rho^3 - z_1 z_2 (1 + \rho^2) + (z_1^2 + z_2^2 - 1)\rho}{(1 - \rho^2)^2}, \tag{13}$$

and the stochastic variable $\boldsymbol{Z} = (Z_1, Z_2)$, having density $f_{\boldsymbol{Z}}$, has moments of all orders since the marginals are standard normal. Therefore, $\text{E}|u(\boldsymbol{Z}, \rho)| < \infty$, so (II) is satisfied. Further, Andrews (1992) shows that uniform continuity of $\partial^2 L_n(\rho)/\partial\rho^2$ as well as Lipschitz continuity of $|\partial^2 L_n(\rho, \boldsymbol{z})/\partial\rho^2 - \partial^2 L(\rho_0, \boldsymbol{z})/\partial\rho^2|$ suffice for stochastic equicontinuity as required in condition (III). The argument in Appendix A.1 goes through also here.

Using a one-term Taylor expansion of the local score function $\partial L_n(\widehat{\rho}, \boldsymbol{z})/\partial\rho$, and following Schervish (1995), p. 422, in writing 0 as $o_P\left((nh_n^2)^{-1/2}\right)$, we get

$$\partial L_n(\rho_0, \boldsymbol{z})/\partial\rho + B_{n,h}(\widehat{\rho}_n - \rho_0) = o_P\left((nh_n^2)^{-1/2}\right),$$

where $B_{n,h} = \partial^2 L_n(\rho^*, \boldsymbol{z})/\partial\rho^2$, and ρ^* lies between ρ_0 and $\widehat{\rho}$. As $n \to \infty$, this quantity tends to its expectation, which we denote by J_h, and is given by

$$\begin{aligned} J_h = &\int K_{\boldsymbol{h}}(\boldsymbol{y} - \boldsymbol{z}) u^2(\boldsymbol{y}, \rho^*(\boldsymbol{z})) \psi(\boldsymbol{y}, \rho^*(\boldsymbol{z}))\, \mathrm{d}\boldsymbol{y} \\ &- \int K_{\boldsymbol{h}}(\boldsymbol{y} - \boldsymbol{z}) u'(\boldsymbol{y}, \rho^*(\boldsymbol{z}))[f(\boldsymbol{y}) - \psi(\boldsymbol{y}, \rho^*(\boldsymbol{z}))]\, \mathrm{d}\boldsymbol{y}. \end{aligned} \tag{14}$$

The arguments of Hjort and Jones (1996), as well as the consistency of $\widehat{\rho}_n$, can be used to see that,

$$J = \lim_{\boldsymbol{h} \to 0} J_h = u^2(\boldsymbol{z}, \rho_{0,k}) \psi(\boldsymbol{z}, \rho_{0,k}).$$

Further, the variance of $\sqrt{nh^2}\partial L_n(\rho_0, \boldsymbol{Z})/\partial\rho$ approaches M_h as $n \to \infty$, where

$$\begin{aligned} M_h = h_1 h_2 &\int (h_1 h_2)^{-2} K^2(\boldsymbol{h}^{-1}(\boldsymbol{y} - \boldsymbol{z})) u^2(\boldsymbol{y}, \rho_{k,0}(\boldsymbol{z})) f(\boldsymbol{y})\, \mathrm{d}\,\boldsymbol{y} \\ &- h_1 h_2 \left(\int K_{\boldsymbol{h}}(\boldsymbol{y} - \boldsymbol{z}) u(\boldsymbol{y}, \rho_0(\boldsymbol{z})) f(\boldsymbol{y})\, \mathrm{d}\boldsymbol{y} \right)^2. \end{aligned}$$

The second term vanishes as $\boldsymbol{h} \to 0$, so we have in the limit that

$$M = \lim_{\boldsymbol{h} \to 0} M_h = u^2(\boldsymbol{z}, \rho_0(\boldsymbol{z})) f(\boldsymbol{z}) \int K^2(\boldsymbol{y})\, \mathrm{d}\boldsymbol{y}.$$

Following the details of Theorem 7.63 in Schervish (1995), it follows that

$$\sqrt{nh_n^2}\,(\widehat{\rho}_n - \rho_0) \xrightarrow{\mathcal{L}} N(0, M/J^2),$$

provided that the quantity

$$Y_n(\boldsymbol{z}) = \frac{1}{n}\sum_{i=1}^{n} K_{\boldsymbol{h}_n}(\boldsymbol{Z}_i - \boldsymbol{z})u(\boldsymbol{Z}_i, \rho_0) = \frac{1}{n}\sum_{i=1}^{n} V_{ni}, \tag{15}$$

is asymptotically normal, and this follows along the lines of Parzen (1962), which we now proceed to establish.

Of the two terms comprising the local likelihood function (4), only the first depends on data. It follows readily from Theorem 1A of Parzen (1962) that the variance of the summands in (15), all identically distributed as $V_n = K_{\boldsymbol{h}_n}(\boldsymbol{Z} - \boldsymbol{z})u(\boldsymbol{Z}, \rho_0)$, satisfies

$$h_n^2 \mathrm{Var}(V_n) \to f_{\boldsymbol{Z}}(\boldsymbol{z})u^2(\boldsymbol{z}, \rho_0)\int_{-\infty}^{\infty} K^2(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y}. \tag{16}$$

Further, a simple Taylor expansion reveals that

$$\begin{aligned}
\mathrm{E}|V_n|^{2+\delta} &= \int_{-\infty}^{\infty} |K_{\boldsymbol{h}_n}(\boldsymbol{y} - \boldsymbol{z})\, u(\boldsymbol{y}, \rho_0)|^{2+\delta} f(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y} \\
&= \frac{1}{(h_{n1}h_{n2})^{1+\delta}} f_{\boldsymbol{Z}}(\boldsymbol{z})|u(\boldsymbol{z}, \rho_0)|^{2+\delta}\int_{-\infty}^{\infty} |K(\boldsymbol{y})|^{2+\delta}\,\mathrm{d}\boldsymbol{y} \\
&\qquad + \text{ higher order terms.}
\end{aligned} \tag{17}$$

The quantity in (16) is finite because of assumption (iv) in Theorem 2. Further,

$$\frac{\mathrm{E}\,|V_n - \mathrm{E}(V_n)|^{2+\delta}}{n^{\delta/2}\sigma^{2+\delta}(V_n)} = \frac{(h_{n1}h_{n2})^{1+\delta}\mathrm{E}|V_n - \mathrm{E}(V_n)|^{2+\delta}}{(nh_{n1}h_{n2})^{\delta/2}(h_{n1}h_{n2})^{1+\delta/2}\sigma^{2+\delta}(V_n)}, \tag{18}$$

which tends to zero as $n \to \infty$ because of (16), (17) and the second part of assumption (iv), and where, for a stochastic variable X, here and in the sequel we use the notation $\sigma(X) = \mathrm{sd}(X)$. The summands comprising $Y_n(\boldsymbol{z})$ therefore satisfy the Lyapunov, and thus the Lindeberg, condition, so $Y_n(\boldsymbol{z})$ is asymptotically normal.

