Statistical framework from Flutre, Wen, Pritchard and Stephens (PLoS Genetics, 2013)

Timothée Flutre

July 22, 2014

Contents

This document describes the statistical framework with more details and sometimes a slightly different notation, notably inspired by Wen & Stephens (Annals of Applied Statistics, 2014) and Wen (Biometrics, 2014).

1 Likelihood of the whole data set

Let us imagine we measured the expression level of G genes in R tissues from N individuals which we also genotyped at P SNPs. Note that, based on Wen's paper in Biometrics, as the tissues are sampled in the same individuals, we use R for the total number of tissues instead of S (in our case $s = 1$).

Using g to index the genes, with $g \in \{1, \ldots, G\}$, we make two important assumptions about their expression levels. We assume that the set of expression levels $\boldsymbol{Y}_g$ of the g-th gene depends on the genotypes only at its m_g *cis* SNPs (with possibly other covariates). We also assume that all genes are independent conditionally on these genotypes (and other covariates), which is reasonable as we focus on *cis* and ignore *trans* SNPs in this model.

With $\boldsymbol{Y} = (Y_1, \ldots, Y_G)$ and $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_G)$ denoting the complete set of expression levels and predictors (genotypes and other covariates) for all genes, the observed log-likelihood of the whole data set can thus be written as

$$l(\Theta; \boldsymbol{Y}|\boldsymbol{X}) = \log p(\boldsymbol{Y}|\boldsymbol{X}, \Theta) = \sum_{g=1}^{G} \log p(\boldsymbol{Y}_g|\boldsymbol{X}_g, \Theta) \tag{1}$$

where Θ denotes the set of all parameters detailed below.

For the moment we make no assumption as whether or not a gene is expressed in all tissues, as this can be dealt with by preprocessing (if Normal likelihood) or by directly modeling read counts (Poisson likelihood and extensions).

Also, in practice, tissue-specific confounding factors in $\boldsymbol{Y}$ are regressed out beforehand (e.g. via PCA or factor analysis, such as in the PEER software), and the residuals, possibly quantile-normalized (if Normal likelihood), are used as responses.

2 Focus on a single gene-SNP pair

The likelihood for gene g and SNP p is:

$$Y_g|X_p, B_{gp}, X_c, B_{gc}, \Sigma_{gp} \sim \mathcal{MN}_{N\times R}(X_p B_{gp} + X_c B_{gc}, I_N, \Sigma_{gp}) \tag{2}$$

where:

- Y_g is the $N \times R$ matrix of expression levels;
- X_p is the $N \times 1$ matrix of genotypes (assuming the same individuals in all tissues);
- B_{gp} is the unknown $1 \times R$ matrix of genotype effect sizes;
- X_c is the $N \times (1 + Q)$ matrix of known covariates (including a column of 1's for the intercepts);
- B_{gc} is the unknown $(1 + Q) \times R$ matrix of covariate effect sizes (including the μ_s);
- $\mathcal{MN}_{N\times R}$ is the matrix Normal distribution;
- Σ_{gp} is the unknown $R \times R$ covariance matrix of the errors.

For mathematical convenience (especially in the case of multiple SNPs), we vectorize the rows of B_{gp} into $\boldsymbol{\beta}_{gp}$. Here, as we focus on one SNP at a time, we directly have $\boldsymbol{\beta}_{gp} = B_{gp}^T$.

To make inference invariant to linear transformations of the response variables (see Servin & Stephens, 2007), we preferentially model the standardized genotype effect sizes:

$$\forall r \; b_{gpr} = \beta_{gpr}/\sigma_{gpr}$$

where β_{gpr} is the r-th element in $\boldsymbol{\beta}_{gp}$ and σ_{gpr} is the r-th element on the diagonal of Σ_{gp}.

We use the notion of *configuration*, a latent indicator R-dimensional vector $\boldsymbol{\gamma}_{gp}$ such that $\gamma_{gpr} = 1$ means the eQTL is active in tissue r, i.e. $b_{gpr} \neq 0$, whereas $\gamma_{gpr} = 0$ means the eQTL is inactive in tissue r, i.e. $b_{gpr} = 0$. Moreover, we introduce an unknown mean $\bar{b}_{gp}$, to finally get the following "spike-and-slab" prior allowing to borrow information across tissues in which the eQTL is active:

$$b_{gpr}|\gamma_{gpr}, \bar{b}_{gp}, \phi \sim \gamma_{gpr}\mathcal{N}(\bar{b}_{gp}, \phi^2) + (1 - \gamma_{gpr})\delta_0 \tag{3}$$

where δ_0 is a point mass at 0, and

$$\bar{b}_{gp}|\omega \sim \mathcal{N}(0, \omega^2) \tag{4}$$

Whereas the configuration handles qualitative heterogeneity (having an effect or not), the hyperparameters ϕ and ω handles quantitative heterogeneity (having possibly different, non-null effects). By integrating out $\bar{b}_{gp}$, we can see that $\phi^2 + \omega^2$ controls the average magnitude of the effect in any tissue and $\phi^2/(\omega^2 + \phi^2)$ controls the amount of heterogeneity.

