10-704: Information Processing and Learning **Spring 2015**

Lecture 6: January 29

Lecturer: Aarti Singh *Scribe: Shashank Singh*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

6.1 Overview

In the previous lecture we reviewed several entropy estimators for discrete variables (plug-in and LP estimators) and continuous variables (different plugin and Von-Mises estimators). We extended this to estimate mutual information, and then applied this to learning tree graphical models via the Chow-Liu algorithm.

In this lecture, we first discuss a procedure for learning more general graphical models (the PC algorithm) using *conditional* mutual information estimators. We then switch to an unrelated topic: maximum entropy distributions and information geometry/information projection.

6.2 Application: Structure Learning in General Graphical Models

Last time we discussed the Chow-Liu algorithm, which uses mutual information estimation to learn the best tree graphical model representing data. Recall that, in each iteration, the Chow-Liu algorithm greedily adds an edge between the pair of unconnected variables exhibiting the greatest (estimated) mutual information. Because we only add edges between unconnected variables, (unconditional) mutual information suffices. However, to learn general graphical models (allowing multiple paths between nodes), we need measure conditional dependence, for which we can estimate conditional mutual information.

In general, there are exponentially many subsets of p variables on which we might have to condition to learn a graphical model. However, by choosing our conditional independence tests in a certain order, we can reduce the search time and the number of tests needed to be polynomial, if the underlying graphical model is sparse (not too many edges, or limitation on the degree of nodes).

By doing so, the PC algorithm, [1] which uses a conditional independence test as a subroutine, gives an efficient procedure for learning a general graphical model from joint observations of the variables. Intuitively, the PC algorithm begins with a complete graph and repeatedly picks an edge at random, removing it if it can find a set of conditioning variables that make the variables conditionally independent. For each edge, the conditioning set size is gradually varied from 0 to maximum degree to leverage sparsity of edges for faster runtime. See Figure 6.1 for pseudocode.

[1] Peter-Clark Algorithm, named for proposers Peter Spirtes and Clark Glymour [PC00]; see also [KB07] (available `http://jmlr.csail.mit.edu/papers/volume8/kalisch07a/kalisch07a.pdf`) for a more recent coverage and statistical analysis.

Inputs: A set $\mathcal{X} = \{X_1, \ldots, X_p\}$ of p variables
A data set of n joint observations $\{(x_{1,i}, \ldots, x_{p,i})\}_{i=1}^n$
A test T for conditional independence ($T(X_i, X_j, \mathcal{Y}) = TRUE$ iff $X_i \perp X_j | \mathcal{Y}$)
Outputs: An undirected graph $G = (\mathcal{X}, E)$ with $\{X_i, X_j\} \in E$ if and only if $T(X_i, X_j, \mathcal{Y}) = FALSE$ for every $\mathcal{Y} \subseteq \mathcal{X} \setminus \{X_i, X_j\}$

1) Initialize a complete graph $G = (\mathcal{X}, E)$
2) Initialize $\ell = -1$
3) **while** ℓ is less than the maximum degree of G
4) $\quad \ell = \ell + 1$
5) $\quad$ **for each** edge $\{X_i, X_j\} \in E$ with $|\mathcal{N}_G(X_i) \backslash \{X_j\}| \geq \ell$
6) $\quad\quad$ **for each** subset of neighbors $\mathcal{Y} \subseteq \mathcal{N}_G(X_i) \backslash \{X_j\}$ with $|\mathcal{Y}| = \ell$
7) $\quad\quad\quad$ **if** $T(X_i, X_j, \mathcal{Y})$ == TRUE
8) $\quad\quad\quad\quad$ delete edge $\{X_i, X_j\}$ from E
9) $\quad\quad\quad\quad$ **break**
10) $\quad\quad\quad$ **endif**
11) $\quad\quad$ **end for each**
12) $\quad$ **end for each**
13) **end while**

Figure 6.1: Pseudocode of the PC algorithm. $\mathcal{N}_G(X_i)$ denotes the set of neighbors of X_i in G.

