The nonparanormal distribution for undirected graphical models

Eviatar Bach

13 July 2016

Machine Learning Reading Group, UBC Department of Computer Science

Sources

[1] H. Liu, J. Lafferty, and L. Wasserman. "The Nonparanormal: Semiparametric Estimation of High Dimensional Undirected Graphs". In: *Journal of Machine Learning Research* 10 (2009). URL: https://arxiv.org/abs/0903.0649.

[2] H. Liu et al. "High-dimensional semiparametric Gaussian copula graphical models". In: *The Annals of Statistics* 40.4 (2012). URL: http://projecteuclid.org/euclid.aos/1358951383.

[3] K. P. Murphy. *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.

[4] M. Schmidt. *CPSC 540 Lecture Slides*. URL: https://www.cs.ubc.ca/~schmidtm/Courses/540-W16/.

Joint distributions

We want to represent a *continuous* joint probability distribution $p(\mathsf{x}|\theta)$.

Joint distributions

We want to represent a *continuous* joint probability distribution $p(\mathbf{x}|\theta)$.

Using the chain rule,

$$p(\mathbf{x}_{1:V}) = p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_2, x_1)p(x_4 \mid x_1, x_2, x_3) \ldots p(x_V \mid \mathbf{x}_{1:V-1}).$$

Joint distributions

We want to represent a *continuous* joint probability distribution $p(\mathbf{x}|\theta)$.

Using the chain rule,

$$p(\mathbf{x}_{1:V}) = p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_2, x_1)p(x_4 \mid x_1, x_2, x_3) \ldots p(x_V \mid \mathbf{x}_{1:V-1}).$$

To be able to represent large joint distributions, we need to make conditional independence assumptions.

Conditional independence

Conditional independence:

$$X \perp Y \mid Z \iff p(X, Y \mid Z) = p(X \mid Z)p(Y \mid Z)$$

Conditional independence

Conditional independence:

$$X \perp Y \,|\, Z \iff p(X, Y \,|\, Z) = p(X \,|\, Z) p(Y \,|\, Z)$$

One of the most common assumptions is the Markov assumption $\mathbf{x}_{t+1} \perp \mathbf{x}_{1:t-1} \,|\, x_t$.

Conditional independence

Conditional independence:

$$X \perp Y \,|\, Z \iff p(X, Y \,|\, Z) = p(X \,|\, Z) p(Y \,|\, Z)$$

One of the most common assumptions is the Markov assumption $\mathbf{x}_{t+1} \perp \mathbf{x}_{1:t-1} \,|\, x_t$.

Then the joint distribution can be written:

$$p(\mathbf{x}_{1:V}) = p(x_1) \prod_{t=1}^{V} p(x_t \,|\, x_{t-1})$$

Conditional independence:

$$X \perp Y \,|\, Z \iff p(X, Y \,|\, Z) = p(X \,|\, Z) p(Y \,|\, Z)$$

One of the most common assumptions is the Markov assumption $\mathbf{x}_{t+1} \perp \mathbf{x}_{1:t-1} \,|\, x_t$.

Then the joint distribution can be written:

$$p(\mathbf{x}_{1:V}) = p(x_1) \prod_{t=1}^{V} p(x_t \,|\, x_{t-1})$$

Graphical models are a more powerful generalization.

Graphical models

Graphical models allow for more complex dependence assumptions.

Graphical models

Graphical models allow for more complex dependence assumptions.

Directed graphical models assume the property $x_s \perp \mathbf{x}_{\text{pred}(s) \setminus \text{pa}(s)} \mid \mathbf{x}_{\text{pa}(s)}$, where pred($s$) is the node's predecessors (which can be defined in a directed acyclic graph) and pa(s) is the node's parents.

Graphical models

Graphical models allow for more complex dependence assumptions.

Directed graphical models assume the property $x_s \perp \mathbf{x}_{\mathrm{pred}(s) \setminus \mathrm{pa}(s)} \mid \mathbf{x}_{\mathrm{pa}(s)}$, where pred($s$) is the node's predecessors (which can be defined in a directed acyclic graph) and pa(s) is the node's parents.

$$x_1 \longrightarrow x_2 \longrightarrow x_3$$

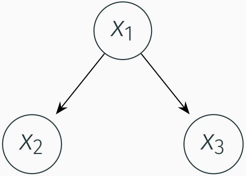
$$p(\mathbf{x}_{1:3}) = p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_2)$$

Graphical models

Graphical models allow for more complex dependence assumptions.

