# Semiparametric Pseudo-Likelihood Estimation in Markov Random Fields

Antonino Freno

# Semiparametric Pseudo-Likelihood Estimation in Markov Random Fields

**Antonino Freno**
INRIA Lille – Nord Europe
antonino.freno@inria.fr

## Abstract

Probabilistic graphical models for continuous variables can be built out of either parametric or nonparametric conditional density estimators. While several research efforts have been focusing on parametric approaches (such as Gaussian models), kernel-based estimators are still the only viable and well-understood option for nonparametric density estimation. This paper develops a semiparametric estimator of probability density functions based on the nonparanormal transformation, which has been recently proposed for mapping arbitrarily distributed data samples onto normally distributed datasets. Pointwise and uniform consistency properties are established for the developed method. The resulting density model is then applied to pseudo-likelihood estimation in Markov random fields. An experimental evaluation on data distributed according to a variety of density functions indicates that such semiparametric Markov random field models significantly outperform both their Gaussian and kernel-based alternatives in terms of prediction accuracy.

## 1 Introduction

When dealing with continuous-valued variables, learning the parameters of probabilistic graphical models from data is much more challenging than in discrete domains. In fact, while the multinomial distribution is an usually adequate choice for estimating conditional probability distributions in the discrete setting, choosing a suitable kind of estimator for (continuous) conditional density functions requires either to assume that the form of the modeled density is known, leading to parametric techniques, or to relax such a parametric assumption, opting for a nonparametric technique [Duda et al., 2001]. The parametric assumption is often limiting, because in real-world applications the true form of the probability density function (pdf) can be rarely assessed *a priori*. On the other hand, nonparametric techniques only make a much weaker assumption concerning the smoothness of the pdf.

While a lot of research has been devoted to parametric graphical models in the machine learning community [Bishop, 2006, Koller and Friedman, 2009], only a few efforts have been devoted to nonparametric (or semiparametric) models. In Bayesian networks (BNs) and Markov random fields (MRFs), nonparametric conditional density estimators (based on kernel methods [Parzen, 1962, Rosenblatt, 1969]) are used for the first time by Hofmann and Tresp [1995, 1997]. A nonparametric technique for learning the structure of BNs is also developed in Margaritis [2005]. However, that method is only aimed at inferring the conditional independencies from data, rather than at learning the overall density function. A semiparametric technique for learning undirected graphs, leading to so-called 'nonparanormal' MRFs (NPMRFs), is proposed by Liu et al. [2009]. The nonparanormal approach consists in mapping the original data points (which are not assumed to satisfy any given distributional form) onto a different set of points, which are assumed to follow a multivariate normal distribution. The graph is then estimated from the transformed dataset using the graphical lasso algorithm [Friedman et al., 2008], which is both computationally efficient and theoretically sound for Gaussian distributions [Ravikumar et al., 2008]. However, the original nonparanormal approach only allows to estimate undirected graphs (rather than densities in the strict sense), i.e. it is not suitable for computing explicitly probability density functions [Liu et al., 2009]. Overcoming such a limitation is the main contribution of this paper, so as

Appearing in Proceedings of the $15^{th}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2012, La Palma, Canary Islands. Volume XX of JMLR: W&CP XX. 

to provide a full-fledged semiparametric (conditional) density estimator. Another attempt of mapping the original dataset into a feature space where data are assumed to be normally distributed is also made by Bach and Jordan [2002], using Mercer kernels.

This paper introduces a semiparametric MRF model for pseudo-likelihood estimation in continuous-valued domains, by developing a novel density estimator based on the nonparanormal mapping. Sec. 2 describes the general statistical framework that the proposed technique is embedded into for the purposes of our application. The nonparanormal density estimator is then presented and analyzed in Sec. 3, proving its asymptotic consistency. In Sec. 4 the estimator is evaluated experimentally on a number of benchmarks. Finally, Sec. 5 summarizes the main contributions of this work and sketches a couple of directions for further research.

## 2 Pseudo-Likelihood and the Quotient-Shape Approach to Conditional Density Estimation

One widely used approach to probabilistic modeling in Markov random fields resorts to the pseudo-likelihood function [Besag, 1975], which has proved to be an efficient yet accurate surrogate for likelihood in the strict sense (which is computationally intractable) in a wide variety of probabilistic models [Strauss and Ikeda, 1990, Hofmann and Tresp, 1997, Richardson and Domingos, 2006, Neville and Jensen, 2007, Freno et al., 2009]. Given the random variables $X_1, \ldots, X_d$, the pseudo-likelihood $p^*$ of any state $x_1, \ldots, x_d$ of those variables is measured as follows:

$$p^*(x_1, \ldots, x_d) = \prod_{i=1}^{d} p(x_i \mid x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d) \tag{1}$$

One convenient property of the pseudo-likelihood measure for application to graphical models (and Markov random fields in particular) is that, as defined by Eq. 1, it reduces to the following function:

$$p^*(x_1, \ldots, x_d) = \prod_{i=1}^{n} p\left(x_i \mid mb(X_i)\right) \tag{2}$$

where $mb(X_i)$ denotes the state of the Markov blanket of $X_i$ [Koller and Friedman, 2009].

In order to exploit the pseudo-likelihood function, we need to specify a technique for estimating the conditional densities involved in the right-hand side of Eq. 2. As a basic rule of probability theory, the conditional density of a (continuous) random variable $X$ given the random variable $Y$ can be derived from a pair of unconditional density functions as follows:

$$p(X \mid Y) = \frac{p(X, Y)}{p(Y)} \tag{3}$$

This means that any problem in conditional density estimation can be straightforwardly reduced to a pair of unconditional pdf estimation problems. Therefore, if our goal is to estimate the conditional density $p(X \mid Y)$, we can address this task by estimating first the (unconditional) density functions $p(X, Y)$ and $p(Y)$, and then by computing their quotient. This approach (which is called the *quotient-shape approach* to conditional density estimation) is the one we adopt in this paper for designing a (semiparametric) conditional density estimation technique, which we use within undirected graphical models, but which is applicable to virtually any kind of probabilistic graphical model [Hofmann and Tresp, 1995, 1997]. Some attempts have also been made in the relevant literature to devise different approaches [Faugeras, 2009], which are currently an active investigation area in multivariate statistics.

