STAT 535 Handout 10

Estimation of graphical models parameters

©Marina Meilă
mmp@stat.washington.edu

1 The estimation problem

Given: $V = \{X_1, X_2, \dots X_n\}$ a set of n discrete variables and $\mathcal{D} = \{(x_1^1 x_2^1 \dots x_n^1), (x_1^2 x_2^2 \dots x_n^2) \dots (x_1^N x_2^N \dots x_n^N)\}$ a **sample** (or **dataset**) of size N.

Assumption: Let $\mathcal{M}$ be a set of graphical models (or belief networks) over V. We assume $\mathcal{D}$ is an i.i.d. sample from some $P \in$.

Wanted: P

The **Maximum Likelihood (ML)** criterion:

$$\boxed{\hat{P} \;=\; \operatorname*{argmax}_{P \in \mathcal{M}} P(\mathcal{D})}$$

or, because the samples $x^i = (x_1^i x_2^i \dots x_n^i)$ are independent

$$\boxed{\hat{P} \;=\; \operatorname*{argmax}_{P \in \mathcal{M}} l(P)}$$

$$\boxed{l(P) \;=\; \textstyle\sum_{i=1}^{N} \log P(x^i)}$$

$l(P)$ is called the **log-likelihood** of the distribution P.

Assume now that the graphical model family $\mathcal{M}$ we are considering is of

Bayes nets. Later we will discuss parameter estimation for Markov Fields.

Each Bayes net in $\mathcal{M}$ is defined by **structure**, i.e. the graph, and **parametrization**, i.e. the set of tables $\{P_{X|Y} \mid X\}$ with $Y = pa(X) \subseteq V$.

Estimating a graphical model means estimating both the structure and the parameters. We shall start by assuming that the structure is fixed and by focusing on the estimation of the parameters.

2 ML Estimation for Bayesian networks

2.1 Estimating the parameters of a multinomial distribution

Assume $V = \{X\}$. Then

$$\begin{aligned} l(P_X) &= \sum_{i=1}^{N} \log P_X(x^i) && (1) \\ &= \sum_{k \in \Omega(X)} N_X(k) \log P_X(k) && (2) \\ &= N \sum_{k \in \Omega(X)} \frac{N_X(k)}{N} \log P_X(k) && (3) \end{aligned}$$

Here we have denoted by

$$N_X(k) = \sum_{i=1}^{N} \delta_{kx^i}$$

the number of times $X = k$ in the sample $\mathcal{D}$.

The ML estimate of P_X is

$$\theta_X^{ML}(k) = \frac{N_X(k)}{N} \quad (4)$$

2.2 Estimating the parameters of a bivariate multinomial distribution

Assume $V = \{X, Y\}$ and the graph is $X \longrightarrow Y$. Define $N_{XY}(k,j) = \sum_{i=1}^{N} \delta_{x^i k}\delta_{y^i j}$ = the number of samples for which $X = k$ and $Y = j$ in $\mathcal{D}$. Then,

$$\begin{aligned} l(P_{XY}) &= \sum_{i=1}^{N} \log P_{XY}(x^i y^i) && (5)\\ &= \sum_{i=1}^{N} \log P_X(x^i) + \sum_{i=1}^{N} \log P_{Y|X}(y^i|x^i) && (6)\\ &= \sum_{k\in\Omega(X)} N_X(k) \log P_X(x^i) + \sum_{k\in\Omega(X)} \sum_{j\in\Omega(Y)} N_{XY}(k,j) \log P_{Y|X}(y^i|x^i) && (7)\\ &= N \sum_{k\in\Omega(X)} \frac{N_X}{N}(k) \log P_X(x^i) + N \sum_{k\in\Omega(X)} \frac{N_X(k)}{N} \sum_{j\in\Omega(Y)} \frac{N_{XY}(k,j)}{N_X(k)} \log P_{Y|X}(y^i|x^i) \end{aligned}$$

$l(P)$ is maximized for P^{ML} given by

$$\begin{aligned} \theta^{ML}_{Y|X}(j|k) &= \frac{N_{XY}(k,j)}{N_X(k)} && (8)\\ \theta^{ML}_{X}(k) &= \frac{N_X(k)}{N} && (9) \end{aligned}$$

