STATS305B – Lecture 12

Jonathon Taylor
Scribed by Michael Howes

02/14/22

Contents

1 Log-linear models

Last lecture we ended with a description of homogeneous association models. These are log linear models where we have multiple categorical features and a Poisson response. We allow for first order and second order effects in the model, but we do not include any terms that are third order or higher. We can represent such a model with a graph G. In the graph G we have a vertex for each feature and an edge between two features if the model includes an interaction term between the two features. For example, consider the following graph

This graph represents the model

$$\begin{aligned}\log(\mu_{ijkl}) = \lambda + \lambda_i^W + la_j^X + \lambda_k^Y + \lambda_l^Z \\ + \lambda_{ij}^{WX} + \lambda_{il}^{WZ} + \lambda_{jk}^{XY} + \lambda_{kl}^{YZ}.\end{aligned}$$

Note the absence of any terms of the form λ_{ik}^{WY} or λ_{jl}^{XZ}. This corresponds to the fact that the $W-Y$ and $X-Z$ edges are absent in the graph. This means that $X \perp\!\!\!\perp Z \mid Y, W$ and $W \perp\!\!\!\perp Y \mid X, Z$. For another example, consider the graph,

This graph corresponds to the model

$$\begin{aligned}\log(\mu_{ijkl}) = \lambda + \lambda_i^W + la_j^X + \lambda_k^Y + \lambda_l^Z \\ + \lambda_{ij}^{WX} + \lambda_{il}^{WZ} + \lambda_{jk}^{WY} + \lambda_{kl}^{YZ}.\end{aligned}$$

In this model we have $X \perp\!\!\!\perp Y, Z \mid W$. Such homogeneous association models are useful for tests of conditional independence since the inclusion and exclusion of edges tell us everything about the conditional independence structure.

1.1 Connection with logistic regression

Consider a multiway table with a binary variable Y. Suppose we are primarily interested in the effect of the other features on Y. If we use $\backslash Y$ to denote the other features, then $Y \mid \backslash Y$ will be distributed according to a logistic model with features $N(Y)$ where $N(Y)$ is the set of neighbors of Y in the graph representing the model. Be varying the feature of interest Y, the log linear model will give us many logistic models. If the variables are categorical rather than binary, then the conditional model is a baseline multinomial logistic model rather than a binary model. The notation gets more complicated, and we have to estimate more parameters, but the concepts are the same.

2 Pseudo-likelihood

2.1 Motivation

Expanding on the previous observation, consider an $I \times J \times K$ table with three variables X, Y, Z. The homogeneous association model can be written as

$$\begin{aligned}\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z \\ + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}.\end{aligned}$$

We have $\lambda \in \mathbb{R}$, $\lambda^X \in \mathbb{R}^I$, $\lambda^Y \in \mathbb{R}^J$, $\lambda^Z \in \mathbb{R}^K$, $\lambda^{XY} \in \mathbb{R}^{I\times J}$, $\lambda^{XZ} \in \mathbb{R}^{I\times K}$ and $\lambda^{YZ} \in \mathbb{R}^{J\times K}$. And we introduce the constraint that the coefficients each zero whenever one or more of the indices satisfy $i = I, j = J$ or $k = K$. As noted above, we can create a baseline multinomial model with any of X, Y or Z as the response. If X is the response, then I is the baseline category and

$$\begin{aligned}\frac{\mathbb{P}(X = i|Y = j, Z = k)}{\mathbb{P}(X = I|Y = j, Z = k)} &= \frac{\exp(\lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ})}{\exp(\lambda + \lambda_I^X + \lambda_j^Y + \lambda_k^Z + \lambda_{Ij}^{XY} + \lambda_{Ik}^{XZ} + \lambda_{jk}^{YZ})} \\ &= \frac{\exp(\lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ})}{\exp(\lambda + \lambda_j^Y + \lambda_k^Z + \lambda_{jk}^{YZ})} \\ &= \exp(\lambda_i^X + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ}).\end{aligned}$$

