1 Bayesian Inference

When using Bayesian statistics, one of the often used tools is integrals. However, it usually is a problem as they are not amenable, except in some particular cases, and need to be approximated. Markov Chain Monte Carlo (MCMC) algorithms are the most used and are fairly quick and accurate when working on reasonably small datasets. When the dataset dimensions grow, however, the MCMC algorithms become really time-consuming up to not being computable.

When performing MCMC inference, likelihoods and sometimes gradients need to be calculated at each iterations, the cost of these calculations increases with the number of parameters. Moreover, the more dimensions our problem has, the less exact our approximations become, which leads to more iterations to keep the precision needed. For the algorithm to end, all the parameters need to have converged, which means all parameters need to be checked and stored, which is close to impossible when their number is really high.

Another way to perform Bayesian inference is using deterministic methods. It consists in turning the inference problem into an optimization problem. Variational inference belongs to these kind of methods.

1.1 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 $q(\boldsymbol{\theta})$. One way to do so is the variational inference, which 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[**?**], it is the most common divergence measure used in statistics and machine learning. It is described as such:

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

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 tricky 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 KL 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 KL divergence can be complicated depending on the density p we want to approximate and the densities family $\mathcal{D}$ we want q to be apart 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 KL divergence is the same as maximizing $\mathcal{L}(q)$. The difference lays in the complexity of the problems, minimizing the KL 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 $\mathcal{L}(q)$.

1.2 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

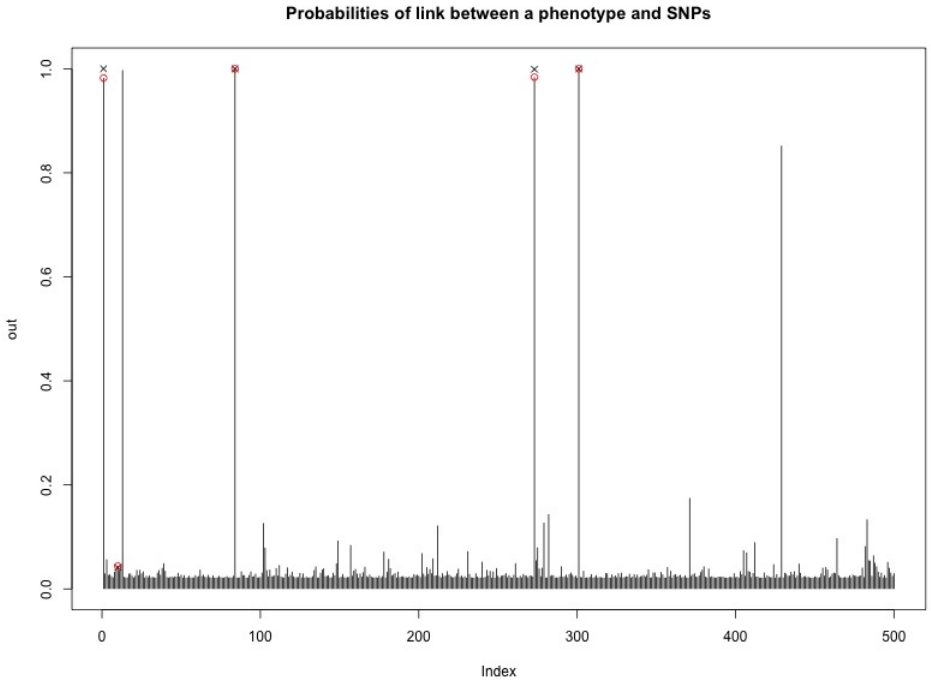
by a distinct factor in the variational density.

