STATS305B – Lecture 13

Jonathon Taylor
Scribed by Michael Howes

02/16/22

Contents

1 A connection between the log-linear model and Gaussian data

Suppose we have a vertex set of binary variables $V = \{X_1, \ldots, X_m\}$ and a graph $G = (V, E)$ describing the model's interactions. This model has an intensity parameter γ and parameters for the vertexes and edges $\Theta \in \mathbb{R}^{m \times m}$ a symmetric matrix with $\Theta_{ij} = 0$ for all $(i, j) \notin E$. The likelihood for this model is

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

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

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

There is a continuous analogy of this model. Suppose we have i.i.d. data $X_i \sim \mathcal{N}(0, \Sigma)$ where the covariance matrix Σ is unknown. The likelihood can be written as

$$\log L(\Sigma) = -\frac{n}{2} \log |\det(\Theta)| - \frac{1}{2} \operatorname{Tr}(\Theta S),$$

where $\Theta = \Sigma^{-1}$ and $S = \sum_{i=1}^n X_i X_i^T = X^T X$. Thus, this Gaussian model is analogous to the graphical model before. One important difference is that in the Gaussian model we know the explicit normalizing constant which in this case equals $\frac{n}{2}\log(|\det(\Theta)|)$. Like in the log-linear case we can penalize the off-diagonal terms to select interactions. This is done with the *graphical LASSO* defined as

$$\widehat{\Theta}_\lambda = \underset{\Theta}{\operatorname{argmin}} \left\{ -\frac{n}{2}\log(|\det(\Theta)|) - \frac{1}{2}\operatorname{Tr}(\Theta S) + \lambda \left\|\Theta_{\text{off-diagonal}}\right\|_1 \right\}.$$

2 Lindsey's method

Lindsey's method is a way to use a log linear model to do density estimation.

2.1 Discrete data

First, suppose that we have samples $X_1^D, \ldots, X_n^D$ discrete random variable taking values in $1, \ldots, k$. We can define new random variables

$$N_i = \#\{1 \le s \le n : X_s^D = i\}.$$

If n is very large, it is sensible to use the model

$$N_i \sim \mathsf{Poisson}(\mu_i).$$

We can then make a log-linear model table for the parameters λ_i. Some examples would be

- The saturated model: $\log(\mu_i) = \lambda + \lambda_i$, λ_i unconstrained.
- Ordinal model: $\log(\mu_i) = \lambda + \beta i$.
- Ordinal model with basis functions $\log(\mu_i) = \lambda + \sum_{j=1}^p \beta_j h_j(i)$.

The essence of Lindsey's method is to discretize a continuous random variable and then fit a log-linear model.

2.2 Continuous data

Suppose we have samples $X_1, \ldots, X_n$ from a continuous distribution with a density f supported on $[0, 1]$. Fix $\Delta > 0$ a small number and define

$$X_s^\Delta = i \iff X_s \in ((i-1)\Delta, i\Delta].$$

Thus, X_s^Δ is a discrete random variable recording which bin X_s belongs to. We can then define $N_i = \#\{1 \le s \le n : X_s^\Delta = i\}$ and use a model $N_i \sim \mathsf{Poisson}(\mu_i)$. Note that

$$\mathbb{P}(X_s^\Delta \in ((i-1)\Delta, i\Delta]) = \int_{(i-1)\Delta}^{i\Delta} f(x)dx \approx \Delta f(i\Delta).$$

Thus, we expect $\mu_i \approx n\Delta f(i\Delta)$ and so $\log(\mu_i) = \log(n\Delta) + \log(f(i\Delta))$. If we use an ordinal model for N_i, then we have

$$\log(\mu_i) = \lambda_0 + \lambda_1 i.$$

Setting $\lambda_0 + \lambda_1 i = \log(n\Delta) + \log(f(i\Delta))$ gives $\lambda_0 = \log(n\Delta)$ and for $x \in ((i-1)\Delta, i\Delta]$,

$$\begin{aligned} f(x) &\approx f(i\Delta) \\ &\propto e^{\lambda_1 i} \\ &= e^{\frac{\lambda_1}{\Delta}\Delta i} \\ &\approx e^{\beta_1 x}, \end{aligned}$$

where $\beta_1 = \frac{\lambda_1}{\Delta}$. This allows us to approximate the density $f(x)$ with functions of the form

$$e^{\beta_0+\beta_1 x}\mathbf{1}_{[0,1]}(x).$$

We take the discretized data and fit a log-linear model of the form $\log(\mu_i) = \log(n\Delta) + \lambda_0 + \lambda_1 i$ to get a fitted value $\widehat{\lambda}_1$, and then we use $\widehat{\beta}_1 = \frac{\widehat{\lambda}_1}{\Delta}$. We also define $\widehat{\beta}_0 = -\log\left(\int_0^1 e^{\widehat{\beta}_1 x}dx\right)$. The turn $\log(n\Delta)$ can be incorporated into the glm by using the parameter `offset` is R's `glm`. Note that it is necessary to include an intercept as this contains information about the normalizing coefficient $\widehat{\beta}_0$.

