Bayesian graphical models summary

Luca Bracone

October 28, 2022

1 Introduction

We are given a set of points $y_1, \ldots, y_n$ and imagine they are realisations of a random distribution $p(y)$. We would like to know how $p(y)$ looks like. A commonly used method is the following: suppose that there is a set of possible random distribution functions $P(y)$ in which $p(y)$ exists. Then we assume that there exists a set Θ and a function $\Theta \to P(y)$ which we call a *parametrisation*. Usually, Θ is a simpler set to study than $P(y)$. Then, statistics is concerned with using the observed $y_1, \ldots, y_n$ to draw a $\theta \in \Theta$ such that $\theta \mapsto p(y)$ is as "close" as possible to the "true" $p(y)$.

Unlike in the usual statistical context, we view the parameter θ as being itself random with some distribution $p(\theta)$. We define the joint probability of y and θ as being a function $p(y, \theta)$ such that

$$p(\theta) = \int p(y, \theta) dy$$
$$\text{and } p(y) = \int p(y, \theta) d\theta$$

Then we define the conditional distribution "of y given θ" as

$$p(y|\theta) = \frac{p(y, \theta)}{p(\theta)}.$$

Using this property twice we obtain Bayes' theorem which allows us to "invert" and obtain the distribution of θ given the observed y. Since we only observe a finite number of y this is only an approximate distribution, we will talk about uncertainty later on.

$$p(\theta|y) = \frac{p(y, \theta)}{p(y)} = \frac{p(\theta)p(y|\theta)}{p(y)}.$$

In practice, the factor $p(y)$ is nothing to worry about since we can obtain it by integrating away θ in the following way, $p(y) = \int p(y|\theta)p(\theta)d\theta$.

2 Bayesian hierarhcical models

Suppose we observe values $y_{ij} \sim P(\theta_j)$ where $j = 1, \ldots, J$ and $i = 1, \ldots, n_j$. So, in other words we have a certain number of observations, each belonging to some group, and the values within each group are i.i.d with parameter θ_j. We will discuss here two methods to analyze such data that will motivate the use of hierarchical models.

It might at first seem appealing to ignore the differences between groups, this means setting all the θ_j to be equal to eachother and now we can perform usual maximum likelihood estimation. Using this method may cause a problem if there is a group with only a few outlying observations. If this happens, the outlying group will have an estimate that is far from its observations.

Then another idea might be to treat each group as a separate estimation, as though they have nothing to do with eachother. In some cases, this is justified, but it poses another problem for our outlying group with only a few observations. In this case, since that group only has a few observations we will get estimates with a large variance.

Finally, the idea of hierarhcical models is to see the first two methods as two extremes: "the θ_j are the same" and "the θ_j are the same". Instead we will imagine that the θ_j are themselves distributed $\theta_j \overset{\text{i.i.d.}}{\sim} Q(\phi)$. For some fixed parameter ϕ.

3 Graphical models

We observe a set of points $y_1, \ldots, y_n$. In particular the covariance for these points $\text{cov}(y_i, y_j)$ for $i, j = 1, \ldots, n$ is unknown and not necessarily zero. We would like to obtain estimates for $p(y_i | y_1, \ldots, \cancel{y_i}, \ldots, y_n)$.

3.1 Spike and slab prior for graphical models

Let $x \in \mathbb{R}^p$ be random vector distributed under the following hierarchical model

$$\begin{aligned} x|\Omega &\sim \text{N}_p(0, \Omega^{-1}) \\ \Omega|\delta &\sim \text{SNS}(\delta, v_0^2, v_1^2, \lambda) \\ \delta|\pi &\sim \text{BER}(\pi) \\ \pi &\sim \text{BETA}(a, b) \end{aligned}$$

where $\text{N}_p(0, \Omega^{-1})$ is the multivariate normal distribution with mean 0 and covariance matrix Ω^{-1}, BER is the product of Bernouilli-distributed variables with mean π for j, k, BETA is the usual beta distribution and finally SNS is the so-called spike-and-slab distribution which has the following density

$$p(\Omega|\delta) = C^{-1} \prod_{j<k} \text{N}(\omega_{jk}|0, v_{\delta_{jk}}^2) \prod_j \text{EXP}\left(\omega_{jj}|\frac{\lambda}{2}\right).$$

With C being some proportionality constant that depends on $\delta, v_0, v_1, \lambda$. Finally, let $X \in \mathbb{R}^{n \times p}$ be a matrix whose rows are i.i.d. observations of x.