And likewise if Y is the response we have

$$\frac{\mathbb{P}(Y = j|X = i, Z = k)}{\mathbb{P}(Y = J|X = i, Z = k)} = \exp(\lambda_j^Y + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ}).$$

And if Z is the response, then we have

$$\frac{\mathbb{P}(Z = k|X = i, Y = j)}{\mathbb{P}(Z = K|X = i, Y = j)} = \exp(\lambda_k^Z + \lambda_{ij}^{XZ} + \lambda_{jk}^{YZ}).$$

Thus, by fitting the three logistic regression models we can estimate the main effects and the interactions. For the main effects, there is exactly one model that estimates each vector $\lambda^X, \lambda^Y, \lambda^Z$. But for the interaction terms, we have two different models estimating the same parameter. For example, when the parameter λ_{ij}^{XY} can be estimated using the model M_X that has X as the response of the model M_Y that has Y as the response. These models give us estimate $\widehat{\lambda}_{ij}^{XY}(M_X)$ and $\widehat{\lambda}_{ij}^{XY}(M_Y)$. In general, we expect $\widehat{\lambda}_{ij}^{XY}(M_X) \neq \widehat{\lambda}_{ij}^{XY}(M_Y)$. This means we need some way to combine the two estimates.

2.2 Definition of the pseudo-likelihood

Recall that $\widehat{\lambda}_{ij}^{XY}(M_X)$ and $\widehat{\lambda}_{ij}^{XY}(M_Y)$ are fit by maximizing the likelihood under the two different models. To get a single estimate of λ_{ij}^{XY} and "pool" the different likelihoods and then maximize this

pooled likelihood. This leads us to the idea of the *pseudo-likelihood*. For the $I \times J \times K$ table, the pseudo-likelihood L_{pseudo} is given by

$$\log L_{\text{pseudo}}(\Lambda|X,Y,Z) := \log L^X_{\text{baseline}}(\Lambda|X,Y,Z) + \log L^Y_{\text{baseline}}(\Lambda|X,Y,Z) + \log L^Z_{\text{baseline}}(\Lambda|X,Y,Z),$$

where Λ contains all the parameters of L^A_{baseline} is the likelihood for the baseline logistic model with A as the response. The pseudo-likelihood can generalize easily to more than three variables. Suppose that we have a homogeneous association model represented by a graph $G = (V, E)$. Meaning that V is the set of variables and the interaction terms described by the set of edges E. For a variable $A \in V$, let $N(A)$ be the set of variables that are neighbors of A in the graph G. The pseudo-likelihood of this model is given by

$$\log L_{\text{pseudo}}(\Lambda|V) = \sum_{A \in V} L^A_{\text{baseline}}(\Lambda|N(A)).$$

The pseudo-likelihood is easier to compute than the exact likelihood. This means that is easier to use optimization techniques such as gradient descent on the pseudo-likelihood rather than exact likelihood. Unfortunately, the pseudo-likelihood is not a likelihood, and so we can not use the asymptotic theory of the MLE to do inference.

2.3 Exact likelihood

The claim is that the pseudo-likelihood is computationally easier than the exact likelihood, but why is this true? The complexity comes in when we have many variables. Consider the binary case when we have a set of variables $V = \{X_1, \ldots, X_m\}$ and a set of interaction edges E. Since we have binary data, the number of parameters if $1 + |V| + |E|$. We have one parameter for overall intensity, one parameter for the main effect of each variable and one parameter for each interaction specified by the edges in E. Let γ be the parameter for overall intensity. The edge and vertex parameters can be specified by a symmetric matrix $\Theta \in \mathbb{R}^{m \times m}$ with the constraint $\Theta_{ij} = \Theta_{ji}$ and $\Theta_{ij} = 0$ for all $i \neq j$ such that $(i, j) \notin E$. If we observe a whole table of counts, the exact likelihood is