Remark: Both the Chow-Liu algorithm for trees and PC algorithm for general graphical models can also be used to recover directed acyclic graphs (DAGs). The undirected graph returned by both algorithms corresponds to the skeleton (undirected graph obtained by ignoring directionality) of a DAG. Directionality can only be obtained up to an equivalence class of DAGs, i.e. there exist post-processing steps which can generate one DAG out of the equivalence class of DAGs all of which can imply the same set of conditional independence relations (for details, see [PC00,KB07]).

6.3 Maximum Entropy Density Estimation

Motivation: We often consider the uniform or Gaussian distributions to be good priors because they seem intuitively to be non-informative. This notion can be formalized in sense that, uniform and Gaussian are maximum entropy distributions: of all distributions satisfying certain constraints, they have the greatest entropy. The uniform distribution arises when we constrain the support of the distribution. The Gaussian appears when we constrain the first two moments (mean and variance). Both distributions belong to the exponential family, which we will show is the family of solutions to the following optimization problem:

$$\begin{aligned} & \max_{p \in \mathcal{P}(\mathcal{X})} H(p) \\ \text{subject to} \quad & \mathbb{E}_{X \sim p}[f_i(X)] = \alpha_i, \qquad i \in \{1, \ldots, n\} \\ \text{and} \quad & \mathbb{E}_{X \sim p}[g_j(X)] \leq \beta_j, \qquad j \in \{1, \ldots, m\}, \end{aligned} \tag{6.1}$$

where $\mathcal{P}(\mathcal{X})$ is the set of probability densities on a sample space $\mathcal{X}$, [2] each $f_i, g_j : \mathbb{R} \to \mathbb{R}$, and each $\alpha_i, \beta_j \in \mathbb{R}$. This problem is natural in the following sense: if we estimate properties of a distribution from data, a reasonable estimate of the distribution is the maximum entropy distribution with those properties. Theorem 6.1 parameterizes the solutions to this problem:

[2] The particular base measure μ on $\mathcal{X}$ is not important for the theory, though, in applications, this must of course be specified, as shown in the examples.

Theorem 6.1 *The density $p^* \in \mathcal{P}(\mathcal{X})$ solving the optimization problem 6.1 is in the exponential family*

$$\mathcal{E}(\mathcal{X}) := \left\{ p : \mathcal{X} \to \mathbb{R}^+ : p(x) = \exp\left(-1 - \lambda_0 + \sum_{i=1}^{n} \lambda_i f_i(x) + \sum_{j=1}^{m} \lambda_{n+j} g_j(x) \right), \qquad \forall x \in \mathbb{R} \right\},$$

for some $\vec{\lambda} \in \mathbb{R}^{1+n+m}$, with $\lambda_{n+1}, \ldots, \lambda_{n+m} \geq 0$, which ensure that p^* satisfied the constraints. Furthermore, any $p^* \in \mathcal{E}(\mathcal{X})$ is a maximum entropy distribution (optimizes 6.1), for some set of linear constraints.

Proof: *Step 1.* We first show, somewhat informally, that any maximum entropy distribution is in $\mathcal{E}(\mathcal{X})$. [3] If we rewrite the objective as minimizing $-H(p)$, then the Lagrangian $\mathcal{L} : \mathcal{P}(\mathcal{X}) \times [0, \infty)^{1+n+m} \to \mathbb{R}$ is

$$\mathcal{L}(p, \vec{\lambda}) = -H(p) + \lambda_0 \int_{\mathcal{X}} p(x)\, dx + \sum_{i=1}^{n} \lambda_i \int_{\mathcal{X}} p(x) f_i(x)\, dx + \sum_{j=1}^{n} \lambda_{n+j} \int_{\mathcal{X}} p(x) g_j(x)\, dx$$

The λ_0 term comes from the implicit constraint $\int_{\mathcal{X}} p(x)\, dx = 1$, since p is a probability density. Technically there are terms in here depending on α_i and β_j, but they fall off when we take the derivative with respect to p. Also just to be fully correct here, the lagrange parameters for the equality constraints should not be constrained to be non-negative.

