1 Summary to make sure I understand what happens

We want to approximate $p(z|x)$ where z are the parameters, x the observed values.

We want to to find $q(z)$ that minimizes $KL(q||p)$, hence maximizes $ELBO(q)$.

We consider $q = \prod q_j(z_j))$ (mean-field approx.) (parameters independent)

Depending on the starting point of the CAVI algorithm to calculate the $q_j(z_j)$, we may arrive at a optimum local and not global. So we vary the starting points in hope to reach the global. Then we perform a "sort of" BMA on the K models we obtain.

We want to estimate :

$$p(M_k|x) = \frac{p(x|M_k)p(M_k)}{\sum_j p(x|M_j)p(M_j)}$$

So we have to find the $ELBO$ for each of the models M_k which will give us $p(x|M_k)$. We can consider the M_k to be equiprobable so all $p(M_k)$ are the same.

2 Situation

We are looking to estimate the relationship between Single Nucleotide Polymorphisms (SNPs) and phenotypes. For one phenotype t, we have :

$$y_{n\times 1} = x_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1}, \epsilon \sim N\left(0, \sigma_\epsilon^2 I_n\right) \tag{1}$$

where β represents the relation between SNP s and phenotype t. $\beta_{st} = 0$ if there is no link between SNP s and phenotype t.

The fitting problem is complicated because it is a *small n - large p* situation, *i.e.* $p \gg n$. Which means that when one wants to fit the parameters to the data, the estimation is not suitable. We first need to diminish the number of parameters to estimate in the model by making assumptions.

We define γ_{st} as a indicator for β_{st} :

$$\gamma_{st} = \begin{cases} 1 & \text{if } \beta_{st} \neq 0, \\ 0 & \text{if } \beta_{st} = 0 \end{cases} \tag{2}$$

So there we can say that there is a relation between SNP s and phenotype t if and only if $\gamma_{st} = 1$.
We have some properties on γ_{st} and β_{st} :

$$\gamma_{st} = \arg \max_{\gamma\in\{0,1\}^p} p(M_\gamma|y) \tag{3}$$

where M_γ is the model including/excluding each p candidates according to γ. Now, to calculate this optimum, one must go through 2^p models. In our situation (*small n - large p*), the computation cost gets really high and it is preferable to reduce the parameters dimensions before trying to find the optimum.

We now define ω_s such that :

$$\gamma_{st}|\omega_s \sim \text{Bern}(\omega_s) \tag{4}$$

where $\omega_s \sim \text{Beta}(a_s, b_s)$, with a_s, b_s to be defined. We can now see that there was $n * p$ $\beta_s t$

We also have :

$$\beta_{st}|\gamma_{st} \sim \gamma_{st} g_\beta + (1 - \gamma_{st})\delta_0 \tag{5}$$

where g_β is an absolute continue density on $\mathbb{R}$ and δ_0 is the Dirac distribution.

This is called a *spike-and-slab* prior, the "spike" δ_0 represents the null effects at zero and the "slab" g_β, the non-null effects.

Now, let's denote z all our parameters that are unknown and that we want to estimate, and x the observed data. We want to determine the density function $p(z|x)$. Depending on the number of parameters, the density function $p(z|x)$

can be really hard to determine. That is why our goal is to diminish the number of parameters to estimate without losing the accuracy of our predictions.

3 Variational Inference

When computing the posterior density of parameters according to observed data, we may want to simplify the computation by approximating the posterior density by a simpler density that does not involve the observed data. 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 an measure of "closeness".

We suppose we have observations x and parameters z, we are looking to approximate the posterior conditional distribution $p(z|x)$. Given a family of densities $\mathcal{D}$ over the parameters, we want to find the distribution $q \in \mathcal{D}$ that is the closest possible to our target distribution $p(z|x)$.

The most prominent divergence measure used in statistics is the Kullback-Leibler (KL) divergence :

$$KL(p||q) := \int q(\theta) \log \left(\frac{q(\theta)}{p(\theta|y)} \right) d\theta \tag{6}$$

It has been introduced by Kulback and Leibler, who described it as a "directed divergence" as it is asymmetric, *i.e.* $KL(p||q) \neq KL(q||p)$.