## 3 Nonparametric Normal Estimation of Probability Density Functions

This section presents a semiparametric method for estimating (conditional) density functions. This technique, which is referred to as *nonparametric normal* (or *nonparanormal*), was introduced by Liu et al. [2009] for learning the structure of (sparse) undirected graphs. However, the main contribution of this paper is to show how the nonparanormal approach can be turned into a general-purpose density estimation method. While the main ideas underlying the nonparanormal method are reviewed in Sec. 3.1, the problem of extending that method to density estimation tasks is addressed in Sec. 3.2. Pointwise and uniform consistency properties of the proposed estimator are then investigated in Sec. 3.3.

### 3.1 Background

In order to present the nonparanormal approach, we first need to recall an important lemma from multivariate calculus, which is commonly known as the *change of variables* theorem:

**Lemma 1.** *Consider two random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$, with domains $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} \subseteq \mathbb{R}^d$ respectively. Suppose that $f : \mathcal{X} \to \mathcal{Y}$ is a one-to-one, differentiable function from $\mathcal{X}$ onto $\mathcal{Y}$. Then, if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are distributed according to density functions $p_{\boldsymbol{X}}(\boldsymbol{x})$ and*

$p_{\boldsymbol{Y}}(\boldsymbol{y})$ *respectively, it follows that*

$$p_{\boldsymbol{Y}}(\boldsymbol{y}) = p_{\boldsymbol{X}}(\boldsymbol{x}) \left| \det \frac{\partial \boldsymbol{x}}{\partial \boldsymbol{y}} \right| \tag{4}$$

*where* $f^{-1} : \mathcal{Y} \to \mathcal{X}$ *is the inverse of* $f$, $\boldsymbol{x} = f^{-1}(\boldsymbol{y})$, *and* $\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{y}}$ *denotes the Jacobian matrix of* $f^{-1}$.

*Proof.* See e.g. Kaplan [1984]. □

The nonparanormal (or *nonparametric normal*) approach is a recently introduced technique for estimating the structure of undirected graphs, based on a Gaussian model, without making any parametric assumption concerning the form of the modeled density [Liu et al., 2009]. Although the previous statement may seem paradoxical, the idea underlying the nonparanormal approach is to map a set of data points (which are not known to be normally distributed) onto a set of data points that can be assumed to follow a normal distribution. Once the density of the normal sample has been estimated using a standard Gaussian model, the density of the points in the original space can then be recovered by applying the change of variables theorem.

First of all, let us define the concept of nonparanormal density:

**Definition 1.** *A random vector* $\boldsymbol{X} = (X_1, \ldots, X_d)$ *with mean* $\boldsymbol{\mu}$ *is said to be nonparanormally distributed if there exists a function* $f$ *such that:*

1. $f(\boldsymbol{X}) = (f_1(X_1), \ldots, f_d(X_d))$, *where* $f_i(X_i)$ *is one-to-one and differentiable (for* $1 \leq i \leq d$*);*
2. *the random vector* $\boldsymbol{Y} = f(\boldsymbol{X})$ *is distributed normally with mean* $\boldsymbol{\mu}$ *and covariance matrix* $\boldsymbol{\Sigma}$.

Given Definition 1, we can prove the following lemma (which is stated by Liu et al. [2009] without proof):

**Lemma 2.** *If the distribution of a random vector* $\boldsymbol{X} = (X_1, \ldots, X_d)$ *is nonparanormal with mapping* $f(\boldsymbol{X}) = (f_1(X_1), \ldots, f_d(X_d))$, *then the density of* $\boldsymbol{X}$ *is given by*

$$p_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{e^{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^{\intercal}\, \boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})}}{(2\pi)^{d/2} \det(\boldsymbol{\Sigma})^{1/2}} \prod_{i=1}^{d} \left| \frac{\mathrm{d}}{\mathrm{d}\, x_i} f_i(x_i) \right| \tag{5}$$

*where* $\boldsymbol{y} = f(\boldsymbol{x})$, $\boldsymbol{\mu}$ *is the mean vector of both* $\boldsymbol{X}$ *and* $\boldsymbol{Y}$, *and* $\boldsymbol{\Sigma}$ *is the covariance matrix of* $\boldsymbol{Y}$.

*Proof.* Let $p_{\boldsymbol{Y}}$ denote the (normal) density function of $\boldsymbol{Y}$. Then, Lemma 1 implies that

$$\begin{aligned} p_{\boldsymbol{X}}(\boldsymbol{x}) &= p_{\boldsymbol{Y}}(\boldsymbol{y}) \left| \det \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} \right| \\ &= \frac{e^{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^{\intercal}\, \boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})}}{(2\pi)^{d/2} \det(\boldsymbol{\Sigma})^{1/2}} \left| \det \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} \right| \end{aligned} \tag{6}$$

Since the value of each $f_i$ only depends on $x_i$, the Jacobian matrix $\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}$ is diagonal. Therefore, the absolute value of the Jacobian determinant is given by

$$\left| \det \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} \right| = \prod_{i=1}^{d} \left| \frac{\mathrm{d}}{\mathrm{d}\, x_i} f_i(x_i) \right| \tag{7}$$

