Statistical inference of eQTL sharing among a high number of tissues

Timothée Flutre, Sarah Urbut, Xiaoquan Wen, Matthew Stephens

September 19, 2014

This document describes the "type" model, an extension of the "config" model in details in the document "config_model.tex/pdf" available in the private repository paper-eQtlBma on GitHub.

1 Motivation

- Pending question from last time: Where does the MLE of the covariance in effects come from with summary stats of different tissues? (i.e., in our Config model document, the off-dagonal elements of $\hat{V}_{gp}$) See SS.ex in the posterior.effect.size simulation?
- **Answer (per William)** This is indeed an issue. In the meta-anlysis case, this is actually always the case because we have no information about the covariance in standard errors of effect sizes from the data. However, the use of a prior which assumes some cross-tissue covariance (i.e., ω is non-zero) captures some of the " true " underlying tissue covariance. From the AoAS paper below:

  Note that the standard error of the fixed effect, ζ, is really just the sum of the individual study-specific standard error of the effects and each tissue-specific prior covariance term. Obviously, if the data is acquired from multiple outside studies, we would have no information about the effect size covariance. But more generally, we know that the residual matrix $\hat{\mathbf{\Sigma}}$ is relatively sparse (as also induced in our tutorial), then the effect-size covariance matrix $\hat{\boldsymbol{V}}$(simply $\mathbf{X}'\mathbf{X} \times \hat{\mathbf{\Sigma}}$, will also be relatively sparse. *So is the only information we have about shared effects is in the prior which we induce.?* No – again, the information about shared effect comes from the grid weight estimation which, if the effects are shared among tissues, will tend to put more weight on larger values of ω. The sparse covariance matrix of standard errors simply suggests that the errors of the estimates are not correlated among tissues.

  Similarly, I ascertained that the reason we can separate this into the product of the $\widehat{\mathrm{BF}}^{\mathrm{ES}}_{\mathrm{fix}}$ and $\widehat{\mathrm{BF}}^{\mathrm{ES}}_{\mathrm{meta}}$ is because we assume the studies and the corresponding residuals are independent, which is effectively our situation if I assume the above correctly. Thus we are no longer interested in the multivariate distribution of effects across tissues, i.e., the vector β, but rather the product of the univariate effects, correct? I think this is the assumption we make by setting ω^2 to 0 in v1. *So are we making inference on a multivariate β in the BF or the product of many univariate (as in AoAS)*

$$\text{BF}^{\text{ES}}(\phi,\omega) = \lim \frac{\int J_{H_a} \prod_s P(\tau_s) d\tau_1 \cdots d\tau_S}{\int J_{H_0} \prod_s P(\tau_s) d\tau_1 \cdots d\tau_S} \tag{A.6}$$

$$\zeta^2 = \frac{1}{\sum_s (\delta_s^2 + \phi^2)^{-1}}, \tag{A.11}$$

Applying (A.14) results in

(A.22)

$$\text{BF}^{\text{ES}}(\phi,\omega) = \sqrt{\frac{\zeta^2}{\zeta^2+\omega^2}} \exp\left(\frac{\mathcal{T}_{\text{ES}}^2}{2}\frac{\omega^2}{\zeta^2+\omega^2}\right) \cdot \prod_s \left(\sqrt{\frac{\delta_s^2}{\delta_s^2+\phi^2}} \exp\left(\frac{T_s^2}{2}\frac{\phi^2}{\delta_s^2+\phi^2}\right)\right)$$

The reason that we are still interested in the joint inference even as a product of univariate if the n
Matthew wrote that:

- *By modeling the sharing of active eQTLs among tissues, this framework increases power to detect eQTLs that are present in more than one tissue compared with tissue-by-tissue analyses that examine each tissue separately" - the type model (without correlations) is still modelllng the *sharing* of eQTLs among tissues (i.e presence/absence of eQTLs among tissues). It just isn't modeling the similarity of the effects.*