Equivalently, we can write this prior as a multivariate Normal:

$$\boldsymbol{b}_{gp}|U_{gp0} \sim \mathcal{N}_R(\mathbf{0}, U_{gp0}) \tag{5}$$

where, following Wen (2014), U_{gp0} is parametrized as $(\Gamma_{gp}, \Delta_{gp})$:

$$p(U_{gp0}) = p(\Delta_{gp}|\Gamma_{gp}) \Pr(\Gamma_{gp}) \tag{6}$$

so that Γ_{gp} is a binary matrix consisting of entry-wise non-zero indicators and is identical in size and layout to U_{gp0}, and Δ_{gp} is an indexed set of numerical values quantifying each non-zero entry in Γ_{gp}. The skeleton Γ_{gp} has $\boldsymbol{\gamma}_{gp}$ on the diagonal. Each off-diagonal entry $\Gamma_{gp,ij}$ is equal to 1 has long as diagonal elements $\Gamma_{gp,ii}$ and $\Gamma_{gp,jj}$ are both equal to 1.

In the current application, we choose:

$$U_{gp0} \,|\, \boldsymbol{\gamma}_{gp} = \mathbf{1}, \phi, \omega \;\; = \begin{pmatrix} \phi^2 + \omega^2 & \cdots & \omega^2 \\ \vdots & \ddots & \vdots \\ \omega^2 & \cdots & \phi^2 + \omega^2 \end{pmatrix} \tag{7}$$

In terms of notation, the 0 in U_{gp0} indicates the prior. See section on posterior effect sizes for U_{gp1}.

In practice, we use a known grid for prior variances, i.e. L pairs of values (ϕ_l, ω_l) leading to a mixture of multivariate Normals:

$$\boldsymbol{b}_{gp}|\boldsymbol{\lambda}, \boldsymbol{U}_{gp0} \sim \sum_{l=1}^{L} \lambda_l \, \mathcal{N}_R(\mathbf{0}, U_{gp0l}) \tag{8}$$

See Wen & Stephens (2014) and Wen (2014) for the priors on the other parameters, as well as the analytical approximation (Laplace method) to calculate the Bayes factor testing any active configuration against the global null.

3 Hierarchical model for all gene-SNP pairs

In order to detect eQTLs and identify in which tissue(s) they are active, we present in this document a generative model for the whole data set. This model is hierarchical with explicit gene and SNP levels. Moreover, it borrows information between tissues as well as between genes.

For gene g, we use a latent binary indicator z_g to denote if there is any eQTL in its *cis*-region for at least one tissue:

$$\Pr(z_g = 1) = 1 - \pi_0. \tag{9}$$

In Flutre *et al*, permutations were used to estimate this parameter, π_0, via Storey's method as implemented in the `qvalue` package in R. Note that the EBF and QBF procedures from Wen (arXiv 2013) can also be used.

We use a latent indicator m_g-dimensional vector $\boldsymbol{v}_g$ to denote the true eQTL (i.e. "the" eQTN) conditional on $z_g = 1$, and let v_{gp} denote the p-th entry of $\boldsymbol{v}_g$ ($p \in \{1, \ldots, m_g\}$). To be more specific, we attempt here to find eQTNs rather than eQTLs. In Flutre *et al*, the "one *cis* eQTL per gene" assumption restricts $\boldsymbol{v}_g$ to have at most one entry equal to 1, with the remaining entries being 0. See Wen (Biometrics, 2014) for the variable selection model relaxing this assumption (but only for fine mapping as it uses MCMC). Accordingly,

$$\Pr(\boldsymbol{v}_g = \boldsymbol{0} | z_g = 0) = 1,$$

and

$$\Pr(v_{gp} = 1 | z_g = 1) = \nu_p, \tag{10}$$

for which we also make the simplifying assumption that $\nu_p = \frac{1}{m_g}$, i.e. all *cis* SNPs are equally likely, *a priori*, to be the eQTL for gene g. But see Veyrieras *et al* to parametrize ν_p in terms of genomic annotations (although in the single-tissue case).

We use a latent indicator J-dimensional vector $\boldsymbol{c}_{gp}$ to denote the actual configuration, where $J = 2^R - 1$. In case the SNP is not an eQTL,

$$\Pr(\boldsymbol{c}_{gp} = \boldsymbol{0} | v_{gp} = 0) = 1. \tag{11}$$

Otherwise, we assume the gene-SNP pair to be in the j-th configuration with prior probability

$$\Pr(c_{gpj} = 1 | v_{gp} = 1) = \eta_j \tag{12}$$

with the constraints $\forall j \ \eta_j \geq 0$ and $\sum_j \eta_j = 1$. All column vectors $\boldsymbol{c}_{gp}$ for all m_g SNPs are gathered into the set C_g.

