The Expxorcist

Nonparametric Graphical Models Via Conditional Exponential Densities

Pradeep Ravikumar
Machine Learning Department
School of Computer Science
Carnegie Mellon University

Nonparametric Density Estimation

Let F be some distribution, with density $f \in \mathcal{F}$.

Given data $X_1, \ldots, X_n \sim F$.

Non-parametric Density Estimation: estimate f given $\{X_i\}_{i=1}^n$ given infinite-dimensional space $\mathcal{F}$.

Nonparametric Density Estimation

Let F be some distribution, with density $f \in \mathcal{F}$.

Given data $X_1, \ldots, X_n \sim F$.

Non-parametric Density Estimation: estimate f given $\{X_i\}_{i=1}^n$ given infinite-dimensional space $\mathcal{F}$.

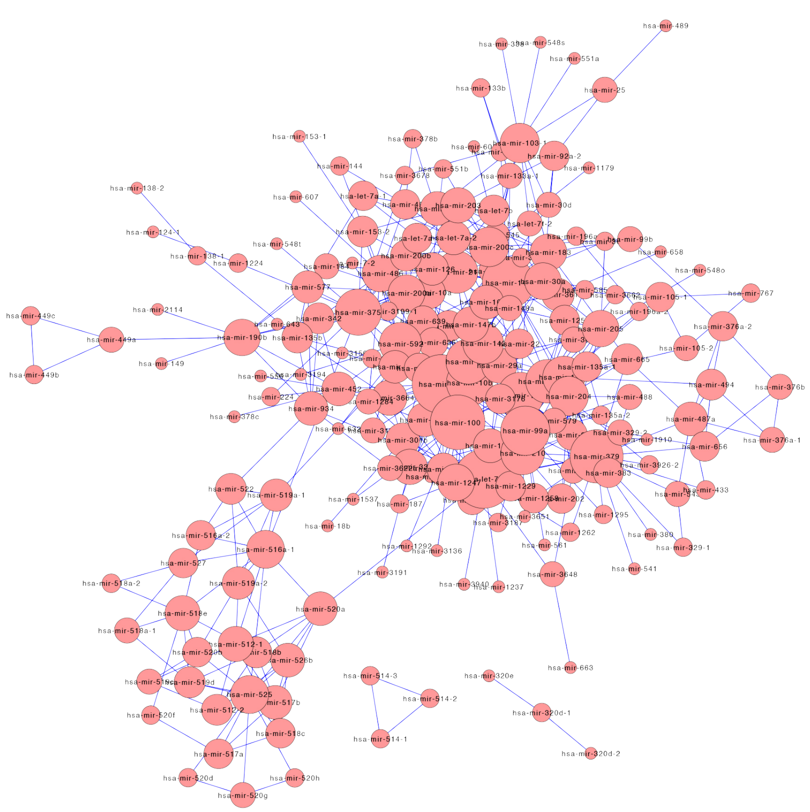
- preferably: making as few assumptions about $\mathcal{F}$ as possible

Why Density Estimation?

- An important “**unsupervised learning**” problem
  - density summarizes the data without any supervision
- Can perform **probabilistic reasoning**
  - how likely is some future event given evidence so far (e.g. how likely is it to have large value for invasive diagnostic test given other symptoms)
  - given joint density over all variables, can compute conditional probabilities of variables of interest given values of other variables
- Given density, can **compute functionals** of interest
  - entropy, moments, …

Why Density Estimation?

