Generative Model of Flutre, Wen, Pritchard and Stephens (2013)

Sarah Urbut

July 14, 2014

This document describes the generative model of Flutre *et al.* with more details and sometimes a slightly different notation.

1 Data

Let's imagine we measured the expression level of G genes in S subgroups, e.g. tissues, from N individuals which we also genotyped at P genetic variants, typically SNPs. The individuals are not necessarily present in all subgroups.

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 only depends on the genotypes 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).

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

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

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 subgroups, as this can be dealt with by preprocessing or by choosing a Poisson-like distribution.

Also, in practice, subgroup-specific confounding factors in $\boldsymbol{Y}$ are regressed out beforehand, and the residuals, possibly quantile-normalized, are used as responses.

2 Likelihood of a gene-SNP pair

The likelihood for gene g and SNP p is:

$Y_g \sim \mathcal{N}_{N \times S}(X_p B_{gp} + X_c B_{gc}, I_N, \Sigma_{gv})$

where:

- Y_g is the $N \times S$ matrix of expression levels;
- $\mathcal{N}_{N \times S}$ is the matrix-variate Normal distribution;
- X_p is the $N \times 1$ matrix of genotypes;
- B_{gp} is the $1 \times S$ matrix of genotypic 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 $(1 + Q) \times S$ matrix of covariate effect sizes (including the μ_s);
- Σ_{gv} is the $S \times S$ covariance matrix of the errors.

For the uninitiated, let's consider how we might simulate the data with a specified matrix of SNPs and configurations. First, let's consider the error matrix $\boldsymbol{\Sigma}_{gv}$. For every level of gene expression measured for a set of n (let's assume unrelated individuals), we have an $N \times S$ matrix describing the residual errors remanning after the deterministic portion of the model has explained a proportion of the variation in gene expression measured on S subgroups. Here, the S subgroups represent tissue types, but we can imagine a situation in which we consider the effect of a particular SNP on disease risk . In attempting to simulate these levels of gene expression, we recognize that within a given individual i, the errors may be correlated. That is, there may be some covariance of gene expression between tissues for a given individual. We know that a vector of normals is distributed according to is the $S \times S$ covariance matrix of the errors, $\boldsymbol{\Sigma}_{error}$. We know that the multivariate normal distribution of a k-dimensional random vector $X = [X_1, X_2, , X_k]$ can be written in the following notation:

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{1}$$

In our case, this means that the vectors of residual errors for a given individual follows the multivariate normal distribution. Because we assume that the individuals are unrelated, we assume that $(\epsilon_{1i}, \ldots, \epsilon_{Si})$ are independent and identically distributed as $\mathcal{N}(0, \boldsymbol{\Sigma}_g v)$. Now all that's left is to simulate $\boldsymbol{\Sigma}_{gv}$, the covariance matrix of the errors. For the moment, in the simulation data we

imagine that every gene possesses the same correlation matrix between tissues, and because the individuals are i.i.d their levels of gene expression between tissues are also distributed according to the same parameter. In order to simulate $\mathbf{\Sigma}_{gv}$, we take advantage of the following:

In Bayesian statistics, in the context of the multivariate normal distribution, the Wishart distribution is the conjugate prior to the precision matrix $\omega = \sigma^{-1}$, where σ_{gv} is the covariance matrix. If we are interested in simulating $\mathbf{\Sigma}_{gv}$, a reasonable choice is thus the inverse Wishart distribution.

Thus $\mathbf{X} = \mathbf{\Sigma}_{gv}$ has an inverse Wishart distribution $\mathbf{X} \sim \mathcal{W}^{-1}(\mathbf{\Sigma}^{-1}, \nu)$

In our case, we use an independent inverse Wishart prior, with parameters m (a positive scalar) and $\mathbf{H}$ (a positive-denite $r \times r$ matrix):

$\mathbf{\Sigma}_{gv} \sim IW(\nu\mathbf{H}$,m).

Here, ν is a function of the number of individuals minus the number of covariates , H is the diagonal matrix of the identity matrix of the number of subgroups, and m is also a function of the number of individuals.

