Nonparametric Forest Density Estimation

John Lafferty

Department of Statistics &
Department of Computer Science
University of Chicago
Collaborators

Anupam Gupta
CMU CS

Han Liu
Johns Hopkins

Larry Wasserman
CMU Stat/ML

Min Xu
CMU ML
High dimensional density estimation

Given sample $X_1, \ldots, X_n$ from a $d$-dimensional distribution, estimate density $p(x_1, \ldots, x_d) \geq 0,$

$$\int_{\mathbb{R}^d} p(x) \, dx = 1$$

- Want to make weakest possible assumptions
- Statistically and computationally intractable in general
- Need structural and/or distributional assumptions
Graph of a random vector $X = (X_1, \ldots, X_d)$:

Vertices corresponding to variables $X_1, \ldots, X_d$. Omits edge between nodes $X_i$ and $X_j$ if and only if $X_i$ and $X_j$ are conditionally independent given the other variables.

- Prediction
- Interpretation
Nonparametric Graphical Models

General family of densities with independence graph $G$:

$$p(x) \propto \exp \left\{ \sum_{\text{cliques } C} f_C(x_C) \right\}$$

- “Parameters” are the functions $f_C(x_C)$
- Computationally intractable to compute normalizer
- Formulating tractable subfamilies is major problem.
Examples

S&P 500 data

Human gene data (HapMap)
• Setup for forest density estimation
• Risk consistency results
• Tree-restricted forests
• Fast rates for forest selection
• Fast rates for semiparametric graphs
• Experiments
A distribution is supported on a forest $F$ with edge set $E(F)$ if

$$p_F(x) = \prod_{(i,j) \in E} \frac{p(x_i, x_j)}{p(x_i)p(x_j)} \prod_{k=1}^{d} p(x_k)$$

- Need to estimate bivariate and univariate marginals
- Can be fully nonparametric
- We do not assume the true graph is a forest
Some Previous Work

- Chow and Liu (1968)
- Bach and Jordan (2003)
- Chechetka and Guestrin (2008)
- Tan et al. (2009)

Our focus: Statistical properties in high dimensions
For a forest-structured density,

\[ p_F(x) = \prod_{(i,j) \in E} \frac{p(x_i, x_j)}{p(x_i)p(x_j)} \prod_{k=1}^{d} p(x_k) \]

\[ \mathbb{E} [\log p_F(X)] = \sum_{(i,j) \in E} I(X_i, X_j) + \text{constant} \]

- For \( p(x_i, x_j) \) known, best tree obtained by Kruskal’s algorithm
- In high dimensions, full spanning tree will overfit
Step 1: Constructing a Full Tree

- Compute kernel density estimates

\[
\hat{p}_{n_1}(x_i, x_j) = \frac{1}{n_1} \sum_{s \in \mathcal{D}_1} \frac{1}{h^2} K \left( \frac{X_i^{(s)} - x_i}{h} \right) K \left( \frac{X_j^{(s)} - x_j}{h} \right)
\]

- Estimate mutual informations

\[
\hat{I}_{n_1}(X_i, X_j) = \frac{1}{m^2} \sum_{k=1}^{m} \sum_{\ell=1}^{m} \hat{p}_{n_1}(x_{ki}, x_{\ell j}) \log \frac{\hat{p}_{n_1}(x_{ki}, x_{\ell j})}{\hat{p}_{n_1}(x_{ki}) \hat{p}_{n_1}(x_{\ell j})}
\]

- Run Kruskal’s algorithm (Chow-Liu) on edge weights
Step 2: Pruning the Tree

- Heldout risk

\[ \hat{R}_{n_2}(p_F) = - \sum_{(i,j) \in E(F)} \int_{X_i \times X_j} \hat{p}_{n_2}(x_i, x_j) \log \frac{p(x_i, x_j)}{p(x_i) p(x_j)} \, dx_i dx_j \]

- Selected forest given by

\[ \hat{k} = \arg \min_{k \in \{0, \ldots, d-1\}} \hat{R}_{n_2} \left( \hat{p}_{\hat{F}^{(k)}_{n_1}} \right) \]

where \( \hat{F}^{(k)}_{n_1} \) is forest obtained after \( k \) steps of Kruskal
Step 2: Alternative Forest Selection

Theoretically optimal procedure asymptotically:

Run Kruskal’s algorithm directly on heldout edge weights

\[
\hat{w}_{n_2}(i, j) = \frac{1}{n_2} \sum_{s \in D_2} \log \frac{\hat{p}_{n_1}(X_i^{(s)}, X_j^{(s)})}{\hat{p}_{n_1}(X_i^{(s)}) \hat{p}_{n_1}(X_j^{(s)})}
\]
Assumptions