- An important functional is the **conditional independence graph**
  - one node for each variable, and no edge between two variables if they are conditionally independent given other variables
- Conditional Independence Graph provides a much sparser “dependency graph” than correlation or thresholded correlation; connotes a more “direct” dependence



- MicroRNA network learnt from The Cancer Genome Atlas (TCGA) Breast Cancer Level II Data

Example: Kernel Density Estimation

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

is consistent for a broad class of density classes

Running Function Class: Sobolev order 2

$$\mathcal{F}_2(c) := \left\{ f : \int |f^{(2)}(x)|^2 dx < c^2 \right\}$$

Let $R_f(\widehat{f}_n) := \mathbb{E}_f \int (\widehat{f}_n(x) - f(x))^2 \, dx.$

$\inf_{\widehat{f}_n} \sup_{f \in \mathcal{F}_2(c)} R_f(\widehat{f}_n) \asymp n^{-4/5}.$

Running Function Class: Sobolev order 2

$$\mathcal{F}_2(c) := \left\{ f : \int |f^{(2)}(x)|^2 dx < c^2 \right\}$$

Let $R_f(\widehat{f}_n) := \mathbb{E}_f \int (\widehat{f}_n(x) - f(x))^2 \, dx$.

$\inf_{\widehat{f}_n} \sup_{f \in \mathcal{F}_2(c)} R_f(\widehat{f}_n) \asymp n^{-4/5}$.

Achieved by kernel density estimation (for appropriate setting of bandwidth)

Running Function Class: Sobolev order 2

- But in higher dimensions, where x is d-dimensional:

Let $R_f(\widehat{f}_n) := \mathbb{E}_f \int (\widehat{f}_n(x) - f(x))^2 \, dx$.

$\inf_{\widehat{f}_n} \sup_{f \in \mathcal{F}_2(c)} R_f(\widehat{f}_n) \asymp n^{-4/(4+d)}$.

(also achieved by
kernel density estimation)

Running Function Class: Sobolev order 2

- But in higher dimensions, where x is d-dimensional:

Let $R_f(\widehat{f}_n) := \mathbb{E}_f \int (\widehat{f}_n(x) - f(x))^2 \, dx.$

$\inf_{\widehat{f}_n} \sup_{f \in \mathcal{F}_2(c)} R_f(\widehat{f}_n) \asymp n^{-4/(4+d)}.$

(also achieved by kernel density estimation)

For risk $R(\widehat{f}_n) \leq \epsilon$, need number of samples $n \geq C \; \left(\frac{1}{\epsilon}\right)^{1+\frac{d}{4}}$

no. of samples required scales **exponentially** with dimension d

Non-parametric Density Estimation

- For lower sample complexity, need to impose some "structure" on the density function
- Typically, we impose this structure on the logistic transform of the density $\eta(x)$ s.t.

$$f(x) = \frac{\exp(\eta(x))}{\int_x \exp(\eta(x))dx}$$

Non-parametric Density Estimation

- Estimate logistic transform \eta(x) from data
  - can impose constraints without worrying about positivity and normalizability
  - still has the same exponential sample complexity

Common Structural Assumptions: RKHS

Assume $\eta(x)$ lies in a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}_k$ with respect to some kernel function $k(\cdot, \cdot)$.

Silverman 82, Gu, Qiu 93, Canu, Smola 06

- still has exponential sample complexity
- also has computational caveat of how to solve infinite-dimensional estimation problem
    - finite-dimensional approximations of function spaces (but with no statistical guarantees)

Common Assumptions: ANOVA Decomposition

$$\eta(x) = \sum_{s} \eta_s(x_s) + \sum_{(s,t)} \eta_{st}(x_s, x_t) + \ldots$$

Gu et al, 13, Sun et al,15

- sample complexity analyses unavailable
- computationally motivated finite-dimensional approximations of function spaces (with no statistical guarantees)

Common Structural Assumptions

- Setting aside statistical i.e. sample complexity analyses, these require computationally motivated approximations

1. Finite-dimensional approximations of infinite-dimensional function space of logistic transform \eta(x)

2. Surrogate likelihoods, since log-likelihood is intractable due to normalization constant

$$\int \exp(\eta(x))\, dx$$

Expxorcist

- Makes the structural assumption:

$$\eta(x) = \sum_s \theta_s B_s(x_s) + \sum_{st} \theta_{st} B_s(x_s)\, B_t(x_t)$$

