Statistical inference of eQTL sharing among a high number of tissues

Timothée Flutre, Sarah Urbut, Xiaoquan Wen, Matthew Stephens

September 11, 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 offidagonal elements of $\hat{V}_{gp}$) See SS.ex in the posterior.effect.size simulation?
- *Most importantly, we assume that, given a type, the activity of an eQTL in tissue r is independent from its activity in other tissues*
- 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.

$$\boldsymbol{b}_{gp} | U_0 \sim \mathcal{N}_R(\mathbf{0}, U_0) \tag{1}$$

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

$$p(U_0) = p(\Delta_{gp} | \Gamma_{gp}) \mathsf{P}(\Gamma_{gp}) \tag{2}$$

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.

- 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?
- In the config model, given a configuration $\boldsymbol{\gamma}$,

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

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:

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

Since the fact that there are non-zero entries on the diagonal means that the convariance is non-zero and beta is a multivariate normal, doesn't b—q mean that the effects are *not independent conditional on type*? So are we always left with the BF conditional on q, or do we ever integrate out q?

I see in solution 2, we are effectively doing as in the supplement of the AOAS paper (i.e., equation A.6)

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,

$$\mathsf{P}(\boldsymbol{t}_{gp} = \mathbf{0}|v_{gp} = 0) = 1. \tag{4}$$

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

$$\mathsf{P}(t_{gpk} = 1|v_{gp} = 1) = \pi_k \tag{5}$$

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,

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

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

$$\mathsf{P}(\gamma_{gpr} = 1|t_{gpk} = 1) = q_{kr}, \tag{7}$$

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 $\boldsymbol{\Gamma}_g$?