- Marginal densities lie in Hölder class with exponent $\beta$
  \[ |p^*(x) - P_{p^*,x_0}^{(\beta)}(x)| \leq L \|x - x_0\|_2^\beta, \quad \forall x \in \mathcal{B}(x_0, (\log n/n)^{1/(2\beta+2)}) \subseteq \mathbb{R}^2 \]

- Bounded densities, decaying to zero not too quickly:
  \[ c_1 \gamma_n \leq \inf_{x} p^*(x_i, x_j) \leq \sup_{x} p^*(x_i, x_j) \leq c_2 \]
  $\mu$-almost surely, where $\gamma_n^2 = \tilde{\Omega} \left( n^{-\beta/2(\beta+1)} \right)$.

- Kernel functions form bounded VC classes

Related technical assumptions in Giné and Guillou (2002) and Rigollet and Vert (2009)
Excess Risk Rates

Theorem. Let \( \hat{p}_{F_d^{(k)}} \) be the forest density obtained by running Kruskal’s algorithm on the heldout data. Then

\[
R(\hat{p}_{F_d^{(k)}}) - \min_F R(\hat{p}_F) = O_P \left( (k^* + \hat{k}) \sqrt{\frac{\log n + \log d}{n^{\beta/(1+\beta)}}} + d \sqrt{\frac{\log n + \log d}{n^{2\beta/(1+2\beta)}}} \right)
\]

where \( k^* \) is the number of edges in \( F^* = \arg \min_F R(\hat{p}_F) \).
Proof Sketch

Establish exponential concentration in sup-norm:

\[ \mathbb{P}\left( \sup_{u \in \mathcal{X}^2} \left| \hat{p}_n(u) - \mathbb{E}\hat{p}(u) \right| > \epsilon \right) \leq L \exp \left\{ -Cn h_2^2 \epsilon^2 \right\} \]

together with control of bias

\[ \sup_{(x_i, x_j) \in \mathcal{X}_i \times \mathcal{X}_j} \left| \mathbb{E}\hat{p}(x_i, x_j) - p^*(x_i, x_j) \right| \leq L h_2^\beta C_K \]

and a union bound to get

\[ \max_{(i,j)} \sup_{(x_i, x_j) \in \mathcal{X}_i \times \mathcal{X}_j} \left( \frac{\hat{p}(x_i, x_j)}{p^*(x_i, x_j)} - 1 \right) = O_P \left( \sqrt{\frac{\log n + \log d}{n^{\beta/(\beta+1)}}} \right) . \]

with bandwidths \( h_2 \asymp \left( \frac{\log n}{n} \right)^{\frac{1}{2+2\beta}} \).
This leads to

$$\sup_{F \in \mathcal{F}_d^{(k)}} |\hat{R}(\hat{p}_F) - R(p^*_F)| \approx \sup_{F \in \mathcal{F}_d^{(k)}} |R(\hat{p}_F) - \hat{R}(\hat{p}_F)|$$

$$= O_P \left( k \sqrt{\frac{\log n + \log d}{n^{\beta/(\beta+1)}}} + d \sqrt{\frac{\log n + \log d}{n^{2\beta/(1+2\beta)}}} \right)$$

and the result comes from chaining these bounds together.
**t-Restricted Forests**

- Asymptotically, Kruskal’s algorithm gives optimal forest
- Overfits for finite data
- One approach: Testing edge weights
- Alternative: Cross-validate over large set of forests

*F* is a *t*-restricted forest if each tree of *F* has $\leq t$ edges.

- Useful for interpretation: Small *t* yields tree-based clusters
Theorem. Maximum weight $t$-restricted forest is NP-hard, for $t \geq 6$.

Proof constructs a reduction from **Exact 3-Cover**: Given a set $X$ and a family $S$ of 3-element subsets, find a subfamily $S' \subset S$ such that every element of $X$ occurs in exactly one member of $S'$.

- Construct graph with special tree-shaped subgraphs called *gadgets*, each corresponding to a 3-element subset in $S$
- Show how finding a maximum weight $t$-restricted forest would recover a solution to X3C
Approximation Algorithm

1. Sort edges in decreasing order of weight
2. Greedily pick a set of edges such that
   (a) The degree of any node is at most $t$
   (b) No cycles are formed
3. Run a dynamic program to select the optimal subforest $\hat{F}_t$ that is $t$-restricted.

**Theorem.** This is a $\frac{1}{4}$-approximation algorithm:

$$\hat{W}(\hat{F}_t) \geq \frac{1}{4} \hat{W}(F^*_t)$$
Proof Sketch

- The family of graphs satisfying (a) degree $\leq t$ and (b) acyclic forms a 2-independence family.