- Why “expxorcist”
    - follows “ghostbusting” naming trend for non-parametric densities: non-paranormal, and non-paranormal skeptic (Gaussian Copulas)
    - uses conditional exponential densities (clarified shortly)
- Computational tractable estimator
- Strong statistical guarantees (n^{-4/5} convergence rate for risk)

Conditional Densities

Joint Density:

$$f(X) \propto \exp\left(\sum_{s\in V}\theta_s B_s(x_s) + \sum_{(s,t)\in E}\theta_{st} B_s(x_s)\, B_t(x_t) + \sum_{s\in V} C_s(x_s)\right)$$

where $\prod_{s\in V}\exp(C_s(x_s))$ is a given product base measure.

Node-conditional Density:

$$f(X_s|X_{-s}) \propto \exp\left(B_s(x_s)\left(\theta_s + \sum_{t\in N(s)}\theta_{st} B_t(x_t)\right) + C_s(x_s)\right)$$

- node-conditional density has exponential family form
    - with sufficient statistics B_s(.)
    - natural parameter that is a linear function of sufficient statistics of other node-conditional densities

Node-conditional Densities

Theorem (Yang, Ravikumar, Allen, Liu 15):
The set of node-conditional densities:

$$f(X_s|X_{-s}) \propto \exp\left(B_s(x_s)\left(\theta_s + \sum_{t\in N(s)} \theta_{st} B_t(x_t)\right) + C_s(x_s)\right)$$

are all consistent with a unique joint density:

$$f(X) \propto \exp\left(\sum_{s\in V} \theta_s B_s(x_s) + \sum_{(s,t)\in E} \theta_{st} B_s(x_s)\, B_t(x_t) + \sum_{s\in V} C_s(x_s)\right)$$

Node-conditional Densities

A more general set of node-conditional densities:

$$f(x_s|x_{-s}) \propto \exp(h(x_s, x_{-s}) + C_s(x_s))$$

need not be consistent with a unique joint density.

Arnold et al. 01, Berti et al. 14, ...

Conditional Density of Exponential Family Form

General conditional density:

$$f(x_s|x_{-s}) \propto \exp(h(x_s, x_{-s}) + C_s(x_s))$$

Conditional density of **exponential family form**:

$$f(x_s|x_{-s}) \propto \exp(B_s(x_s)\, E_s(x_{-s}) + C_s(x_s))$$

Thus, conditional density of exponential family form has its logistic transform that factorizes:

$$h(x_s, x_{-s}) = B_s(x_s)\, E_s(x_{-s})$$

Node-conditional Densities

Theorem (Yang, Ravikumar, Allen, Liu 15):
The set of node-conditional densities:

$$f(x_s|x_{-s}) \propto \exp(B_s(x_s)\, E_s(x_{-s}) + C_s(x_s))$$

are all consistent with a joint density **iff**:

$$E_s(x_{-s}) = \theta_s + \sum_{t \in V} \theta_{st} B_t(x_t)$$

and the resulting unique joint density has the form:

$$f(X) \propto \exp\left(\sum_{s \in V} \theta_s B_s(x_s) + \sum_{(s,t) \in E} \theta_{st} B_s(x_s)\, B_t(x_t) + \sum_{s \in V} C_s(x_s) \right)$$

Thus the **expxorcist** class of densities follows without loss of much generality, in particular, if we make the very general assumption that the node-conditional densities have “exponential family form”

Estimation of Expxorcist Densities

- Expxorcist node-conditional densities are consistent with a unique expxorcist joint density
- We reduce joint density estimation to a set of node-conditional density estimation problems

We estimate:

$$f(X_s|X_{-s}) \propto \exp\left(B_s(x_s)\left(\theta_s + \sum_{t\in N(s)} \theta_{st} B_t(x_t)\right) + C_s(x_s)\right)$$

assuming, for identifiability that:

$\int B_s(x_s)dx_s = 0, \int B_s^2(x_s)dx_s = 1, \text{ and } \theta_s \geq 0.$

Estimation of Expxorcist Densities

Let:

$$\mathcal{L}_s(B; \mathbb{X}_n) = \frac{1}{n} \sum\nolimits_{i=1}^{n} \left\{ -B_s(X_s^{(i)}) \left(1 + \sum\nolimits_{t \in V \setminus s} B_t(X_t^{(i)}) \right) + A(X_{-s}^{(i)}; B) \right\},$$

With some re-parameterization, ell_1 regularized node-conditional MLE can be written as:

$$\min_{B} \; \mathcal{L}_s(B; \mathbb{X}_n) + \lambda_n \sum\nolimits_{t \in V} \sqrt{\int_{\mathcal{X}_t} B_t(X)^2 dX}$$
$$\text{s.t.} \; \int_{\mathcal{X}_t} B_t(X) dX = 0 \quad \forall t \in V.$$

Estimation of Expxorcist Densities

Suppose we are given a uniformly bounded orthonormal basis $\{\phi_k(\cdot)\}_{k=0}^{\infty}$ for the function space of $\{B_s(\cdot)\}_{s\in V}$.

Expansion of B_t(.) in terms of this basis yields:

$$B_t(X) = \sum\nolimits_{k=1}^{m} \alpha_{t,k}\phi_k(X) + \rho_{t,m}(X) \quad \text{where} \quad \rho_{t,m}(X) = \alpha_{t,0}\phi_0(X) + \sum\nolimits_{k=m+1}^{\infty} \alpha_{t,k}\phi_k(X).$$

Then the infinite-dimensional problem earlier, can be approximated as:

$$\min_{\alpha_{\mathbf{m}}} \mathcal{L}_{s,m}(\alpha_{\mathbf{m}}; \mathbb{X}_n) + \lambda_n \sum_{t\in V} \|\alpha_{\mathbf{t},\mathbf{m}}\|_2,$$

where $\alpha_{\mathbf{t},\mathbf{m}} = \{\alpha_{t,k}\}_{k=1}^{m}$, $\alpha_{\mathbf{m}} = \{\alpha_{\mathbf{t},\mathbf{m}}\}_{t\in V}$ and $\mathcal{L}_{s,m}$ is defined as

$$\mathcal{L}_{s,m}(\alpha_{\mathbf{m}}; \mathbb{X}_n) = \frac{1}{n}\sum_{i=1}^{n}\left\{-\sum_{k=1}^{m}\alpha_{s,k}\phi_k(X_s^{(i)})\left(1 + \sum_{t\in V\setminus\{s\}}\sum_{l=1}^{m}\alpha_{t,l}\phi_l(X_t^{(i)})\right) + A(X_{-s}^{(i)}; \alpha_{\mathbf{m}})\right\}.$$

Expxorcist Estimation

$$\min_{\alpha_{\mathbf{m}}} \mathcal{L}_{s,m}(\alpha_{\mathbf{m}}; \mathbb{X}_n) + \lambda_n \sum_{t \in V} \|\alpha_{\mathbf{t},\mathbf{m}}\|_2,$$

where $\alpha_{\mathbf{t},\mathbf{m}} = \{\alpha_{t,k}\}_{k=1}^{m}$, $\alpha_{\mathbf{m}} = \{\alpha_{\mathbf{t},\mathbf{m}}\}_{t \in V}$ and $\mathcal{L}_{s,m}$ is defined as

$$\mathcal{L}_{s,m}(\alpha_{\mathbf{m}}; \mathbb{X}_n) = \frac{1}{n} \sum_{i=1}^{n} \left\{ -\sum_{k=1}^{m} \alpha_{s,k} \phi_k(X_s^{(i)}) \left(1 + \sum_{t \in V \setminus \{s\}} \sum_{l=1}^{m} \alpha_{t,l} \phi_l(X_t^{(i)}) \right) + A(X_{-s}^{(i)}; \alpha_{\mathbf{m}}) \right\}.$$