Finally, based on a grid of L pairs of values (ω_l^2, ϕ_l^2), we use a latent L-dimensional vector $\boldsymbol{d}_{gp}$ to indicate the actual pair of variances for the prior on the standardized genotype effect sizes. The l-th entry of this indicator vector is denoted by d_{gpl}, for which we assume prior probability

$$\Pr(\boldsymbol{d}_{gp} = \boldsymbol{0} | v_{gp} = 0) = 1 \tag{13}$$

and

$$\Pr(d_{gpl} = 1 | v_{gp} = 1) = \lambda_l \tag{14}$$

with the constraints $\forall l \ \lambda_l \geq 0$ and $\sum_l \lambda_l = 1$. All column vectors $\boldsymbol{d}_{gp}$ for all m_g SNPs are gathered into the set D_g.

4 Augmented likelihood

In the maximum likelihood framework, for any gene g, we treat latent variables $z_g, \boldsymbol{v}_g, C_g$ and D_g as missing data. Let $\boldsymbol{z} = (z_1, \ldots, z_G)$, $\boldsymbol{V} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_G)$, $\boldsymbol{C} = (C_1, \ldots, C_G)$ and $\boldsymbol{D} = (D_1, \ldots, D_G)$ denote the complete collection of latent variables.

Let $\Theta = (\pi_0, \eta_1, \ldots, \eta_K, \lambda_1, \ldots, \lambda_L)$ denote the set of parameters. Based on the hierarchical model described in the previous section, we can now write out the augmented log-likelihood as follows,

$$l_a(\Theta; \boldsymbol{Y}, \boldsymbol{z}, \boldsymbol{V}, \boldsymbol{C}, \boldsymbol{D} | \boldsymbol{X}) = \sum_g \log p(Y_g, z_g, \boldsymbol{v}_g, C_g, D_g | \boldsymbol{X}_g, \Theta)$$

Expanding the term inside the sum:

$$\begin{aligned}\log p(Y_g, z_g, \boldsymbol{v}_g, C_g, D_g|\boldsymbol{X}_g, \Theta) &= \log p(z_g|\Theta) + \log p(Y_g, \boldsymbol{v}_g, C_g, D_g|\boldsymbol{X}_g, \Theta, z_g) \\ &= (1-z_g)\log\pi_0 + z_g\log(1-\pi_0) \\ &+ (1-z_g)\log p(Y_g|\boldsymbol{X}_g, z_g=0) + z_g \log p(Y_g, \boldsymbol{v}_g, C_g, D_g|\boldsymbol{X}_g, \Theta, z_g=1) \\ &= (1-z_g)\log\pi_0 + z_g\log(1-\pi_0) + \log p(Y_g|\boldsymbol{X}_g, z_g=0) \\ &+ z_g \log\frac{p(Y_g, \boldsymbol{v}_g, C_g, D_g|\boldsymbol{X}_g, \Theta, z_g=1)}{p(Y_g|\boldsymbol{X}_g, z_g=0)}\end{aligned}$$

The ratio corresponds to the Bayes factor for gene g:

$$\begin{aligned}\log \mathrm{BF}_g &= \log\frac{p(Y_g, \boldsymbol{v}_g, C_g, D_g|\boldsymbol{X}_g, \Theta, z_g=1)}{p(Y_g|\boldsymbol{X}_g, z_g=0)} \\ &= \sum_p v_{gp}\log\nu_p + \sum_p v_{gp}\log\frac{p(Y_g, \boldsymbol{c}_{gp}, \boldsymbol{d}_{gp}|\boldsymbol{X}_{gp}, \Theta, v_{gp}=1)}{p(Y_g|\boldsymbol{X}_g, z_g=0)}\end{aligned}$$

The ratio inside the sum corresponds to the Bayes factor for gene g and SNP p:

$$\begin{aligned}\log \mathrm{BF}_{gp} &= \log\frac{p(Y_g, \boldsymbol{c}_{gp}, \boldsymbol{d}_{gp}|\boldsymbol{X}_{gp}, \Theta, v_{gp}=1)}{p(Y_g|\boldsymbol{X}_g, z_g=0)} \\ &= \sum_j c_{gpj}\log\eta_j + \sum_l d_{gpl}\log\lambda_l \\ &+ \sum_j\sum_l c_{gpj}d_{gpl}\log\frac{p(Y_g|\boldsymbol{X}_{gp}, \Theta, c_{gpj}=1, d_{gpl}=1)}{p(Y_g|\boldsymbol{X}_g, z_g=0)}\end{aligned}$$

The last ratio corresponds to the Bayes factor for gene g and SNP p in configuration j with prior variances l, and can be analytically approximated as described in Wen (2014):

$$\mathrm{BF}_{gpjl} = \frac{p(Y_g|\boldsymbol{X}_{gp}, \Theta, c_{gpj}=1, d_{gpl}=1)}{p(Y_g|\boldsymbol{X}_g, z_g=0)}$$