  So then the advantage of the analysis is that if we consider the ABF for a model in which we have information on 5 tissues, and the sample size for one tissue is smaller, the evidence that a SNP is active in the smaller-sample size tissue will be augmented by the additional information from the more-easily ascertained tissue, analogous to the configuration model. The difference is that we ignore the sharing of effect, and so it is not necessary to pull the integrals back together (as in AoAS paper equation A.6, and see my 'Understanding the Config Model.pdf'). **This also assumes that not only the tissues but the effect sizes are independent among tissues, because we use $\hat{\beta}_r$ as a summary for Y and thus if the gene expression readings for each element of Y are i.i.d., so are the $\hat{\beta}_r$?**

- *N*ote that the first version of the "type" model (currently implemented in the eQtlBma package) drastically simplifies this sum by only considering the consistent configuration.

  So is this analogous to assuming that in:

$$\begin{aligned} \text{BF}_{gpkl}(\boldsymbol{q}_k, \omega_l, \phi_l) &= \frac{p(Y_g|\boldsymbol{X}_{gp}, t_{gpk}=1, d_{gpl}=1)}{p(Y_g|\boldsymbol{X}_{gp}, z_g=0)} \\ &= \frac{\int p(Y_g|\boldsymbol{X}_{gp}, \boldsymbol{\mu}, \boldsymbol{b}, \Sigma)\, p(\boldsymbol{\mu}, \Sigma)\, p(\boldsymbol{b}|\boldsymbol{q}_k, \omega_l, \phi_l)\, \mathrm{d}\boldsymbol{b}\, \mathrm{d}\boldsymbol{\mu}\, \mathrm{d}\Sigma}{\int p(Y_g|\boldsymbol{X}_{gp}, \boldsymbol{\mu}, \boldsymbol{b}=\boldsymbol{0}, \Sigma)\, p(\boldsymbol{\mu}, \Sigma)\, \mathrm{d}\boldsymbol{\mu}\, \mathrm{d}\Sigma} \end{aligned} \tag{1}$$

  both Σ (the covariance in residuals errors) and the the corresponding $\mathbf{X}'\mathbf{X} \times \hat{\boldsymbol{\Sigma}}$ (i.e., the true covariance in effects) are diagonal and non-singular (i.e., every diagonal entry complete) matrices,

allowing us to factor the BF and use study-specific error approximations? Note that we write:

In the scale-invariance formulation, we have $\boldsymbol{b} = diag(\Sigma)^{-\frac{1}{2}}\boldsymbol{\beta}$*, so that* $\boldsymbol{b}|\boldsymbol{q}_k \sim \mathcal{N}_R(\boldsymbol{0}, U_k)$ *induces the prior* $\boldsymbol{\beta}|\boldsymbol{q}_k \sim \mathcal{N}_R(\boldsymbol{0}, W_k)$ *where* $W_k = diag(\Sigma)^{-\frac{1}{2}} U_k diag(\Sigma)^{-\frac{1}{2}}$*.*

thus cementing my questions that if we assume the residuals between tissues are independent, so are the effect-sizes! From page 5: *note that we also have to assume that the errors are independent between tissues*): I notice that different from the AoAS paper, we integrate over one τ rather than a τ specific for each tissue - but where do we get the information in the likelihood to reinforce the covariance in effects (i.e., $\delta^2 = \frac{1}{\boldsymbol{g}^T\boldsymbol{g} - N\bar{g}^2}$)

Ahh no! The covariance in effects is reflected in similar β_s, which is captured by larger weights λ_l on the grid which puts a lot of weight on non-zero shared effects (i.e., big ω) and perhaps small ϕ, suggesting that most tissues are derived from some underlying mean effects with little variance. In fact, the tissue expression could be perfectly correlated while the matrix of residuals sum of squares (and the corresponding residual standard error) could be very sparse - these two things are not related. My problem has been in conflating correlated residual standard error with correlated gene expression - variance reflects the deviation from the mean present among the data, while the residual sum of squares merely reflects deviation from the fit. There could be a perfect fit (i.e., very low RSS) that captures huge variability in the data. Just because the 'errors are independent between tissues' and thus the standard errors of effect size are independent between tissues does not mean that the effects are uncorrected - rather, their relative error in estimation is uncorrelated between tissues