□

Based on Lemma 2, the crucial problem for the nonparanormal approach is how to estimate the functions $f_1(X_1), \ldots, f_d(X_d)$. The technique developed by Liu et al. [2009] prescribes to estimate the value of each $f_i$ as

$$\hat{f}_i(x) = \hat{\mu}_i + \hat{\sigma}_i \hat{h}_i(x) \tag{8}$$

where $\hat{\mu}_i$ and $\hat{\sigma}_i$ are the sample mean and standard deviation of variable $X_i$, and $\hat{h}_i(x)$ is defined as follows:

$$\hat{h}_i(x) = \Phi^{-1}(\widehat{F}_i(x)) \tag{9}$$

In Eq. 9, $\Phi^{-1}$ is the inverse of the standard normal cumulative distribution function (cdf) $\Phi$, given by

$$\Phi(x) = \frac{1}{2} \left[ 1 + \operatorname{erf}\left( \frac{x}{\sqrt{2}} \right) \right] \tag{10}$$

where $\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, \mathrm{d}\, t$. On the other hand, $\widehat{F}_i$ is the so-called *truncated* estimator of the empirical cdf $F_i^E$ of $X_i$ [Dixon, 1960]. Let $n$ be the number of data points and $\delta_n$ be a truncation parameter. The truncated estimator of $F_i^E$ is then defined as

$$\widehat{F}_i(x) = \begin{cases} \delta_n & \text{if } F_i^E(x) < \delta_n \\ F_i^E(x) & \text{if } \delta_n \leq F_i^E(x) \leq 1 - \delta_n \\ 1 - \delta_n & \text{if } 1 - \delta_n < F_i^E(x) \end{cases} \tag{11}$$

where, if $\Theta$ denotes the Heaviside step function, the value of $F^E(x)$ is given by

$$F^E(x) = \frac{1}{n} \sum_{j=1}^{n} \Theta(x - x_j) \tag{12}$$

The suggested setting for the truncation parameter is given by

$$\delta_n = \frac{1}{4 n^{1/4} \sqrt{\pi \log n}} \tag{13}$$

The choice specified in Eq. 13 is reported to result in a generally satisfying behavior of the nonparanormal estimator, especially in the high-dimensional setting [Liu et al., 2009].

As defined in Eq. 11, the truncated estimator $\widehat{F}_i(x)$ is discontinuous. This prevents us from computing the derivatives contained in Eq. 5. In other words, the approach described thus far is not yet suitable as a thorough density estimation technique. On the other

hand, it is sufficient instead for estimating the structure of the undirected graph underlying the density $p_{\boldsymbol{Y}}$, as this structure is conveyed by the precision matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$. In fact, if $\boldsymbol{\Omega}_{i\,j} = 0$, then the nodes $X_i$ and $X_j$ will not be adjacent in the graph of a MRF representing $p_{\boldsymbol{Y}}$ [Lauritzen, 1996]. Now, one key result proved by [Liu et al., 2009] establishes that, if $\boldsymbol{X}$ is nonparanormal with mapping $\boldsymbol{Y} = f(\boldsymbol{X})$, then $X_i$ is independent of $X_j$ given a subset $\mathcal{S}_{\boldsymbol{X}}$ of $\{X_1, \ldots, X_d\}$ if and only if $Y_i$ is independent of $Y_j$ given the set $\mathcal{S}_{\boldsymbol{Y}} = \{Y_k : X_k \in \mathcal{S}_{\boldsymbol{X}}\}$. This result can be used for learning the structure of MRFs. To this aim, the guiding idea is that the precision matrix $\boldsymbol{\Omega}$ of the random vector $\boldsymbol{Y}$ fixes not only the graph of a MRF for $p_{\boldsymbol{Y}}$, but also the graph of a MRF for $p_{\boldsymbol{X}}$. However, since our interest lies in exploiting the nonparanormal approach for the sake of (conditional) density estimation, we move beyond the strategy presented above, trying to fit the approach to our overall goal.

### 3.2 Multilogistic Estimation of Cumulative Distribution Functions

A differentiable approximation $F_i^*(x)$ of $F_i^E(x)$ can be obtained by using the logistic function:

$$F_i^*(x) = \frac{1}{n}\sum_{j=1}^{n} \frac{1}{1 + \exp\left(-\frac{x - x_j}{h}\right)} \tag{14}$$

where $h$ is a parameter controlling the logistic smoothness. The approximation is justified by the fact that the Heaviside step function (employed in Eq. 12) can be defined as follows:

$$\Theta(x) = \lim_{h \to 0} \frac{1}{1 + \exp\left(-\frac{x}{h}\right)} \tag{15}$$

We refer to the cdf estimator defined in Eq. 14 as the *multilogistic estimator* of (univariate) distribution functions. Given the multilogistic estimator, we replace $\hat{h}_i(x)$ by $\hat{h}_i^*(x)$:

$$\hat{h}_i^*(x) = \Phi^{-1}\big(F_i^*(x)\big) \tag{16}$$

An approximate value of the derivatives referred to in Eq. 5 is then given by:

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}\,x}\hat{f}_i(x) = \\ &= \frac{\hat{\sigma}_i\sqrt{2\pi}}{n\,h\,e^{-\operatorname{erf}^{-1}\left(2\,F_i^*(x)-1\right)^2}} \sum_{j=1}^{n} \frac{e^{-\frac{x-x_j}{h}}}{\left(1 + e^{-\frac{x-x_j}{h}}\right)^2} \end{aligned} \tag{17}$$

If the value specified in Eq. 17 (which is derived in the Appendix) is substituted into Eq. 5, the resulting model can be straightforwardly used as a conditional density estimator based on the quotient-shape approach.

### 3.3 Consistency Results

Before establishing consistency results for the proposed density estimation technique, it is useful to recall an important theoretical property of the nonparanormal mapping. One lemma proved by Liu et al. [2009] shows that, if each function $f_i$ is monotone and differentiable, then the nonparanormal is a Gaussian copula [Sklar, 1959, Nelsen, 2006] such that the density of the (nonparanormally distributed) vector $\boldsymbol{X}$ is given by Eq. 5. This means that, in order to estimate a multivariate nonparanormal distribution, the crucial problem is to estimate the univariate cumulative distribution functions of the variables $X_1, \ldots, X_d$. Therefore, the key to understanding the consistency properties of the proposed density estimator is to elucidate the consistency properties of each (univariate) estimator $F_i^*$.