Having established asymptotic normality for each $\widehat{\rho}_k$ (and proven Theorem 2), we extend the argument above to the p-variate, and thus $d = p(p-1)/2$-parameter, case; let $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_d)$ be the vector of local correlations, let $\boldsymbol{u}(\boldsymbol{z}, \boldsymbol{\rho}_0) = (u_1(\boldsymbol{z}, \boldsymbol{\rho}_0), \ldots, u_d(\boldsymbol{z}, \boldsymbol{\rho}_0))$ be the vector of score functions, defined before as $u_k(\boldsymbol{z}, \boldsymbol{\rho}) = \partial\psi(\boldsymbol{z}, \boldsymbol{\rho})/\partial\rho_k$, and, finally, note that $\boldsymbol{Y}_n(\boldsymbol{z}) = n^{-1}\sum_{i=1}^{n}\boldsymbol{V}_{ni}$ is now a stochastic vector, so that $\boldsymbol{Y}_n(\boldsymbol{z}) = \{Y_{nk}(\boldsymbol{z})\}_{k=1}^{d}$ and $\boldsymbol{V}_{ni} = \{V_{nik}\}_{k=1}^{d}$.

We proceed to show that

$$\sum_k t_k Y_{nk}(x) \xrightarrow{\mathcal{L}} \sum_k t_k Z_k^*, \tag{19}$$

where $\boldsymbol{t} = (t_1, \ldots, t_d)$ and $\boldsymbol{Z}^* = (Z_1^*, \ldots, Z_d^*)$ are an arbitrary vector of constants, and a jointly normally distributed vector respectively. Asymptotic normality of the vector $\boldsymbol{Y}_n(\boldsymbol{z})$ then follows from the Cramèr-Wold device (Billingsley (2008), p. 383). First, if $\boldsymbol{t}\boldsymbol{Y}_n(\boldsymbol{z})$ is asymptotically normal at all, it must converge to $\boldsymbol{t}\boldsymbol{Z}^*$ because of Slutsky's theorem and the asymptotic normality of each of the Y_{nk}. The normality of $\boldsymbol{t}\boldsymbol{Y}_n(\boldsymbol{z})$ follows immediately from the one-dimensional case by writing $W_{ni} = \sum_{k=1}^{d} t_k V_{nik}$ so that $\sum_{k=1}^{d} t_k Y_{nk}(\boldsymbol{z}) = \sum_{i=1}^{n} W_{ni}$, where all summands are identically distributed as $W_n = \sum_{k=1}^{d} t_k K_{\boldsymbol{h}_n}(\boldsymbol{Z} - \boldsymbol{z})u_k(\boldsymbol{z}, \boldsymbol{\rho_0}) = \sum_{k=1}^{d} t_k V_{nk}$. Jensen's inequality implies $|\sum_{k=1}^{d} Z_k|^{2+\delta} \leq d^{1+\delta}\sum_{k=1}^{d}|Z_k|^{2+\delta}$, and so

$$\frac{\mathrm{E}\,|W_n - \mathrm{E}(W_n)|^{2+\delta}}{n^{\delta/2}\sigma^{2+\delta}(W_n)} = \frac{\mathrm{E}\,|\sum t_k V_{nk} - \mathrm{E}(\sum t_k V_{nk})|^{2+\delta}}{n^{\delta/2}\sigma^{2+\delta}(\sum t_k V_{nk})}$$
$$\leq d^{1+\delta} \sum_{k=1}^{d} \frac{|t_k|^{2+\delta}(h_{n1}h_{n2})^{1+\delta}\mathrm{E}|V_{nk} - \mathrm{E}(V_{nk})|^{2+\delta}}{(nh_{n1}h_{n2})^{\delta/2}(h_{n1}h_{n2})^{1+\delta/2}\sigma^{2+\delta}(\sum t_k V_{nk})}. \tag{20}$$

Recall that all variables are on the same Gaussian scale, and that all bandwidths tend to zero at the same rate. Therefore, it does not matter which bandwidths we use in the above expression. Further, the variance in the denominator of (20) stays away from zero because of (16). Following the same reasoning as in the univariate case (18), the Lyapunov condition is satisfied for the W_n, implying (19), and so the vector $\boldsymbol{Y}_n(x)$ is jointly asymptotically normal.

It remains to show that the asymptotic covariance matrix is diagonal. Indeed, the covariance between two local correlation estimates *with no common index* goes to zero as n^{-1}. If they share a common index, one can go through the arguments below once again and see that their covariance $\mathrm{Cov}(\widehat{\rho}_{ij}, \widehat{\rho}_{jk})$ tends to zero as $(nh_n)^{-1}$. Both rates are negligible compared to $(nh_n^2)^{-1}$.