Putting everything back together:

$$\begin{aligned}l_a(\Theta; \boldsymbol{Y}, \boldsymbol{z}, \boldsymbol{V}, \boldsymbol{C}, \boldsymbol{D}|\boldsymbol{X}) &= \sum_g (1-z_g)\log\pi_0 + \sum_g z_g\log(1-\pi_0) + \sum_g \log p(Y_g|\boldsymbol{X}_g, z_g=0) \\ &+ \sum_{g,p} z_g v_{gp}\log\frac{1}{m_g} + \sum_{g,p,j} z_g v_{gp}c_{gpj}\log\eta_j + \sum_{g,p,l} z_g v_{gp}d_{gpl}\log\lambda_l \\ &+ \sum_{g,p,j,l} z_g v_{gp}c_{gpj}d_{gpl}\log\mathrm{BF}_{gpjl}\end{aligned} \tag{15}$$

5 EM algorithm

The EM algorithm searches for the maximum likelihood estimate of Θ, by iteratively performing an expectation (E) step and a maximization (M) step of the following objective function, noted Q:

$$Q(\Theta|\boldsymbol{Y}, \boldsymbol{X}, \Theta^{(i)}) = \mathbb{E}_{\boldsymbol{z},\boldsymbol{V},\boldsymbol{C},\boldsymbol{D}|\boldsymbol{Y},\boldsymbol{X},\Theta}[l_a(\Theta)|\boldsymbol{Y}, \boldsymbol{X}, \Theta^{(i)}] \tag{16}$$

Starting from randomly-initialized parameters $\Theta^{(0)}$, in the E-step for the $(i+1)^{\text{th}}$ iteration, we evaluate the objective function (16), i.e. the conditional expectation over the latent variables of the augmented log-likelihood (15) given the observed data $\boldsymbol{X}$ and $\boldsymbol{Y}$ and the current estimates of the parameters $\Theta^{(i)}$.

$$\begin{aligned}
\mathbb{E}[z_g|\boldsymbol{Y},\boldsymbol{X},\Theta^{(i)}] &= \Pr(z_g=1|\boldsymbol{Y},\boldsymbol{X},\Theta^{(i)}) \\
&= \frac{\Pr(z_g=1|\Theta^{(i)})\; p(\boldsymbol{Y}|\boldsymbol{X},\Theta^{(i)},z_g=1)}{p(\boldsymbol{Y}|\boldsymbol{X},\Theta^{(i)})} \\
&= \frac{\Pr(z_g=1|\Theta^{(i)})\; \prod_{g'} p(\boldsymbol{Y}_{g'}|\boldsymbol{X}_{g'},\Theta^{(i)},z_g=1)}{\prod_{g'} p(\boldsymbol{Y}_{g'}|\boldsymbol{X}_{g'},\Theta^{(i)})} \\
&= \frac{\Pr(z_g=1|\Theta^{(i)})\; p(Y_g|\boldsymbol{X}_g,\Theta^{(i)},z_g=1)}{p(Y_g|\boldsymbol{X}_g,\Theta^{(i)})} \\
&= \frac{(1-\pi_0^{(i)})\mathrm{BF}_g^{(i)}}{\pi_0^{(i)}+(1-\pi_0^{(i)})\mathrm{BF}_g^{(i)}}.
\end{aligned} \tag{17}$$

where

$$\begin{aligned}
\mathrm{BF}_g^{(i)} &= \frac{p(Y_g|\boldsymbol{X}_g,\Theta^{(i)},z_g=1)}{p(Y_g|\boldsymbol{X}_g,z_g=0)} \\
&= \sum_{p,j,l} \frac{1}{m_g}\eta_j^{(i)}\lambda_l^{(i)}\mathrm{BF}_{gpjl}.
\end{aligned}$$

Similarly,

$$\mathbb{E}[z_g v_{gp}|\boldsymbol{Y},\boldsymbol{X},\Theta^{(i)}] = \frac{(1-\pi_0^{(i)})\frac{1}{m_g}\mathrm{BF}_{gp}^{(i)}}{\pi_0^{(i)}+(1-\pi_0^{(i)})\mathrm{BF}_g^{(i)}}, \tag{18}$$

where

$$\begin{aligned}
\mathrm{BF}_{gp}^{(i)} &= \frac{p(Y_g|\boldsymbol{X}_{gp},\Theta^{(i)},z_g=1,v_{gp}=1)}{p(Y_g|\boldsymbol{X}_g,z_g=0)} \\
&= \sum_{j,l} \eta_j^{(i)}\lambda_l^{(i)}\mathrm{BF}_{gpjl}.
\end{aligned}$$

