Maximum likelihood in log-linear models

Steffen Lauritzen, University of Oxford

Graphical Models, Lecture 4, Michaelmas Term 2010

October 22, 2010

Let $\mathcal{A}$ denote an arbitrary set of subsets of V. A density f (or function) *factorizes* w.r.t. $\mathcal{A}$ if there exist functions $\psi_a(x)$ which depend on x_a only and

$$f(x) = \prod_{a \in \mathcal{A}} \psi_a(x).$$

Similar to factorization w.r.t. graph, but $\mathcal{A}$ are not necessarily complete subsets of a graph.

The set of distributions $\mathcal{P}_{\mathcal{A}}$ which factorize w.r.t. $\mathcal{A}$ is the *hierarchical log–linear model* generated by $\mathcal{A}$.

To avoid redundancy, it is common to assume the sets in $\mathcal{A}$ to be incomparable in the sense that no subset in $\mathcal{A}$ is contained in any other member of $\mathcal{A}$. $\mathcal{A}$ is the *generating class* of the log–linear model.

For any generating class $\mathcal{A}$ we construct the dependence graph $G(\mathcal{A}) = G(\mathcal{P}_\mathcal{A})$ of the log–linear model $\mathcal{P}_\mathcal{A}$.

Since the pairwise Markov property has to hold for all members of $\mathcal{P}_\mathcal{A}$, it has at least to hold for all positive members. *The dependence graph is determined by the relation*

$$\alpha \sim \beta \iff \exists a \in \mathcal{A} : \alpha, \beta \in a.$$

For sets in $\mathcal{A}$ are clearly complete in $G(\mathcal{A})$ and therefore *distributions in $\mathcal{P}_\mathcal{A}$ do factorize according to $G(\mathcal{A})$*. On the other hand, any graph with fewer edges would not suffice.

They are thus also global, local, and pairwise Markov w.r.t. $G(\mathcal{A})$.

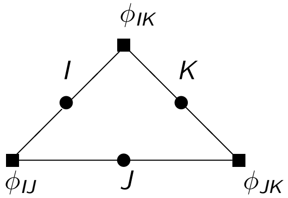
As a generating class defines a dependence graph $G(\mathcal{A})$, the reverse is also true.

The set $\mathcal{C}(\mathcal{G})$ of *cliques* (maximal complete subsets) of $\mathcal{G}$ is a generating class for the log–linear model of distributions which factorize w.r.t. $\mathcal{G}$.

If the dependence graph completely summarizes the restrictions imposed by $\mathcal{A}$, i.e. if

$$\mathcal{A} = \mathcal{C}(G(\mathcal{A})),$$

$\mathcal{A}$ is *conformal*.



The *factor graph* of $\mathcal{A}$ is the bipartite graph with vertices $V \cup \mathcal{A}$ and edges define by

$$\alpha \sim a \iff \alpha \in a.$$

Using this graph even non-conformal log–linear models admit a simple visual representation.

Data in list form

Consider a sample $X^1 = x^1, \ldots, X^n = x^n$ from a distribution with probability mass function p. We refer to such data as being in *list form*, e.g. as

case	Admitted	Sex
1	Yes	Male
2	Yes	Female
3	No	Male
4	Yes	Male
⋮	⋮	⋮

Contingency Table

Data often presented in the form of a *contingency table* or *cross-classification*, obtained from the list by sorting according to category:

	Sex	
Admitted	Male	Female
Yes	1198	557
No	1493	1278

The numerical entries are *cell counts*

$$n(x) = |\{\nu : x^{\nu} = x\}|$$

and the total number of observations is $n = \sum_{x \in \mathcal{X}} n_(x)$.

Assume now $p \in \mathcal{P}_{\mathcal{A}}$ but otherwise unknown. The likelihood function can be expressed as

$$L(p) = \prod_{\nu=1}^{n} p(x^{\nu}) = \prod_{x \in \mathcal{X}} p(x)^{n(x)}.$$

In contingency table form the data follow a multinomial distribution

$$P\{N(x) = n(x), x \in \mathcal{X}\} = \frac{n!}{\prod_{x \in X} n(x)!} \prod_{x \in \mathcal{X}} p(x)^{n(x)}$$

but this only affects the likelihood function by a constant factor.