Directed graphical models assume the property $x_s \perp \mathbf{x}_{\mathrm{pred}(s) \setminus \mathrm{pa}(s)} \mid \mathbf{x}_{\mathrm{pa}(s)}$, where pred($s$) is the node's predecessors (which can be defined in a directed acyclic graph) and pa(s) is the node's parents.

$$p(\mathbf{x}_{1:3}) = p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_2)$$



$$p(\mathbf{x}_{1:3}) = p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1)$$

Undirected graphical models

Undirected graphical models have the *global Markov property*: $x_s \perp \mathbf{x}_{\setminus(\mathrm{mb}(s)\cup\{s\})} \mid \mathbf{x}_{\mathrm{mb}(s)}$, where $\mathrm{mb}(s)$ is the *Markov blanket* of s, its neighbours in the graph

Undirected graphical models

Undirected graphical models have the *global Markov property*: $x_s \perp \mathbf{x}_{\setminus(\mathrm{mb}(s) \cup \{s\})} \mid \mathbf{x}_{\mathrm{mb}(s)}$, where $\mathrm{mb}(s)$ is the *Markov blanket* of s, its neighbours in the graph

Some distributions can only be represented by directed graphical models, some with only undirected

Undirected graphical models

Undirected graphical models have the *global Markov property*: $x_s \perp \mathbf{x}_{\setminus(\mathrm{mb}(s) \cup \{s\})} \mid \mathbf{x}_{\mathrm{mb}(s)}$, where $\mathrm{mb}(s)$ is the *Markov blanket* of s, its neighbours in the graph

Some distributions can only be represented by directed graphical models, some with only undirected

In *pairwise Markov random fields*, a potential $\psi_{ij}(x_i, x_j)$ is associated with each edge $(i, j) \in \mathcal{E}$, and the joint distribution is

$$p(\mathbf{x}) \propto \prod_{(i,j) \in \mathcal{E}} \psi_{ij}(x_i, x_j).$$

Gaussian Markov random fields

Gaussian Markov random fields are pairwise Markov random fields with a Gaussian joint distribution.

Gaussian Markov random fields

Gaussian Markov random fields are pairwise Markov random fields with a Gaussian joint distribution.

The pairwise potentials are also Gaussian:

$$\begin{aligned} p(\mathbf{x}) &\propto \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) \\ &\propto \exp\left(-\frac{1}{2}\mathbf{x}^T\Sigma^{-1}\mathbf{x} + \mathbf{x}^T\underbrace{\Sigma^{-1}\boldsymbol{\mu}}_{\boldsymbol{\eta}}\right) \\ &\propto \exp\left(-\frac{1}{2}\sum_{i=1}^{d}\sum_{j=1}^{d}x_i x_j \boldsymbol{\Sigma}_{ij}^{-1} + \sum_{i=1}^{d}x_i\eta_i\right) \\ &= \left(\prod_{i=1}^{d}\prod_{j=1}^{d}\underbrace{\exp\left(-\frac{1}{2}x_i x_j\boldsymbol{\Sigma}_{ij}^{-1}\right)}_{\psi_{ij}(x_i,x_j)}\right)\left(\prod_{i=1}^{d}\underbrace{\exp(x_i\eta_i)}_{\psi_i(x_i)}\right) \end{aligned}$$

Motivation for non-Gaussian models

A multivariate Gaussian joint distribution is a significant restriction.

Motivation for non-Gaussian models

A multivariate Gaussian joint distribution is a significant restriction.

We can get more flexibility using the *nonparanormal* (nonparametric normal) distribution. Here we estimate a transformation f_j for each variable, and assume that the *transformed* data is jointly Gaussian.

Motivation for non-Gaussian models

A multivariate Gaussian joint distribution is a significant restriction.