So for every individual, we draw a vector of residual errors from the multivariate normal distribution with a covariance matrix drawn from the Inverse Wishart Distribution. We then stack these vectors of residual errosinto an $N \times S$ matrix, which is the same as drawing one $N \times S$ matrix from the matrix-variate normal distribution with the $S \times S$ covariance matrix $\mathbf{\Sigma}_{gv}$ and the $N \times N$ covariance matrix as the Identity matrix $\mathbf{I}_{nxn}$. If we assume that the the measurements on tissues are uncorrelated - perhaps they come from different individuals, we assume that it is an identity matrix, and thus specified with coverr=0. Thus

$$\mathbf{E} \sim MN(\mathbf{0}, \mathbf{I}, \mathbf{I}) \tag{2}$$

To summarize:

When the tissue samples are taken from the same individuals we allow that the observations on the same individual may be correlated with one another. Specifically, let $E := (\boldsymbol{e}_1 \cdots \boldsymbol{e}_s)$ denote the $N \times S$ matrix of residual errors, the we assume it to follow a matrix-variate normal (MN) distribution, i.e.,

$$\mathbf{E} \sim \mathrm{MN}(0, I, \mathbf{\Sigma}_{GV}). \tag{3}$$

That is, the vectors $(\epsilon_{1i}, \ldots, \epsilon_{Si})$ are independent and identically distributed as $\mathcal{N}(0, \Sigma)$.

Now we need to consider how we will simulate the effect size $\boldsymbol{\beta}$

A key component of our Bayesian model is the distribution $p(\boldsymbol{\beta}|\gamma, \theta)$, where θ denotes hyper-parameters that are to be specified or estimated from the data. (In the main text we used $p(\beta|\gamma, \theta)$ to simplify exposition, but we actually work with the standardized effects $\boldsymbol{\beta}$.) Of course, if $\gamma_s = 0$ then $b_s = 0$ by definition.

So it remains to specify the distribution of the remaining b_s values for which $\gamma_s = 1$.

We use the distribution from which provides a flexible way to model the heterogeneity of genetic effects of an eQTL in multiple tissues.

Specifically,

consider a distribution $p(\boldsymbol{\beta}|\phi, \omega, \gamma)$, with two hyper-parameters, ϕ, ω, in which the non-zero effects are normally distributed about some mean $\bar{b}$, which itself is normally distributed:

$$b_s|\bar{b}, \gamma_s = 1 \sim \mathcal{N}(\bar{b}, \phi^2), \tag{4}$$

and

$$\bar{b} \sim \mathcal{N}(0, \omega^2). \tag{5}$$

Note that $\phi^2 + \omega^2$ controls the variance (and hence the expected absolute size) of b_s, and $\phi^2/(\phi^2 + \omega^2)$ controls the heterogeneity (indeed, $\omega^2/(\phi^2 + \omega^2)$ is the correlation of $b_s, b_{s'}$ for different subgroups $s \neq s'$). If $\phi^2 = 0$ then this model corresponds to the "fixed effects" model in which the effects in all subgroups are equal (e.g. [**?**]).

If we consider

$$b_s|\bar{b}, \gamma_s = 1 \sim \mathcal{N}(\bar{b}, \phi^2), \tag{6}$$

and we wish to determine the covariance matrix of the $b_s|\gamma_{1=1}, \gamma_{2=1}$, we can integrate out $\bar{b}$ using the following trick:

$$\begin{aligned} E[\mathrm{E}[X \mid \mathcal{Y}, \mathcal{Z}] \mid \mathcal{Y}] &= \mathrm{E}[X \mid \mathcal{Y}]. \\ \mathrm{E}[\mathrm{E}[b_{s1}, b_{s2} \mid \bar{b}, \gamma] \mid \gamma] &= \mathrm{E}[b_{s1}, b_{s2} \mid \gamma]. \\ \mathrm{E}[\bar{b}^2 \mid \gamma] &= \omega^2 \end{aligned}$$

Similarly,

$$\begin{aligned} E[\mathrm{E}[b_{s1}^2 \mid \bar{b}, \gamma] \mid \gamma] &= \mathrm{E}[b_{s1}^2 \mid \gamma]. \\ \mathrm{E}[\phi^2 + \bar{b}^2 \mid \gamma] &= \omega^2 + \phi^2 \end{aligned}$$

Thus we can draw the s-component vector of effect sizes for a given gene-SNP pair with a specified effect from a multivariate normal,

$$\mathbf{b_s}|\gamma_s \sim \mathcal{N}_f(0, \mathbf{\Sigma_g}) \tag{7}$$

Where $\mathbf{\Sigma_g}$is an $S \times S$ matrix with $\omega^2 + \phi^2$ on the diagonal ω on the off diagonal.

Then, to simulate data assuming at most one eQTN per gene, we draw a simulated configuration if the gene is indeed affected by an eQTN, and the gene expression Y is simulated from a simple linear regression as:

$Y_g \sim \mathcal{N}_{N \times S}(X_p B_{gp} + X_c B_{gc}, I_N, \Sigma_{gv})$