A preliminary result we are going to prove is that, given a random variable $X$, the multilogistic estimator $F^*$ of the cumulative distribution function of $X$ (as defined by Eq. 14) results in a kernel estimate $\hat{p}$ of the pdf of $X$, as explained by the following lemma:

**Lemma 3.** *If $X$ is a random variable with probability density function $p$ and cumulative distribution function $F$, then the multilogistic estimator $F^*$ of $F$ is equivalent to a kernel estimator $\hat{p}$ of $p$ with bandwidth $h$.*

*Proof.* Since $p(x) = \frac{\mathrm{d}}{\mathrm{d}\,x} F(x)$, the density $\hat{p}(x)$ corresponding to the estimate $F^*(x)$ can be expressed as follows (as shown in derivation 23):

$$\begin{aligned} \hat{p}(x) &= \frac{\mathrm{d}}{\mathrm{d}\,x} F^*(x) \\ &= \frac{1}{n\,h}\sum_{j=1}^{n} \frac{\exp\left(-\frac{x-x_j}{h}\right)}{\left(1 + \exp\left(-\frac{x-x_j}{h}\right)\right)^2} \\ &= \frac{1}{n\,h}\sum_{j=1}^{n} K\left(\frac{x - x_j}{h}\right) \end{aligned} \tag{18}$$

where $K$ is defined as:

$$K(t) = \frac{e^{-t}}{(1 + e^{-t})^2} \tag{19}$$

This means that $F^*(x)$ amounts to a kernel estimator $\hat{p}(x)$ of the density of $X$, with kernel function $K$ and bandwidth $h$ (where $h$ is exactly the parameter controlling the smoothness of the logistic functions employed in $F^*$). □

The importance of this result lies in the fact that, based on Lemma 3, the multilogistic estimator developed in this paper inherits in a straightforward manner the consistency properties of kernel density estimators.

We first establish a pointwise consistency property for the proposed estimator:

**Theorem 1.** *Let $X$ be a random variable with probability density function $p$ and cumulative distribution function $F$, and let $\hat{p}(x)$ be the kernel density estimate resulting from a multilogistic estimator $F^*$ of $F$. Then, if the bandwidth $h$ of the estimator is such that $h \to 0$ and $n\,h \to \infty$ as $n \to \infty$, it follows that $\frac{\mathrm{d}}{\mathrm{d}\,x}F^*(x) \to \frac{\mathrm{d}}{\mathrm{d}\,x}F(x)$ in probability as $n \to \infty$.*

*Proof.* It is shown by Parzen [1962] that, for any density function $p$ which is continuous at $x$, if a kernel estimate $\hat{p}(x)$ with kernel function $K$ is such that $h \to 0$ and $n\,h \to \infty$ as $n \to \infty$, then $\hat{p}(x) \to p(x)$ in probability as $n \to \infty$ provided that $K$ is a bounded Borel function satisfying the following conditions: $(i)$ $\int_{-\infty}^{\infty}|K(t)|\,\mathrm{d}\,t < \infty$; $(ii)$ $\int_{-\infty}^{\infty}K(t)\,\mathrm{d}\,t = 1$; $(iii)$ $|t\,K(t)| \to 0$ as $|t| \to \infty$. Such conditions are satisfied by the kernel function defined in Eq. 19 (as they are by a wide variety of kernels [Silverman, 1986]). Based on Lemma 3, this means that, if $h \to 0$ and $n\,h \to \infty$ as $n \to \infty$, then $\frac{\mathrm{d}}{\mathrm{d}\,x}F^*(x) \to \frac{\mathrm{d}}{\mathrm{d}\,x}F(x)$ in probability as $n \to \infty$. □

Uniform consistency also holds for the multilogistic estimator, under conditions that are only slightly stronger than the ones required for pointwise consistency:

**Theorem 2.** *Let $X$ be a random variable with probability density function $p$ and cumulative distribution function $F$, and let $\hat{p}$ be the kernel density estimator resulting from a multilogistic estimator $F^*$ of $F$. Then, if the bandwidth $h$ of the estimator is such that $h \to 0$ and $(n\,h)^{-1}\log n \to 0$ as $n \to \infty$, it follows that $\sup_x \left|\frac{\mathrm{d}}{\mathrm{d}\,x}F^*(x) - \frac{\mathrm{d}}{\mathrm{d}\,x}F(x)\right| \to 0$ almost surely as $n \to \infty$.*

*Proof.* It is shown by Bertrand-Retali [1978] and Silverman [1978, 1980] that, for any density function $p$ which is uniformly continuous on $(-\infty, \infty)$, if a kernel estimate $\hat{p}(x)$ with kernel function $K$ is such that $h \to 0$ and $(n\,h)^{-1}\log n \to 0$ as $n \to \infty$, then $\sup_x |\hat{p}(x) - p(x)| \to 0$ almost surely as $n \to \infty$ provided that $K$ is a (bounded) function with bounded variation satisfying the following conditions: $(i)$ $\int_{-\infty}^{\infty}|K(t)|\,\mathrm{d}\,t < \infty$; $(ii)$ $\int_{-\infty}^{\infty}K(t)\,\mathrm{d}\,t = 1$; $(iii)$ the set of discontinuities of $K$ has Lebesgue measure zero. Again, such conditions are satisfied by the kernel function defined in Eq. 19 (as well as by many other kernels [Silverman, 1978]). Therefore, Lemma 3 implies that, if $h \to 0$ and $(n\,h)^{-1}\log n \to 0$ as $n \to \infty$, then $\sup_x \left|\frac{\mathrm{d}}{\mathrm{d}\,x}F^*(x) - \frac{\mathrm{d}}{\mathrm{d}\,x}F(x)\right| \to 0$ almost surely as $n \to \infty$. □

As a consequence of Theorem 2, we can state the following corollary:

**Corollary 1.** *Let $X$ be a random variable with cumulative distribution function $F$, and let $F^*$ be a multilogistic estimator of $F$. Then, if the bandwidth $h$ of the estimator is such that $h \to 0$ and $(n\,h)^{-1}\log n \to 0$ as $n \to \infty$, it follows that $\sup_x |F^*(x) - F(x)| \to 0$ almost surely as $n \to \infty$.*

*Proof.* The corollary follows straightforwardly from Theorem 2. □

## 4 Experimental Evaluation

The aim of this section is to evaluate the accuracy of the semiparametric pseudo-likelihood estimation technique presented thus far at modeling the distribution of (multivariate) data featuring nonlinear dependencies between the variables plus non-Gaussian random noise. In particular, the idea is to sample a number of datasets from synthetic distributions, where the latter are generated in such a way as to make it unlikely that any particular parametric assumption (such as normality) may be satisfied. We can then exploit the produced data for pattern classification, so as to compare the prediction accuracy of semiparametric (nonparanormal) MRFs both to parametric (Gaussian) and nonparametric (kernel-based) MRFs, respectively. After briefly summarizing the main ideas underlying the data generation process, Sec. 4.1 states the basic properties of the used datasets, while the results of the experiments are reported in Sec. 4.2.

### 4.1 Datasets

In order to sample pattern-classification datasets featuring nonlinear correlations between pairs of variables in each class, suitable distributions are defined by generating random BNs. The BN is created by generating $(i)$ a random (directed acyclic) graph, $(ii)$ a set of functions (with random parameters) characterizing the dependence of every node on each one of its parents in the graph, and $(iii)$ a set of functions (with randomly assigned parameters) defining the probability density of each node. While a complete description of the data generation technique is given by Freno et al. [2010], Fig. 1 provides some examples of the distributions that can be generated using such method. In the plotted examples, the underlying distributions (featuring cubic correlation functions and beta densities) are organized in a DAG $(\mathcal{V}, \mathcal{E})$ such that $\mathcal{V} = \{X, Y\}$ and $\mathcal{E} = \{(X, Y)\}$. Notice how the beta function associated with variable $Y$ produces nearly uniform density over the support of the distribution for Fig. 1a, while it generates noise which is peaked toward the

lower/higher extreme in Fig. 1b/1c, or toward both extremes in Fig. 1d. On the other hand, the parameters of the polynomial functions are able to determine nearly-quadratic dependencies (Figs. 1a, 1b), cubic dependencies (Fig. 1c), and nearly-linear dependencies (class 2 in Figs. 1d–1f). The employed data generation technique is capable of producing a relatively wide variety of pattern classification problems: easy problems, where the classes are linearly separable (such as in Fig. 1f); moderately difficult tasks, where the classes overlap to a significant extent but a linear separation might be settled for with relatively good results (such as in Fig. 1a); fairly hard problems, where patterns drawn from different classes are neither linearly separable nor belonging to neatly separated regions of the feature space (such as in Figs. 1b–1e).

The described data generation technique is applied to the present experimental setting in the following way: (*i*) four datasets (CB1 through CB4) are generated using random cubic functions for the variable correlations and random beta densities for the variable distributions; (*ii*) four datasets (SE1 through SE4) are generated using random logistic functions for the variable correlations and random exponential densities for the variable distributions; (*iii*) four datasets (LG1 through LG4) are generated using random linear functions for the variable correlations and random Gaussian densities for the variable distributions. While datasets CB1–CB4 and SE1–SE4 offer a benchmark featuring a wide range of nonlinear variable correlations and non-Gaussian probability densities, datasets LG1–LG4 provide instead a baseline for evaluating two important issues. First, we want to verify whether the nonparanormal technique proposed in this paper can be expected to be at least as accurate as a given parametric technique whenever the latter makes the correct assumption concerning the form of the modeled distribution, which is the case for Gaussian MRFs (GMRFs) with the LG1–LG4 datasets. Second, we want to assess how significant the loss of prediction accuracy is for GMRFs (with respect to nonparanormal and kernel-based MRFs) whenever the related parametric assumption is instead violated by the data, as compared to the accuracy achieved when the normality assumption is satisfied. All datasets used in the experiments contain a total number of 500 patterns, equally split into two classes. Table 1 summarizes the main properties of each dataset, also indicating the dimensionality of each one.

Table 1: General properties of the synthetic datasets used in the experimental evaluation (where $d$ is the number of random variables).

| Dataset | Correlations | Densities | $d$ |
|---|---|---|---|
| **CB1** | cubic | beta | 12 |
| **CB2** | cubic | beta | 14 |
| **CB3** | cubic | beta | 17 |
| **CB4** | cubic | beta | 18 |
| **SE1** | logistic | exponential | 11 |
| **SE2** | logistic | exponential | 12 |
| **SE3** | logistic | exponential | 13 |
| **SE4** | logistic | exponential | 14 |
| **LG1** | linear | Gaussian | 7 |
| **LG2** | linear | Gaussian | 13 |
| **LG3** | linear | Gaussian | 16 |
| **LG4** | linear | Gaussian | 17 |

### 4.2 Results

The prediction accuracy of the NPMRF model developed in this paper is compared to the accuracy achieved by GMRFs [Koller and Friedman, 2009] and kernel-based MRFs (KMRFs) [Hofmann and Tresp, 1997]. In KMRFs, the graph is estimated using the structure learning algorithm proposed by Hofmann and Tresp [1997], based on a maximum pseudo-likelihood strategy. In GMRFs and NPMRFs, structure learning is performed instead by means of the graphical lasso technique [Friedman et al., 2008, Liu et al., 2009], while conditional densities are modeled for the resulting graphical structures by Gaussian and nonparanormal estimators respectively. To the best of our knowledge, the learning algorithms considered for GMRFs, KMRFs, and NPMRFs are the state of the art emerging from the literature on continuous MRFs.