We can get more flexibility using the *nonparanormal* (nonparametric normal) distribution. Here we estimate a transformation f_j for each variable, and assume that the *transformed* data is jointly Gaussian.

The nonparanormal distribution was introduced in 2009 by Liu, Lafferty, and Wasserman [1]. To understand it we need to first go over copulas.

Copulas: introduction

Start with random vector $(X_1, X_2, \ldots, X_d)$. We only assume that each variable X_i has a continuous CDF $F_i(x) = \mathbb{P}(X_i \leq x)$.

Copulas: introduction

Start with random vector $(X_1, X_2, \ldots, X_d)$. We only assume that each variable X_i has a continuous CDF $F_i(x) = \mathbb{P}(X_i \leq x)$.

Then consider the vector $\mathbf{U} = (U_1, U_2, \ldots, U_d) = (F_1(X_1), F_2(X_2), \ldots, F_d(X_d))$. Notice that we are "feeding back" each variable into *its own CDF*.

Copulas: introduction

Start with random vector $(X_1, X_2, \ldots, X_d)$. We only assume that each variable X_i has a continuous CDF $F_i(x) = \mathbb{P}(X_i \leq x)$.

Then consider the vector $\mathbf{U} = (U_1, U_2, \ldots, U_d) = (F_1(X_1), F_2(X_2), \ldots, F_d(X_d))$. Notice that we are "feeding back" each variable into *its own CDF*.

$\mathbf{U}$ has uniform marginals (each U_i is uniformly distributed on $[0, 1]$). Why?

U has uniform marginals: proof

Consider the CDF of U_i:

$$\begin{aligned}
\mathbb{P}(U_i \leq u) &= \mathbb{P}(F_i(X_i) \leq u) \\
&= \mathbb{P}(X_i \leq F_i^{-1}(u)) \\
&= F_i(F_i^{-1}(u)) \\
&= u
\end{aligned}$$

U has uniform marginals: proof

Consider the CDF of U_i:

$$\begin{aligned}\mathbb{P}(U_i \leq u) &= \mathbb{P}(F_i(X_i) \leq u) \\ &= \mathbb{P}(X_i \leq F_i^{-1}(u)) \\ &= F_i(F_i^{-1}(u)) \\ &= u\end{aligned}$$

This is the CDF of a uniform random variable on $[0, 1]$. Thus **U** has uniform marginals.

Copulas

Define the *copula* C of $(X_1, X_2, \dots, X_d)$ as the joint CDF of $\mathbf{U}$:

$$
\begin{aligned}
C(u_1, u_2, \dots, u_d) &= \mathbb{P}(U_1 \le u_1, U_2 \le u_2, \dots, U_d \le u_d) \\
&= \mathbb{P}(X_1 \le F_1^{-1}(u_1), X_2 \le F_2^{-1}(u_2), \dots, X_d \le F_d^{-1}(u_d))
\end{aligned}
$$

Copulas

Define the *copula* C of $(X_1, X_2, \dots, X_d)$ as the joint CDF of $\mathbf{U}$:

$$\begin{aligned} C(u_1, u_2, \dots, u_d) &= \mathbb{P}(U_1 \le u_1, U_2 \le u_2, \dots, U_d \le u_d) \\ &= \mathbb{P}(X_1 \le F_1^{-1}(u_1), X_2 \le F_2^{-1}(u_2), \dots, X_d \le F_d^{-1}(u_d)) \end{aligned}$$

Given any multivariate CDF H, we can see that

$$\begin{aligned} H(x_1, x_2, \dots, x_d) &= \mathbb{P}(X_1 \le x_1, X_2 \le x_2, \dots, X_d \le x_d) \\ &= C(F_1(x_1), F_2(x_2), \dots, F_d(x_d)). \end{aligned}$$

In fact, *any multivariate distribution (not only those with continuous marginals) can be expressed in terms of its marginals and copula!*

Copulas

Define the *copula* C of $(X_1, X_2, \ldots, X_d)$ as the joint CDF of $\mathbf{U}$:

$$\begin{aligned} C(u_1, u_2, \ldots, u_d) &= \mathbb{P}(U_1 \le u_1, U_2 \le u_2, \ldots, U_d \le u_d) \\ &= \mathbb{P}(X_1 \le F_1^{-1}(u_1), X_2 \le F_2^{-1}(u_2), \ldots, X_d \le F_d^{-1}(u_d)) \end{aligned}$$

