1 Situation

We want 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 situation 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 :

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

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 $p(z|x)$.

2 Variational Inference

When working with Bayesian variable selection, one usually wants to compute the posterior marginal distributions, *i.e.* the inference. To approximate

some complicated to compute densities, one of the options is to use variational inference. It is a way to approximate by a more tractable distribution the conditional density of latent variables given observed variables.
We suppose we have n observations x and m parameters z, we are looking for the 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 minimizes the Kullback-Leiber divergence.

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

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 closer in KL to the desired distribution. 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.

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

2.1 Evidence Lower Bound

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

$$KL(q(z)||p(z|x)) = \mathbb{E}\left[\log q(z)\right] - \mathbb{E}\left[\log p(z|x)\right] \tag{8}$$
$$= \mathbb{E}\left[\log q(z)\right] - \mathbb{E}\left[\log p(z,x)\right] + \log p(x) \tag{9}$$

We define the evidence lower bound (ELBO) :

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

We have :

$$\log p(x) = \underbrace{KL(q||p)}_{\geq 0} + ELBO(q) \Rightarrow \log p(x) \geq ELBO(q) \tag{11}$$

Hence, minimizing the KL divergence is equivalent to maximizing the ELBO.

3 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.

$$q(z) = \prod_{j=1}^{m} q_j(z_j) \tag{12}$$

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.

3.1 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 :

$$q_j^*(z_j) \propto \exp\left[\mathbb{E}_{-j}\left[\log p(z_j|z_{-j}, x)\right]\right] \tag{13}$$

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

as we supposed the parameters are independent.

IN : $p(x, z)$, data set x
(OUT : $q(z) = \prod q_j(z_j)$
INIT : $q_j(z_j)$
WHILE : $ELBO$ not converging :
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 : $ELBO(q) = \mathbb{E}\left[\log p(z, x)\right] - \mathbb{E}\left[\log \text{ıı}(\digamma)\right]$
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. We will use Bayesian Model Averaging to try to find the global optimum.

4 Bayesian Model Averaging

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 :

$$p(\Delta|D) = \sum_{k=1}^{K} p(\Delta|M_k) \cdot p(M_k|D) \tag{15}$$

where :

$$p(M_k|D) = \frac{p(D|M_k) \cdot p(M_k)}{\sum_{i=1}^{K} p(D|M_i) \cdot p(M_i)} \tag{16}$$

and :

$$p(D|M_k) = \int \underbrace{(D|\theta_k, M_k)}_{\text{likelihood}} \cdot p(\theta_k|M_k) d\theta_k \tag{17}$$

In our case, we consider the solution given by the algorithm CAVI $q(z)$, which is a local optimum. To reach the global optimum, we will start the algorithm with different parameters. That way, we will obtain different optima and hopwefully, reach the global optimum. Then, we do a weighted average to find a plausible distribution.

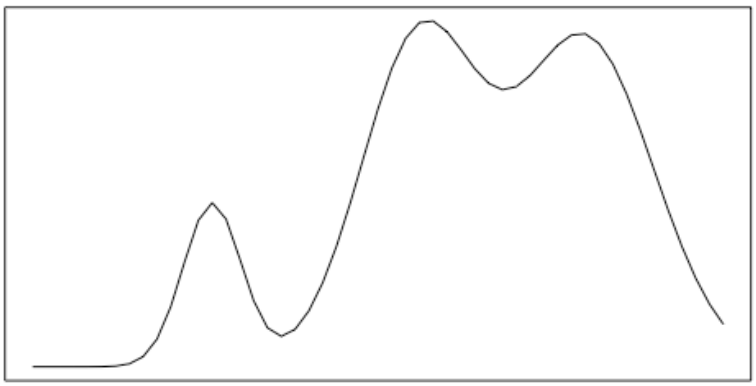

FIGURE 1 – Depending on the starting parameters for *CAVI* algorithm, it is possible to reach a local optimum that is not global. When using different starting points, the global optimum is reachable.

5 Mixing everything