Assume without loss of generality that we have four variables $Z_1, \ldots, Z_4$ with joint density $f_{\boldsymbol{Z}}(\boldsymbol{z})$ and that we estimate the local correlations $\widehat{\rho}_{12}$ and $\widehat{\rho}_{34}$ according to the scheme described in Section 2. Again, we identify the parameters with single indices, so that we in this case have $\boldsymbol{\rho} = (\rho_1, \rho_2)$. They are estimated independently from each other by maximising the local likelihood functions $L_{n,1}(\rho_1, Z_1, Z_2)$ and $L_{n,2}(\rho_2, Z_3, Z_4)$, as defined by Equation (4). Taylor-expanding the estimation equations $L_{n,1} = 0$ and $L_{n,2} = 0$ about the population values $\rho_{1,0}$ and $\rho_{2,0}$ respectively, yields

$$0 = \begin{pmatrix} \partial L_{n,1}(\widehat{\rho}_1)/\partial\rho_1 \\ \partial L_{n,2}(\widehat{\rho}_2)/\partial\rho_2 \end{pmatrix} = \begin{pmatrix} S_1(\widehat{\rho}_1) \\ S_2(\widehat{\rho}_2) \end{pmatrix}$$
$$= \begin{pmatrix} S_1(\rho_{1,0}) \\ S_2(\rho_{2,0}) \end{pmatrix} + \begin{pmatrix} \partial S_1(\rho_1^*)/\partial\rho_1 & 0 \\ 0 & \partial S_2(\rho_2^*)/\partial\rho_2 \end{pmatrix} \begin{pmatrix} \widehat{\rho}_1 - \rho_{1,0} \\ \widehat{\rho}_2 - \rho_{2,0} \end{pmatrix},$$

where ρ_k^* again lies between $\widehat{\rho}_k$ and $\rho_{k,0}$. More compactly, we write

$$(nh^2)^{1/2}(\widehat{\boldsymbol{\rho}} - \boldsymbol{\rho}_0) = -\boldsymbol{U}^{-1}(\boldsymbol{\rho}^*)(nh_n^2)^{1/2}\boldsymbol{S}(\boldsymbol{\rho}_0)$$

where $\boldsymbol{U}$ is the diagonal matrix of derivatives. The non-zero elements in $\boldsymbol{U}$ converge, as $n \to \infty$ and $h \to 0$, to the quantities J_1 and J_2, which we have seen to be

$$J_k = u_k^2(\boldsymbol{z}_k, \rho_{0,k})\psi(\boldsymbol{z}_k, \rho_{0,k}), \qquad k = 1, 2.$$

Denote by $\boldsymbol{M}_h$ the covariance matrix of $\sqrt{nh^2}\boldsymbol{S}(\boldsymbol{\rho}_0)$. The diagonal elements of $\boldsymbol{M_h}$ are given by

$$M_k = u_k^2(\boldsymbol{z}_k, \rho_{k.0}(\boldsymbol{z}_k))f_k(\boldsymbol{z}_k) \int K^2(\boldsymbol{y}_k)\,\mathrm{d}\boldsymbol{y}_k.$$

The off-diagonal element in $\boldsymbol{M_h}$ is $O(h^2)$, because

$$\boldsymbol{M}_{\boldsymbol{h}}^{(1,2)} = \boldsymbol{M}_{\boldsymbol{h}}^{(2,1)}$$
$$= h^2 \int K_{\boldsymbol{h}}(\boldsymbol{y}_1 - \boldsymbol{z}_1)K_{\boldsymbol{h}}(\boldsymbol{y}_2 - \boldsymbol{z}_2)u_1(\boldsymbol{y_1}, \rho_{1,0}(\boldsymbol{z_1}))u_2(\boldsymbol{y_2}, \rho_{2,0}(\boldsymbol{z_2}))f_Z(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y}$$
$$- \text{a higher order term,}$$

Writing $\boldsymbol{J}_h = \text{diag}(J_{\boldsymbol{h},1}, J_{\boldsymbol{h},2})$, where $J_{\boldsymbol{h},k}$ was defined in (14), we collect these results and write the covariance matrix of $\sqrt{n\boldsymbol{h}^2}(\widehat{\rho}_1, \widehat{\rho}_2)^T$ in terms of its asymptotic order;

$$\begin{aligned}\boldsymbol{J}_{\boldsymbol{h}}^{-1}\boldsymbol{M}_{\boldsymbol{h}}(\boldsymbol{J}_{\boldsymbol{h}}^{-1})^T &\sim \begin{pmatrix} J_{\boldsymbol{h},1}^{-1} & 0 \\ 0 & J_{\boldsymbol{h},1}^{-1} \end{pmatrix} \begin{pmatrix} M_{\boldsymbol{h},1} & h^2 \\ h^2 & M_{\boldsymbol{h},2} \end{pmatrix} \begin{pmatrix} J_{\boldsymbol{h},1}^{-1} & 0 \\ 0 & J_{\boldsymbol{h},2}^{-1} \end{pmatrix} \\ &\to \begin{pmatrix} M_1/J_1^2 & 0 \\ 0 & M_2/J_2^2 \end{pmatrix},\end{aligned}$$

as $h \to 0$, indicating that the asymptotic covariance between $\widehat{\rho}_1$ and $\widehat{\rho}_2$ tends to zero as n^{-1}. The same procedure must be repeated in order to establish $\text{Cov}(\widehat{\rho}_{ij}, \widehat{\rho}_{jk}) = O((nh)^{-1})$.

A.3 Proof of Lemma 1

By inspecting the preceding proof of Theorems 2 and 3, we see that Lemma 1 holds if the asymptotic distribution of $Y_n(\boldsymbol{z})$ in (15) remains unchanged when we replace the marginally standard normal observations $\boldsymbol{Z}_n$ with their pseudo-observations $\widehat{\boldsymbol{Z}}_n$ as defined by (2). Apart from the factor $u(\cdot)$, this is exactly the same expression as analysed in Proposition 3.1 by Geenens et al. (2014), so we proceed to show that this difference does not alter their proof in any other way than a little more complicated algebraic expressions.