- Non-convex
- But can compute a local minimum efficiently using alternating minimization and proximal gradient descent

Statistical Guarantees

- Theorem (Suggala, Kolar, Ravikumar 17):

Under some regularity conditions, any local minimum of the expxorcist density estimation problem $\widehat{\alpha}_{\mathbf{m}}$ satisfies:

$$\|\widehat{\alpha}_{\mathbf{m}} - \alpha^*_{\mathbf{m}}\|_2 \leq C\sqrt{d}\,\|\nabla \mathcal{L}_{sm}(\alpha^*_{\mathbf{m}})\|_\infty,$$

where $d :=$ maximum node-degree of conditional independence graph of multivariate expxorcist density.

Statistical Guarantees

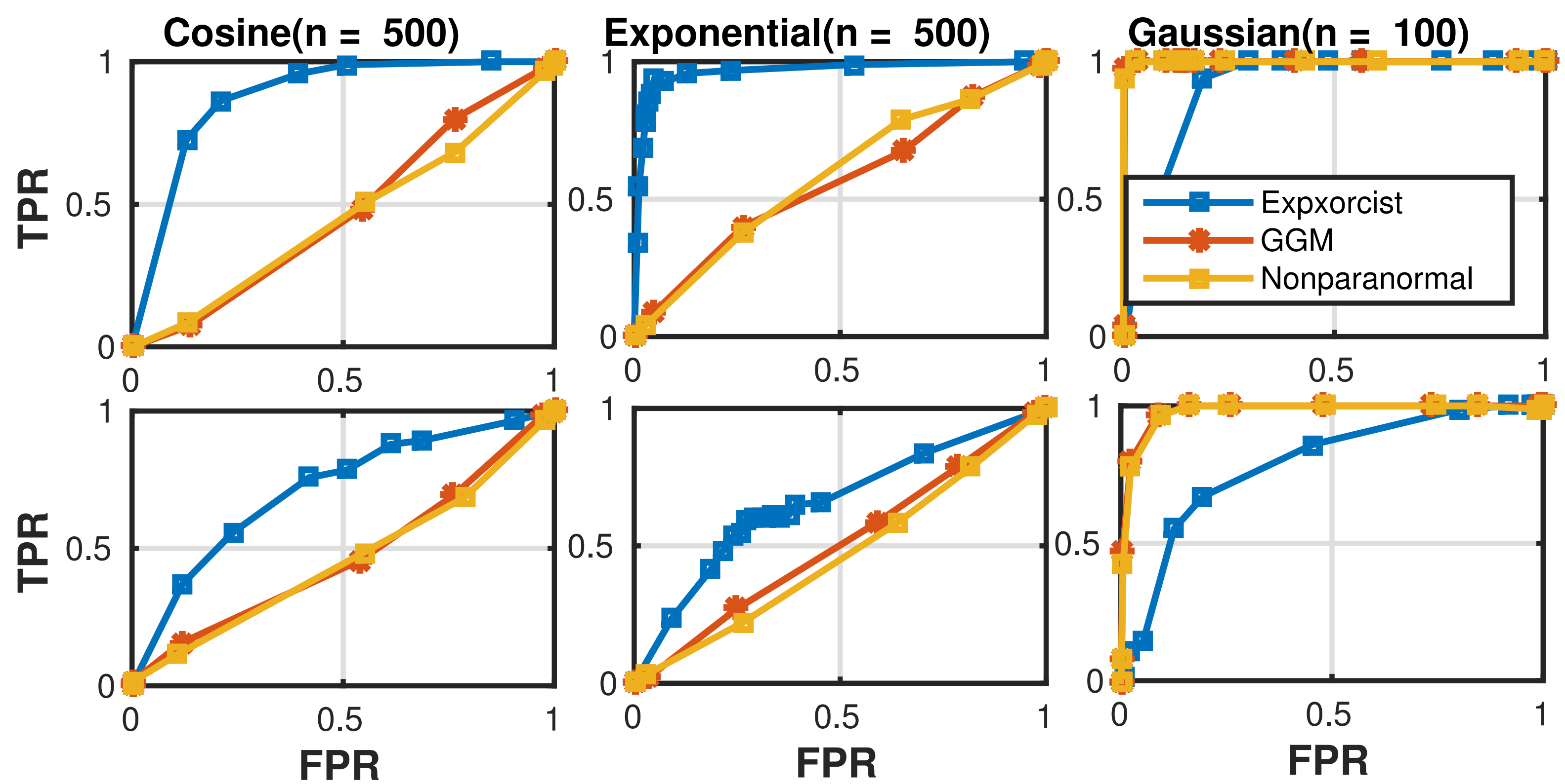
- Corollary (Suggala, Kolar, Ravikumar 17):