Given any multivariate CDF H, we can see that

$$\begin{aligned} H(x_1, x_2, \ldots, x_d) &= \mathbb{P}(X_1 \le x_1, X_2 \le x_2, \ldots, X_d \le x_d) \\ &= C(F_1(x_1), F_2(x_2), \ldots, F_d(x_d)). \end{aligned}$$

In fact, *any multivariate distribution (not only those with continuous marginals) can be expressed in terms of its marginals and copula!*

This is Sklar's Theorem, which also gives uniqueness results for continuous marginals.

The nonparanormal distribution

A random vector $(X_1, X_2, \ldots, X_d)$ has a nonparanormal distribution $NPN(\boldsymbol{\mu}, \Sigma, \{f_j\}_{j=1}^d)$ if there exists a set of functions $\{f_j\}_{j=1}^d$ such that $(f_1(X_1), f_2(X_2), \ldots, f_d(X_d)) \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$.

A random vector $(X_1, X_2, \ldots, X_d)$ has a nonparanormal distribution $NPN(\boldsymbol{\mu}, \Sigma, \{f_j\}_{j=1}^d)$ if there exists a set of functions $\{f_j\}_{j=1}^d$ such that $(f_1(X_1), f_2(X_2), \ldots, f_d(X_d)) \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$.

As a copula, with $\Phi_{\mu,\boldsymbol{\Sigma}}$ the CDF of a multivariate Gaussian $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ and Φ the CDF of the standard normal,

$$F(x_1, x_2, \ldots, x_d) = \Phi_{\mu,\boldsymbol{\Sigma}} \left(\Phi^{-1}(F_1(x_1)), \Phi^{-1}(F_2(x_2)), \ldots, \Phi^{-1}(F_d(x_d))\right)$$

The nonparanormal distribution

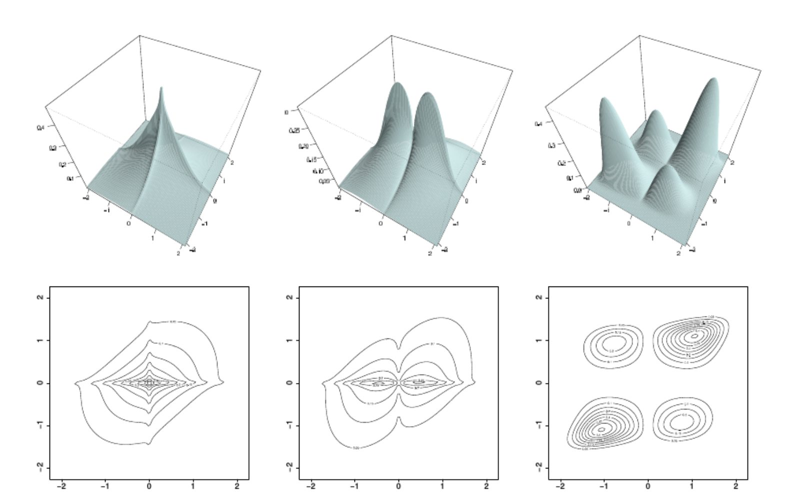
A random vector $(X_1, X_2, \ldots, X_d)$ has a nonparanormal distribution $NPN(\boldsymbol{\mu}, \Sigma, \{f_j\}_{j=1}^d)$ if there exists a set of functions $\{f_j\}_{j=1}^d$ such that $(f_1(X_1), f_2(X_2), \ldots, f_d(X_d)) \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$.