We have assumed the bivariate kernel function to be the product of two univariate kernels, so write in this section $K(\boldsymbol{z}) = K(z_1)K(z_2)$, even though that is a slight abuse of notation. Further, and following the notation of Geenens et al. (2014), write

$$\begin{aligned}J_{\boldsymbol{z},h}(\boldsymbol{v}) = K\left(\frac{z_1 - \Phi^{-1}(v_1)}{h}\right) K\left(\frac{z_2 - \Phi^{-1}(v_2)}{h}\right) \\ \times u(\Phi^{-1}(\boldsymbol{v})),\end{aligned}$$

where $\boldsymbol{v} = (v_1, v_2) \in (0,1)^2$. Thus, writing $k(z) = K'(z)$ and $u_i(\boldsymbol{z}) = \partial u(\boldsymbol{z})/\partial z_i$, we have

$$\begin{aligned}\frac{\partial J_{\boldsymbol{z},h}}{\partial v_1} &= k\left(\frac{z_1 - \Phi^{-1}(v_1)}{h}\right) K\left(\frac{z_2 - \Phi^{-1}(v_2)}{h}\right) \frac{u(\Phi^{-1}(\boldsymbol{v}))}{h\phi(\Phi^{-1}(v_1))} \\ &\quad + K\left(\frac{z_1 - \Phi^{-1}(v_1)}{h}\right) K\left(\frac{z_2 - \Phi^{-1}(v_2)}{h}\right) \frac{u_1(\Phi^{-1}(\boldsymbol{v}))}{\phi(\Phi^{-1}(v_1))}, \\ \frac{\partial J_{\boldsymbol{z},h}}{\partial v_2} &= K\left(\frac{z_1 - \Phi^{-1}(v_1)}{h}\right) k\left(\frac{z_2 - \Phi^{-1}(v_2)}{h}\right) \frac{u(\Phi^{-1}(\boldsymbol{v}))}{h\phi(\Phi^{-1}(v_1))} \\ &\quad + K\left(\frac{z_1 - \Phi^{-1}(v_1)}{h}\right) K\left(\frac{z_2 - \Phi^{-1}(v_2)}{h}\right) \frac{u_2(\Phi^{-1}(\boldsymbol{v}))}{\phi(\Phi^{-1}(v_1))},\end{aligned}$$

which means that the expressions for $R_n(\boldsymbol{z})$, $B_{n,1}(\boldsymbol{z})$ and $B_{n,2}(\boldsymbol{z})$, as defined by Geenens et al. (2014), in our case have four terms instead of just one, resulting from the multiplications of $\frac{\partial J_{\boldsymbol{z},h}}{\partial v_1}$ and $\frac{\partial J_{\boldsymbol{z},h}}{\partial v_2}$. We will not write any more details here, because that will necessitate a much bigger body of notation. Straightforward algebra, however, exploiting the boundedness of $K(\boldsymbol{z})$, $k(\boldsymbol{z})$ (from assumption (iv)), and the boundedness of $u(\boldsymbol{z})$ (defined in (13)) for a fixed $\boldsymbol{z}$, will reveal that each of our terms is smaller than a constant times the corresponding term in Geenens et al. (2014), which is enough to prove the result.

A.4 Proof of Theorem 4

It follows from Lemma 1 and the delta method that $\widehat{f}_{\boldsymbol{Z}}(\boldsymbol{z})$ is asymptotically normal. It remains to show that the asymptotic normality still holds after the final back-transformation (7), with suitable estimates for the marginal density and distribution functions. Under assumption (viii), the normalised estimates $\sqrt{nh^2}\left(\widehat{f}_i(x_i) - f_i(x_i)\right)$ and $\sqrt{nh^2}\left(\widehat{F}_i(x_i) - F_i(x_i)\right)$ both converge in distribution to the constant 0, which again implies $\widehat{f}_i(x_i) - f_i(x_i) = o_P(1)$, and $\widehat{F}_i(x_i) - F_i(x_i) = o_P(1)$. It follows that

$$\begin{aligned}\phi\left(\Phi^{-1}(\widehat{F}_i(x_i))\right) &\\ = \phi\left(\Phi^{-1}(F_i(x_i))\right) &+ \phi'\left(\Phi^{-1}(F_i(x_i))\right)\left[\Phi^{-1}\right]'(F_i(x_i))(\widehat{F}_i(x_i) - F_i(x_i)) \\ &+ \text{higher order terms},\end{aligned}$$

where the second term is $o_P(1)$ in all x such that $F(x) \in (0,1)$. We can then write

$$\begin{aligned}\frac{\widehat{f}_i(x_i)}{\phi\left(\Phi^{-1}(\widehat{F}_i(x_i))\right)} &= \frac{f_i(x_i) + o_P(1)}{\phi\left(\Phi^{-1}(F_i(x_i))\right)} + \left(1 - \frac{o_P(1)}{\phi\left(\Phi^{-1}(F_i(x_i))\right)} + \cdots\right) \\ &= \frac{f_i(x_i)}{\phi\left(\Phi^{-1}(F_i(x_i))\right)} + o_P(1),\end{aligned}$$

from which it follows that

$$\prod_{i=1}^{p} \frac{\widehat{f}_i(x_i)}{\phi\left(\Phi^{-1}(\widehat{F}_i(x_i))\right)} = \prod_{i=1}^{p} \frac{f_i(x_i)}{\phi\left(\Phi^{-1}(F_i(x_i))\right)} + o_P(1).$$

By Slutsky's Theorem, we have that $\widehat{f}(\boldsymbol{x})$ as defined by equation (7) is asymptotically normal. The expression for the asymptotic variance of the density estimate follows from Theorem 3, Lemma 1, and the delta method applied to the asymptotic covariance matrix of the local correlations (10), using the function (1).

A.5 Proof of Theorem 5

Stone (1990) provides large sample theory for logspline density estimates. The asymptotic bias is shown to be asymptotically negligible provided that the true density is twice continuously differentiable. He shows further that $\widehat{f}_i$ is asymptotically normal with asymptotic variance of order $O(n^{-(0.5+\epsilon)})$, where $\epsilon \in (0, 1/2)$ is a tuning parameter that controls the rate at which new nodes are added in the logspline model. Stone (1990) develops theory for compactly supported densities only, but we proceed using a truncation argument to show that the asymptotic normality holds equally well for densities f_i satisfying $f_i = o(|z|^{-(5/2+\gamma)})$, where $\gamma = 2\epsilon/(1-2\epsilon)$ (see below) is close to zero if ϵ is small. If $\epsilon \to 1/2$, then $k \to \infty$, which is intuitively reasonable because the number of nodes will increase very slowly, meaning that the probability of extreme observations beyond the smallest and largest node must necessarily be small.