The likelihood function

$$L(p) = \prod_{x \in \mathcal{X}} p(x)^{n(x)},$$

is continuous as a function of the ($|\mathcal{X}|$-dimensional vector) unknown probability distribution p.

Since the *closure* $\overline{\mathcal{P}_{\mathcal{A}}}$ is compact (bounded and closed), *L attains its maximum on $\overline{\mathcal{P}_{\mathcal{A}}}$.*

Unfortunately, $\mathcal{P}_{\mathcal{A}}$ is not closed by itself so limits of factorizing distributions do not necessarily factorize.

The maximum of the likelihood function may not necessarily on $\mathcal{P}_{\mathcal{A}}$ itself, so it is necessary in general to include the boundary points.

Indeed, it is also true that *L has a unique maximum over $\overline{\mathcal{P}_{\mathcal{A}}}$,* which we shall now show.

For simplicity, we only establish uniqueness within $\mathcal{P}_{\mathcal{A}}$. The proof is indirect, but quite simple.

Assume $p_1, p_2 \in \mathcal{P}_{\mathcal{A}}$ with $p_1 \neq p_2$ and

$$L(p_1) = L(p_2) = \sup_{p \in \mathcal{P}_{\mathcal{A}}} L(p). \tag{1}$$

Define

$$p_{12}(x) = c\sqrt{p_1(x)p_2(x)},$$

where $c^{-1} = \{\sum_x \sqrt{p_1(x)p_2(x)}\}$ is a normalizing constant.

Then $p_{12} \in \mathcal{P}_{\mathcal{A}}$ because

$$\begin{aligned} p_{12}(x) &= c\sqrt{p_1(x)p_2(x)} \\ &= c\prod_{a\in\mathcal{A}}\sqrt{\psi_a^1(x)\psi_a^2(x)} = \prod_{a\in\mathcal{A}}\psi_a^{12}(x), \end{aligned}$$

where e.g. $\psi_a^{12} = c^{1/|\mathcal{A}|}\sqrt{\psi_a^1(x)\psi_a^2(x)}$.

The Cauchy–Schwarz inequality yields

$$c^{-1} = \sum_x \sqrt{p_1(x)p_2(x)} < \sqrt{\sum_x p_1(x)}\sqrt{\sum_x p_2(x)} = 1$$

i.e. *we have $c > 1$.*

Hence

$$
\begin{aligned}
L(p_{12}) &= \prod_x p_{12}(x)^{n(x)} \\
&= \prod_x \left\{ c\{\sqrt{p_1(x)p_2(x)} \right\}^{n(x)} \\
&= c^n \prod_x \sqrt{p_1(x)}^{n(x)} \prod_x \sqrt{p_2(x)}^{n(x)} \\
&= c^n \sqrt{L(p_1)L(p_2)} \\
&> \sqrt{L(p_1)L(p_2)} = L(p_1) = L(p_2),
\end{aligned}
$$

which contradicts (1). Hence we conclude $p_1 = p_2$.

The extension to $\overline{\mathcal{P}_{\mathcal{A}}}$ is almost identical. It just needs a limit argument to establish $p_1, p_2 \in \overline{\mathcal{P}_{\mathcal{A}}} \Rightarrow p_{12} \in \overline{\mathcal{P}_{\mathcal{A}}}$.

The maximum likelihood estimate $\hat{p}$ of p is the unique element of $\overline{\mathcal{P}_{\mathcal{A}}}$ which satisfies the system of equations

$$n\hat{p}(x_a) = n(x_a), \forall a \in \mathcal{A}, x_a \in \mathcal{X}_a. \tag{2}$$

Here $g(x_a) = \sum_{y:y_a=x_a} g(y)$ is the *a-marginal* of the function g.
The system of equations (2) expresses the *fitting of the marginals* in $\mathcal{A}$.
It can be seen as an instance of the fact that in an exponential family (log-linear $\sim$ exponential), the MLE is found by equating the sufficient statistics (marginal counts) to their expectation.