- An optimal forest $F^{**}$ satisfying these constraints satisfies $W(F^{**}) \geq W(F^*_t)$ where $F^*_t$ is an optimal $t$-restricted forest.

- By Haussman, Korte and Jenkyns (1980), greedy is a 2-approximation algorithm.

- Root each tree at a degree-one node, and label edges odd or even, according to distance from root. Odd or even edges form stars that have at least half the weight.
Singh and Lau (1997) give an algorithm based on rounded linear programming that yields a forest with degrees $\leq t + 1$ satisfying

$$W(F_{SL}) \geq W(F^{**}) \geq W(F^*_t)$$

(Ph.D. thesis at CMU Tepper, resolved conjecture of Goemans...)

Now apply the dynamic program, which has a 2-approximation guarantee.
Risk Guarantees

Population edge weights are

\[ W(i, j) = \mathbb{E} \left( \log \frac{\hat{p}(X_i, X_j)}{\hat{p}(X_i) \hat{p}(X_j)} \right) \]

**Theorem.** Let \( \hat{F}_t \) be the forest constructed using a \( c \)-approximation algorithm, and let \( F_t^* \) be the optimal forest (both w.r.t. finite sample edge weights \( \hat{w}_{n_1} = \hat{l}_{n_1} \)). Then

\[ W(\hat{F}_t) \geq \frac{1}{c} W(F_t^*) + O_P \left( (k^* + \hat{k}) \sqrt{\frac{\log n + \log d}{n^{\beta/(1+\beta)}}} \right) \]

where \( \hat{k} \) and \( k^* \) are the number of edges in \( \hat{F}_t \) and \( F_t^* \).
More Efficient Forest Estimation

Recent improvement:

- Use an *undersmoothed* kernel density estimator
- Can show concentration of mutual information
- Achieves parametric rate of convergence, $O(n^{-1/2})$
- Optimal tree and forest estimation
Details: Optimal Mutual Information Estimation

- Assume $p(x_1, x_2)$ from a 2nd-order Hölder class
- $0 < \kappa_1 < \min_x p(x) \leq \max_x p(x) \leq \kappa_2 < \infty$
- $\tilde{p}_h(x) = T_{\kappa_1, \kappa_2}(\hat{p}_h(x))$
- $H(\tilde{p}) = -\int \tilde{p}_h(x) \log \tilde{p}_h(x) \, dx$
A calculation shows that
\[
\sup_{x_1', \ldots, (x_j)', \ldots, x_n} |H(\tilde{p}_h) - H(\tilde{p}_h')| \leq \frac{8c}{n}
\]

By McDiarmid’s inequality,
\[
P \left( |H(\tilde{p}_h) - \mathbb{E}H(\tilde{p}_h)| > \epsilon \right) \leq 2 \exp \left( - \frac{n\epsilon^2}{32c^2} \right)
\]

After removing boundary bias by reflection,
\[
\sup_p \left| \mathbb{E}H(\tilde{p}_h) - H(p) \right| \leq \frac{C}{\sqrt{n}} \quad \text{with} \quad h \asymp 1/n^{1/4}
\]

Putting pieces together
\[
\sup_p P \left( \left| \mathbb{E}H(\tilde{p}_h) - H(p) \right| > \epsilon \right) \leq 2 \exp \left( - \frac{n\epsilon^2}{128c^2} \right)
\]
General nonparametric graphical model takes the form

\[ p(x) \propto \exp \left\{ \sum_{\text{cliques } C} f_C(x_C) \right\} \]

- Computationally intractable to compute normalizer
- Formulating tractable subfamilies is major problem.
Gaussian Case

- If $X \sim N(\mu, \Sigma)$ then there is no edge between $X_i$ and $X_j$ if and only if
  $$\Omega_{ij} = 0$$
  where $\Omega = \Sigma^{-1}$.

- Given
  $$X^1, \ldots, X^n \sim N(\mu, \Sigma).$$

- For $n > p$, let
  $$\hat{\Omega} = \hat{\Sigma}^{-1}$$
  and test
  $$H_0 : \Omega_{ij} = 0 \quad \text{versus} \quad H_1 : \Omega_{ij} \neq 0.$$
Penalized log-likelihood:

\[-\ell(\Omega) + \lambda \sum_{j \neq k} |\Omega_{jk}|\]

where

\[\ell(\Omega) = \frac{1}{2} (\log |\Omega| - \text{tr}(\Omega S))\]

(maximized over \(\mu\)).