In order to exploit the models for pattern classification, we take for each dataset $\mathcal{D}$ the two subsets $\mathcal{D}_1$ and $\mathcal{D}_2$, where all patterns in $\mathcal{D}_i$ belong to class $\omega_i$. For each model, we learn two class-specific versions, training each version on the respective set of data points. Patterns in the test set are then classified as follows. For each $\omega_i$, we estimate the posterior probability $P(\omega_i \mid x)$ that a pattern $\boldsymbol{x}$ belongs to class $\omega_i$:

$$P(\omega_i \mid \boldsymbol{x}) = \frac{p(\boldsymbol{x} \mid \omega_i)P(\omega_i)}{p(\boldsymbol{x})} \tag{20}$$

where $p(\boldsymbol{x} \mid \omega_i)$ is the pseudo-likelihood of the model learned for $\omega_i$ given $\boldsymbol{x}$, $P(\omega_i)$ is the prior probability of class $\omega_i$ (estimated as $\frac{|\mathcal{D}_i|}{|\mathcal{D}|}$), and $p(\boldsymbol{x}) = \sum_j p(\boldsymbol{x} \mid \omega_j)P(\omega_j)$. Given the posterior probability of each class, we attach to $\boldsymbol{x}$ the label with the highest probability, based on a maximum a posteriori strategy. The results of the experiments are reported in Table 2, where values are averaged by 5-fold cross-validation. In order to assess the statistical significance of the results, Table 2 also provides $p$-values for the paired $t$-test between NPMRFs and each one of the other two

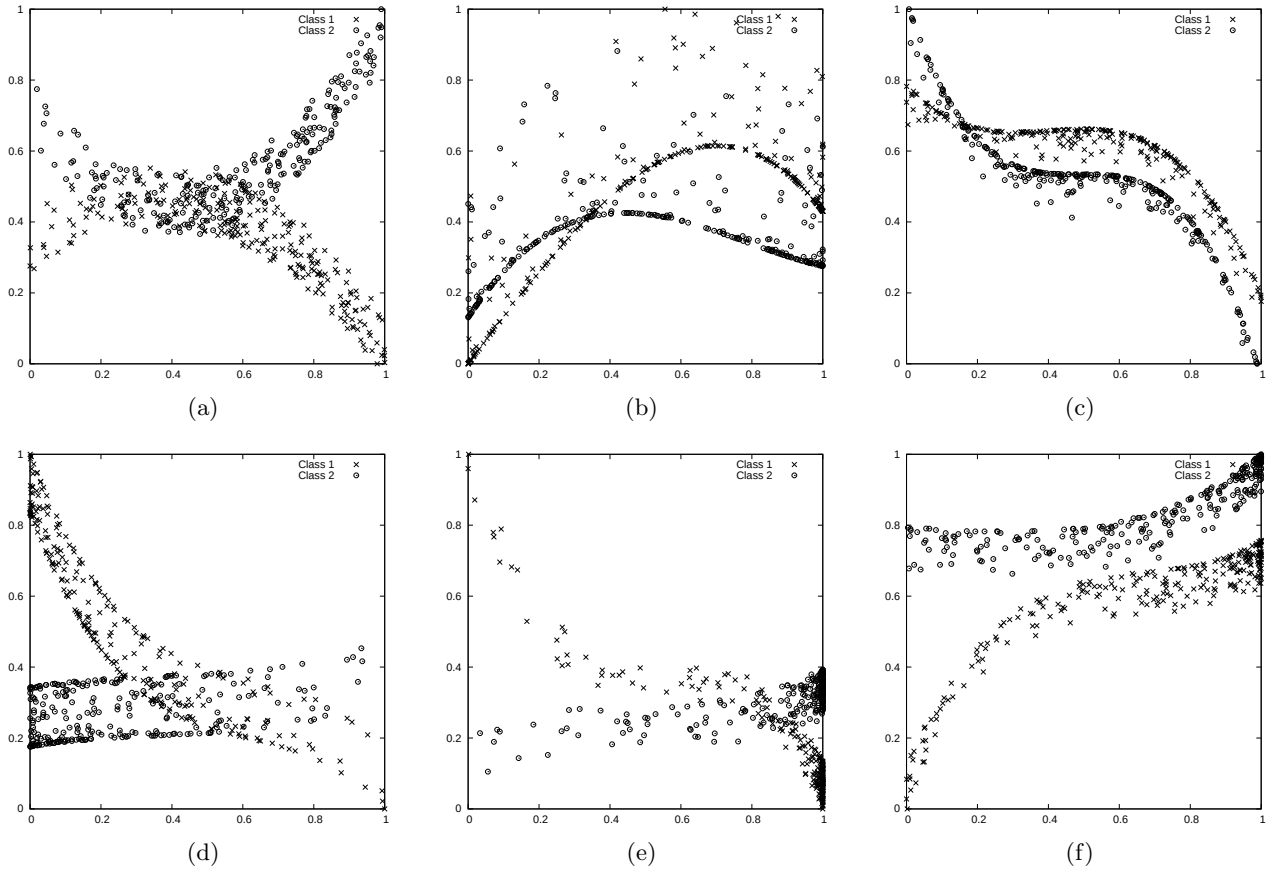


Figure 1: Randomly generated bivariate distributions for pattern classification tasks. $X$ and $Y$ are distributed according to random beta densities, while the dependence of $Y$ on $X$ is shaped by a random cubic function.

models.