As a copula, with $\Phi_{\mu,\boldsymbol{\Sigma}}$ the CDF of a multivariate Gaussian $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ and Φ the CDF of the standard normal,

$$F(x_1, x_2, \ldots, x_d) = \Phi_{\mu,\boldsymbol{\Sigma}}\left(\Phi^{-1}(F_1(x_1)), \Phi^{-1}(F_2(x_2)), \ldots, \Phi^{-1}(F_d(x_d))\right)$$

The dependence information is encoded in the precision matrix $\Omega = \Sigma^{-1}$: $X_i \perp X_j \,|\, \mathsf{X}_{\setminus\{i,j\}} \iff \Omega_{ij} = 0$

Example densities



Liu, Lafferty, and Wasserman [1]

Estimating a nonparanormal from data

We want to *learn the joint distribution and graph structure from data.*

Estimating a nonparanormal from data

We want to *learn the joint distribution and graph structure from data.*

Estimating the marginals $\{F_j\}_{j=1}^d$ gives us the transformations $\{f_j\}_{j=1}^d$, since

$$F_j(x) = \mathbb{P}(X_j \leq x) = \mathbb{P}(f_j(X_j) \leq f_j(x)) = \Phi\left(\frac{f_j(x) - \mu_j}{\sigma_j}\right),$$

and thus

$$f_j(x) = \mu_j + \sigma_j \Phi^{-1}(F_j(x)).$$

Estimating a nonparanormal from data

We want to *learn the joint distribution and graph structure from data.*

Estimating the marginals $\{F_j\}_{j=1}^d$ gives us the transformations $\{f_j\}_{j=1}^d$, since

$$F_j(x) = \mathbb{P}(X_j \leq x) = \mathbb{P}(f_j(X_j) \leq f_j(x)) = \Phi\left(\frac{f_j(x) - \mu_j}{\sigma_j}\right),$$

and thus

$$f_j(x) = \mu_j + \sigma_j \Phi^{-1}(F_j(x)).$$

Given n data points $X^{(1)}, X^{(2)}, \ldots, X^{(n)}$, F_j can be estimated using the empirical CDF:

$$\hat{F}_j(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}[X_j^{(i)} \leq t]$$

Estimating a nonparanormal from data: continued

The mean is estimated as the sample mean $\boldsymbol{\mu}_n$. See Liu et al. [2] for estimators of the correlation matrix.

Estimating a nonparanormal from data: continued

The mean is estimated as the sample mean $\boldsymbol{\mu}_n$. See Liu et al. [2] for estimators of the correlation matrix.

An ℓ_1-regularized estimator for Ω, to encourage graph sparsity, can be computed using the graphical lasso.

Estimating a nonparanormal from data: continued

The mean is estimated as the sample mean $\boldsymbol{\mu}_n$. See Liu et al. [2] for estimators of the correlation matrix.

An ℓ_1-regularized estimator for Ω, to encourage graph sparsity, can be computed using the graphical lasso.

Consistency results for Ω with respect to the Frobenius norm and the ℓ_2 norm: if the data was generated from a nonparanormal with precision matrix Ω, as $n \to \infty$, $\|\hat{\Omega} - \Omega\| \to 0$.

Estimating a nonparanormal from data: continued

The mean is estimated as the sample mean $\boldsymbol{\mu}_n$. See Liu et al. [2] for estimators of the correlation matrix.

An ℓ_1-regularized estimator for Ω, to encourage graph sparsity, can be computed using the graphical lasso.

Consistency results for Ω with respect to the Frobenius norm and the ℓ_2 norm: if the data was generated from a nonparanormal with precision matrix Ω, as $n \to \infty$, $\|\hat{\Omega} - \Omega\| \to 0$.

Achieves the same rate of convergence as the Gaussian model. The authors advocate it as a drop-in replacement.

Notes

This is a *semiparametric* model: the parametric part is estimating $\boldsymbol{\mu}$ and Σ, the nonparametric part is estimating $\{f_j\}_{j=1}^d$.

Notes

This is a *semiparametric* model: the parametric part is estimating $\boldsymbol{\mu}$ and Σ, the nonparametric part is estimating $\{f_j\}_{j=1}^{d}$.

For most recent details on implementation and convergence, see Liu et al. [2].

Notes

This is a *semiparametric* model: the parametric part is estimating $\boldsymbol{\mu}$ and Σ, the nonparametric part is estimating $\{f_j\}_{j=1}^d$.

For most recent details on implementation and convergence, see Liu et al. [2].

The R package huge implements undirected graph estimation with the nonparanormal distribution.