CS281A/Stat241A: Statistical Learning Theory

GLIMS (contd..) and Completely Observed Graphical Models (10/12/04)

Lecturer: Michael I. Jordan *Scribes: David Chu and Dilip Antony Joseph*

1 GLIMS review

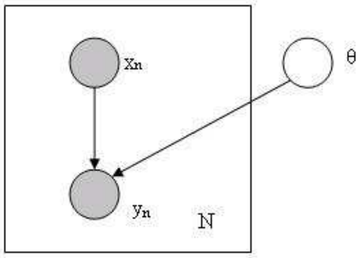

Figure 1: Graphical model

We review graphical models of the form shown in Figure 1. The relationship among variables X and Y and model parameters is:

$$\begin{array}{l} \theta \searrow \\ \qquad \zeta \xrightarrow{f} \mu \xrightarrow{\psi} \eta \xrightarrow{exponential family} y \\ x \nearrow \end{array} \tag{1}$$

The exponential family takes the form:

$$p(y|x,\theta) = h(y)\exp\{\eta^T T(y) - A(\eta)\} \tag{2}$$

Recall $\mu = E[Y|X]$. ψ is determined by the exponential family. But we have a choice in choosing f, the *response function*, where $\mu = f(\theta^T x)$. If we use the canonical response function $f = \psi^{-1}$, then $\eta = \theta^T x$ and

$$p(y|x,\theta) = h(y)\exp\{\theta^T x T(y) - A(\theta^T x)\} \tag{3}$$

Example: The Bernoulli distribution, $y \in \{0,1\}$ and

$$\begin{aligned} B(\mu) &= \mu^y (1-\mu)^{1-y} \\ &= \exp\left\{ y \log\left(\frac{\mu}{1-\mu}\right) + \log(1-\mu)\right\} \end{aligned} \tag{4}$$

This yields $\eta = \psi(\mu) = \log \frac{\mu}{1-\mu}$ and $\mu = f(\eta) = \frac{1}{1+e^{-\eta}} = \frac{1}{1+e^{-\theta^T X}}$ which is a logistic regression.

Example: The Poisson distribution

$$p(y) = \frac{\mu^y e^{-\mu}}{y!} \tag{5}$$

$$= \frac{1}{y!} \exp\{y \log(\mu) - \mu\} \tag{6}$$

This yields $\eta = \log \mu$ and $\mu = e^{\eta}$. If we use the canonical response function, then $\mu = e^{\eta} = e^{\theta^T X}$.

To solve, first substitute $e^{\theta^T X}$ for μ into $p(y)$. Second, take the product over the set of data:

$$\prod_{n=1}^{N} p(y_n|x_n, \theta) \tag{7}$$

Writing out the maximum likelihood estimation of θ:

$$l(\theta) = \sum_{n=1}^{N} p(y_i|x_i, \theta) \tag{8}$$

$$= \sum_{i=1}^{N} \theta^T x_n T(y_n) - A(\theta^T x_n) \tag{9}$$

$$\bigtriangledown_\theta l(\theta) = \sum_{n=1}^{N} (x_n T(y_n) - \frac{dA}{d\eta_n} \frac{d\eta_n}{d\theta}) \tag{10}$$

$$= \sum_{n=1}^{N} x_n T(y_n) - \mu_n x_n \tag{11}$$

$$= \sum_{n=1}^{N} (T(y_n) - \mu_n) x_n \tag{12}$$

Example: Linear Regression, $T(y_n) = y_n$ and $\mu_n = \theta^T x_n$:

$$\bigtriangledown_\theta l(\theta) = (y_n - \theta^T x_n) x_n \tag{13}$$

Example: Logistic Regression, $T(y_n) = y_n$ and $\mu_n = \frac{1}{1+e^{-\theta^T x_n}}$. Therefore,

$$\bigtriangledown_\theta l(\theta) = (y_n - \frac{1}{1 + e^{-\theta^T x_n}}) x_n \tag{14}$$

Note: The canonical response function is used mainly for mathematical convenience. The problems we find in the real world do not usually cooperate with the canonical response function. However, this is a good way to start analyzing the problem.