And

$$\mathbb{E}[z_g v_{gp} c_{gpj}|\boldsymbol{Y},\boldsymbol{X},\Theta^{(i)}] = \frac{(1-\pi_0^{(i)})\frac{1}{m_g}\eta_j^{(i)}\sum_l \lambda_l^{(i)}\mathrm{BF}_{gpjl}}{\pi_0^{(i)}+(1-\pi_0^{(i)})\mathrm{BF}_g^{(i)}}, \tag{19}$$

$$\mathbb{E}[z_g v_{gp} d_{gpl}|\boldsymbol{Y},\boldsymbol{X},\Theta^{(i)}] = \frac{(1-\pi_0^{(i)})\frac{1}{m_g}\lambda_l^{(i)}\sum_j \eta_j^{(i)}\mathrm{BF}_{gpjl}}{\pi_0^{(i)}+(1-\pi_0^{(i)})\mathrm{BF}_g^{(i)}}. \tag{20}$$

In the M-step for the $(i+1)^{\text{th}}$ iteration, we estimate a new set of parameters, $\Theta^{(i+1)}$, by maximizing the objective function (16), i.e. the conditional expectation over the latent variables of the augmented log-likelihood (15) given the observed data $\boldsymbol{X}$ and $\boldsymbol{Y}$ and the current estimates of the parameters $\Theta^{(i)}$.

In particular, for π_0,

$$\frac{\partial Q}{\partial \pi_0}(\pi_0) = \sum_g \left[(1-\mathbb{E}[z_g|\boldsymbol{Y},\boldsymbol{X},\Theta^{(i)}])\times\frac{1}{\pi_0}\right] - \sum_g \left[\mathbb{E}[z_g|\boldsymbol{Y},\boldsymbol{X},\Theta^{(i)}]\times\frac{1}{1-\pi_0}\right],$$

$$\frac{\partial Q}{\partial \pi_0}(\pi_0^{(i+1)}) = 0 \Leftrightarrow \pi_0^{(i+1)} = \frac{1}{G}\sum_g (1 - \mathbb{E}[z_g|\boldsymbol{Y}, \boldsymbol{X}, \Theta^{(i)}]),$$

which gives

$$\pi_0^{(i+1)} = \frac{1}{G}\sum_g \frac{\pi_0^{(i)}}{\pi_0^{(i)} + (1-\pi_0^{(i)})\mathrm{BF}_g^{(i)}}. \tag{21}$$

Now for the grid points, using a Lagrange multiplier, L_a, to enforce the constraints,

$$\frac{\partial Q}{\partial \lambda_l}(\lambda_l) = \sum_{g,p}\left[\mathbb{E}[z_g v_{gp} d_{gpl}|\boldsymbol{Y}, \boldsymbol{X}, \Theta^{(i)}] \times \frac{1}{\lambda_l}\right] - L_a,$$

$$\frac{\partial Q}{\partial \lambda_l}(\lambda_l^{(i+1)}) = 0 \Leftrightarrow \lambda_l^{(i+1)} = \frac{1}{L_a}\sum_{g,p} \mathbb{E}[z_g v_{gp} d_{gpl}|\boldsymbol{Y}, \boldsymbol{X}, \Theta^{(i)}],$$

which gives

$$\lambda_l^{(i+1)} = \frac{\sum_{g,p,j} \frac{\frac{1}{m_g}\eta_j^{(i)}\mathrm{BF}_{gpjl}}{\pi_0^{(i)}+(1-\pi_0^{(i)})\mathrm{BF}_g^{(i)}}\ \lambda_l^{(i)}}{\sum_{l'}\left(\sum_{g,p,j} \frac{\frac{1}{m_g}\eta_j^{(i)}\mathrm{BF}_{gpjl'}}{\pi_0^{(i)}+(1-\pi_0^{(i)})\mathrm{BF}_g^{(i)}}\ \lambda_{l'}^{(i)}\right)}. \tag{22}$$

where the $(1-\pi_0^{(i)})$ simplified but the $1/m_g$ are kept as the SNP prior could easily become SNP-specific via the usage of external information such as genome annotations.

Finally, for the configurations,

$$\eta_j^{(i+1)} = \frac{\sum_{g,p,l} \frac{\frac{1}{m_g}\lambda_l^{(i)}\mathrm{BF}_{gpjl}}{\pi_0^{(i)}+(1-\pi_0^{(i)})\mathrm{BF}_g^{(i)}}\ \eta_j^{(i)}}{\sum_{j'}\left(\sum_{g,p,l} \frac{\frac{1}{m_g}\lambda_l^{(i)}\mathrm{BF}_{gpjl}}{\pi_0^{(i)}+(1-\pi_0^{(i)})\mathrm{BF}_g^{(i)}}\ \eta_{j'}^{(i)}\right)}. \tag{23}$$

We initiate the EM algorithm by setting $\Theta^{(0)}$ to some initial values, random or not, and run iterations until some pre-defined convergence threshold is met. In practice, we monitor the monotonic increase of the observed log-likelihood (1) between successive iterations, as guaranteed by the EM algorithm, and stop as the increment becomes sufficiently small (e.g. 0.05).