2.3 Estimating the parameters of a graphical model

For a fixed DAG structure G, the distribution P can be expressed as:

$$P_V = \prod_{X\in V: Y=\mathrm{pa}(X)} P_{X|Y} \qquad (10)$$

The log-likelihood is expressed as:

$$\begin{aligned} l(P_V) &= \sum_{i=1}^{N} \log \prod_{X\in V: Y=\mathrm{pa}(X)} P_{X|Y}(x^i|y^i) && (11)\\ &= \sum_{X\in V: Y=\mathrm{pa}(X)} \sum_{i=1}^{N} \log P_{X|Y}(x^i|y^i) && (12) \end{aligned}$$

$$= \sum_{X \in V: Y = \mathrm{pa}(X)} \sum_{k \in \Omega(X)} \sum_{j \in \Omega(Y)} N_{XY}(k,j) \log P_{X|Y}(k|j) \quad (13)$$

$$= N \sum_{X \in V: Y = \mathrm{pa}(X)} \sum_{j \in \Omega(Y)} \left[\frac{N_Y(j)}{N} \sum_{k \in \Omega(X)} \frac{N_{XY}(k,j)}{N_Y(j)} \log P_{X|Y}(k|j) \right] \quad (14)$$

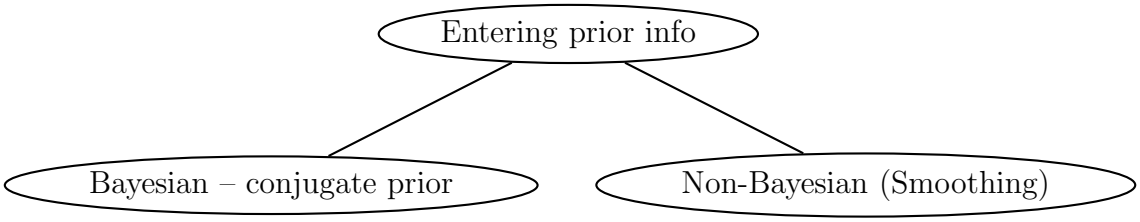
By a similar argument as above, we have

$$\theta^{ML}_{X|Y}(k|j) \; = \; \frac{N_{XY}(k,j)}{N_Y(j)} \quad (15)$$

for all $X \in V$.

3 Bayesian Estimation of Bayes network parameters

The Maximum Likelihood estimate is not always the best choice for the esimate of a set of parameters. Sometimes we have prior knowledge that we want to take into account, and sometimes the data is too scarce (especially when the number of parent configurations is large) and ML overfits.



The ML paradigm assumes that the model is estimated using the data only and excluding other sources of knowledge about parameters or model structure. But, if prior knowledge exists and if it can be represented as a probability distribution over the space of models, then one can use the Bayesian formulation of learning to combine the two sources of information.

In the Bayesian framework, the main object of interest is the posterior distribution over models given the observed data $Pr[P|\mathcal{D}]$. By Bayes' formula,

the posterior is proportional to

$$Pr[P|\mathcal{D}] \;\propto\; Pr[P,\mathcal{D}] \;=\; Pr[P]\prod_{x} \in VP(x) \tag{16}$$

In the above $Pr[P]$ represents the prior distribution over the class of models; the second factor is $P(\mathcal{D})$, the likelihood of the data given the model P.

The probability of an observation x is obtained by *model averaging*

$$Pr[x] \;=\; \int P(x)Pr[P|\mathcal{D}]\,dP \tag{17}$$

It is worth noting that, except for a few special cases that we will discuss further neither $Pr[x]$ nor the posterior $Pr[P|\mathcal{D}]$ are representable in closed form. A common approach then is to approximate the posterior distribution around its mode(s) for example by the Laplace approximation. Another approach is to replace the integration in equation (17) by a finite sum over a set $\mathcal{M}$ of models with high posterior probability.

$$Pr[x] \;=\; \sum_{P\in\hat{\mathcal{M}}} P(x)Pr[P|\mathcal{D}] \tag{18}$$

This approximation is equivalent to setting $Pr[P|\mathcal{D}]$ to 0 for all the models not in $\hat{\mathcal{M}}$. Consequently, the normalization constant in the above formula is computed over $\hat{\mathcal{M}}$ only.

Finally, if we are to choose one model only to summarize the posterior distribution $Pr[P|\mathcal{D}]$ then a natural choice is the mean of the distribution. As it will be shown the mean can sometimes be expressed as the MAP estimate under a certain parameterization.

3.1 The Dirichlet prior

We will introduce now an important subclass of priors for parameters called *Dirichlet* priors. The Dirichlet prior is the conjugate prior of the *multinomial* distribution.

Let z be a discrete random variable taking r values and let $\theta_j = P_z(j),\ j = 1, \ldots r$. Then, the probability distribution of an i.i.d. sample of size N from P_z is given by

$$P(z^{1,\ldots N}) \ = \ \prod_{j=1}^{r} \theta_j^{N_j} \tag{19}$$

where $N_j,\ j = 1, \ldots r$ represent the number of times the value j is observed in and are called the *sufficient statistics* of the data. The sample itself is said to obey a *multinomial* distribution.