Table 2 supports the following interpretation. First, when the normality assumption happens to be satisfied, NPMRFs are generally equivalent to GMRFs in terms of prediction accuracy. On the contrary, in these cases KMRFs are relatively unreliable, since their accuracy is often lower than the accuracy achieved by GMRFs and NPMRFs. Second, whenever the involved parametric assumption is violated by the given distribution (CB1–CB4 and SE1–SE4 datasets), semiparametric MRFs are dramatically superior to parametric MRFs. Moreover, although in such cases KMRFs are usually more accurate than GMRFs, NPMRFs regularly outperform their kernel-based alternative to a significant extent. Since the multilogistic cdf estimator employed in the nonparanormal model reveals a very tight connection with kernel density estimation (as explained in Sec. 3.3), one plausible reason for the advantage of NPMRFs over KMRFs is given by the fact that the Gaussian copula allows the nonparanormal approach to break down a multivariate estimation problem into its univariate counterparts, which are generally easier to be dealt with. Overall, the results suggest that the semiparametric MRF model developed in this paper is a much more flexible pseudo-likelihood estimation technique than its Gaussian and kernel-based alternatives.

## 5 Conclusions and Future Work

In this paper, a novel semiparametric technique for estimating probability density functions has been introduced, based on complementing the nonparanormal framework with the multilogistic cdf estimator. On the one hand, pointwise and uniform consistency results have been proved for the multilogistic estimator. On the other hand, the developed technique has been successfully applied to the problem of learning Markov random fields from data, using the pseudo-likelihood approach. In particular, a number of pattern classification benchmarks show that semiparametric MRFs are a generally accurate and flexible model for data distributed according to a variety of density functions, featuring both linear and nonlinear variable correlations.

While the proposed technique overcomes one limita-

Table 2: Recognition accuracy (average ± standard deviation) measured by 5-fold cross-validation on the CB1–CB4, SE1–SE4, and LG1–LG4 datasets for GMRFs, KMRFs, and NPMRFs respectively. For GMRFs and KMRFs, $p$-values from the paired $t$-test against NPMRFs are reported in brackets (with bold font indicating that the $p$-value is less than 0.05).

| | **Recognition Accuracy (%)** | | | | |
|---|---|---|---|---|---|
| **Dataset** | **GMRF** | | **KMRF** | | **NPMRF** |
| **CB1** | 70.6 ± 4.27 | (**0.001**) | 79.0 ± 2.68 | (**0.002**) | 87.6 ± 3.00 |
| **CB2** | 91.6 ± 3.13 | (0.098) | 88.8 ± 1.32 | (**0.002**) | 95.8 ± 1.16 |
| **CB3** | 50.6 ± 1.35 | (**0.001**) | 55.8 ± 3.96 | (**0.030**) | 64.8 ± 4.48 |
| **CB4** | 53.6 ± 2.72 | (**0.001**) | 62.2 ± 5.91 | (0.111) | 66.4 ± 2.93 |
| **SE1** | 66.6 ± 5.31 | (0.861) | 71.6 ± 5.85 | (0.355) | 67.2 ± 7.30 |
| **SE2** | 59.4 ± 3.66 | (**0.000**) | 56.8 ± 3.96 | (**0.014**) | 67.4 ± 3.38 |
| **SE3** | 63.6 ± 3.97 | (**0.029**) | 62.4 ± 7.55 | (0.089) | 73.0 ± 6.06 |
| **SE4** | 63.6 ± 6.56 | (**0.004**) | 72.6 ± 6.40 | (**0.016**) | 80.8 ± 3.54 |
| **LG1** | 84.6 ± 2.33 | (0.099) | 59.0 ± 3.52 | (**0.000**) | 83.8 ± 2.31 |
| **LG2** | 90.6 ± 4.49 | (0.085) | 90.6 ± 1.95 | (**0.001**) | 82.2 ± 3.12 |
| **LG3** | 97.8 ± 1.72 | (0.541) | 79.2 ± 5.41 | (**0.001**) | 97.4 ± 1.85 |
| **LG4** | 79.0 ± 1.41 | (**0.000**) | 71.2 ± 5.19 | (**0.016**) | 81.8 ± 1.60 |

tion of the original nonparanormal approach, which was only suitable for estimating undirected graphs (rather than providing a full model of density functions), one challenge for future research is to evaluate how the developed density estimator compares to kernel-based estimators outside the specific setting of pseudo-likelihood estimation, and of graphical models in general. In fact, while the collected evidence suggests that NPMRFs are generally more accurate than GMRFs and KMRFs, an interesting question is how the developed semiparametric estimator would behave if plugged into other kinds of graphical models, or if used as a standalone density estimation technique.

**Acknowledgments.**

This work has been partially supported by a grant from the French National Research Agency (ANR-09-EMER-007). The author is grateful to Ilaria Castelli, Duccio Papini, Franco Scarselli, and Edmondo Trentin for their comments on previous drafts of the paper.

## Appendix

The value of $\frac{\mathrm{d}}{\mathrm{d}\,x}\hat{f}_i(x)$ is given by:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\,x}\hat{f}_i(x) &= \frac{\mathrm{d}}{\mathrm{d}\,x}\left(\hat{\mu}_i + \hat{\sigma}_i\hat{h}_i^*(x)\right)\\ &= \hat{\sigma}_i\,\frac{\mathrm{d}}{\mathrm{d}\,x}\hat{h}_i^*(x)\\ &= \hat{\sigma}_i\,\frac{\mathrm{d}}{\mathrm{d}\,x}\Phi^{-1}\big(F_i^*(x)\big)\\ &= \hat{\sigma}_i\,\frac{\mathrm{d}}{\mathrm{d}\,F_i^*(x)}\Phi^{-1}\big(F_i^*(x)\big)\,\frac{\mathrm{d}}{\mathrm{d}\,x}F_i^*(x)\end{aligned} \tag{21}$$

We first derive $\Phi^{-1}\big(F_i^*(x)\big)$ with respect to $F_i^*(x)$:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\,F_i^*(x)}\Phi^{-1}(F_i^*(x)) &= \frac{1}{\frac{\mathrm{d}}{\mathrm{d}\,\Phi^{-1}\left(F_i^*(x)\right)}\Phi\left(\Phi^{-1}(F_i^*(x))\right)}\\ &= \frac{\sqrt{2\pi}}{\exp\left(-\frac{1}{2}\left(\Phi^{-1}(F_i^*(x))\right)^2\right)}\\ &= \frac{\sqrt{2\pi}}{\exp\left(-\frac{\left(\sqrt{2}\,\mathrm{erf}^{-1}(2F_i^*(x)-1)\right)^2}{2}\right)}\\ &= \frac{\sqrt{2\pi}}{\exp\left(-\mathrm{erf}^{-1}\big(2\,F_i^*(x)-1\big)^2\right)}\end{aligned} \tag{22}$$