Note: We were being frequentist all this while. If we are Bayesian, we can put a prior on θ. But $p(\theta|x_n, y_n)$ usually does not have a closed form except for the case of a linear model.

Summary: We can now estimate parameters for GLIMs of the form shown in Figure 2. Plug the appropriate equations into the general on-line algorthm with step size ρ:

$$\theta^{(t+1)} \leftarrow \theta^{(t)} + \rho(y_n - \mu_n) x_n.$$

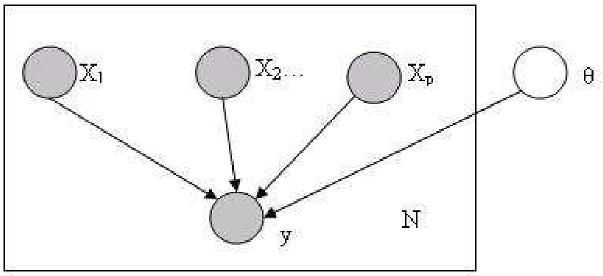

Figure 2: A more complex graphical model.

2 Completely Observed Graphical Models

Earlier we looked at simple graphical models involving a node and its parents. Now we analyze more complex graphical models like the one in Figure 3. We first look at the case of directed graphical models.

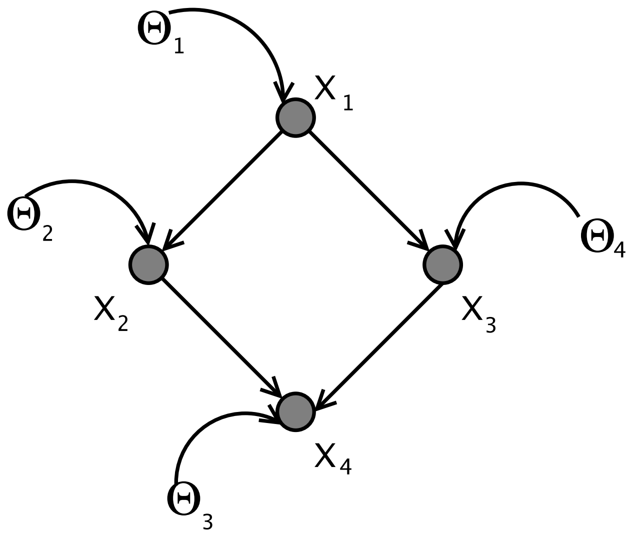

Figure 3: A more complex graphical model.

The joint probability distribution of the graphical model in Figure 3 can be factorized as

$$p(x_1, x_2, x_3, x_4|\theta) = p(x_1|\theta_1)p(x_2|x_1, \theta_2)p(x_3|x_1, \theta_3)p(x_4|x_2, x_3, \theta_4) \tag{15}$$

All nodes X_1, X_2, X_3, X_4 are observed. Our problem is to estimate the parameter vector $\theta = (\theta_1, \theta_2, \theta_3, \theta_4)$.

First, consider a single data point (each data point is a 4-tuple). We find the maximum likelihood estimate for θ as follows:

$$\begin{aligned} l(\theta) &= \log(p(x_1, x_2, x_3, x_4|\theta)) \qquad (16) \\ &= \log(p(x_1|\theta_1)p(x_2|x_1, \theta_2)p(x_3|x_1, \theta_3)p(x_4|x_2, x_3, \theta_4)) \end{aligned}$$

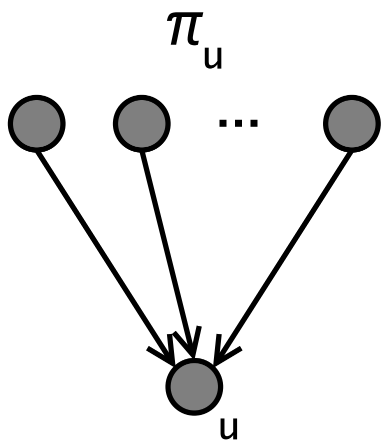

Figure 4: Family of nodes

$$= \log(p(x_1|\theta_1)) + \log(p(x_2|x_1, \theta_2)) + \log(p(x_3|x_1, \theta_3)) + \log(p(x_4|x_2, x_3, \theta_4))$$