Proof: Assume $p^* \in \mathcal{P}_\mathcal{A}$ is a solution to the equations (2). That p^* maximizes the likelihood function follows from the calculation below, where $p \in \mathcal{P}_\mathcal{A}$ is arbitrary and $\phi_a = \log \psi_a$:

$$\begin{aligned}
\log L(p) &= \sum_{x\in\mathcal{X}} n(x) \log p(x) = \sum_{x\in\mathcal{X}} n(x) \sum_{a\in\mathcal{A}} \phi_a(x) \\
&= \sum_{a\in\mathcal{A}} \sum_{x\in\mathcal{X}} n(x)\phi_a(x) \\
&= \sum_{a\in\mathcal{A}} \sum_{x_a\in\mathcal{X}_a} \sum_{y:y_a=x_a} n(y)\phi_a(y) \\
&= \sum_{a\in\mathcal{A}} \sum_{x_a\in\mathcal{X}_a} n(x_a)\phi_a(x),
\end{aligned}$$

as $n(x_a) = \sum_{y:y_a=x_a} n(y)$.

Further we get

$$\begin{aligned}\log L(p) &= \sum_{a\in\mathcal{A}}\sum_{x_a\in\mathcal{X}_a} n(x_a)\phi_a(x)\\ &= \sum_{a\in\mathcal{A}}\sum_{x_a\in\mathcal{X}_a} np^*(x_a)\phi_a(x)\\ &= \sum_{a\in\mathcal{A}}\sum_{x\in\mathcal{X}} np^*(x)\phi_a(x)\\ &= \sum_{x\in\mathcal{X}} np^*(x)\log p(x).\end{aligned}$$

Thus, for any $p \in \mathcal{P}_{\mathcal{A}}$ we have established that

$$\log L(p) = \sum_{x\in\mathcal{X}} np^*(x)\log p(x).$$

This is in particular also true for p^*. The information inequality now yields

$$\begin{aligned}\log L(p) &= \sum_{x\in\mathcal{X}} np^*(x)\log p(x)\\ &\leq \sum_{x\in\mathcal{X}} np^*(x)\log p^*(x) = \log L(p^*).\end{aligned}$$

The case of $p^* \in \overline{\mathcal{P}_{\mathcal{A}}}$ needs an additional limit argument.

To show that the equations (2) indeed have a solution, we simply describe a convergent algorithm which solves it. This cycles (repeatedly) through all the a-marginals in $\mathcal{A}$ and fit them one by one.

For $a \in \mathcal{A}$ define the following *scaling* operation on p:

$$(T_a p)(x) \leftarrow p(x) \frac{n(x_a)}{np(x_a)}, \quad x \in \mathcal{X}$$

where $0/0 = 0$ and $b/0$ is undefined if $b \neq 0$.

Fitting the marginals

The operation T_a *fits the a-marginal* if $p(x_a) > 0$ when $n(x_a) > 0$:

$$\begin{aligned} n(T_a p)(x_a) &= n \sum_{y:y_a=x_a} p(y) \frac{n(y_a)}{np(y_a)} \\ &= n \frac{n(x_a)}{np(x_a)} \sum_{y:y_a=x_a} p(y) \\ &= n \frac{n(x_a)}{np(x_a)} p(x_a) = n(x_a). \end{aligned}$$

Consequently, we have $T_a^2 = T_a$. No reason to do it twice.

Make an ordering of the generators $\mathcal{A} = \{a_1, \ldots, a_k\}$. Define S by a full cycle of scalings

$$Sp = T_{a_k} \cdots T_{a_2} T_{a_1}.$$

Define the iteration

$$p_0(x) \leftarrow 1/|\mathcal{X}|, \quad p_n = Sp_{n-1}, n = 1, \ldots.$$

It then holds that

$$\lim_{n\to\infty} p_n = \hat{p}$$

where $\hat{p}$ is the unique maximum likelihood estimate of $p \in \overline{\mathcal{P}_\mathcal{A}}$, i.e. the solution of the equation system (2).