Denote by J_n the number of nodes that are used to fit the logspline model to $f(z)$ based on iid observations $Z_1, \ldots, Z_n$. Stone (1990) assume $J_n = o(n^{0.5-\epsilon})$. Construct a sequence L_n, $n = 1, 2, \ldots$ that is $o(J_n)$, and define a new stochastic variable by truncation;

$$Z^{(n)} = Z_n 1_{|Z_n| \leq L_n} + U_n 1_{|Z_n| > L_n},$$

where U_n is uniformly distributed on $[-L_n, L_n]$. Let $c_n = \int_{|z|\leq L_n} f(z)\,\mathrm{d}z$, then the density of $Z^{(n)}$ is given by

$$f^{(n)}(z) = c_n^{-1} f(z) 1_{|z|\leq L_n}.$$

The index i in $f_i = f_i(z_i)$ is not important here, and will be omitted. Let $\widehat{f}_n$ be the logspline estimate of f based on $Z_1, \ldots, Z_n$, and let $\widehat{f}_n^{(n)}$ be the logspline estimate of $f^{(n)}$ based on $Z_1^{(n)}, \ldots, Z_n^{(n)}$. We wish to show that $\sqrt{n^{0.5+\epsilon}}(\widehat{f}(z) - f(z))$ is asymptotically normal, and make the following decomposition:

$$\begin{aligned}
&\sqrt{n^{0.5+\epsilon}}(\widehat{f}(z) - f(z)) \\
&\quad = \sqrt{n^{0.5+\epsilon}}\left(\left\{\widehat{f}(z) - \widehat{f}^{(n)}(z)\right\} + \left\{\widehat{f}^{(n)}(z) - f^{(n)}(z)\right\} + \left\{f^{(n)}(z) - f(z)\right\} \right). \quad (21)
\end{aligned}$$

The first parenthesis converges in probability to zero provided that the tails of f are not too heavy. To see this, assume that there exists a $z_0 > 0$ such that $f(z) < M_1|z|^{-s}$ for all $|z| > z_0$ and some constant M_1. It follows from elementary calculus that

$$\begin{aligned}
P\left(\left|\widehat{f}(z) - \widehat{f}^{(n)}(z)\right| > 0\right) &\leq 1 - (P(|Z| \leq L_n))^n \\
&\leq 1 - (1 - M_2 L_n^{1-s})^n,
\end{aligned}$$

for a new constant M_2. We have from a Taylor expansion that the right hand side is $O(n^{(1/2-\epsilon)(1-s)+1})$ since $L_n = o(J_n) = o(n^{0.5-\epsilon})$, and so balancing this with the convergence rate in the normal approximation, $n^{1/4+\epsilon/2}$, yields $s^* = 5/2 + 2\epsilon/(1-2\epsilon)$ as the limiting value for s. Thus $\gamma = 2\epsilon/(1-2\epsilon)$.

It follows easily that the third parenthesis in (21) converges to zero if we assume that $f = o(|z|^{-s^*})$.

To see that the middle parenthesis in (21) is asymptotically normal, choose an arbitrary constant $T > 0$. Then there exists a positive integer N such that $[-T, T] \subset [-L_n, L_n]$ for all $n > N$. Make a new decomposition:

$$\begin{aligned}
&\sqrt{n^{0.5+\epsilon}}(\widehat{f}^{(n)}(z) - f^{(n)}(z)) \\
&= \sqrt{n^{0.5+\epsilon}}\left(\left\{\widehat{f}^{(n)}(z) - \widehat{f}^{(T)}(z)\right\} + \left\{\widehat{f}^{(T)}(z) - f^{(T)}(z)\right\} + \left\{f^{(T)}(z) - f^{(n)}(z)\right\} \right) \quad (22)
\end{aligned}$$

If we let $n \to \infty$ we have on the right hand side, and for any T according to the theory by Stone (1990), an asymptotically normally distributed variable in the middle. The first and third parentheses in (22) can be made arbitrarily small by choosing a large enough T. It follows from Slutsky's Theorem that the logspline estimates of the marginal densities are asymptotically normal with convergence rate $\sqrt{n^{1/2+\epsilon}}$, provided their tails are thinner than those of $|z|^{-s^*}$.

The exact same argument as above goes through for the marginal distribution and quantile functions as well, with convergence rate being equal to the usual $\sqrt{n}$.

References

Kjersti Aas, Claudia Czado, Arnoldo Frigessi, and Henrik Bakken. Pair-copula constructions of multiple dependence. *Insurance: Mathematics and Economics*, 44(2):182–198, 2009.

Donald W.K. Andrews. Generic uniform convergence. *Econometric Theory*, 8(2):241–257, 1992.

Adelchi Azzalini. The skew-normal distribution and related multivariate families. *Scandinavian Journal of Statistics*, pages 159–188, 2005.

Tim Bedford and Roger M Cooke. Vines: A new graphical model for dependent random variables. *Annals of Statistics*, pages 1031–1068, 2002.

Geir Drage Berentsen and Dag Tjøstheim. Recognizing and visualizing departures from independence in bivariate data using local Gaussian correlation. *Statistics and Computing*, 24(5):785–801, 2014.

Geir Drage Berentsen, Bård Støve, Dag Tjøstheim, and Tommy Nordbø. Recognizing and visualizing copulas: an approach using local Gaussian approximation. *Insurance: Mathematics and Economics*, 57:90–103, 2014.

Patrick Billingsley. *Probability and Measure*. John Wiley & Sons, 2008.

Adrian W Bowman. An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, 71(2):353–360, 1984.

Jerome H Friedman, Werner Stuetzle, and Anne Schroeder. Projection pursuit density estimation. *Journal of the American Statistical Association*, 79(387):599–608, 1984.

Gery Geenens. Probit transformation for kernel density estimation on the unit interval. *Journal of the American Statistical Association*, 109(505):346–358, 2014.

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

Christian Genest and Johan Segers. On the covariance of the asymptotic empirical copula process. *Journal of Multivariate Analysis*, 101(8):1837–1845, 2010.

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

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009.

Nils Lid Hjort and Ingrid K Glad. Nonparametric density estimation with a parametric start. *The Annals of Statistics*, 23(3):882–904, 1995.

Nils Lid Hjort and M.C. Jones. Locally parametric nonparametric density estimation. *The Annals of Statistics*, 24(4):1619–1647, 1996.

Jenq-Neng Hwang, S-R Lay, and Alan Lippman. Nonparametric multivariate density estimation: a comparative study. *Signal Processing, IEEE Transactions on*, 42(10): 2795–2810, 1994.