Given this, we would like to obtain values for Ω, δ, π that maximise the log posterior joint distribution $\log(p(\Omega, \delta, \pi|X))$. After a few manipulations, the posterior joint distribution can be decomposed as

$$p(\Omega, \delta, \pi|X) = p(\Omega|\delta)p(\delta|\pi)p(X|\Omega)p(\pi)p(X)^{-1}$$

and in practice the factor $p(X)^{-1}$ is left out since for us it is constant, and so it has no influence on the maximisation. Using the definitions and the decomposition above we find that $\log(p(\Omega, \delta, \pi|X))$ equals

$$\begin{aligned}\text{constants} + \sum_{j<k} -\log(v_0^2(1-\delta_{jk}) + v_1^2\delta_{jk}) - \frac{\omega_{jk}^2}{2}\frac{1}{v_0^2(1-\delta_{jk}) + v_1^2\delta_{jk}} - \sum_j \frac{\lambda}{2}\omega_{jj} \\ + \sum_{j<k} \delta_{jk}\log(\pi) + (1-\delta_{jk})\log(1-\pi) \\ + (a-1)\log(\pi) + (b-1)\log(1-\pi) + \frac{n}{2}\log\det(\Omega) - \frac{1}{2}\operatorname{tr}(X^tX\Omega). \quad (1)\end{aligned}$$

The derivative with respect to π of eq. (1) is

$$\sum_{j<k} \frac{\delta_{jk}}{\pi} - \frac{1-\delta_{jk}}{1-\pi} + \frac{a-1}{\pi} + \frac{1-b}{1-\pi}$$

which is equal to zero when

$$\hat{\pi} = \frac{a - 1 + \sum_{j<k}\delta_{jk}}{a + b - 2 + \frac{p(p-1)}{2}}.$$

The derivative of eq. (1) with respect to δ_{jk} for some specific values of j, k is

$$-\frac{v_1^2 - v_0^2}{v_0^2(1-\delta_{jk}) + v_1^2\delta_{jk}} - \frac{\omega_{jk}^2}{2}\frac{v_1^2 - v_0^2}{(v_0^2(1-\delta_{jk}) + v_1^2\delta_{jk})^2} + \log\left(\frac{\pi}{1-\pi}\right)$$

And it is equal to zero when

$$\hat{\delta}_{jk} = \frac{\omega_{jk}^2(v_1^2 - v_0^2) - 2\left[(v_1^2 - v_0^2)\log\left(\frac{\pi}{1-\pi}\right)\right]}{2(v_1^2 - v_0^2)\left[1 + (v_1^2 - v_0^2)\log\left(\frac{\pi}{1-\pi}\right)\right]}.$$

Finally the derivative of eq. (1) with respect to ω_{jk} for some specific values of j, k is

$$-\frac{\omega_{jk}}{v_0^2(1-\delta_{jk}) + v_1^2\delta_{jk}} + \frac{\operatorname{cof}(\Omega)_{jk}}{2\det(\Omega)} - \frac{1}{2}\tilde{x}_j^t\tilde{x}_k$$

where $cof(\Omega)_{jk}$ is the jk-cofactor of Ω, $\tilde{x}_j$ and $\tilde{x}_k$ are the j-th and k-th column of X respectively.

As we can see, optimisation using these equations is intractable so we will use an EM approach.

3.2 Exepectation Maximisation for graphical Bayesian models

Following the approach in [ZT17] instead of maximising the expression in eq. (1) we decide to iteratively maximise its expectation over δ. So now, eq. (1) becomes

$$Q(\Omega, \pi | \Omega^{(l)}, \pi^{(l)}) = \mathbb{E}_{\delta | \Omega^{(l)}, \pi^{(l)}} \left(\log[p(\Omega, \delta, \pi | X)] \middle| \Omega^{(l)}, \pi^{(l)} \right)$$

References

[ZT17] Li Zehang Richard and McCormick Tyler H. "An Expectation Conditional Maximization approach for Gaussian graphical models". In: (Sept. 2017). DOI: 10.48550/arXiv.1709.06970. eprint: 1709.06970v3. URL: https://arxiv.org/abs/1709.06970v3.