6.1 Overview

In the previous lecture we reviewed several entropy estimators for discrete variables (plugin 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 reduce (by a polynomial factor) the number of tests needed so as to be feasible for a moderate number of variables.

By doing so, the PC algorithm, 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 the variables conditionally independent. See Figure 6.1 for pseudocode.

---

1 Peter-Clark Algorithm, named for proposers Peter Spirtes and Clark Glymour [PC00]; see also [KB07] (available \url{http://jmlr.csail.mit.edu/papers/volume8/kalisch07a/kalisch07a.pdf}) for a more recent coverage.
Inputs: A set $\mathcal{X} = \{X_1, \ldots, X_n\}$ 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 $(T(X_i, X_j, \mathcal{Y}) = \text{TRUE} \iff X_i \perp X_j | \mathcal{X}$, for any $\mathcal{Y} \subseteq \mathcal{X}$)

Outputs: An undirected graph $G = (\mathcal{X}, E)$ with $\{X_i, X_j\} \in E$ if and only if $T(X_i, X_j, \mathcal{Y}) = \text{FALSE}$
for every $\mathcal{Y} \subseteq \mathcal{X}$ with $X_i$ and $X_j$ connected in the graph induced by $(\mathcal{X}, E \setminus \{X_i, X_j\})$ on $\mathcal{X} \setminus \mathcal{Y}$.

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

Figure 6.1: Pseudocode of the PC algorithm. $N_G(X_i)$ denotes the set of neighbors of $X_i$ in $G$.

### 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 maximum entropy distributions: of all distributions satisfying certain constraints, they have the greatest entropy. The particular base measure $\mu$ 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

$$
\max_{p \in \mathcal{P}(\mathcal{X})} H(p)
$$

subject to

$$
\mathbb{E}_{X \sim p}[f_i(X)] = \alpha_i, \quad i \in \{1, \ldots, n\}
$$

and

$$
\mathbb{E}_{X \sim p}[g_j(X)] \leq \beta_j, \quad j \in \{1, \ldots, m\},
$$

where $\mathcal{P}(\mathcal{X})$ is the set of probability densities on a sample space $\mathcal{X}$, 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 be the maximum entropy distribution with those properties. Theorem 6.1 parameterizes the solutions to this problem:

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

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

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

---

2The particular base measure $\mu$ on $\mathcal{X}$ is not important for the theory, though, in applications, this must of course be specified, as shown in the examples.
**Proof:** Step 1. We first show, somewhat informally, that any maximum entropy distribution is in $\mathcal{E}(\mathcal{X})$.\footnote{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.} If we rewrite the objective as minimizing $-H(p)$, then the Lagrangian $L : \mathcal{P}(\mathcal{X}) \times (\mathbb{R}^{n+1} \times [0, \infty)^m) \to \mathbb{R}$ is

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

(the $\lambda_0$ term comes from the implicit constraint $\int_{\mathcal{X}} p(x) \, dx = 1$, since $p$ is a probability density).

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 $\lambda \in \mathbb{R}^{1+n+m}$,

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

Solving for $p^*(x)$ gives

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

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

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

where $\Box$ follows from the constraints and $\Box$ 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 $E_{X \sim p}[1_A(X)] = 1$ for some $A \subseteq \mathcal{X}$ with $0 < \mu(A) < \infty$ for some base measure $\mu$. Then, for some $\lambda_0, \lambda_1 \in \mathbb{R}$,

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

which is clearly uniform over $A$, and solving for $\lambda_0, \lambda_1$ 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 $E_{X \sim p}[1_{[0, \infty)}(X)] = 1$ and the mean $E_{X \sim p}[X] = \mu$. Then, for some $\lambda_0, \lambda_1, \lambda_2 \in \mathbb{R}$,

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

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 $E_{X \sim p}[X] = \mu$ and the variance $E_{X \sim p}[(X - \mu)^2] = \sigma$. Then, for some $\lambda_0, \lambda_1, \lambda_2 \in \mathbb{R}$,

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

Since $E_{X \sim p}[X] = \mu, \lambda_1 = 0$. $p^*$ then has the form of a Gaussian, and it follows that $p^*(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2}\right)$ (solving for $\lambda_0$ 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 an application.

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 (in certain cases) as

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

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., $p^* := \arg \min_{p \in Q} D(p||p_0)$. 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 $Z_\lambda$ is a normalization constant. Next lecture, we will study this via information geometric tools.

### References
