Statistical analysis on genomic data

William van Rooij - EPFL

16th April 2019

Contents

1 Introduction

$$y_{n\times q} = x_{n\times p}\ \beta_{p\times q} + \epsilon_{n\times q},\ \epsilon_t \sim \mathcal{N}(0, \tau_t^{-1} I_n)$$

2 Variational inference

When computing the posterior density of parameters $\boldsymbol{\theta}$ according to observed data $\boldsymbol{y}$, variational inference simplifies the computation by approximating the posterior density $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ with a simpler density $q(\boldsymbol{\theta})$. It gives an approximation of the posterior distribution as a result of an optimization problem that minimizes a measure of "closeness" as objective function.

We suppose we have observations y and parameters $\boldsymbol{\theta}$, we are looking to determine the posterior distribution of the parameters conditional on the observations $p(\boldsymbol{\theta} \mid \boldsymbol{y})$. Given a family of densities $\mathcal{D}$ over the parameters, we want to find the distribution $q \in \mathcal{D}$ that minimizes the "closeness" measure compared to $p(\boldsymbol{\theta} \mid \boldsymbol{y})$.

Variational inference minimizes the Kullback–Leibler divergence as a "closeness" measure. Introduced in 1951 by Kullback and Leibler[1], it is the most common divergence measure used in statistics and machine learning. It is described as such:

$$\mathrm{KL}(q \parallel p) := \int q(\boldsymbol{\theta}) \log \left(\frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta} \mid \boldsymbol{y})} \right) \mathrm{d}\boldsymbol{\theta}.$$

It is described as a "directed divergence" as it is asymmetric, *i.e.* $\mathrm{KL}(p \parallel q) \neq \mathrm{KL}(q \parallel p)$.

Determining the family $\mathcal{D}$ can be difficult as we need the family to be simple enough to be optimized efficiently, but flexible enough for the approximation $q \in \mathcal{D}$ to be close to $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ w.r.t the Kullbach–Leibler divergence. The approximation will then be:

$$q^*(\boldsymbol{\theta}) = \underset{q(\boldsymbol{\theta}) \in \mathcal{D}}{\arg\min}\, \mathrm{KL}\left(q(\boldsymbol{\theta}) \parallel p(\boldsymbol{\theta} \mid \boldsymbol{y})\right).$$

Minimizing the Kullbach–Leibler divergence can be complicated depending on the density p that we want to approximate and the densities family $\mathcal{D}$ we want q to be part of. We can decompose the KL divergence as follows:

$$\begin{aligned} \mathrm{KL}\left(q(\boldsymbol{\theta}) || p(\boldsymbol{\theta} \mid \boldsymbol{y})\right) &= \mathbb{E}\left[\log q(\boldsymbol{\theta})\right] - \mathbb{E}\left[\log p(\boldsymbol{\theta} \mid \boldsymbol{y})\right] \\ &= \mathbb{E}\left[\log q(\boldsymbol{\theta})\right] - \mathbb{E}\left[\log p(\boldsymbol{y}, \boldsymbol{\theta})\right] + \log p(\boldsymbol{y}). \end{aligned}$$

We introduce the evidence lower bound (ELBO):

$$\begin{aligned} \mathcal{L}(q) &= \mathbb{E}\left[\log p(\boldsymbol{\theta}, \boldsymbol{y})\right] - \mathbb{E}\left[\log q(\boldsymbol{\theta})\right] \\ &= \int q(\boldsymbol{\theta}) \log \frac{p(\boldsymbol{y}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})} \mathrm{d}\boldsymbol{\theta}. \end{aligned}$$

When decomposing the KL divergence, we obtain:

$$\mathrm{KL}(q \parallel p) = \log(p) - \mathcal{L}(q).$$

This means that the KL divergence is the difference between the marginal log-likelihood with no effect on the optimization and a function : $\mathcal{L}(q)$. So minimizing the Kullbach–Leibler divergence is the same as maximizing $\mathcal{L}(q)$. The difference lays in the complexity of the problems, minimizing the Kullbach–Leibler divergence is not tractable, but maximizing $\mathcal{L}(q)$ admits a closed form when the family of densities $\mathcal{D}$ is well chosen. In such a case, we prefer to use $\mathcal{L}(q)$ as an objective function.