The Dirichlet distribution is defined over the domain of $\theta_{1,\ldots r}$ and depends on r real parameters $N'_{1,\ldots r} > 0$ by

$$D(\theta_{1,\ldots r}; N'_{1,\ldots r}) \ = \ \frac{\Gamma(\sum_j N'_j)}{\prod_j \Gamma(N'_j)} \prod_j \theta_j^{N'_j - 1} \tag{20}$$

In the above $\Gamma()$ represents the Gamma function defined by

$$\Gamma(p) \ = \ \int_0^\infty t^{p-1} e^{-t} dt. \tag{21}$$

For any nonnegative integer n

$$\Gamma(n+1) \ = \ n! \tag{22}$$

The importance of the Dirichlet distribution in connection with a multinomially distributed variable resides in the following fact: If the parameters θ of the multinomial distribution have as prior a Dirichlet distribution with parameters $N'_j,\ j = 1, \ldots r$, then, after observing a sample with sufficient statistics $N_j,\ j = 1, \ldots r$, the posterior distribution of θ is a Dirichlet distribution with parameters $N'_j + N_j,\ j = 1, \ldots r$. This justifies denoting the distribution's parameters by N'. One popular alternative parameterization for the Dirichlet distribution is given by:

$$N' \ = \ \sum_{j=1}^{r} N'_j \tag{23}$$

$$P'_j \ = \ \frac{N'_j}{N'} \quad j = 1, \ldots r \tag{24}$$

$$\tag{25}$$

Note that in this parameterization the means of the parameters θ_j are equal to P'_j. We say that the Dirichlet distribution is a conjugate prior for the class of multinomial distributions. The property of having conjugate priors is characteristic for the exponential family of distributions.

3.1.1 The Dirichlet prior in natural coordinates

The multinomial distribution was represented before as defined by the parameters θ_j, $j = 1, \ldots r$. But there are infinitely many ways to parametrize the same distribution. For each set of parameters $y_{1,\ldots r_y}$ the corresponding representation of the Dirichlet prior results from the well known change of variable formula

$$D(y_{1,\ldots r_y};\, N'_{1,\ldots r-1}) \;=\; D(\theta_{1,\ldots r}(y_{1,\ldots r_y}); N'_{1,\ldots r}).|\frac{\partial \theta}{\partial y}| \tag{26}$$

with $|\frac{\partial y}{\partial \theta}|$ representing the absolute value of the determinant of the Jacobian of $\theta(y)$. Note that because of the presence of this factor, the maximum of $D(.; N'_{1,\ldots r})$ has both a different value and a different position in each parametrization. This dependence of the parametrization is one fundamental drawback of MAP estimation which justifies the Bayesian and approximate Bayesian approaches mentioned above.

By contrast, the mean of $f(y)D(y; N')$ of any measurable function f over any measurable set is independent of the parametrization. In particular, the mean of the Dirichlet distribution is independent of the parametrization and equal to

$$E[\theta_j] \;=\; \frac{N'_j}{\sum_{j'=1}^{r} N'_{j'}} \quad j = 1, \ldots r \tag{27}$$

Of special interest is the so called *natural* parametrization of the multinomial, defined by $r-1$ unconstrained parameters ϕ:

$$\phi_i \;=\; \log \frac{\theta_i}{\theta_r} \quad i = 1, \ldots r-1. \tag{28}$$

The parameters ϕ take values in $(-\infty, \infty)$ when $\theta_{1,\ldots r} > 0$. The reverse transformation from ϕ coordinates to θ coordinates is defined by:

$$\theta_r = \frac{1}{1 + \sum_{i=1}^{r-1} e^{\phi_i}} \tag{29}$$

$$\theta_j = \frac{e^{\phi_j}}{1 + \sum_{i=1}^{r-1} e^{\phi_i}}, \quad j = 1, \ldots r-1 \tag{30}$$

$$\tag{31}$$

The Jacobian of this transformation[1] is (expressed in θ variables)

$$\left| \frac{\partial \phi}{\partial \theta} \right| = \theta_1 \theta_2 \ldots \theta_r \tag{32}$$

In the natural parametrization, the Dirichlet distribution is expressed as

$$D(\phi_{1,\ldots r-1};\, N'_{1,\ldots r}) = \frac{\Gamma(\sum_{j=1}^{r} N'_j)}{\prod_{j=1}^{r} \Gamma_(N'_j)} \prod_{i=1}^{r-1} \left(\frac{e^{\phi_i}}{1 + \sum_{j=1}^{r-1} e^{\phi_j}} \right)^{N'_i} \frac{1}{(1 + \sum_{j=1}^{r-1} e^{\phi_j})^{N'_r}} \tag{33}$$