Known as the *IPS*-algorithm or *IPF*-algorithm, or as a variety of other names. Implemented e.g. (inefficiently) in *R* in `loglin` with front end `loglm` in MASS.

Key elements in proof:

1. If $p \in \overline{\mathcal{P}_{\mathcal{A}}}$, so is $T_a p$;
2. T_a is continuous at any point p of $\overline{\mathcal{P}_{\mathcal{A}}}$ with $p(x_a) \neq 0$ whenever $n(x_a) = 0$;
3. $L(T_a p) \geq L(p)$ so likelihood always increases;
4. $\hat{p}$ is the unique fixpoint for T (and S);
5. $\overline{\mathcal{P}_{\mathcal{A}}}$ is compact.

A simple example

	Admitted		
Sex	Yes	No	S-marginal
Male	1198	1493	2691
Female	557	1278	1835
A-marginal	1755	2771	4526

Admissions data from Berkeley. Consider $A \perp\!\!\!\perp S$, corresponding to $\mathcal{A} = \{\{A\}, \{S\}\}$.
We should fit A-marginal and S-marginal iteratively.

Initial values

	Admitted		
Sex	Yes	No	S-marginal
Male	1131.5	1131.5	2691
Female	1131.5	1131.5	1835
A-marginal	1755	2771	4526

Entries all equal to $4526/4$. Gives initial values of np_0.

Fitting S-marginal

	Admitted		
Sex	Yes	No	S-marginal
Male	1345.5	1345.5	2691
Female	917.5	917.5	1835
A-marginal	1755	2771	4526

For example

$$1345.5 = 1131.5 \frac{2691}{1131.5 + 1131.5}$$

and so on.

Fitting A-marginal

	Admitted		
Sex	Yes	No	S-marginal
Male	1043.46	1647.54	2691
Female	711.54	1123.46	1835
A-marginal	1755	2771	4526

For example

$$711.54 = 917.5 \frac{1755}{917.5 + 1345.5}$$

and so on.

Algorithm has converged, as both marginals now fit!

Normalised to probabilities

	Admitted		
Sex	Yes	No	S-marginal
Male	0.231	0.364	0.595
Female	0.157	0.248	0.405
A-marginal	0.388	0.612	1

Dividing everything by 4526 yields $\hat{p}$.
It is overkill to use the IPS algorithm as there is an explicit formula, as we shall see next time.

- *L attains its maximum on* $\overline{\mathcal{P}_{\mathcal{A}}}$.

- *L attains its maximum on $\overline{\mathcal{P}_{\mathcal{A}}}$.*
- *The maximizer $\hat{p} \in \overline{\mathcal{P}_{\mathcal{A}}}$ of L is unique.*

- *L attains its maximum on $\overline{\mathcal{P}_{\mathcal{A}}}$.*
- *The maximizer $\hat{p} \in \overline{\mathcal{P}_{\mathcal{A}}}$ of L is unique.*
- *The maximizer $\hat{p}$ is determined as the unique solution within $\overline{\mathcal{P}_{\mathcal{A}}}$ of the equations*

$$n\hat{p}(x_a) = n(x_a), \forall a \in \mathcal{A}, x_a \in \mathcal{X}_a.$$

- *L attains its maximum on $\overline{\mathcal{P}_\mathcal{A}}$.*
- *The maximizer $\hat{p} \in \overline{\mathcal{P}_\mathcal{A}}$ of L is unique.*
- *The maximizer $\hat{p}$ is determined as the unique solution within $\overline{\mathcal{P}_\mathcal{A}}$ of the equations*

$$n\hat{p}(x_a) = n(x_a), \forall a \in \mathcal{A}, x_a \in \mathcal{X}_a.$$

- *The maximizer is the limit of the convergent repeated fitting of marginals*

$$T_a : p(x) \leftarrow p(x)n(x_a)/\{np(x_a)\}, x \in \mathcal{X}, a \in \mathcal{A}.$$