Using Jensen's inequality, we can show that $\mathcal{L}(q)$ is a lower bound for the marginal log-likelihood, which is why we call it the evidence lower bound, or variational lower bound.

$$\begin{aligned}
\log p(\boldsymbol{y}) &= \log \int p(\boldsymbol{y}, \boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}, \\
&= \log \int \frac{p(\boldsymbol{y}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})} q(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}, \\
&\geq \int q(\boldsymbol{\theta}) \log \left\{ \frac{p(\boldsymbol{y}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})} \right\} \mathrm{d}\boldsymbol{\theta}, \\
&= \mathcal{L}(q).
\end{aligned}$$

Hence, $\log p(\boldsymbol{y}) \geq \mathcal{L}(q)$, justifying the name "lower bound" for $\mathcal{L}(q)$.

2.1 Mean-field approximation

The complexity of the optimization problem is directly bound to the complexity of the family of densities $\mathcal{D}$ we want $q(\boldsymbol{\theta})$ to be apart of. We introduce the mean-field variational family, where the parameters are mutually independent and are governed by a distinct factor in the variational density.

We introduce $\{\theta_j\}_{j=1}^{J}$ a partition of $\boldsymbol{\theta}$, if $q \in \mathcal{D}$ and $\mathcal{D}$ a mean-field variational family, then:

$$q(\boldsymbol{\theta}) = \prod_{j=1}^{J} q_j(\theta_j)$$

We determine the variational factors $q_j(\theta_j)$ by maximizing $\mathcal{L}(q_j)$. Hence, the variational family does not directly represent the observed data, they are both linked through the optimization of the ELBO.

To visualise the mean-field approximation, we consider a two dimensional Gaussian distribution, represented in clear in Figure 1. The mean-field approximation of the posterior distribution is represented by the barred circle. We can see that the mean of the approximation is the same as the real mean, but the covariance doesn't match the covariance of the real posterior.

We have transformed, using the ELBO and the mean-field approximation our problem into a optimization problem. We now need a way to solve this problem. In the following section, we will be looking at the coordinate ascend mean-field variational inference.

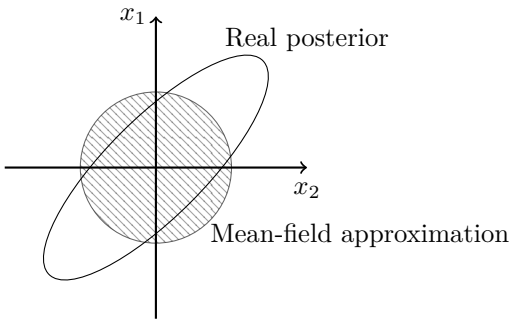

Figure 1: Visualisation of mean-field approximation to a two-dimensional Gaussian posterior. The correlations in the mean-field approximation do not represent the correlations of the real posterior.

2.2 Coordinate ascent algorithm

The coordinate ascent mean-field variational inference (CAVI) is one of the most common used to solve this kind of optimization problem. The algorithm iterates on the parameters of the mean-field approximation, optimizing them one at the time. It yields a local optimum for the ELBO.

The CAVI algorithm is based on the following result:

Lemma 2.1 *If we fix $q_l(\theta_l)$, $l \neq j$, then the optimal $q_j^*(\theta_j)$ verifies:*

$$q_j^*(\theta_j) \propto \exp\left\{\mathbb{E}_{-j}\left[\log p(\theta_j \mid \boldsymbol{\theta}_{-j}, \boldsymbol{y})\right]\right\}.$$

Based on this result, it updates one parameter θ_j at the time while the others are fixed. The CAVI algorithm stops when the ELBO varies less than a determined threshold ε.