We try to optimize the family of densities over latent variables, parametrized by variational parameters. Finding the best suitable family is finding the best settings of parameters closest to the desired distribution w.r.t. KL. We are looking for $\mathcal{D}$ flexible enough for the approximation $q \in \mathcal{D}$ to be close $p(z|x)$ w.r.t. the KL divergence but simple enough for efficient optimization.

$$q^*(z) = \arg \min_{q(z) \in \mathcal{D}} KL(q(z)||p(z|x)) \tag{7}$$

Now, minimizing the KL divergence can be complicated depending on the density we want to approximate and the densities family $\mathcal{D}$ we want q to be apart of. To ease the calculations, we will use the evidence lower bound.

3.1 Evidence Lower Bound (ELBO)

Assume $\mathcal{D}$ a density family, $q(z) \in \mathcal{D}$ a candidate approximation for $p(z|x)$. We have :

$$KL(q(z)||p(z|x)) = \mathbb{E}\left[\log q(z)\right] - \mathbb{E}\left[\log p(z|x)\right] \tag{8}$$

$$= \mathbb{E}\left[\log q(z)\right] - \mathbb{E}\left[\log p(z,x)\right] + \log p(x) \tag{9}$$

We call the evidence lower bound (ELBO) :

$$ELBO(q) = \mathbb{E}\left[\log p(z,x)\right] - \mathbb{E}\left[\log q(z)\right] \tag{10}$$

which means that :

$$KL(q||p) = \log p - ELBO(q) \tag{11}$$

Now, we can see that when optimizing on q, the term $\log p$ has no influence on the optimum. Hence, minimizing the KL divergence w.r.t. q is equivalent to maximizing the ELBO w.r.t. q.

We have :

$$\log p(x) = \underbrace{KL(q||p)}_{\geq 0} + ELBO(q) \Rightarrow \log p(x) \geq ELBO(q) \tag{12}$$

4 Mean-Field Approximation

When approximating the density of the parameters $q(z)$, keeping in perspective the goal to diminish the complexity of the problem. One can assume that the parameters are independent and governed by a distinct factor in variable density. The goal is to simplify the complexity of the calculations and diminish the time for computation. This is called the mean-field approximation.

$$q(z) = \prod_{j=1}^{m} q_j(z_j) \tag{13}$$

The mean-field approximation does not compute the correlations between two parameters and the marginal variances of approximations under represents those of the targets. If we approximate p with q, the mean-field approximation penalizes more placing mass in q where p has less mass and penalizes less the inverse.

4.1 Coordinate Ascent Mean-Field Variable Inference (CAVI)

The complete conditional of z_j is $p(z_j|z_{-j}, x)$. If we fix $q(z_l)$, $\forall l \neq j$ we have :

$$q_j^*(z_j) \propto \exp\left[\mathbb{E}_{-j}\left[\log p(z_j|z_{-j}, x)\right]\right] \tag{14}$$

$$\propto \exp\left[\mathbb{E}_{-j}\left[\log p(z_j, z_{-j}, x)\right]\right] \tag{15}$$

as we supposed the parameters are independent.

IN : $p(x, z)$, data set x, tolerance tol
OUT : $q(z) = \prod q_j(z_j)$
INIT : $ELBO(q) \leftarrow -\infty$,
REPEAT :
FOR : $j \in \{1, \ldots, m\}$
SET : $q_j(z_j) \propto \exp\left[\mathbb{E}_{-j}\left[\log p(z_j|z_{-j}, x)\right]\right]$
COMPUTE : $ELBO^{old}(q) \leftarrow ELBO(q)$
$ELBO(q) = \mathbb{E}\left[\log p(z, x)\right] - \mathbb{E}\left[\log q(z)\right]$
UNTIL : $|ELBO(q) - ELBO^{old}(q)| < tol$
RETURN : $q(z)$

This algorithm yields a local optimum, not necessarily a global optimum. However, we suppose that a global optimum exists and we can reach it through the previous algorithm from a particular initialisation. We will use Bayesian Model Averaging to try to find the global optimum.