$$\log(L(\gamma, \Theta, X)) = \gamma \cdot 2^m + \text{Tr}(\Theta S) - \text{sum}(N) \log \left(\sum_{x \in \{0,1\}^m} \exp(\gamma + x^T \Theta x) \right),$$

where N is a vector of counts for each of the 2^n assignments of variables and

$$S_{ij} = \begin{cases} (1,1) \text{ entry of the } X_i \times X_j \text{ marginal table} & \text{if } i \neq j, \\ \text{the number of 1's in the } X_i \text{ marginal table} & \text{if } i = j. \end{cases}$$

If we have a matrix $B \in \{0,1\}^{\text{sum}(N) \times m}$ where $B_{i,j} = 1$ if and only if the i^{th} observation has a 1 for X_j, then

$$S = B^T B \in \mathbb{R}^{m \times m}.$$

For a moderately large value of m, the normalizing constant is very difficult to compute since it is a sum over exponentially many terms.

2.4 Selecting interactions

If we do not have a graphical model in mind but believe that some interaction terms are zero, then we can use the LASSO to select parameters. We work with the full marginal homogeneous model, but optimize the penalized pseudo-likelihood,

$$L_{\text{pseudo}}(\gamma, \Theta|X) + \lambda \left\| \Theta_{\text{off diagonal}} \right\|_1.$$

That is, we add a LASSO penalty to the interaction coefficients and leave the main effects unpenalized.

2.5 Relation to Gibbs sampling

We know that the distribution of $X_i|X_{-i} = X|N(X_i)$ by the assumptions of the graphical model. Thus, the posterior densities used in Gibb's sampling are proportional to the pseudo-likelihood to times the prior.

2.6 Nuisance parameters

The "pieces" of the pseudo-likelihood are conditional likelihoods. Conditional likelihoods are useful for eliminating nuisance parameters. Whenever a model uses an exponential family, we can condition on sufficient statistics to eliminate nuisance parameters. Consider for example, Gaussian linear regression. In this model, we have

$$\begin{aligned} -\log L(\beta, \sigma^2|X, Y) &= \frac{1}{2\sigma^2} \|Y - X\beta\|_2^2 + \frac{n}{2} \log(2\pi\sigma^2) \\ &= \frac{1}{2\sigma^2} \|Y\|_2^2 - \frac{1}{\sigma^2} \beta^T (X^T Y) + C(\sigma^2, \beta), \end{aligned}$$

for some constant $C(\sigma^2, \beta)$. This is an exponential model with sufficient statistic $X^T Y$ and $\|Y\|_2^2$. The natural parameters are $\frac{\beta}{\sigma^2}$ and $-\frac{1}{2\sigma^2}$. The MLE estimate of σ^2 is $\widehat{\sigma}^2_{MLE} = \frac{1}{n} \|(I - H)Y\|_2^2$ where $H = X(X^T X)^{-1} X^T$ is the hat matrix for X. We could also estimate σ^2 be using the conditional likelihood. We are interested in the distribution of $\|Y\|_2^2 \, |X^T Y$ under (σ^2, β). Since $\|Y\|_2^2 = \|(I - H)Y\|_2^2 + \|HY\|_2^2$ and $\|HY\|_2^2$ is a function of $X^T Y$, the conditional likelihood is the likelihood of $\|(I - H)Y\|_2^2$. We know that $\frac{1}{\sigma^2} \|(I - H)Y\|_2^2$ has χ^2_{n-p} distribution and does not depend on β. Thus, the conditional likelihood has no β dependence and the MLE of the conditional likelihood is

$$\widehat{\sigma}^2_{cMLE} = \frac{1}{n - p} \|(I - H)Y\|_2^2 ,$$

which is a better than the MLE. In particular, the conditional MLE is unbiased for σ^2 and the regular MLE is not.