At every iteration, $\mathcal{L}(q)$ is guaranteed to increase. The CAVI yields a local optimum depending on the initialization of the $q_j(\theta_j)$, $j = 1, \ldots, J$. Having different initializations could yield different optimums. These different optimums correspond to different sets of parameters $\boldsymbol{\theta}_k$, $k = 1, \ldots, K$

3 Multimodality

Bayesian model averaging is a solution to the inference problem with multiple competing models. It consists of weighting the different models in a weighted average with the probability that the data corresponds to such a model. If multiple models

Algorithm 1: Coordinate ascent variational inference (CAVI)

input : $p(\boldsymbol{y}, \boldsymbol{\theta})$, dataset y tolerance ε

output : $q(\boldsymbol{\theta}) = \prod_{j=1}^{J} q_j(\theta_j)$

initialize: $q_j(\theta_j)$

repeat

  for $j \in \{1, \ldots, J\}$ **do**

    set $q_j(\theta_j) \propto \exp\{\mathbb{E}_{-j}[\log p(\theta_j \mid \boldsymbol{\theta}_{-j}, \boldsymbol{y})]\}$

  $\mathcal{L}^{\mathsf{old}}(q) \leftarrow \mathcal{L}(q)$;

  $\mathcal{L}(q) \leftarrow \mathbb{E}[\log p(\boldsymbol{\theta}, \boldsymbol{y})] - \mathbb{E}[\log q(\boldsymbol{\theta})]$

until $|\mathcal{L}^{old}(q) - \mathcal{L}(q)| < \varepsilon$;

return $q(\boldsymbol{\theta})$

are strongly probable of corresponding to the data, it can be seen in the result of the Bayesian model averaging. The more the model corresponds to the observed data, the more it will stand out in the result.

Assume the data $\boldsymbol{y}$ could correspond to multiple models M_k, $k = 1, \ldots, K$, and Δ is the quantity of interest. We have the posterior distribution:

$$p(\Delta \mid \boldsymbol{y}) = \sum_{k=1}^{K} p(\Delta \mid M_k, \boldsymbol{y})\, p(M_k \mid \boldsymbol{y}). \tag{3.1}$$

This corresponds to a weighted average of the posterior distribution under each of the considered models with weights corresponding to the posterior models probabilities. The posterior probability for model M_k is given by:

$$p(M_k \mid \boldsymbol{y}) = \frac{p(\boldsymbol{y} \mid M_k)\, p(M_k)}{\sum_{j=1}^{K} p(\boldsymbol{y} \mid M_j)\, p(M_j)}, \tag{3.2}$$

where $p(\boldsymbol{y} \mid M_k)$ is the likelihood of model M_k, and $p(M_k)$ is the prior probabilities of the model M_k. It can, for example, depend on the complexity of the model, to favour the simpler models, or, if we consider the models to be equiprobable, it would be equal to $p(M_k) = 1/K$, $k = 1, \ldots, K$.

We know that the ELBO and the KL divergence are related and that minimizing the KL divergence is equivalent to maximizing the ELBO, and that they verify:

$$\mathrm{KL}(q \parallel p) = \log p(\boldsymbol{y}) - \mathcal{L}(q).$$

As we minimized the KL div, we can use $\mathcal{L}(q)$ as an approximation for $\log p(\boldsymbol{y})$ in Equation 3.2.

Now, instead of $p(\Delta \mid \boldsymbol{y})$ in Equation 3.1, we might be interested in approximating:

$$\mathbb{E}[\Delta \mid \boldsymbol{y}] = \sum_{k=1}^{K} \mathbb{E}[\Delta \mid M_k, \boldsymbol{y}]\ p(M_k \mid \boldsymbol{y}).$$

The same way we did in Equation 3.1, we calculate $p(M_K \mid \boldsymbol{y})$ with Equation 3.2.

4 Implementation

4.1 Model description

We denote $X = (X_1, \ldots, X_p)$ the SNPs and $y = (y_1, \ldots, y_q)$ the traits. The SNPs are strongly correlated. Our goal is to estimate the association between SNP s and trait t. To do so, we consider y as the response matrix and X as the candidate predictors, of the linear model where each response y_t is linearly related with the predictors X and has a residual precision $\tau_t \sim \text{Gamma}(\eta_t, \kappa_t)$, *i.e.*:

$$\boldsymbol{y}_{n\times q} = \boldsymbol{x}_{n\times p}\, \boldsymbol{\beta}_{p\times q} + \boldsymbol{\epsilon}_{n\times q},\ \boldsymbol{\epsilon}_t \sim \mathcal{N}(0, \tau_t^{-1} I_n)$$

We introduce $\boldsymbol{\gamma}_{p\times q}$, a binary matrix

References

[1] S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, (22):79–86, 1951.