M.C. Jones. Simple boundary correction for kernel density estimation. *Statistics and Computing*, 3(3):135–146, 1993.

Clive R Loader. Local likelihood density estimation. *The Annals of Statistics*, 24(4): 1602–1618, 1996.

Thomas Mikosch. Copulas: Tales and facts. *Extremes*, 9(1):3–20, 2006.

Thomas Nagler and Claudia Czado. Evading the curse of dimensionality in multivariate kernel density estimation with simplified vines. *arXiv preprint arXiv:1503.03305*, 2015.

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

Håkon Otneim, Hans Arnfinn Karlsen, and Dag Tjøstheim. Bias and bandwidth for local likelihood density estimation. *Statistics & Probability Letters*, 83(5):1382–1387, 2013.

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

Dimitris N Politis and Joseph P Romano. Multivariate density estimation with general flat-top kernels of infinite order. *Journal of Multivariate Analysis*, 68(1):1–25, 1999.

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/`.

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, 1995.

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

Bernard W Silverman. *Density Estimation for Statistics and Data Analysis*, volume 26. CRC press, 1986.

Abe Sklar. *Fonctions de répartition à n dimensions et leurs marges*. Université Paris 8, 1959.

Stefan Sperlich, Dag Tjøstheim, and Lijian Yang. Nonparametric estimation and testing of interaction in additive models. *Econometric Theory*, 18(02):197–251, 2002.

Charles J Stone. An asymptotically optimal window selection rule for kernel density estimates. *The Annals of Statistics*, 12(4):1285–1297, 1984.

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.

Mervyn Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):111–147, 1974.

Bård Støve, Dag Tjøstheim, and Karl Ove Hufthammer. Using local Gaussian correlation in a nonlinear re-examination of financial contagion. *Journal of Empirical Finance*, 25:62–82, 2014.

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

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Matt P. Wand and M.C. Jones. Multivariate plug-in bandwidth selection. *Computational Statistics*, 9(2):97–116, 1994.

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

Paper III

3.3 Non-parametric estimation of conditional density functions: A new method

Håkon Otneim and Dag Tjøstheim

III

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 estmator 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.

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.

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 the 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 are 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). Li and Racine (2007, 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 case 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 prove consistency and asymptotic normality 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 $\boldsymbol{x}$;

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

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

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

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 $Z_i = \Phi^{-1}(F_i(X_i))$.

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. In (5), each correlation $\rho_{ij}(\boldsymbol{z})$ depends on the coordinates of the entire $\boldsymbol{z}$-vector. In order to circumvent the curse of dimensionality, we restrict $\rho_{ij}(\boldsymbol{z})$ 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 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 neighbour technique in order to obtain adaptive bandwidths, but simulation results suggest that, of the two, the global bandwidth selector performs better. Indeed, 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),

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

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

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

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 $\boldsymbol{z_2} = (z_{k+1}, \ldots, z_p)$. 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 $f_{\boldsymbol{X}_1|\boldsymbol{X}_2=\boldsymbol{x}_2}$ essentially 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 $f_i(x_i)$, $i = 1, \ldots, k$. 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 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

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

as $n \to \infty$, 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 $\boldsymbol{h} = \boldsymbol{h}_n \to 0$, 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 neighbourhoods. 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 neighbourhood-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 (and 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 modelling 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. 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,*

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

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,*

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

where

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

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

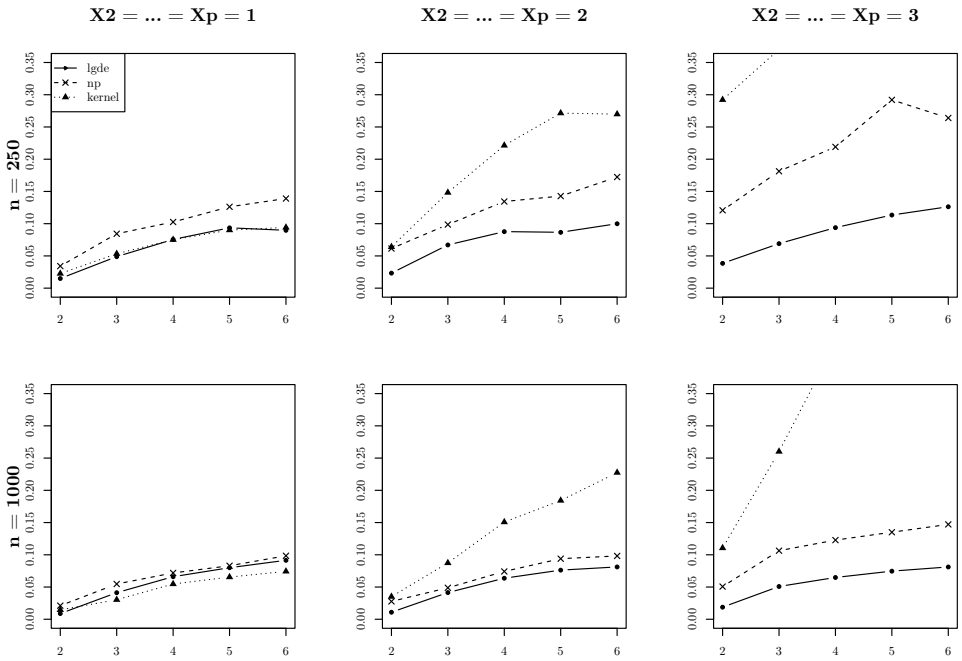


Figure 1: The integrated squared error of conditional density estimates of $f_{X_1|X_2,\ldots,X_p}$ as a function of p, generated from a density with exponential margins and a Joe copula with Kendall's Tau equal to 0.6.