We see that each term in Equation 16 contains only one of θ_1, θ_2, θ_3 and θ_4. Thus to find the maximum likelihood (ML) estimate of θ_1, it is necessary to only consider the term $\log(p(x_1|\theta_1))$, i.e.

$$\begin{aligned} \hat{\theta}_{1,ML} &= \arg\max_{\theta} l(\theta) \\ &= \arg\max_{\theta} \log(p(x_1|\theta_1)) \end{aligned}$$

In general, ML estimation decouples into separate estimation problems for each *family.* in a fully observed, directed graphical model. A family (see Figure 4) consists of a node and its parents. This means that ML estimation can be done locally at each node in a directed graphical model. Note that this does not hold in general for undirected graphical models.

Each of the local ML estimation problems (for example, $p(x_4|x_2, x_3, \theta)$) may be solved as a GLIM. Thus, we now have a tool to solve real problems. Once we have estimated all the parameters, we can use the inference algorithms we studied earlier in the course to find the marginals and other interesting properties of the model.

2.1 Notation

We consider the case where each of the nodes X_i take only a finite number of values. Let $\mathcal{G}(\mathcal{V}, \mathcal{E})$ represent a single replica of the graphical model (as in Figure 3). $X_\mathcal{V}$ denotes the set of all nodes in $\mathcal{G}(\mathcal{V}, \mathcal{E})$. The joint probability distribution can be written as follows:

$$p(x_\mathcal{V}|\theta) = \prod_{u \in \mathcal{V}} p(x_u|x_{\pi_u}, \theta_u) \tag{17}$$

$\mathcal{G}^{(N)}$ represents the graph consisting of the N replicas of $\mathcal{G}(\mathcal{V}, \mathcal{E})$, where N is the number of data tuples. Figure 5 shows this equivalence. Note that we have assumed i.i.d replicates.

The observed data is represented as follows:

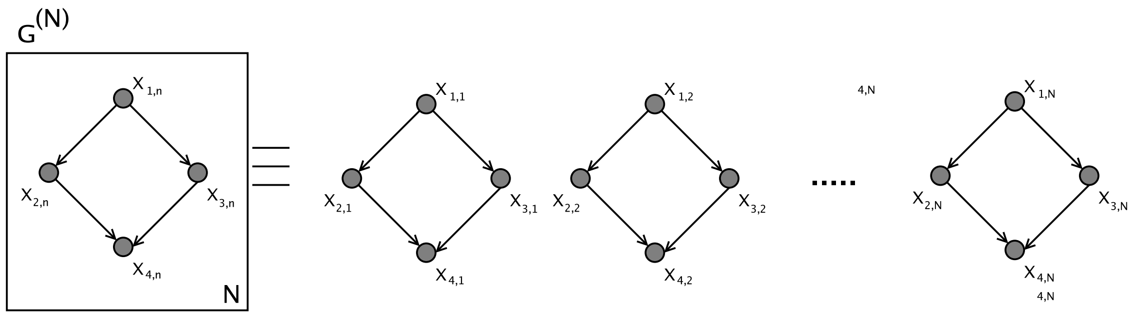

Figure 5: Replicating graph $\mathcal{G}^{(N)}$

$$\mathcal{D} = (x_{\mathcal{V},1}, x_{\mathcal{V},2}, \ldots, x_{\mathcal{V},i}, \ldots, x_{\mathcal{V},N}) \tag{18}$$

where $x_{\mathcal{V},i}$ corresponds to the values of all nodes in replicate i.