A remarkable property of the natural parametrization is that its mode coincides in position with the mean. To see this, it suffices to equate to 0 the partial derivatives of the Dirichlet distribution w.r.t the ϕ parameters. After some calculations, one obtains

$$\frac{e^{\phi_i}}{1 + \sum_{i'=1}^{r-1} e^{\phi_{i'}}} = \frac{N'_i}{\sum_{j=1}^{r} N'_j} \quad i = 1, \ldots r-1 \tag{34}$$

or equivalently

$$\theta_j = \frac{N'_i}{N'} \quad j = 1, \ldots r \tag{35}$$

3.2 Dirichlet priors for graphical models

Heckerman, Geiger and Chickering 2005 (HGC) show that the assumptions of *likelihood equivalence* which says that data should not help discriminate

[1] To have a square matrix, we express $\theta_r = 1 - \theta_1 - \ldots - \theta_{r-1}$ and take the partial derivatives of $\phi_{1:r-1}$ w.r.t $\theta_{1:r-1}$.

between structures which represent the same probability distribution, *parameter modularity* which says that the parameters corresponding to an edge of the tree should have the same prior every time the edge is present in the tree and *parameter independence* which says that in any directed tree parametrization the parameters of each edge are independent of anything else. These combined with some weak technical assumptions[2] imply that the parameter prior is Dirichlet.

HGC also show that the likelihood equivalence constrains the Dirichlet priors for all the parameter sets to share a common equivalent sample size N'.

Alternatively, one can normalize the counts and express the Dirichlet prior over all trees as a table of fictitious marginal probabilities P'_Y for each subset Y of variables plus an *equivalent sample size* N' that gives the strength of the prior.

Note that two of the assumptions, likelihood equivalence and parameter modularity, make sense only if we consider multiple graph structures. I.e we should take them into account only if we are also estimating model structure. If we are not, then some of the constraints on the (Dirichlet) prior can be relaxed, as for example the constraint of having the same prior strenght N'.

The *uninformative* prior given by

$$P'_{X|Y}(xy) \;=\; \frac{1}{r_X} \tag{36}$$

is valid since it represents the set of pairwise marginals of the uniform distribution over $\Omega(V)$[3].

If the Dirichlet prior is represented in the natural parameters and the empirical distribution is P, with sample size N, then, from the fact that the Dirichlet prior is a conjugate prior, it follows that finding the MAP parame-

[2]These technical assumptions assumptions amount to the positivity of the joint prior.

[3]This prior is called the BDeu prior in HGC. We note in passing that the uninformative prior denoted there as the K2 metric is not a valid prior from the point of view of equations (**??**).

ters is equivalent to finding the ML parameters for

$$\tilde{P} = \frac{1}{N+N'}(N'P' + NP). \tag{37}$$

Consequently, the parameters of the optimal model will be

$$\theta^{Bayes}_{X|Y} = \frac{\tilde{P}_{XY}}{\tilde{P}_Y} \tag{38}$$

and, according to the previous section and equation (35) they will also represent the mean of the posterior distribution. Moreover, using the parameter independence assumption, we can conclude that the optimal distribution P^{Bayes} itself is the mean of the posterior distribution given the structure.

4 Smoothing

Smoothing is essentially adjusting the ML estimates of discrete probabilities in the case of little data. There are very many smoothing methods, some developed for very special situations and some with more general applicability. We will discuss two classes of smoothing methods:

- **shrinkage/backing off** which applies to a hierarchy of discrete conditional distributions
- **discounting** which applies to an discrete distribution taken separately and involves essentially taking probability mass off the observed values of a variable to spread it to the values with 0 counts

4.1 Discounting

4.1.1 Laplace or Dirichlet smoothing

The simplest (and least useful) discounting method is the Laplace smoothing. It is equivalent with a Dirichlet prior with all fictitious counts equal to 1 (i.e

we assume we have seen one more example of each value in Ω_X). The method can be summarized as: add 1 to each count and renormalize.

$$\theta_i^{Laplace} = \frac{N_i + 1}{N + r_X} \quad \text{for } i = 1 : r_X \tag{39}$$

