CS281A/Stat241A: Statistical Learning Theory

ML Estimation of Completely Observed Data - Undirected Graphical Models

Lecturer: Michael I. Jordan *Scribes: Kamalika Chaudhuri, Edward De Guzman*

1 Introduction

We saw in the last class that the problem of maximum likelihood estimation of parameters for completely observed data in directed graphical models conveniently decouples into separate maximum likelihood estimation problems for each node and its parents. Unfortunately, as we shall see in this class, this does not happen for general undirected graphical models. This is because the normalization factor Z in undirected models couples the parameters, which complicates the parameter estimation problem. For the general case of maximum likelihood parameter estimation in an undirected graph, we use an iterative algorithm called the IPF or Iterative Proportional Fitting. However, the problem does decouple for a special class of undirected graphical models called the decomposable models. Before we go into the details of this, let us first introduce some notation.

2 Notation

The notation that we use while dealing with undirected graphical models is very similar to the notation we used for directed graphical models. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be an undirected graph, and suppose our observed data is N I.I.D replicates of $\mathcal{G}$. That is,

$$\mathcal{D} = \left\{ X_{\mathcal{V},1}, X_{\mathcal{V},2}, \cdots, X_{\mathcal{V},N} \right\} \tag{1}$$

Here $X_{\mathcal{V},n}$ denotes the n^{th} replicate of all the variables in the graph; in general for given $C \subset \mathcal{V}$, $X_{C,n}$ denotes the n^{th} replicate of subset C. Then we have

$$p(X_{\mathcal{V}}|\theta) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \Psi_C(X_C) \tag{2}$$

where $\mathcal{C}$ denotes the set of all cliques in the graph $\mathcal{G}$. Here $\theta = \{\Psi_C(X_C) : C \in \mathcal{C}\}$ is the vector of all parameters, and Z is a normalisation factor defined by

$$Z = \sum_X \prod_{C \in \mathcal{C}} \Psi_C(X_C) \tag{3}$$

Like we did in the case of directed graphical models, we define the marginal counts as

$$m(X_{\mathcal{V}}) = \sum_{n=1}^{N} \delta(X_{\mathcal{V}}, X_{\mathcal{V},n}) \tag{4}$$

$$m(X_C) = \sum_{X_{\mathcal{V} \setminus C}} m(X_{\mathcal{V}}) \tag{5}$$

Note that

$$N = \sum_{X_\mathcal{V}} m(X_\mathcal{V}) \tag{6}$$

3 Maximum Likelihood Estimation

Given this notation, we can write the likelihood as

$$\begin{aligned} p(\mathcal{D}|\theta) &= \prod_{n=1}^{N} p(X_{\mathcal{V},n}|\theta) \\ &= \prod_{n=1}^{N} \prod_{X_\mathcal{V}} p(X_\mathcal{V}|\theta)^{\delta(X_\mathcal{V}, X_{\mathcal{V},n})} \end{aligned} \tag{7}$$

The log-likelihood can be computed as

$$\begin{aligned} \ell(\theta|\mathcal{D}) &= \sum_n \sum_{X_\mathcal{V}} \delta(X_\mathcal{V}, X_{\mathcal{V},n}) \log p(X_\mathcal{V}|\theta) \\ &= \sum_{X_\mathcal{V}} m(X_\mathcal{V}) \log p(X_\mathcal{V}|\theta) \\ &= \sum_{X_\mathcal{V}} m(X_\mathcal{V}) \log \big(\frac{1}{Z} \prod_{C \in \mathcal{C}} \Psi_C(X_C)\big) \\ &= \sum_{X_\mathcal{V}} m(X_\mathcal{V}) \sum_{C \in \mathcal{C}} \log \Psi_C(X_C) - \sum_{X_\mathcal{V}} m(X_\mathcal{V}) \log Z \\ &= \sum_{C \in \mathcal{C}} \sum_{X_C} \sum_{X_{\mathcal{V} \setminus C}} m(X_\mathcal{V}) \log \Psi_C(X_C) - N \log Z \\ &= \sum_{C \in \mathcal{C}} \sum_{X_C} \log \Psi_C(X_C) \sum_{X_{\mathcal{V} \setminus C}} m(X_\mathcal{V}) - N \log Z \\ &= \sum_{C \in \mathcal{C}} \sum_{X_C} m(X_C) \log \Psi_C(X_C) - N \log Z \end{aligned} \tag{8}$$

To compute the maximum likelihood estimate of the parameter vector θ, we need to compute the derivative of the log-likelihood with respect to the $\Psi_C(X_C)$'s. The derivative of the first term in the above expression

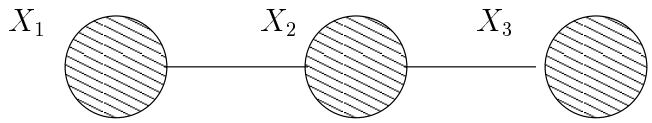

Figure 1: A Simple Decomposable Graphical Model

gives us $\frac{m(X_C)}{\Psi_C(X_C)}$. Taking the derivative of the second term, we get

$$\begin{aligned}
\frac{\partial \log Z}{\partial \Psi_C(X_C)} &= \frac{1}{Z}\frac{\partial}{\partial \Psi_C(X_C)}[\sum_{\tilde{X}} \prod_{D\in\mathcal{C}} \Psi_D(\tilde{X}_D)] \\
&= \frac{1}{Z}\sum_{\tilde{X}} \delta(\tilde{X}_C, X_C)\frac{\partial}{\partial \Psi_C(X_C)}[\prod_{D\in\mathcal{C}} \Psi_D(\tilde{X}_D)] \\
&= \frac{1}{Z}\sum_{\tilde{X}} \delta(\tilde{X}_C, X_C)\prod_{D\in\mathcal{C}, D\neq C} \Psi_D(\tilde{X}_D) \\
&= \frac{1}{Z}\sum_{\tilde{X}} \delta(\tilde{X}_C, X_C)\frac{1}{\Psi_C(X_C)}\prod_{D\in\mathcal{C}} \Psi_D(\tilde{X}_D) \\
&= \sum_{\tilde{X}} \delta(\tilde{X}_C, X_C)\frac{1}{\Psi_C(X_C)}p(\tilde{X}|\theta) \\
&= \frac{p(X_C|\theta)}{\Psi_C(X_C)} \qquad (9)
\end{aligned}$$

