Maximum Likelihood Estimation of the Latent Class Model through Model Boundary Decomposition

Daniel Moss

February 14, 2020

1 Introduction

The *binary latent class model* is an instance of a model with incomplete data. While the EM algorithm is commonly used in such a setup, it comes with no guarantee of reaching the global optimum (i.e. the actual MLE). Thus, we study the MLE from a theoretical point of view.

By a *model* we mean a collection of candidate laws $\mathcal{P}$. In the discrete setting, we are always fully parametric so misspecification is not a concern (that is, we can write down a parametric model capturing all possibilities).

In our setting, we denote by $\mathcal{M}_{n,r}$ the model corresponding to n binary observed nodes $X = (X_1, X_2, \ldots, X_n) \in \{1,2\}^n$, which are conditionally independent given an unobserved variable $Z \in \{1, \ldots, r\}$. We slightly abuse notation by letting this set define the set of p.m.f.s for the model, which are technically densities with respect to the counting measure on the (finite) observation space. The following diagram, taken from [**?**], gives a graphical representation of $\mathcal{M}_{5,2}$.

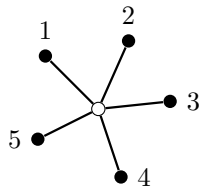

FIGURE 1: *The star graph model with* 5 *leaves. The internal vertex represents an unobserved random variable.*

We can parametrise this model by the p.m.f. $\lambda \in \Delta_{r-1}$ of the latent variable, where $\Delta_{k-1} \in \mathbb{R}^k$ is the $(k-1)-$dimensional simplex embedded into

k-dimensional space, and the conditional laws of the X_i, which are denoted

$$A^{(i)} = \begin{pmatrix} a_{11}^{(i)} & a_{12}^{(i)} \\ \vdots & \vdots \\ a_{r1}^{(i)} & a_{r2}^{(i)} \end{pmatrix}, \quad i = 1, \dots, n,$$

where the matric is stochastic (i.e. row sum is zero). We then write

$$\theta = (\lambda, A^{(1)}, \dots, A^{(n)})$$

for the full parameter, and the parametrisation is given by $\theta \mapsto p(\theta) \in \Delta_{2^n-1}$ where

$$\Pr(X_1 = i_1, \dots X_n = i_n; \theta) = p_{i_1,\dots,i_n}(\theta) = \sum_{k=1}^{r} \lambda_i \prod_{j=1}^{n} a_{ki_j}^{(j)} \tag{1}$$

The map is injective (when $n \geq 3$? Kruskal/Allman09?) so these parameters are identifiable.[1] Moreover, this parametrisation shows that the p.m.f. is a binary tensor of *nonnegative rank* at most r, since it is the sum of r nonnegative simple tensors (ones of the form $T_{ijk} = u_i v_k w_k$ etc.)

We will be primarily interested in the setting $r = 2$, and so we write $\mathcal{M}_n$=$\mathcal{M}_{n,2}$ for convenience.

It is clear that $\mathcal{M}_n \subset \Delta_{2^n-1}$. Suppose now we are interested in the problem of maximum likelihood estimation in this model. Write

$$U = (u_{i_1,\dots,i_n})_{i_1,\dots,i_n \in \{1,2\}}$$

for the count data from an observation[2] of size N. Thus the sum of all elements in U is equal to N. Then the likelihood is given by

$$L_N(X^1, \dots, X^N) = \sum_{i_1,\dots,i_n \in \{1,2\}} u_{i_1,\dots,i_n} \log p_{i_1,\dots,i_n}(\theta)$$

It is straightforward to see that the MLE (for the p.m.f.) over the whole space Δ_{2^n-1} is given by the proportions of count data. However, the inclusion $\mathcal{M}_n \subset \Delta_{2^n-1}$ is strict and thus this need not be the MLE over the model we consider. In the case where the count data does not define a p.m.f. in $\mathcal{M}_n$, the MLE over this model will be found on the boundary (why?).

In order to understand the structure of $\mathcal{M}_n$, we give a *semi-algebraic description* of the set as a subset of the simplex (that is, one described by equalities and inequalities).

[1] Even if they aren't, does it matter? Are we trying to infer the latent structure or just the p.m.f. of the $X_1, \dots, X_n$. It seems to me that latent models may be used to understand more deeply the structure of the joint law, but in this setting they have a secondary use of allowing certain algebraic decompositions.

[2] One observation X^i amounts to observing a copy of the vector $X = (X_1, \dots, X_n)$

This description is given by Theorem 2 in [**?**], which says that for the $r = 2$ case, tensors of non-negative rank at most two are exactly those which are *supermodular* and which have *flattening rank* at most two. The supermodularity conditions will be referred to as the *inequality conditions* (are these the same thing as requiring signed MTP_2? c.f. [**?**] page 9). To define the flattening rank, we define a matrix flattening of a binary tensor P as the $2^{|\Gamma|} \times 2^{n-|\Gamma|}$ matrix obtained by placing suffices corresponding to $\Gamma \subset \{1, \ldots, n\}$ along the rows, and others along the columns. For example, if $\Gamma = \{1, 2\}$ and $n = 5$ we get

$$\begin{pmatrix} p_{11111} & p_{11211} & p_{11121} & p_{11221} & p_{11112} & p_{11212} & p_{11122} & p_{11222} \\ p_{21111} & p_{21211} & p_{21121} & p_{21221} & p_{21112} & p_{21212} & p_{21122} & p_{21222} \\ p_{12111} & p_{12211} & p_{12121} & p_{12221} & p_{12112} & p_{12212} & p_{12122} & p_{12222} \\ p_{22111} & p_{22211} & p_{22121} & p_{22221} & p_{22112} & p_{22212} & p_{22122} & p_{22222} \end{pmatrix}$$

The flattening rank is then the maximal rank of any such flattenings. These constraints are known as the *equality* constraints (because all three-minors must vanish).

Armed with this semi-algebraic description, the authors in [**?**] then go on to decompose the boundary of $\mathcal{M}_n$ ('boundary straficiation') into certain collections (five types) of components of different dimension. These boundary strata correspond to certain 'context-specific' conditional independencies, meaning (I think) that if the MLE lies on this strata then we are additionally maxing a sort of 'most likely structure' comment. An example of one such strata 'type 5' is the case of full (unconditional) independence. The remainder of the chapter in which this decomposition result is stated is then dedicated to proving this result.

2 Ising model

[Relation to $\mathcal{M}_{n,r}$?] The *Ising model* is defined by

Where $V = \{1, \ldots, n\}$. In order for these distributions to be signed MTP_2, we require the following inequalities to hold:

or permutations thereof by swaping the label of the observed variabls (thereby changing some choice of 2 of the three $\geq$ to $\leq$). We can fit this model, either through GLM (Poisson family) and then checking the inequalities, or by using the technique in the paper (cite). Then we can check if the *equality* conditions hold, which would mean that the fitted distribution when we assume our observations arise from an Ising model are members of the $\mathcal{M}_n$ parametric family (the latent parameters may then be identified for $n \geq 3$ if so desired).

The purpose of this approach is to fit an $\mathcal{M}_n$ model by making the simplifying assumption that the observations come from a more computationally convenient signed MTP_2 distribution, and then see if fitting this model indeed fits the particular signed MTP_2 distribution in which we are interested.t

What is $C^\wedge$? [Dual cone]

Relation between exponential family models and latent variable models? Both signed MTP2, rank condition? Refs linking the two?

References