2.3 Basis functions

We don't have to restrict ourselves to a linear ordinal model. Suppose we have continuous functions $h_j : [0,1] \to \mathbb{R}$ for $j = 1, \ldots, p$. Consider the log-linear model

$$\log(\mu_i) = \lambda_0 + \sum_{j=1}^{p} \lambda_j h_j(i\Delta),$$

where, as before, $N_i \sim \mathsf{Poisson}(\mu_i)$ is the number of times X_s lies in the bin $(\delta(i-1), i\delta]$. We can fit this log-linear model to get coefficients $\widehat{\lambda}$. This then gives a density estimate

$$\widehat{f}(x) \propto \exp\left(\sum_{j=1}^{p} \widehat{\lambda}_j h_j(x)\right)\mathbf{1}_{[0,1]}(x).$$

2.4 Other domains

If our density is supported on a bounded interval $[a, b]$, then we can scale the bins and use the same procedure. For a density with unbounded support we can add two infinite bins of the form $(-\infty, a]$ and $[b, \infty)$ for some $a << 0$ and $b >> 0$.

2.5 Multivariate

Suppose we have a joint density $h_{X,Y}$. We can use a grid to discretize (X, Y) and then fit an analogous log-linear model to estimate the joint density. A natural model to fit one of the form

$$\log(\mu_{ij}) = \lambda + \sum_{k=1}^{p} \lambda_k^X h_k^X(\Delta i) + \sum_{k'=1}^{p'} \lambda_{k'}^Y h_{k'}^Y(\Delta j) + \sum_{k=1}^{p}\sum_{k'=1}^{p'} \lambda_{ij}^{XY} h_k^X(\Delta i) h_{k'}^Y(\Delta j).$$

This ensures that the conditional densities of X and Y are in the same "class" and the densities fit in the previous section.

2.6 Other options

This idea is very general. There are models where the bin width varies with i and where other points in the interval are chosen to evaluate the function. There are endless possibilities.

3 Modelling interactions

We have seen that if we have two binary variables X, Y, their interaction can be described by a single parameter. Likewise, if we have two continuous variables, and we assume they are jointly Gaussian, then a single parameter describes their interaction. For categorical variables, the number of parameters increases. In general, we have,

- X and Y continuous: one interaction parameter.
- X takes I discrete values Y continuous: $I-1$ interaction parameters.
- X continuous and Y takes J discrete values: $J-1$ interaction parameters.
- X takes I discrete values and Y takes J discrete values: $(I-1)\times(J-1)$ interaction parameters.

Suppose we want to automatically select which pairs of variables should have an interaction. Unfortunately the LASSO wouldn't work in this situation. Suppose we have X discrete and Y continuous, then the LASSO could zero out some but not all of that $I-1$ interaction parameters. This isn't what we want. We want a penalty that can zero out a whole vector (one discrete, one continuous) or even a whole matrix (two discrete). It turns out the group LASSO penalty is the way to go!

3.1 Group LASSO

The interaction parameters fall into groups. For each unordered (X,Y) pair we have a group of parameters. This group of parameters is a single number when X and Y are both continuous, a vector when only one is continuous and a matrix when both are discrete. For a group g, let β_g be the parameters in the group g. Define

$$\|\beta_g\|_2 = \begin{cases} |\beta_g| & \text{if } \beta_g \text{ is a number,} \\ \|\beta_g\|_2 & \text{if } \beta_g \text{ is a vector,} \\ \|\beta_g\|_{\text{Frob}} & \text{if } \beta_g \text{ is a matrix.} \end{cases}$$

The notation $\|M\|_{\text{Frob}}$ is the *Frobenious* norm of the matrix M and if defined by

$$\|M\|_{\text{Frob}}^2 = \sum_{i,j} M_{i,j}^2.$$

The *group LASSO penalty* is defined to be

$$\mathcal{P}(\beta) = \lambda \sum_g \omega_g \|\beta_g\|_2 ,$$

where the sum is over all groups and $\lambda, \omega_g \geq 0$ are hyperparameters.

3.2 Zeroing out coefficients

Why does the group LASSO zero out groups coefficients? If we are using the group LASSO to select interactions, then our objective function would be

$$\text{minimize: } \log(L^{\text{pseudo}}(\beta)) + \mathcal{P}(\beta),$$

where $L^{\text{pseudo}}(\beta)$ is the pseudo-likelihood. Consider the simpler problem

$$\widehat{\beta} = \underset{\beta}{\text{argmin}} \frac{1}{2} \|Z-\beta\|_2^2 + \lambda \|\beta\|_2 .$$

The stationary conditions for $\widehat{\beta}$ are

$$Z - \widehat{\beta} = \lambda \widehat{u},$$

where $\widehat{u} \in \partial(\|\cdot\|_2)(\widehat{\beta})$. We have seen before that

$$\partial(\|\cdot\|_2)(\widehat{\beta}) = \begin{cases} \left\{ \frac{\widehat{\beta}}{\|\widehat{\beta}\|_2} \right\} & \text{if } \widehat{\beta} \neq 0, \\ \\ \{\widehat{u} : \|\widehat{u}\|_2 \leq 1\} & \text{if } \widehat{\beta} = 0. \end{cases}$$