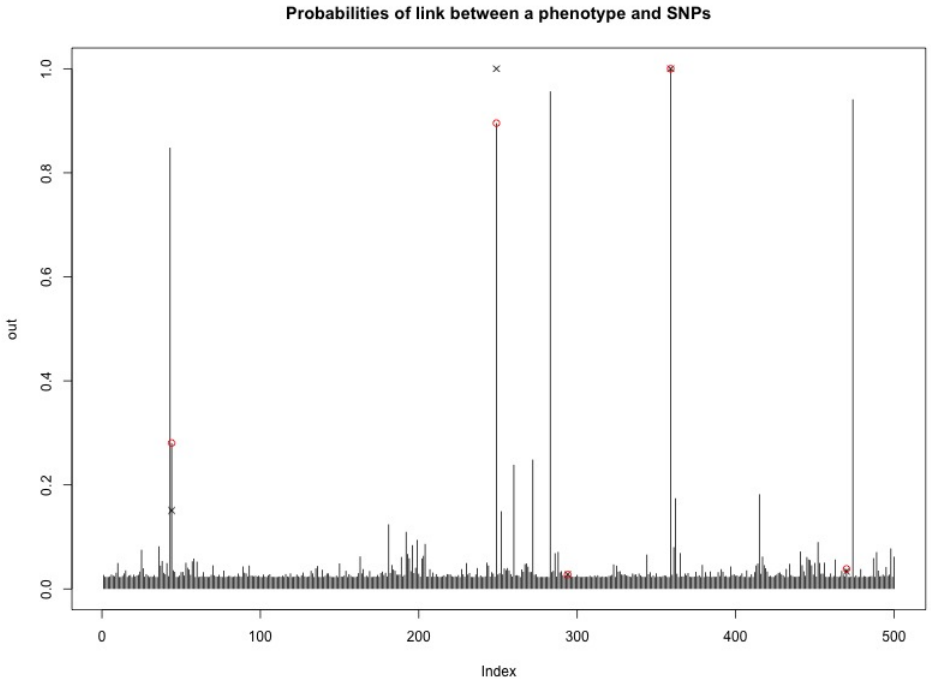
Let's call $\{\theta_j\}_{j=1}^{J}$ a partition of $\boldsymbol{\theta}$, if $q \in \mathcal{D}$ and $\mathcal{D}$ a mean-field variational family, then:

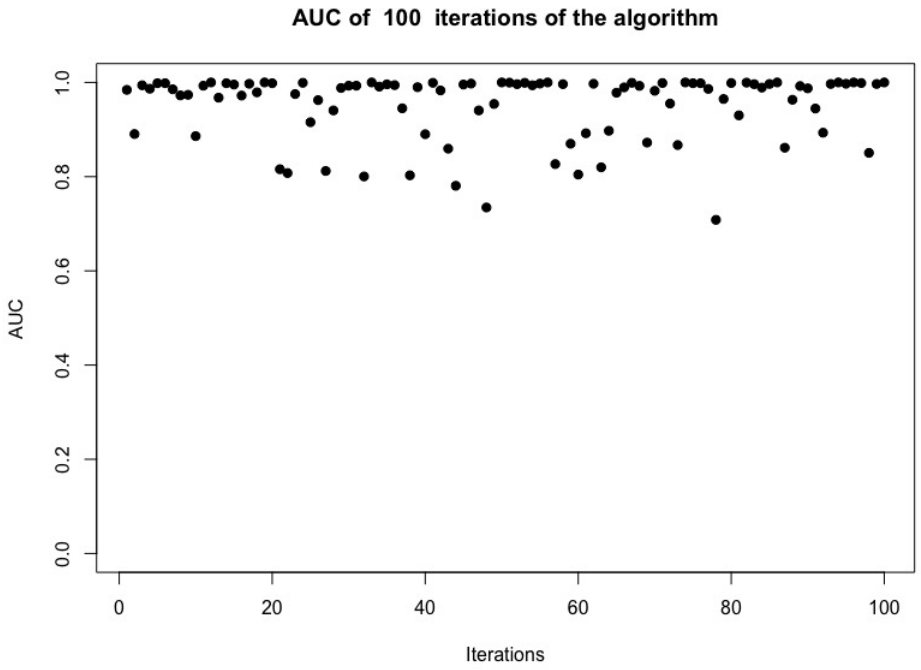
$$q(\boldsymbol{\theta}) = \prod_{j=1}^{J} q_j(\theta_j)$$

We determine the variational factors $q_j(z_j)$ 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.