- Why consider tissues jointly if shared effect is zero? **We can still learn about the consistency of effects (i.e., the weight on the consistent config) and the sample size help.**
- *Most importantly, we assume that, given a type, the activity of an eQTL in tissue r is independent from its activity in other tissues*
- Yes: given a type, the Bayes factor for a vector (i.e., across all tissues) of expression is the product of the BF in each tissues. Because β_s is always drawn conditional on $\gamma_{gar} = 1$, once we have this product, we must integrate out γ_{gppr}, which crucially can be done in a univariate manner since we no longer have to consider the joint probability of a full configuration but rate rthe tissue-specific probability of being on or off (i.e., q_{kr}). This avoids also the problem of only considering the consistent configuration.

$$\begin{aligned}
\mathrm{BF}_{gpkl}(\boldsymbol{q}_k, \omega_l, \phi_l) &= \frac{\int p(\bar{b}_{gpl})\ \prod_r p(\boldsymbol{y}_{gr}|\boldsymbol{X}_{gp}, t_{gpk} = 1, d_{gpl} = 1, \bar{b}_{gpl})\ \mathrm{d}\bar{b}_{gpl}}{\prod_r p(\boldsymbol{y}_{gr}|\boldsymbol{X}_{gp}, z_g = 0)} \\
&= \frac{\int p(\bar{b}_{gpl})\ \prod_r [q_{kr} p(\boldsymbol{y}_{gr}|\boldsymbol{X}_{gp}, d_{gpl} = 1, \bar{b}_{gpl}, \gamma_{gpr} = 1) + (1 - q_{kr}) p(\boldsymbol{y}_{gr}|\boldsymbol{X}_{gp}, \gamma_{gpr} = 0)]\ \mathrm{d}\bar{b}_{gpl}}{\prod_r p(\boldsymbol{y}_{gr}|\boldsymbol{X}_{gp}, z_g = 0)} \\
&= \int_{-\infty}^{+\infty} f(\bar{b}_{gpl}, \boldsymbol{q}_k, \omega_l, \phi_l)\ \mathrm{d}\bar{b}_{gpl}
\end{aligned} \tag{2}$$

- However, does this still allow us to exploit the shared effect among tissues? For example, in the **Likeilhood of the Whole Data Set pdf**, equation (5) and (6) show that the prior covariance matrix is indexed by configurations.

$$\boldsymbol{b}_{gp}|U_0 \sim \mathcal{N}_R(\mathbf{0}, U_0) \tag{3}$$

where, following Wen (2014), U_0 is parametrized as $(\Gamma_{gp}, \Delta_{gp})$:

$$p(U_0) = p(\Delta_{gp}|\Gamma_{gp})\mathsf{P}(\Gamma_{gp}) \tag{4}$$

so that Γ_{gp} is a binary matrix consisting of entry-wise non-zero indicators and is identical in size and layout to U_0, 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.

Yes - the shared effect among tissues will still be captured in the M step of the EM algorithm which will have larger BF when the effect sizes are similar (i.e., ω is non-zero) and the corresponding grid weight.

- Doesn't saying that given a type, the activity of an eQTL in tissue is independent from its activity in other tissues directly negate the covariance in effects?

  No - the phrase refers to the fact that given a type, the binary indicator γ_r is independent of the other tissues, but the effects are still correlated conditional on γ_r (per William) with covariance matrix W. We integrate over all γ below.

- In the config model, given a configuration $\boldsymbol{\gamma}$, where we use an unknown mean $\bar{b}_{gp}$ to borrow information across tissues in which the eQTL is active.

$$b_{gpr}|\gamma_{gpr}, \bar{b}_{gp}, \phi \sim \gamma_{gpr}\mathcal{N}(\bar{b}_{gp}, \phi^2) + (1 - \gamma_{gpr})\delta_0 \tag{5}$$

By setting $\omega^2 = 0$, aren't we effectively curtailing our ability to "borrow information across tissues in which the eQTL is active"?