4.1.2 Witten-Bell discounting – probability of a new value

We look at the observation sequence as a binary process: either we observe a value of X that was observed before, or we observe a new one. Assuming that of the total of r_X possible values r_0 were observed and $r_X - r_0$ were unobserved, the probability of observing a new value is $p_0 = \frac{r_0}{N}$. We extrapolate by setting the total probability of the yet unseen values of X to p_0. The other probabiliy estimates are renormalized accordingly, yielding

$$\theta_i^{WB} = \begin{cases} \frac{N_i}{N}\frac{1}{1+p_0} = \frac{N_i}{N+r_0} & N_i > 0 \\ \frac{1}{r_X - r_0}\frac{1}{1+p_0} = \frac{1}{r_X - r_0}\frac{r_0}{N+r_0} & N_i = 0 \end{cases} \tag{40}$$

While Laplace smoothing and Ney-Essen smoothing (below) apply to any set of observations, Witten-Bell smoothing makes sense only for the case when some N_i counts are zero. If all $N_i > 0$ then W-B smoothing does nothing (i.e $p_0 = 0$).

WB smoothing has no parameter to choose. In this sense it presents an advantage w.r.t Ney-Essen; it removes the subjective user choice and replaces it with a principled method for estimating the mass of the unseen values.

4.1.3 Ney-Essen discounting – shave off some mass from every value

In this method, a fixed amount $\delta \in (0, 1)$ is substracted from every non-zero count. The total amount is then equally distributed to *all* counts. This simple method works surprisingly well in practice.

$$D = \sum_i \min(N_i, \delta) \quad (41)$$

$$\theta_i^{NE} = \frac{N_i - \min(N_i, \delta) + D/r_X}{N} \quad (42)$$

This method is both general and flexible, and has the following advantage over Laplace smoothing. For r_X large and r_0 small, i.e when the probability mass is concentrated on a few values, D will be small too, because $D \leq \delta r_0$, and consequently the mass $D/r_X \leq \delta r_0/r_X$ distributed to the unobserved values will be small too. On the other hand, if the N observation are distributed over a large number r_0 of values, then we have reason to believe that the unobserved values also have larger probability of appearing. This is exaclty what Ney-Essen smoothing does in this case: since D will be much larger than in the previous case, and thus the unseen values will receive more mass.

While Laplace and WB always reduce the θ_i estimates (w.r.t to the ML estimates θ_i^{ML} for the observed values in order to provide for the unobserved values, the NE smoothing *may increase* the non-zero zero θ_i's whose counts N_i are below δ. The intuition behind this is that any value i with $N_i < \delta$ should be considered a rare value and should be treated in the same way, no matter how many observations it actually has. Indeed, when $\delta > 1$, the final θ_i is the same for all values that have $N_i = 0$ and $N_i = 1$.

A rule of thumb for choosing the smoothing parameter is $\delta \propto \frac{1}{N}$. The motivation for this fact is that if in N observations a value is observed once, its probability is about $1/N$ (and if it's observed zero times, its probability is order $1/N$ or smaller).

4.2 Back-off or shrinkage – mixing with simpler models

This method smooths the probability estimates of a more detailed model (i.e having more parameters) by the probabilities of a coarser model (i.e simpler, with fewer parameters). The latter are more reliable having better

data support. Under the name of **shrinkage** this class of methods has been known in statistics for a while as standard variance reduction methods.

For instance, a model over word trigrams (billions of parameters) can be backed off with a model over bigrams (only millions of parameters) which in turn can be backed off with a model over unigrams (only thousand parameters). Hence, a model in which a variable has two parents is backed off with models with 1 and 0 parents (more independencies). One can also resort to back-off models which bin the variable values. For example, bi- and tri-gram models over words can be backed off by bi- or tri-gram models over parts of speech.

Backing off represents mixing the detailed model of interest with the coarser models. For example, in any Bayes net with $pa(X) = \{Y_1, Y_2, \ldots\}$ we can do the following smoothing

$$\theta_{X|Y_1Y_2\ldots} = \lambda_1\theta^{ML}_{X|Y_1Y_2\ldots} + \lambda_2\theta^{ML}_{X|Y_1} + \lambda_2\theta^{ML}_{X|Y_2} + \ldots + \lambda_4\theta^{ML}_{X} + \lambda_5\theta^{ML}_{\tilde{X}|\tilde{Y}_1\tilde{Y}_2\ldots} \quad (43)$$

In the above, $\tilde{X}, \tilde{Y}_j$ represent "coarsened" or "binned" versions of Y_j and the λ_k's are coefficients that sum to one in order to ensure that the result is a valid distribution over Ω_X.

The coefficients are estimated by cross-validation. More sophisticated methods are possible, and used for e.g. language models, whereby the λ coefficients, very many, depend on the parents, and their values depend on the number of observations of a certain parent configuration. Under these conditions, instead of cross-validation, one uses the EM algorithm (on the hold-out set) to estimate the λ's.