2 Results


Probabilities of link between a phenotype and SNPs
out
0.0
0.2
0.4
0.6
0.8
1.0
0
100
200
300
400
500
Index


Probabilities of link between a phenotype and SNPs
out
0.0
0.2
0.4
0.6
0.8
1.0
0
100
200
300
400
500
Index


AUC of 100 iterations of the algorithm
AUC
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
Iterations

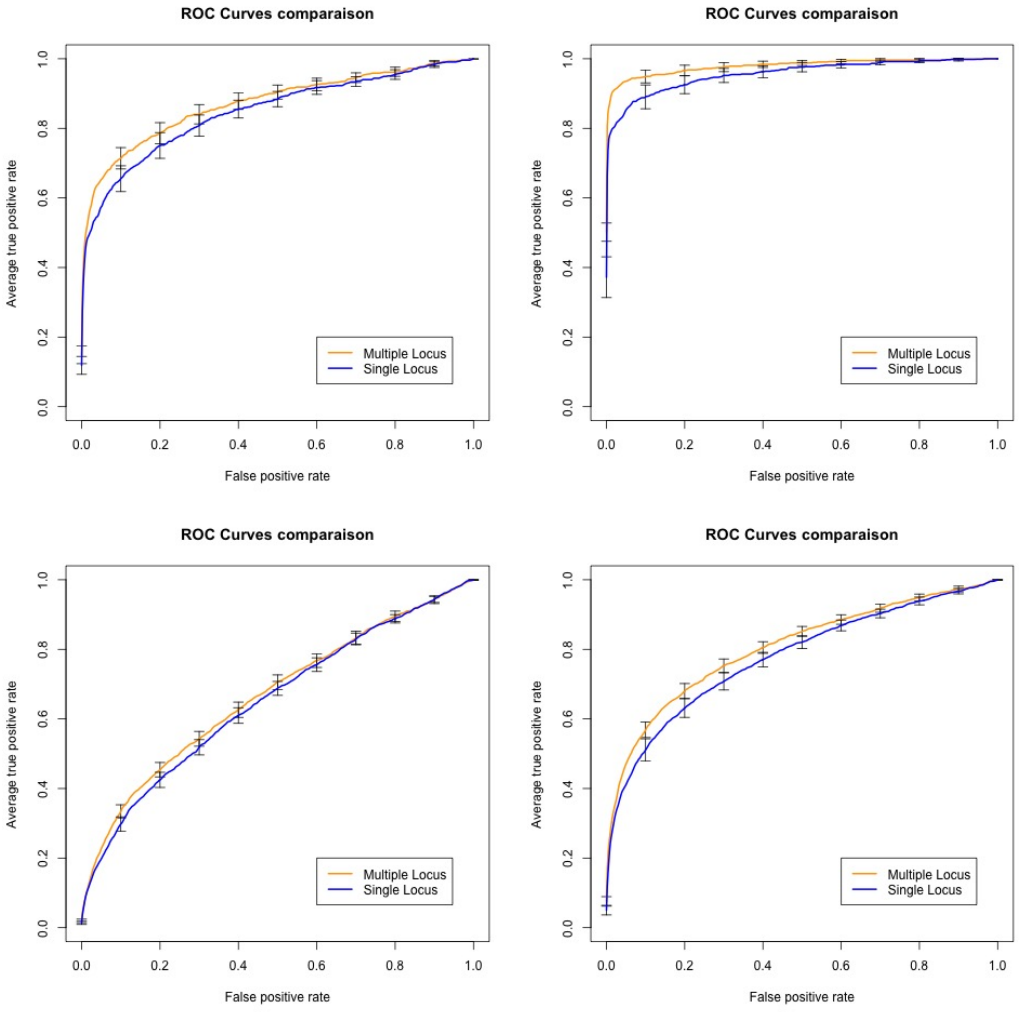


Figure 1: Top: $p_0 = 15$, Left: Max tot. PVE= 0.5, Bottom: $p_0 = 50$, Right: Max tot. PVE= 0.8