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 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 Li and Racine (2007), which we denote by

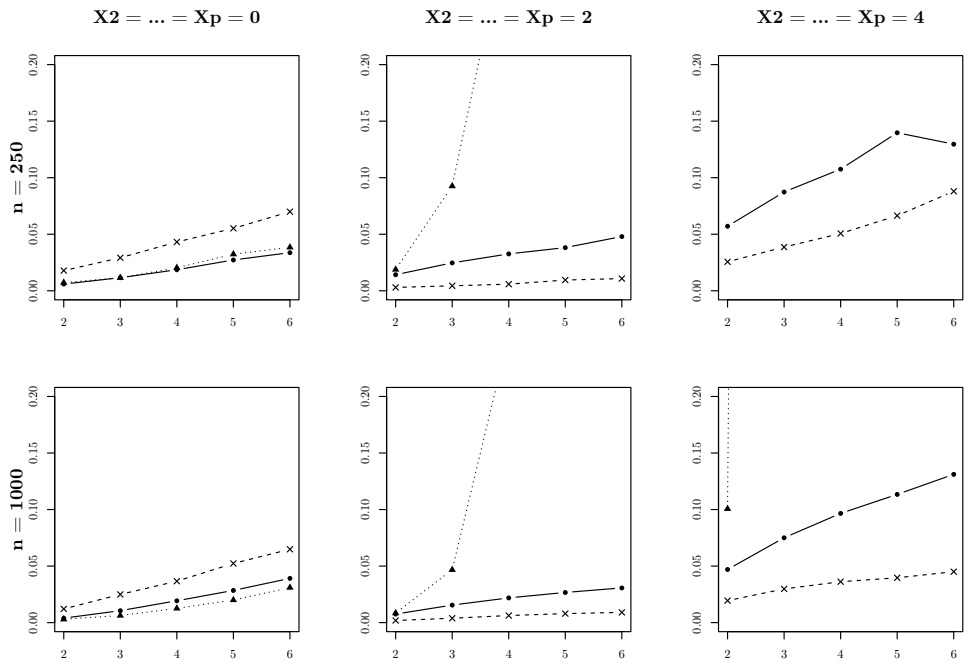


Figure 2: The integrated squared error of conditional density estimates of $f_{X_1|X_2,\ldots,X_p}$ as a function of p, generated from the multivariate t-distribution with 4 degrees of freedom.

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 behaviour 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 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 behaviour 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 behaviour of $f_{X_1|X_2,\ldots,X_p}$ in the tail.

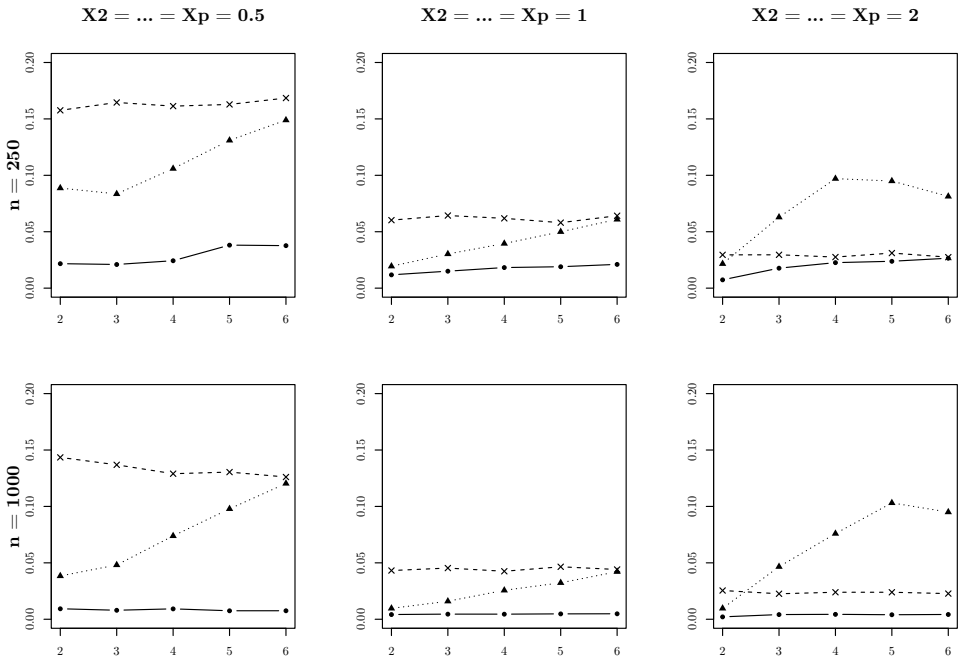


Figure 3: The integrated squared error of conditional density estimates of $f_{X_1|X_2,\ldots,X_p}$ as a function of p, generated from a density in which the first two variables are marginally log-normal with a $t(10)$-copula, and the rest of the variables are multivariate $t(5)$-distributed, independently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

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 Li and Racine (2007) 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 Li and Racine (2007), which smooths irrelevant variables away automatically. In our next example, however, most 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

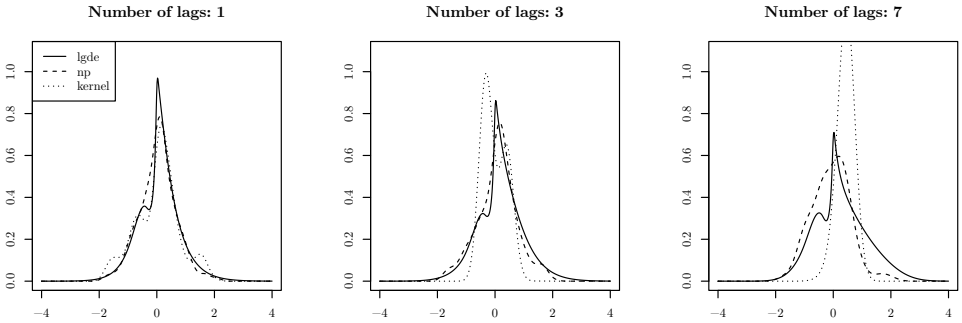


Figure 4: Estimate of the conditional density of the US log-returns conditioned on $X_{t-1} = \cdots = X_{t-k} = -1$ with $k = 1, 3, 7$ respectively.