Thus, if $\|Z\|_2 \leq \lambda$, then we can take $\widehat{\beta} = 0$ and $\widehat{u} = \frac{Z}{\lambda}$. If $\|Z\|_2 > \lambda$, then we have $Z = \left(1 + \frac{\lambda}{\|\widehat{\beta}\|_2}\right)\widehat{\beta}$. And thus,

$$\|Z\|_2 = \|\widehat{\beta}\|_2 + \lambda.$$

Which gives $\|\widehat{\beta}\|_2 = \|Z\|_2 - \lambda$ and hence

$$\widehat{\beta}_2 = \frac{\|\widehat{\beta}\|_2}{\|\widehat{\beta}\|_2 + \lambda} Z = \frac{\|Z\|_2 - \lambda}{\|Z\|_2} Z.$$

Thus, we have

$$\begin{aligned}
\widehat{\beta} &= \operatorname*{argmin}_{\beta} \frac{1}{2} \|Z - \beta\|_2^2 + \lambda \|\beta\|_2 \\
&= \begin{cases} 0 & \text{if } \|Z\|_2 \leq \lambda, \\ \frac{\|Z\|_2 - \lambda}{\|Z\|_2} Z & \|Z\|_2 > \lambda. \end{cases} \\
&= \max\left\{\frac{\|Z\|_2 - \lambda}{\|Z\|_2}, 0\right\} Z.
\end{aligned}$$

Thus, for sufficiently large λ, $\widehat{\beta} = 0$. And for small λ, $\widehat{\beta}$ points in the same direction as Z (so no coefficients are zeroed out). This shows why we expect the group LASSO to zero out entire groups of coefficients but otherwise leave the coefficients as non-zero. The map,

$$Z \mapsto \max\left\{\frac{\|Z\|_2 - \lambda}{\|Z\|_2}, 0\right\} Z,$$

is the proximal map for the group LASSO. Thus, like the regular LASSO, the group LASSO can be fit using proximal gradient descent.

4 Matched pairs

Consider a square contingency table where the row names and the column names agree. Some data are naturally described by such table. For example,

1. Voter affiliation recorded at two different times.
2. Presence/absence of a disease in two twins.
3. Student grades on two different pieces of assessment.

Such a table satisfies *marginal homogeneity* if $\pi_{i+} = \pi_{+i}$ for all $1 \leq i \leq I$. We will consider different ways of testing marginal homogeneity and fitting models that satisfy marginal homogeneity.

4.1 Binary tables

If we have 2×2 table, then the assumption of marginal homogeneity is determined by the parameter $\delta = \pi_{+1} - \pi_{1+}$ with marginal homogeneity holding if and only if $\delta = 0$. We can estimate δ with $\widehat{\delta} = \widehat{\pi}_{+1} - \widehat{\pi}_{1+}$. Under the null $\delta = 0$, the statistic

$$\widehat{SE}(\widehat{\delta}) = \sqrt{\frac{N_{12} + N_{21}}{N_{++}^2}},$$

is an estimate for the standard error of $\widehat{\delta}$. This allows us to test the null hypothesis and construct confidence intervals for δ. The test based on this estimate of the standard error is called *McNemar's test*.

4.2 Cochran–Mantel–Haenszel test

Suppose we have $n = N_{++}$ observations in our 2×2 contingency table. We could create a $2 \times 2 \times n$ table where in the i^{th} 2×2 table we record the response of individual i. Let X be the row variable, Y be the column variable and let Z be the third variable that records the index of the observation. Marginal homogeneity is equivalent to $X \perp\!\!\!\perp Y | Z$. That is, conditional on the individual observation, X and Y are independent. The Cochran–Mantel–Haenszel test is similar to Fisher's exact test. Consider the i^{th} 2×2 table and suppose we condition on the marginals N_{1+i} and N_{+1i}. Given these marginals, the values in the table are determined by N_{11i} and, under H_0,

$$N_{11+} \sim \mathsf{Hypergeometric}.$$

We can define $\mu_i = \mathbb{E}_{H_0}[N_{11+} | N_{1+i}, N_{+1i}]$ and $\sigma_i^2 = \mathrm{Var}_{H_0}(N_{11+} | N_{1+i}, N_{+1i})$. The *Cochran–Mantel–Haenszel test statistic* is

$$z = \frac{\sum_{i=1}^n N_{11+} - \mu_i}{\sqrt{\sum_{i=1}^n \sigma_i^2}}.$$

Under H_0, $z \overset{n \to \infty}{\sim} \mathcal{N}(0, 1)$. The Cochran–Mantel–Haenszel test generalizes to $I \times I$ tables. Conditioned on the marginals, the entries of the table have a multivariate hypergeometric distribution under H_0.