Suppose $B_s(\cdot) \in \mathcal{F}_2(c)$, $\{\phi_k\}_{k=0}^{\infty}$ be the trigonometric basis of $\mathcal{F}_2(c)$, and let $d :=$ maximum node-degree of conditional independence graph of multivariate expxorcist density. Then under some regularity conditions, any local minimum expxorcist node-conditional density estimate $\widehat{f}_n$ satisfies:

$$R_f(\widehat{f}_n) \asymp d^3 \, (\log p)^4 \, n^{-4/5}.$$

- One-dimensional non-parametric rate, dependence on dimension p is logarithmic

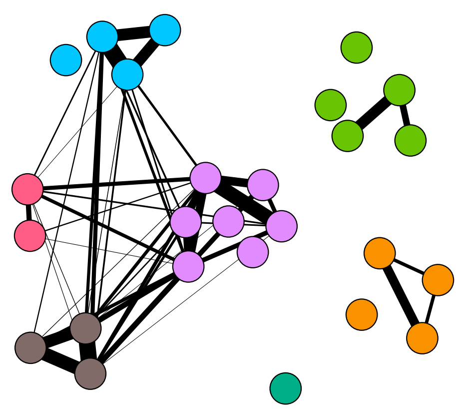
ROC Plots for estimating Conditional Independence Graph



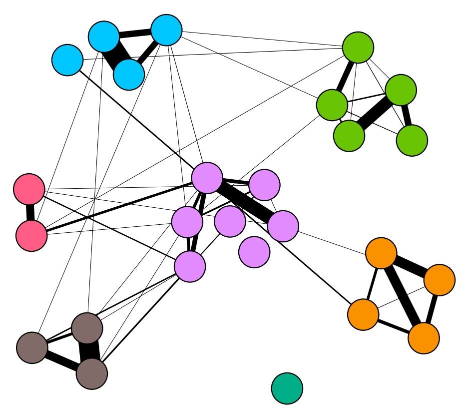
Top: chain graphs, Bottom: grid graphs
Columns correspond to different multivariate densities

- Generated synthetic data from multivariate densities with different non-linearities (cosine, exponential, Gaussian)
- Our non-parametric estimator (expxorcist) recovers these adaptively

Futures Intraday Data



Gaussian Copulas



Expxorcist

- Top 26 most liquid instruments (traded at CME)
- 1 minute price returns (from 9 AM - 3 PM Eastern); multimodal, fat tailed
- 895 training, 650 test samples
- Expxorcist can be seen to identify clusters better

Summary

- General non-parametric density estimation has high sample complexity in high dimensions
- Need to impose structural assumptions
- Expxorcist imposes the following very natural non-parametric assumptions:
  - Node-conditonal densities follow "exponential family form" (for unknown sufficient statistics)
  - Conditional Independence Graph of density is sparse
- We propose a computationally practical estimator with strong statistical guarantees:
  - node-conditional density estimation has one-dimensional non-parametric rate