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 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

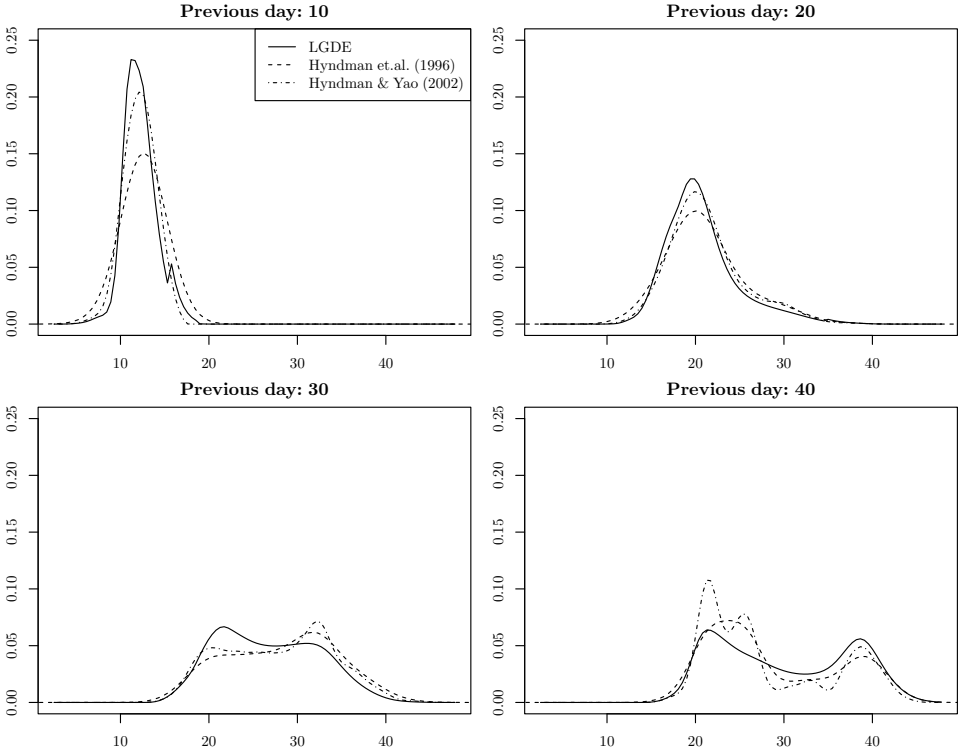


Figure 5: Australian temperature data, with estimated conditional density of the maximum daily air temperature, given a preceding recording of 10, 20, 30 and 40 degrees Celsius respectively.

are conserved. The tails in particular shows great robustness compared to the other two methods, and we believe that this behaviour 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 example 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.

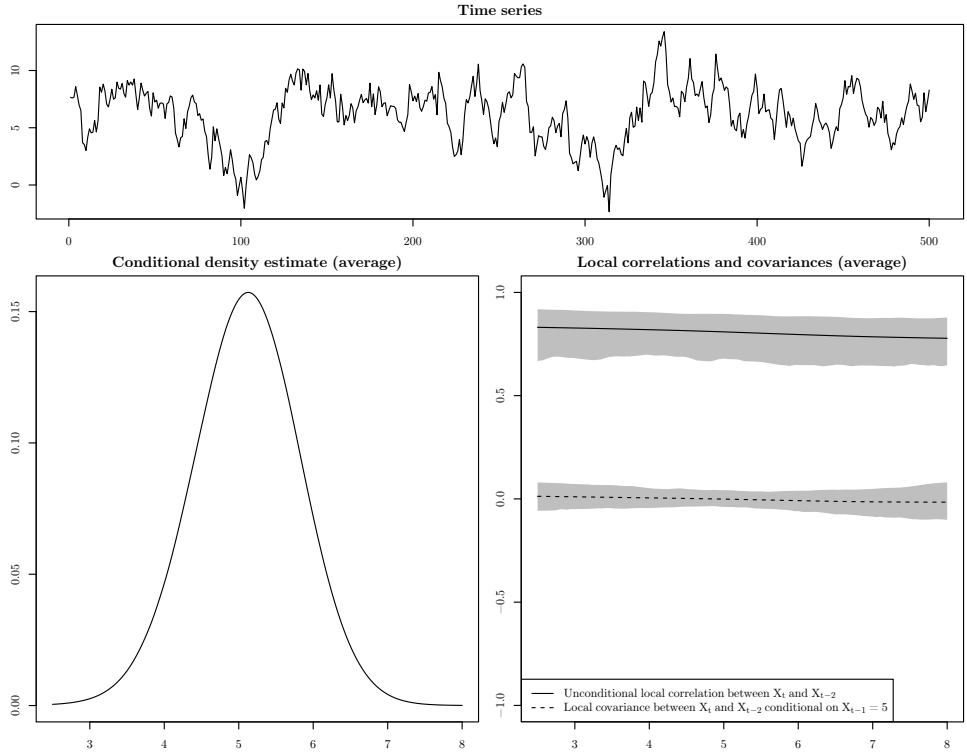


Figure 6: The top panel displays a simulated time series. The lower left panel displays the average of the estimated conditional densities of $X_t|X_{t-2} = 5$, and the lower right panel shows the unconditional diagonal local correlation between X_t and X_{t-2}, as well as the same quantity when conditioned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

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

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

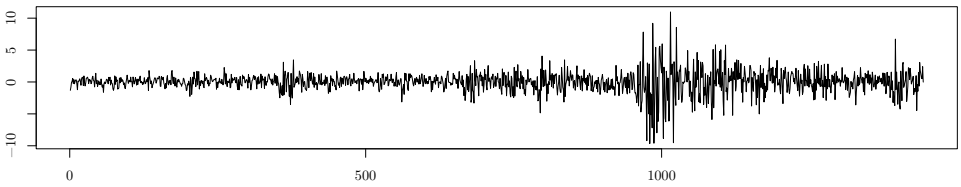


Figure 7: Value of the portfolio over a period of 1442 days.

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

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.

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 in-

dices, 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 Li and Racine (2007), 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, modelling 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 labelled 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

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

where

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

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,

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

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

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

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 $\sqrt{nh^2}(\theta_n - \theta) \xrightarrow{\mathcal{L}} N(0, A)$ and $q : R^n \to R$ has continuous first partial derivatives, then $\sqrt{nh^2}(q(\theta_n) - q(\theta)) \xrightarrow{\mathcal{L}} N(0, \nabla q(\theta)^T A \nabla q(\theta))$ (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):

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

and

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

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.

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.

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.

Qi Li and Jeffrey Scott Racine. *Nonparametric econometrics: theory and practice*. Princeton University Press, 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. *Unpublished manuscript*, 2016.

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.

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.

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