Setting the derivative of the integrand with respect to $p(x)$ equal to 0 gives, for the optimum $p^* \in \mathcal{P}(\mathcal{X})$ and $\vec{\lambda} \in \mathbb{R}^{1+n+m}$,

$$0 = 1 + \log p^*(x) + \lambda_0 + \sum_{i=1}^{n} \lambda_i f_i(x) + \sum_{j=1}^{j} \lambda_{n+j} g_j(x).$$

Solving for $p^*(x)$ gives

$$p^*(x) = \exp\left(-1 - \lambda_0 - \sum_{i=1}^{n} \lambda_i f_i(x) - \sum_{j=1}^{j} \lambda_{n+j} g_j(x). \right),$$

which is the form of an exponential family distribution.

Step 2. We now show any $p^* \in \mathcal{E}(\mathcal{X})$ is a maximum entropy distribution under appropriate constraints. For

[3] Since $H(p)$ is convex in p and the constraints are linear in p, this calculation can be made into a formal proof of optimality using methods from the calculus of variations.

any $p \in \mathcal{P}(\mathcal{X})$, first applying Gibbs Inequality,

$$H(p) = -\int_{\mathcal{X}} p(x) \log \left(\frac{p(x)}{p^*(x)} p(x) \right) dx = -D(p||p^*) - \int_{\mathcal{X}} p(x) \log p^*(x)\, dx \leq -\int_{\mathcal{X}} p(x) \log p^*(x)\, dx \quad (6.2)$$

$$= \int_{\mathcal{X}} p(x) \left(1 + \lambda_0 + \sum_{i=1}^{n} \lambda_i f_i(x) + \sum_{j=1}^{m} \lambda_{n+j} g_j(x) \right) dx \quad (6.3)$$

$$\leq \int_{\mathcal{X}} p(x) \left(1 + \lambda_0 + \sum_{i=1}^{n} \lambda_i \alpha_i + \sum_{j=1}^{m} \lambda_{n+j} \beta_j \right) dx \quad (6.4)$$

$$= \int_{\mathcal{X}} p^*(x) \left(1 + \lambda_0 + \sum_{i=1}^{n} \lambda_i \alpha_i + \sum_{j=1}^{m} \lambda_{n+j} \beta_j \right) dx \quad (6.5)$$

$$= \int_{\mathcal{X}} p^*(x) \left(1 + \lambda_0 + \sum_{i=1}^{n} \lambda_i f_i(x) + \sum_{j=1}^{m} \lambda_{n+j} g_j(x) \right) dx \quad (6.6)$$

$$= \int_{\mathcal{X}} p^*(x) \log p^*(x)\, dx = H(p^*), \quad (6.7)$$

where (6.4) follows from the constraints and (6.6) follows from complementary slackness, since

$$\lambda_i^*(f_i(x) - \alpha_i) = \lambda_{n+j}(g_j(x) - \beta_j) = 0.$$

■

We now give a few examples of maximum entropy distributions under certain constraints.

Example 1 (Uniform): Suppose we constrain the domain $\mathbb{E}_{X\sim p}[1_A(X)] = 1$ for some $A \subseteq \mathcal{X}$ with $0 < \mu(A) < \infty$ for some base measure μ. Then, for some $\lambda_0, \lambda_1 \in \mathbb{R}$,

$$p^*(x) = \exp\left(-1 - \lambda_0 + \lambda_1 1_A(x)\right),$$

which is clearly uniform over A, and solving for λ_0, λ_1 from the constraints gives $p^*(x) = \frac{1_A(x)}{\mu(A)}$, $\forall x \in \mathcal{X}$.