Can be solved with a simple blockwise gradient descent algorithm that iterates over different \(\ell_1\)-penalized regressions (lasso).
A random vector \( X = (X_1, \ldots, X_d)^T \) has a \textit{nonparanormal} distribution

\[
X \sim \text{NPN}(\mu, \Sigma, f)
\]

in case

\[
Z \equiv f(X) \sim N(\mu, \Sigma)
\]

where \( f(X) = (f_1(X_1), \ldots, f_d(X_d)) \).

Joint density

\[
p_X(x) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp \left\{ -\frac{1}{2} (f(x) - \mu)^T \Sigma^{-1} (f(x) - \mu) \right\} \prod_{j=1}^d |f'_j(x_j)|
\]

- Semiparametric Gaussian copula
Examples
The Nonparanormal

- Define $h_j(x) = \Phi^{-1}(F_j(x))$ where $F_j(x) = \mathbb{P}(X_j \leq x)$.
- Let $\Lambda$ be the covariance matrix of $Z = h(X)$. Then

\[
\begin{array}{c|c|c}
X_j & X_k & X_{\text{rest}} \\
\end{array}
\]

if and only if $\Lambda_{jk}^{-1} = 0$.

- Hence we need to:

  1. Estimate $\hat{h}_j(x) = \Phi^{-1}(\hat{F}_j(x))$.
  2. Estimate covariance matrix of $Z = \hat{h}(X)$ using the glasso.
Fast Rates: Rank-Based Estimation

Assuming $X \sim NPN(f, \Sigma^0)$, we have

$$\Sigma^0_{jk} = 2 \sin \left( \frac{\pi}{6} \rho_{jk} \right)$$

where $\rho$ is Spearman’s rho:

$$\rho_{jk} := \text{Corr} \left( F_j(X_j), F_k(X_k) \right).$$

Empirical estimate:

$$\hat{\rho}_{jk} = \frac{\sum_{i=1}^{n} (r^i_j - \bar{r}_j)(r^i_k - \bar{r}_k)}{\sqrt{\sum_{i=1}^{n} (r^i_j - \bar{r}_j)^2 \cdot \sum_{i=1}^{n} (r^i_k - \bar{r}_k)^2}}.$$
Concentration of the Estimator

**Theorem.** With probability at least $1 - 1/n^2$,

$$\max_{jk} \left| \hat{S}^\rho_{jk} - \Sigma^0_{jk} \right| \leq \frac{3 \sqrt{2 \pi}}{2} \sqrt{\frac{\log d + \log n}{n}}.$$

Proof based on Hoeffding inequalities for U-statistics.

*We can thus estimate the covariance at the (minimax optimal) parametric rate.*
Examples: Graphs on the S&P 500

- Data from Yahoo! Finance (finance.yahoo.com).
- Daily closing prices for 452 stocks in the S&P 500 between 2003 and 2008 (before onset of the “financial crisis”).
- Log returns $X_{tj} = \log \left( \frac{S_{t,j}}{S_{t-1,j}} \right)$.
- Winsorized to trim outliers.
- In following graphs, each node is a stock, and color indicates GICS industry.

<table>
<thead>
<tr>
<th>Consumer Discretionary</th>
<th>Consumer Staples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Energy</td>
<td>Financials</td>
</tr>
<tr>
<td>Health Care</td>
<td>Industrials</td>
</tr>
<tr>
<td>Information Technology</td>
<td>Materials</td>
</tr>
<tr>
<td>Telecommunications Services</td>
<td>Utilities</td>
</tr>
</tbody>
</table>
S&P Data: Glasso vs. Nonparanormal

difference

common
Example Neighborhood

Target Corp. (Consumer Discretionary):

- Big Lots, Inc. (Consumer Discretionary)
- Costco Co. (Consumer Staples)
- Family Dollar Stores (Consumer Discretionary)
- Kohl’s Corp. (Consumer Discretionary)
- Lowe’s Cos. (Consumer Discretionary)
- Macy’s Inc. (Consumer Discretionary)
- Wal-Mart Stores (Consumer Staples)
S&P Data: Optimal Spanning Tree(?)
S&P Data: Forest vs. Nonparanormal

difference

common
Gene Data: Forest vs. Gaussian

tree (subset)  Gaussian
Summary

- Two nonparametric graphical models
  - Fully nonparametric bivariate marginals, restricted graphs
  - Fully nonparametric univariate marginals, unrestricted graphs
- For both: Can estimate graph at fast nonparametric rate $1/\sqrt{n}$
- Prefer forest densities:
  - Performs well for prediction
  - Interpretable graphs
- Current work: Conditional density estimation $p(y \mid x)$