From this calculation, we get

$$\frac{\partial \ell}{\partial \Psi_C(X_C)} = \frac{m(X_C)}{\Psi_C(X_C)} - N\frac{p(X_C)}{\Psi_C(X_C)} \qquad (10)$$

This implies that

$$p(X_C) = \frac{1}{N}m(X_C) \qquad (11)$$

Our goal is to estimate $\Psi_C(X_C)$'s from these relations. This is done for general graphs using an iterative algorithm called IPF or Iterative Proportional Fitting.

4 Decomposable Graphs

However, for certain types of undirected graphical models, we can solve for $\Psi_C(X_C)$'s directly; these types of models are called *Decomposable*. Let us look at an example. Consider the undirected graph in Figure(1), which is a Markov Chain on three variables $\{X_1, X_2, X_3\}$. The joint probability model can be written as:

$$p(X_1, X_2, X_3) = \frac{1}{Z}\Psi_{12}(X_1, X_2)\Psi_{23}(X_2, X_3) \qquad (12)$$

By our previous derivation, the necessary conditions that marginals must follow are

$$\begin{aligned}
\tilde{p}(X_1, X_2) &= \frac{1}{N}m(X_1, X_2) \\
\tilde{p}(X_2, X_3) &= \frac{1}{N}m(X_2, X_3)
\end{aligned}$$

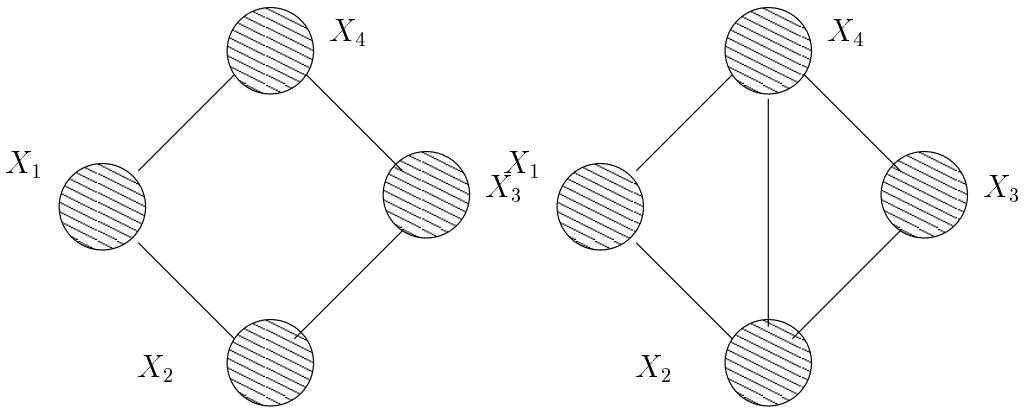

Figure 2: A Non- Decomposable Graph and a Decomposable Graph

Let us make the following guess on the potential functions.

$$\tilde{\Psi}_{12,ML} = \tilde{p}(X_1, X_2) \tag{13}$$

$$\tilde{\Psi}_{23,ML} = \frac{\tilde{p}(X_2, X_3)}{\tilde{p}(X_2)} = \tilde{p}(X_3|X_2) \tag{14}$$

It can be easily verified that this is a good guess; the model marginals sum up correctly, and also $Z = 1$. We cannot however, find such good guesses for all undirected graphical models. The class of models for which we can solve directly is called *Decomposable*. We shall show later, that if a graph is decomposable, it is triangulated and vice versa. Thus decomposability and being triangulated are equivalent properties for undirected graphs.

Figure(2) shows examples of a non-decomposable and a decomposable graph.

5 Iterative Proportional fitting

For parameter estimation in non-decomposable graphical models, we use an iterative algorithm called IPF or Iterated Proportional Fitting. Recall, that our minimisation of the log-likelihood led to the relation

$$\frac{p(X_c)}{\Psi_c(X_c)} = \frac{1}{N}\frac{m(X_c)}{\Psi_c(X_c)} \tag{15}$$

This gives rise to the iterative equation

$$\Psi_c^{(t+1)}(X_c) = \Psi_c^{(t)}(X_c)\frac{\tilde{p}(X_c)}{p^{(t)}(X_c)} \tag{16}$$

where

$$\tilde{p}(X_c) = \frac{1}{N}m(X_c) \tag{17}$$

Here $p^{(t)}(X_c)$ denotes the marginal probability of configuration X_c computed under the current model; this is obtained by running an inference algorithm, such as the sum-product algorithm on the graph, using the current values of the parameters. Thus we have a multiplicative update equation. There are several things to be noted about this algorithm.

1. After updating $\Psi_c(X_c)$, we have
$$p^{(t+1)}(X_c) = \tilde{p}(X_c) \tag{18}$$
That is, the marginals sum up correctly after each iteration.

2. The normalisation constant Z remains the same across iterations.
$$Z^{(t+1)} = Z^{(t)} \tag{19}$$

3. IPF does coordinate ascent, where each coordinate is a potential function $\Psi_c(X_c)$.