Now we choose the notation for *counts* as follows:

$$\begin{aligned} m(x_{\mathcal{V}}) &= \textit{number of times } x_{\mathcal{V}} \textit{ is observed in } \mathcal{D} \\ &= \sum_{n=1}^{N} \delta(x_{\mathcal{V}}, x_{\mathcal{V},n}) \end{aligned} \tag{19}$$

$$\begin{aligned} m(x_C)_{C \subset \mathcal{V}} &= \textit{count of some subset of states} \\ &= \sum_{X_{\mathcal{V} \setminus C}} m(x_{\mathcal{V}}) \end{aligned}$$

The family shown in Figure 4 is denoted as $\phi_u = u \cup \pi_u$. The probability distribution of a node X_u is given by

$$\theta_u(x_{\phi_u}) = p(x_u | x_{\pi_u}, \theta_u) \tag{20}$$

We know that for a fixed set of parents x_{π_u},

$$\sum_{x_u} \theta_u(x_u, x_{\pi_u}) = 1 \tag{21}$$

The probability distribution of all the nodes of a single replica of $\mathcal{G}(\mathcal{V}, \mathcal{E})$ taken together is given by:

$$p(x_{\mathcal{V}} | \theta) = \prod_u \theta_u(x_{\phi_u}) \tag{22}$$

The above equation can be rewritten as follows.

$$p(x_{\mathcal{V},n} | \theta) = \prod_{x_{\mathcal{V}}} p(x_{\mathcal{V}} | \theta)^{\delta(x_{\mathcal{V}}, x_{\mathcal{V},n})} \tag{23}$$

This is just a big switch statement. Here n refers to a particular replicate of $\mathcal{G}^{(N)}$.

2.2 ML Estimation

Let us now do the ML estimation for $\mathcal{G}^{(N)}$ as follows.

$$\begin{aligned} l(\theta) &= \log \prod_{n=1}^{N} \prod_{u \in \mathcal{V}} p(x_{u,n} | x_{\pi_u,n}, \theta_u) \\ &= \sum_n \sum_u \log p(x_{u,n} | x_{\pi_u,n}, \theta_u) \end{aligned}$$

From the above equation, it appears that we need to store all the observed data to calculate the ML estimate. Storing huge amounts of data may not be feasible.

Let us now do the ML estimation in a different way, making use of the notation we defined in the previous section.

$$\begin{aligned} l(\theta) &= \log(\mathcal{D}|\theta) \\ &= \log\left(\prod_n p(x_{\mathcal{V},n}|\theta)\right) \\ &= \sum_n \sum_{x_\mathcal{V}} \delta(x_\mathcal{V}, x_{\mathcal{V},n}) \log(p(x_\mathcal{V}|\theta)) \\ &= \sum_{x_\mathcal{V}} \log(p(x_\mathcal{V}|\theta) m(x_\mathcal{V}) \end{aligned}$$

The above equations show that $l(\theta)$ is just a function of the counts, i.e. the sufficient statistic is the counts of the different $X_\mathcal{V}$.

Continuing, we get

$$\begin{aligned} l(\theta) &= \sum_{x_\mathcal{V}} m(x_\mathcal{V}) \log(\prod_u \theta_u(x_{\phi_u})) \\ &= \sum_{x_\mathcal{V}} m(x_\mathcal{V}) \sum_u \log \theta_u(x_{\phi_u}) \\ &= \sum_u \sum_{x_\mathcal{V}} m(x_\mathcal{V}) \log \theta_u(x_{\phi_u}) \\ &= \sum_u \sum_{x_{\phi_u}} (\sum_{x_{\mathcal{V} \setminus \phi_u}} m(x_\mathcal{V})) \log \theta_u(x_{\phi_u}) \end{aligned}$$

Since $x_{\mathcal{V}\setminus\phi_u}$ and $\theta_u(x_{\phi_u})$ are independent, we get

$$l(\theta) = \sum_u \sum_{x_{\phi_u}} m(x_{\phi_u}) \log \theta_u(x_{\phi_u})$$

Note that each term in the above equation has only a single θ_u. Thus the estimation of each θ_u can be decoupled and done independent of other θ_u. We can also see that we need only $m(x_{\phi_u})$ in the calculation of $l(\theta)$ for each family ϕ_u.

By using Lagrangian multipliers and doing the calculations, we get

$$\hat{\theta}_{u,ML} = \frac{m(x_u, x_{\pi_u})}{m(x_{\pi_u})} \tag{24}$$