5 Bayesian Model Averaging (BAM)

Bayesian Model Averaging is a solution to inference problem with multiple competing models. Suppose data D could correspond to different models M_k, $k = 1, \ldots, K$, and Δ is a quantity of interest. We have the posterior distribution :

$$p(\Delta|D) = \sum_{k=1}^{K} p(\Delta|M_k) \cdot p(M_k|D) \tag{16}$$

where :

$$p(M_k|D) = \frac{p(D|M_k) \cdot p(M_k)}{\sum_{i=1}^{K} p(D|M_i) \cdot p(M_i)} \tag{17}$$

and :

$$p(D|M_k) = \int \underbrace{(D|\theta_k, M_k)}_{\text{likelihood}} \cdot p(\theta_k|M_k) d\theta_k \tag{18}$$

This is basically a weighted average of distributions with the weights representing the probability of each model based on the observed data. Averaging over all models provides a better average predictive ability than using any single model M_j conditional on $\mathcal{M} = \{M_i : i = 1, \ldots, k\}$.

One of the tricky parts of this solution is to find the right models to include in our average. Suppose we have a list of models $\mathcal{M} = \{M_i : i = 1, \ldots, k\}$. If k is large, BAM could have a computational cost large as well. We could do a thinning of the models considered for the averaging. Madigan and Raftery proposed two steps to thin the models without losing accuracy.
The first step consists of taking in count only the models that are fairly close to the best one, *i.e.* only the models belonging to the following $\mathcal{A}'$:

$$\mathcal{A}' = \left\{ M_k : \frac{\max_l \{\text{pr}(M_l|D)\}}{\text{pr}(M_k|D)} \leq C \right\} \tag{19}$$

where C is chosen.

The second step is leaving out the models that are more complicated than another model also in the model list but is less likely to represent the data, *i.e.* we don't consider the models in the following $\mathcal{B}$:

$$\mathcal{B} = \left\{ M_k : \exists M_l \in \mathcal{A}', M_l \subset M_k, \frac{\text{pr}(M_l|D)}{\text{pr}(M_k|D)} > 1 \right\} \tag{20}$$

The sum (16) becomes :

$$p(\Delta|D) = \sum_{M_k \in \mathcal{A}} p(\Delta|D) \cdot p(M_k|D),\ \mathcal{A} := \mathcal{A}' \setminus \mathcal{B} \tag{21}$$

This reduces considerably the number of models taken in consideration for the averaging. When a model is rejected, all of his sub-models are rejected too.

In our situation, we want to approximate $p(z|x)$ by $q(z)$, and using the mean-field approximation, we consider that $q(z) = \prod_{j=1}^{m} q_j(z_j)$. To do so, we use the *CAVI* algorithm and it will give us an optimum for the parameters. The optimum that we obtain is not necessarily a global optimum, as the objective function we are optimizing is not necessarily concave. For example, if we try to find the maxima of the objective function represented in Figure 1, depending on where we start, we might find a local maximum that is not global.

Varying the initialisation of the parameters, we can obtain different optima.

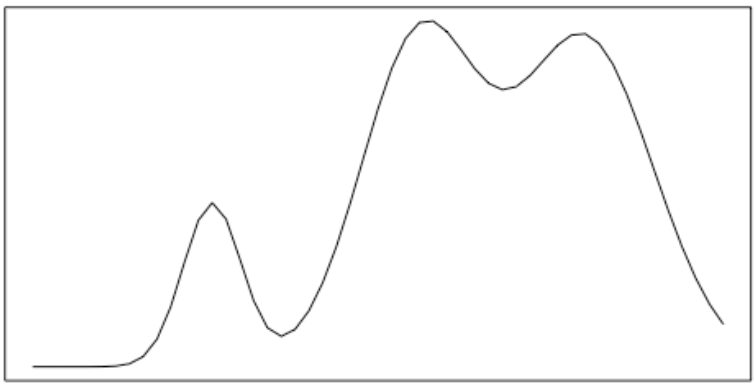

FIGURE 1 – Depending on the starting parameters for *CAVI* algorithm, it is possible to reach a local optimum that is not global. When using different starting points, the global optimum is reachable.

We assume that a global optimum exists and is reachable. We compute multiple optima using multiple starting parameters and we obtain different models, each more or less fitting the observed data, this gives us different sets of parameters that represent models $\mathcal{M} = \{M_k : k = 1, \ldots, K\}$.

We can then perform a variation of Bayesian Model Averaging on our model set $\mathcal{M}$ to find a distribution of the parameters that would give us a set of parameters corresponding to a model close to the one we are looking for.