We can also construct confidence intervals for $\hat{\Theta}$ using the profile likelihood. For example, a $(1-\alpha)\%$ profile likelihood confidence set for π_0 is built as

$$\{\pi_0 : \log p(\boldsymbol{Y}|\boldsymbol{X}, \pi_0, \hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\lambda}}) > \log p(\boldsymbol{Y}|\boldsymbol{X}, \hat{\pi}_0, \hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\lambda}}) - \frac{1}{2}Z^2_{(1-\alpha)}\}, \tag{24}$$

where $\hat{\pi}_0, \hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\lambda}}$ are the MLEs obtained from the EM algorithm.

6 Posteriors on latent variables

Once the hyperparameters have been estimated, several posteriors of interest can be calculated. The first is the posterior probability that the gene has one eQTL in at least one tissue:

$$\begin{aligned}\Pr(z_g = 1|\boldsymbol{Y}, \boldsymbol{X}, \Theta) &= \frac{\Pr(z_g = 1|\Theta)\; p(\boldsymbol{Y}|\boldsymbol{X}, \Theta, z_g = 1)}{p(\boldsymbol{Y}|\boldsymbol{X}, \Theta)} \\ &= \frac{(1-\hat{\pi}_0)\mathrm{BF}_g}{\hat{\pi}_0 + (1-\hat{\pi}_0)\mathrm{BF}_g}\end{aligned} \tag{25}$$

Note that the BF_g's alone can be used in order to estimate π_0 via the EBF procedure (Wen, arXiv 2013).

At the SNP level, the posterior probability that the SNP is "the" eQTL for the gene, in at least one tissue, given that the gene has exactly one eQTL, assuming all cis SNPs are equally likely, is given by:

$$\begin{aligned}\Pr(v_{gp} = 1|\boldsymbol{Y}, \boldsymbol{X}, \Theta, z_g = 1) &= \frac{\frac{1}{m_g} p(Y_g|X_g, \Theta, z_g = 1, v_{gp} = 1)}{p(Y_g|X_g, \Theta, z_g = 1)} \\ &= \frac{\frac{1}{m_g}\mathrm{BF}_{gp}}{\mathrm{BF}_g} = \frac{\mathrm{BF}_{gp}}{\sum_p \mathrm{BF}_{gp}}\end{aligned} \tag{26}$$

In order to make results more comparable with models only considering the SNP level, the posterior probability that the SNP is "an" eQTL for the gene, in at least one tissue, given that all SNPs are independent (and effects are small), can be written as:

$$\Pr(v_{gp} = 1|\boldsymbol{Y}, \boldsymbol{X}, \Theta, z'_g \geq 1) = \frac{\frac{1}{m_g}\mathrm{BF}_{gp}}{\frac{1}{m_g}\mathrm{BF}_{gp} + (1 - \frac{1}{m_g})} \tag{27}$$

We are also interested in the posterior on configurations:

$$\Pr(c_{gpj} = 1|\boldsymbol{Y}, \boldsymbol{X}, \Theta, z_g = 1, v_{gp} = 1) = \frac{\hat{\eta}_j \sum_l \hat{\lambda}_l \mathrm{BF}_{gpjl}}{\mathrm{BF}_{gp}} \tag{28}$$

But of most interest is the posterior probability that the SNP is an eQTL in a given tissue, given that it is "the" eQTL for the gene:

$$\begin{aligned}\Pr(\gamma_{gpr} = 1|\boldsymbol{Y}, \boldsymbol{X}, \Theta, z_g = 1, v_{gp} = 1) &= \sum_{j:\gamma_{gpr}=1|c_{gpj}} \Pr(c_{gpj} = 1|\boldsymbol{Y}, \boldsymbol{X}, \Theta, z_g = 1, v_{gp} = 1) \\ &= \sum_{j:\gamma_{gpr}=1|c_{gpj}} \frac{\hat{\eta}_j \sum_l \hat{\lambda}_l \mathrm{BF}_{gpjl}}{\sum_{j',l} \hat{\eta}_{j'} \hat{\lambda}_l \mathrm{BF}_{gpj'l}}\end{aligned} \tag{29}$$

When reporting the results, it is usually better to report the following marginal posterior:

$$\begin{aligned}\Pr(\gamma_{gpr} = 1|\boldsymbol{Y}, \boldsymbol{X}, \Theta) = \Pr(\gamma_{gpr} = 1|\boldsymbol{Y}, \boldsymbol{X}, \Theta, z_g = 1, v_{gp} = 1) \\ \times \Pr(z_g = 1|\boldsymbol{Y}, \boldsymbol{X}, \Theta)\end{aligned} \tag{30}$$

Note that the above expression doesn't use $\Pr(v_{gp} = 1|\boldsymbol{Y}, \boldsymbol{X}, \Theta, z_g = 1)$ in order to make the marginal posterior comparable with the models that assume all SNPs to be independent (e.g. Li *et al*, arXiv 2013).