We also derive $F_i^*(x)$ with respect to $x$:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\,x}F_i^*(x) &= \frac{\mathrm{d}}{\mathrm{d}\,x}\left(\frac{1}{n}\sum_{j=1}^{n}\frac{1}{1+\exp\left(-\frac{x-x_j}{h}\right)}\right)\\ &= \frac{1}{n}\sum_{j=1}^{n}\frac{\mathrm{d}}{\mathrm{d}\,x}\frac{1}{1+\exp\left(-\frac{x-x_j}{h}\right)}\\ &= -\frac{1}{n}\sum_{j=1}^{n}\frac{\frac{\mathrm{d}}{\mathrm{d}\,x}\exp\left(-\frac{x-x_j}{h}\right)}{\left(1+\exp\left(-\frac{x-x_j}{h}\right)\right)^2}\\ &= \frac{1}{n\,h}\sum_{j=1}^{n}\frac{\exp\left(-\frac{x-x_j}{h}\right)}{\left(1+\exp\left(-\frac{x-x_j}{h}\right)\right)^2}\end{aligned} \tag{23}$$

Given Eqs. 22–23, Eq. 17 follows immediately.

## References


Francis R. Bach and Michael I. Jordan. Learning Graphical Models with Mercer Kernels. In *Advances in Neural Information Processing Systems*, pages 1009–1016, 2002.

M. Bertrand-Retali. Convergence uniforme d'un estimateur de la densité par la méthode du noyau. *Revue Roumaine de Mathématiques Pures et Appliqués*, 23:361–385, 1978.

Julian Besag. Statistical Analysis of Non-Lattice Data. *The Statistician*, 24:179–195, 1975.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York (NY), 2006.

Wilfrid J. Dixon. Simplified Estimation from Censored Normal Samples. *The Annals of Mathematical Statistics*, 31(2):385–391, 1960.

Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. John Wiley & Sons, New York (NY), second edition, 2001.

Olivier P. Faugeras. A Quantile-Copula Approach to Conditional Density Estimation. *Journal of Multivariate Analysis*, 100:2083–2099, 2009.

Antonino Freno, Edmondo Trentin, and Marco Gori. Scalable Pseudo-Likelihood Estimation in Hybrid Random Fields. In J.F. Elder, F. Fogelman-Souli, P. Flach, and M. Zaki, editors, *Proceedings of the 15th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2009)*, pages 319–327. ACM, 2009.

Antonino Freno, Edmondo Trentin, and Marco Gori. Kernel-Based Hybrid Random Fields for Nonparametric Density Estimation. In *19th European Conference on Artificial Intelligence (ECAI 2010)*, pages 427–432. IOS Press, 2010.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse Inverse Covariance Estimation with the Graphical Lasso. *Biostatistics*, 9:432–441, 2008.

Reimar Hofmann and Volker Tresp. Discovering Structure in Continous Variables Using Bayesian Networks. In *Advances in Neural Information Processing Systems*, pages 500–506, 1995.

Reimar Hofmann and Volker Tresp. Nonlinear Markov Networks for Continous Variables. In *Advances in Neural Information Processing Systems*, 1997.

Wilfred Kaplan. *Advanced Calculus*. Addison-Wesley, Reading (MA), third edition, 1984.

Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, Cambridge (MA), 2009.

Steffen L. Lauritzen. *Graphical Models*. Clarendon Press, Oxford (UK), 1996.

Han Liu, John Lafferty, and Larry Wasserman. The Nonparanormal: Semiparametric Estimation of High Dimensional Undirected Graphs. *Journal of Machine Learning Research*, 10:2295–2328, 2009.

Dimitris Margaritis. Distribution-Free Learning of Bayesian Network Structure in Continuous Domains. In *AAAI*, pages 825–830, 2005.

Roger B. Nelsen. *An Introduction to Copulas*. Springer-Verlag, New York (NY), second edition, 2006.

Jennifer Neville and David Jensen. Relational Dependency Networks. *Journal of Machine Learning Research*, 8:653–692, 2007.

Emanuel Parzen. On Estimation of a Probability Density Function and Mode. *Annals of Mathematical Statistics*, 33:1065–1076, 1962.

Pradeep Ravikumar, Garvesh Raskutti, Martin Wainwright, and Bin Yu. Model Selection in Gaussian Graphical Models: High-Dimensional Consistency of $\ell_1$-regularized MLE. In *Advances in Neural Information Processing Systems*, pages 1329–1336, 2008.

Matthew Richardson and Pedro Domingos. Markov Logic Networks. *Machine Learning*, 62:107–136, 2006.

M. Rosenblatt. Conditional Probability Density and Regression Estimators. In P.R. Krishnaiah, editor, *Multivariate Analysis*, volume II, pages 25–31. Academic Press, New York, 1969.

Bernard W. Silverman. Weak and Strong Uniform Consistency of the Kernel Estimate of a Density and Its Derivatives. *The Annals of Statistics*, 6:177–184, 1978.

Bernard W. Silverman. Addendum to Weak and Strong Uniform Consistency of the Kernel Estimate of a Density and Its Derivatives. *The Annals of Statistics*, 8:1175–1176, 1980.

Bernard W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, 1986.

Abe Sklar. Fonctions de répartition à $n$ dimensions et leurs marges. *Publications de l'Institute de Statistique de L'Université de Paris*, 8:229–231, 1959.

David Strauss and Michael Ikeda. Pseudolilkelihood Estimation for Social Networks. *Journal of the American Statistical Association*, 85:204–212, 1990.