But here, according to Solution (1) of the document,

The target distribution $\boldsymbol{b}|\boldsymbol{q}_k$ can thus be approximated by the $\mathcal{N}_R(0, U_k)$ where:

$$U_k = \begin{pmatrix} q_{k1}(\phi_l^2 + q_{k1}\omega_l^2) & \cdots & q_{k1}q_{kR}\omega_l^2 \\ \vdots & \ddots & \vdots \\ \cdots & \cdots & q_{kR}(\phi_l^2 + q_{kR}\omega_l^2) \end{pmatrix}$$

2 Question

In the type-based model, we use a latent indicator K-dimensional vector $\boldsymbol{t}_{gp}$ to denote the actual type. In case the SNP is not an eQTL,

$$\mathsf{P}(\boldsymbol{t}_{gp} = \mathbf{0}|v_{gp} = 0) = 1. \tag{6}$$

Otherwise, we assume the gene-SNP pair belongs to the k-th type with prior probability

$$\mathsf{P}(t_{gpk} = 1|v_{gp} = 1) = \pi_k \tag{7}$$

with the constraints $\forall k\ \pi_k \geq 0$ and $\sum_k \pi_k = 1$. All column vectors $\boldsymbol{t}_{gp}$ for all m_g SNPs are gathered into a $K \times m_g$ matrix T_g.

In the type-based model, we also index all tissues in which the eQTL is active via a latent indicator R-dimensional vector $\boldsymbol{\gamma}_{gp}$, which hence corresponds to its configuration. However, compare to the "config"

model where $\boldsymbol{\gamma}_{gp}$ simply corresponds to the latent variable $\boldsymbol{c}_{gp}$ indexing the configuration, we now put a prior on $\boldsymbol{\gamma}_{gp}$. In case the SNP is not an eQTL,

$$\mathsf{P}(\boldsymbol{\gamma}_{gp} = \mathbf{0} | \boldsymbol{t}_{gp} = \mathbf{0}) = 1. \tag{8}$$

Otherwise, we assume the eQTL is active in the r-th tissue with prior probability depending on the type

$$\mathsf{P}(\gamma_{gpr} = 1 | t_{gpk} = 1) = q_{kr}, \tag{9}$$

for which we could also parametrize the q_{kr} in terms of tissue-specific annotations, e.g. DNase peaks, and therefore have q_{pkr}. Joining the column vectors $\boldsymbol{\gamma}_{gp}$ for all K types, we obtain a latent $R \times K$ matrix Γ_{gp}. All these matrices are gathered into $\boldsymbol{\Gamma}_g = (\Gamma_{g1}, \ldots, \Gamma_{gm_g})$. As a result, this bypasses the need for the J-dimensional vectors $\boldsymbol{c}_{gp}$ where $J = 2^R - 1$ and the corresponding prior probabilities (the η_j's).

- I understand that the $K \times m_g$ matrix T_g represents the identity of each SNP in the columns, so that the column sum is equal to 1 and the row sum is equal to the number of SNPs in the class.
- Furthermore, the $R \times K$ matrix Γ_{gp} represents the tissue-specific patterns of expression for each class across tissues, such that the column sums will represent the number of tissues a SNP of type 'k' might be active and the rows reprsent the total number of types from which a tissue derives activity.

- However, since the matrix Γ_{gp} is the same for all SNPs of a particular type, why is it necessary to stack the Γ_{gp} into $\boldsymbol{\Gamma}_g$?

  Answer The answer is that even if two SNPs are of the same type, suppose that the q vector for a particular type is [0.10 0.95 0.05 0.95 0.95]. One SNP could be [1 1 0 1 1] and one SNP could be [0 1 0 1 1] (although the second case is more likely)