7 Posteriors on genotype effect sizes

Also of interest are the posterior probabilities of the genotype effect sizes per gene-SNP pair in each tissue.

By maximum likelihood in each tissue separately, we can easily obtain the estimates of the standardized genotype effect sizes, $\hat{\boldsymbol{b}}_{gp}$, and their standard errors recorded on the diagonal of an $R \times R$ matrix noted $\hat{V}_{gp} = \mathbb{V}(\hat{\boldsymbol{b}}_{gp})$. Using each pair of tissues, we can also fill the off-diagonal elements of $\hat{V}_{gp}$.

If we now view $\hat{\boldsymbol{b}}_{gp}$ and $\hat{V}_{gp}$ as *observations* (i.e. known), we can "forget" about the original data X_p, X_c and Y_g, and write a new "likelihood" (using only the sufficient statistics):

$\hat{\boldsymbol{b}}_{gp}|\boldsymbol{b}_{gp} \sim \mathcal{N}_R(\boldsymbol{b}_{gp}, \hat{V}_{gp})$

Let us imagine first that the prior on $\boldsymbol{b}_{gp}$ is not a mixture but a single Normal: $\boldsymbol{b}_{gp} \sim \mathcal{N}_R(\boldsymbol{0}, U_{gp0})$. As this prior is conjuguate to the "likelihood" above, the posterior simply is (see Wikipedia):

$\boldsymbol{b}_{gp}|\hat{\boldsymbol{b}}_{gp} \sim \mathcal{N}_R(\boldsymbol{\mu}_{gp1}, U_{gp1})$

where:

- $\boldsymbol{\mu}_{gp1} = U_{gp1}(\hat{V}_{gp}^{-1}\hat{\boldsymbol{b}}_{gp})$;
- $U_{gp1} = (U_{gp0}^{-1} + \hat{V}_{gp}^{-1})^{-1}$.

In practice, we use a mixture for prior (see 8):

$$\boldsymbol{b}_{gp}|\boldsymbol{\lambda}, \boldsymbol{U}_{gp0j} \sim \sum_{l=1}^{L} \lambda_l \, \mathcal{N}_R(\boldsymbol{0}, U_{gp0jl}) \tag{31}$$

for which the hyper-parameters are either fixed ($\boldsymbol{U}_{gp0j}$) or estimated ($\boldsymbol{\lambda}$) as described above using the full hierarchical model and the EM algorithm. Moreover, the posteriors we may want are $\boldsymbol{b}_{gp}|\hat{\boldsymbol{b}}_{gp}, \hat{V}_{gp}, \hat{\Theta}$ (full marginal) or $\boldsymbol{b}_{gp}|\hat{\boldsymbol{b}}_{gp}, \hat{V}_{gp}, \hat{\Theta}_{-\pi_0}, v_{gp} = 1$ (conditional on being the eQTN in at least one tissue, averaging over all configurations and whole grid) or $\boldsymbol{b}_{gp}|\hat{\boldsymbol{b}}_{gp}, \hat{V}_{gp}, \hat{\boldsymbol{\lambda}}, \gamma_{gpr} = 1$ (conditional on being an active eQTN in tissue r, averaging over all configurations in which this tissue is active and whole grid) or $\boldsymbol{b}_{gp}|\hat{\boldsymbol{b}}_{gp}, \hat{V}_{gp}, \hat{\boldsymbol{\lambda}}, c_{gpj} = 1$ (conditional on being the eQTN and in configuration j, averaging over whole grid).

Let us start with the latter:

$$\begin{aligned} p(\boldsymbol{b}_{gp}|\hat{\boldsymbol{b}}_{gp}, \hat{V}_{gp}, \hat{\boldsymbol{\lambda}}, c_{gpj} = 1) &= \frac{p(\boldsymbol{b}_{gp}|\hat{\boldsymbol{\lambda}}, \boldsymbol{U}_{gp0j}) \; p(\hat{\boldsymbol{b}}_{gp}|\boldsymbol{b}_{gp}, \hat{V}_{gp})}{p(\hat{\boldsymbol{b}}_{gp}|\hat{V}_{gp}, \hat{\boldsymbol{\lambda}})} \\ &= \sum_{l=1}^{L} \hat{\lambda}_l \; p_l(\boldsymbol{b}_{gp}) \frac{p_l(\hat{\boldsymbol{b}}_{gp})}{p(\hat{\boldsymbol{b}}_{gp}|\hat{V}_{gp}, \hat{\boldsymbol{\lambda}})} \frac{p(\hat{\boldsymbol{b}}_{gp}|\boldsymbol{b}_{gp}, \hat{V}_{gp})}{p_l(\hat{\boldsymbol{b}}_{gp})} \\ &= \sum_{l=1}^{L} \tilde{\lambda}_l \; \tilde{p}_l(\boldsymbol{b}_{gp}|\hat{\boldsymbol{b}}_{gp}) \end{aligned} \tag{32}$$

where:

- $p_l(\hat{\boldsymbol{b}}_{gp})$ corresponds to the marginal distribution of $\hat{\boldsymbol{b}}_{gp}|d_{gpl} = 1$;
- $\tilde{\lambda}_l = \hat{\lambda}_l \frac{p_l(\hat{\boldsymbol{b}}_{gp})}{p(\hat{\boldsymbol{b}}_{gp})}$ is the probability that $\hat{\boldsymbol{b}}_{gp}$ belongs to component l;

- $\tilde{p}_l(\boldsymbol{b}_{gp}|\hat{\boldsymbol{b}}_{gp}) = \frac{p_l(\boldsymbol{b}_{gp})p(\hat{\boldsymbol{b}}_{gp}|\boldsymbol{b}_{gp})}{p_l(\hat{\boldsymbol{b}}_{gp})}$ is the posterior probability of $\boldsymbol{b}_{gp}$ given that it is from component l.

Using the law of total expectations (as done by Sarah), we get:

$$\mathbb{E}[\hat{\boldsymbol{b}}_{gp}|d_{gpl}=1] = \mathbb{E}[\mathbb{E}[\hat{\boldsymbol{b}}_{gp}|d_{gpl}=1,\boldsymbol{b}_{gp}]] = \mathbb{E}[\boldsymbol{b}_{gp}|d_{gpl}=1] = 0 \tag{33}$$

and using the law of total variances:

$$\mathbb{V}[\hat{\boldsymbol{b}}_{gp}|d_{gpl}=1] = \mathbb{V}[\mathbb{E}[\hat{\boldsymbol{b}}_{gp}|d_{gpl}=1,\boldsymbol{b}_{gp}]] + \mathbb{E}[\mathbb{V}[\hat{\boldsymbol{b}}_{gp}|d_{gpl}=1,\boldsymbol{b}_{gp}]] = U_{gp0l} + \hat{V}_{gp} \tag{34}$$

Therefore the marginal is $\hat{\boldsymbol{b}}_{gp}|d_{gpl}=1 \sim \mathcal{N}_R(\mathbf{0}, U_{gp0l} + \hat{V}_{gp})$.

7.1 Computing the Full Joint Posterior on $\boldsymbol{b}_{gp}$

We recognize that the marginal is $\hat{\boldsymbol{b}}_{gp}|d_{gpl}=1 \sim \mathcal{N}_R(\mathbf{0}, U_{gp0l} + \hat{V}_{gp})$. Therefore,

$$\begin{aligned}
p(\boldsymbol{b}_{gp}|\hat{\boldsymbol{b}}_{gp}, \hat{V}_{gp}, \hat{\boldsymbol{\lambda}}, c_{gpj}=1) &= \sum_{l=1}^{L} \tilde{\lambda}_l \frac{p_l(\boldsymbol{b}_{gp})p(\hat{\boldsymbol{b}}_{gp}|\boldsymbol{b}_{gp})}{p_l(\hat{\boldsymbol{b}}_{gp})} \\
&= \sum_{l=1}^{L} \hat{\lambda}_l \frac{p_l(\hat{\boldsymbol{b}}_{gp})}{p(\hat{\boldsymbol{b}}_{gp})} \frac{p_l(\boldsymbol{b}_{gp})p(\hat{\boldsymbol{b}}_{gp}|\boldsymbol{b}_{gp})}{p_l(\hat{\boldsymbol{b}}_{gp})} \\
&= \sum_{l=1}^{L} \hat{\lambda}_l \frac{p_l(\boldsymbol{b}_{gp})p(\hat{\boldsymbol{b}}_{gp}|\boldsymbol{b}_{gp})}{p(\hat{\boldsymbol{b}}_{gp})} \\
&= \sum_{l=1}^{L} \frac{\hat{\lambda}_l \mathcal{N}_R(\hat{\boldsymbol{b}}_{gp}|\mathbf{0}, U_{gp0l} + \hat{V}_{gp})\mathcal{N}_R(\boldsymbol{b}_{gp}|\boldsymbol{\mu}_{gp1jl}, U_{gp1jl})}{\sum_{l=1}^{L} \hat{\lambda}_l \mathcal{N}_R(\hat{\boldsymbol{b}}_{gp}|\mathbf{0}, U_{gp0l} + \hat{V}_{gp})}
\end{aligned} \tag{35}$$

where:

- $\boldsymbol{\mu}_{gp1jl} = U_{gp1jl}(\hat{V}_{gp}^{-1}\hat{\boldsymbol{b}}_{gp})$;
- $U_{gp1jl} = (\mathbf{I} + U_{gp0jl}\hat{V}^{-1})^{-1}U_{gp0jl}$.

Thus we require the summation over a variety of configuration weights conditional being the eQTN in configuration j