Example 2 (Exponential): Suppose $\mathcal{X} = \mathbb{R}$ and we constrain the domain $\mathbb{E}_{X\sim p}[1_{[0,\infty)}(X)] = 1$ and the mean $\mathbb{E}_{X\sim p}[X] = \mu$. Then, for some $\lambda_0, \lambda_1, \lambda_2 \in \mathbb{R}$,

$$p^*(x) = \exp\left(-1 - \lambda_0 + \lambda_1 1_{[0,\infty)}(x) + \lambda_2 x\right),$$

which is an exponential distribution, and solving for $\lambda_0, \lambda_1, \lambda_2$ from the constraints gives $p^*(x) = \frac{1}{\mu} e^{-x/\mu} 1_{[0,\infty)}$.

Example 3 (Gaussian): Suppose $\mathcal{X} = \mathbb{R}$ and we constrain the mean $\mathbb{E}_{X\sim p}[X] = \mu$ and the variance $\mathbb{E}_{X\sim p}[(X-\mu)^2] = \sigma^2$. Then, for some $\lambda_0, \lambda_1, \lambda_2 \in \mathbb{R}$,

$$p^*(x) = \exp\left(-1 - \lambda_0 + \lambda_1 x + \lambda_2 (x-\mu)^2\right).$$

Since $\mathbb{E}_{X\sim p}[X] = \mu$, $\lambda_1 = 0$. p^* then has the form of a Gaussian, and it follows that $p^*(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2}\right)$ (solving for λ_0 directly here involves some difficult integration).

6.4 Information Geometry and Information Projection

Information geometry studies probability distributions geometrically, considering the family $\mathcal{P}(\mathcal{X})$ of probability densities on a sample space $\mathcal{X}$ as (isomorphic to) a simplex in the space $[0,\infty)^{(|\mathcal{X}|-1)}$ ($\mathcal{X}$ may be

infinite). For example, the family of Bernoulli distributions is isomorphic to the one-dimensional simplex. The next lecture will discuss the geometric view of information geometry; here, we discuss a generalization of the maximum entropy idea.

An important problem in information geometry is *information projection*, a generalization of the maximum entropy problem discussed above. The maximum entropy problem can be viewed as

$$p^* := \arg\max_{p \in Q} H(p) = \arg\min_{p \in Q} -H(p) = \arg\min_{p \in Q} \mathbb{E}_{X \sim p}[\log p(X)] = \arg\min_{p \in Q} D(p||u),$$

where $Q \subseteq \mathcal{P}(\mathcal{X})$ is a constraint set and u is the uniform distribution on $\mathcal{X}$; i.e., p^* is the constrained distribution closest to the uniform in KL-divergence. For a general distribution p_0, we can find the constrained distribution closest to p_0; i.e., $p^* := \arg\min_{p \in Q} D(p||p_0)$. The distribution p_0 can be thought of as our prior belief in what p^* should be, before observing any data i.e. without placing the constraints. Under mild assumptions, the solution is the *Gibbs distribution* (a natural generalization of the exponential family)

$$p^*(x) = p_0(x) \exp\left(-1 - \lambda_0 - \sum_{i=1}^{n} \lambda_i f_i(x)\right) = \frac{p_0(x) e^{\sum_{i=1}^n \lambda_i f_i(x)}}{Z_\lambda},$$

where Z_λ is a normalization constant. Next lecture, we will study this via information geometric tools.

References

[PC00] P. Spirtes, C. Glymour and R. Scheines, *Causation, Prediction, and Search.* The MIT Press, 2nd edition, 2000.

[KB07] M. Kalisch and P. Bühlmann, "Estimating High-Dimensional Directed Acyclic Graphs with the PC-Algorithm," *Journal of Machine Learning Research 8*, 2007, pp. 613–636.