Statistical framework from Flutre, Wen, Pritchard and Stephens (PLoS Genetics, 2013): Halfway to understanding the Bayes Factors

Sarah Urbut

July 23, 2014

Contents

1 Likelihood of the whole data set **1**

2 Focus on a single gene-SNP pair **1**

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

2 Focus on a single gene-SNP pair

The likelihood for gene g and SNP p is:

$$Y_g|X_p, B_{gp}, X_c, B_{gc}, \Sigma_{gp} \sim \mathcal{N}_{N\times R}(X_p B_{gp} + X_c B_{gc}, I_N, \Sigma_{gp}) \tag{1}$$

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);
- $\mathcal{N}_{N\times R}$ 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 $\boldsymbol{\beta}_{gp} = B_{gp}^T$.

The conditional posterior of B: $\mathsf{P}(B|Y, X, \tau) = \frac{\mathsf{P}(B,Y|X,\tau)}{\mathsf{P}(Y|X,\tau)}$

Let's neglect the normalization constant for now. We note that if we expand this expression in full, we get the following:

$$\Pr(\boldsymbol{b_s}|\boldsymbol{Y_s}, \boldsymbol{X_s}, \boldsymbol{\Phi_s^{-1}}) \propto \exp((\boldsymbol{b_s} - \bar{\boldsymbol{b}})^t \boldsymbol{\Phi_s^{-1}} (\boldsymbol{b_s} - \bar{\boldsymbol{b}})) \exp((\boldsymbol{Y_s} - \boldsymbol{X_s b_s})^t (\boldsymbol{Y_s} - \boldsymbol{X_s b_s})) \tag{2}$$

Taking terms out of the exponent and distributing terms, we arrive at:

$$\Pr(\boldsymbol{b_s}|\boldsymbol{Y_s}, \boldsymbol{X_s}, \boldsymbol{\Phi_s^{-1}}) \propto \boldsymbol{b_s}^t \boldsymbol{\Phi_s^{-1}} \boldsymbol{b_s} - \boldsymbol{b_s}^t \boldsymbol{\Phi_s^{-1}} \bar{\boldsymbol{b}} - \bar{\boldsymbol{b}}^t \boldsymbol{\Phi_s^{-1}} \boldsymbol{b_s} + \bar{\boldsymbol{b}}^t \boldsymbol{\Phi_s^{-1}} \bar{\boldsymbol{b}} + \boldsymbol{Y_s}^t \boldsymbol{Y_s} - (\boldsymbol{X_s b_s})^t \boldsymbol{Y_s} - \boldsymbol{Y_s}^t \boldsymbol{X_s b_s} + (\boldsymbol{X_s b_s})^t (\boldsymbol{X_s b_s}) \tag{3}$$

We will first leave out the terms that don't include $\boldsymbol{b_s}$, i.e., $\boldsymbol{Y_s}^t \boldsymbol{Y_s}$ and $\bar{\boldsymbol{b}}^t \boldsymbol{\Phi_s^{-1}} \bar{\boldsymbol{b}}$, and group some terms:

$$\Pr(\boldsymbol{b_s}|\boldsymbol{Y_s}, \boldsymbol{X_s}, \boldsymbol{\Phi_s^{-1}}) \propto \boldsymbol{b_s}^t (\boldsymbol{\Phi_s^{-1}} + \boldsymbol{X_s}^t \boldsymbol{X_s}) \boldsymbol{b_s} - \boldsymbol{b_s}^t (\boldsymbol{\Phi_s^{-1}} \bar{\boldsymbol{b}} - \boldsymbol{X_s}^t \boldsymbol{Y_s})^t - (\boldsymbol{\Phi_s^{-1}} \bar{\boldsymbol{b}} + \boldsymbol{X_s}^t \boldsymbol{Y_s})^t \boldsymbol{b_s} \tag{4}$$

In order to aid our completion of the square, let's add a term that doesn't contain $\boldsymbol{b_s}$ which is a legitimate

$$(\boldsymbol{\Phi_s^{-1}} \bar{\boldsymbol{b}} + \boldsymbol{X_s}^t \boldsymbol{Y_s})^t (\boldsymbol{\Phi_s^{-1}} - \boldsymbol{X_s}^t \boldsymbol{X_s}) (\boldsymbol{\Phi_s^{-1}} \bar{\boldsymbol{b}} + \boldsymbol{X_s}^t \boldsymbol{Y_s}) \tag{5}$$

Now, let's define $\boldsymbol{\Omega_s^{-1}}$ as: $(\boldsymbol{\Phi_s^{-1}} - \boldsymbol{X_s}^t \boldsymbol{X_s})$ and $\boldsymbol{\mu_s}$ as $(\boldsymbol{\Phi_s^{-1}} - \boldsymbol{X_s}^t \boldsymbol{X_s})$ $(\boldsymbol{\Phi_s^{-1}} \bar{\boldsymbol{b}} + \boldsymbol{X_s}^t \boldsymbol{Y_s})$

Then it becomes apparent that we can rewrite $\Pr(\boldsymbol{b_s}|\boldsymbol{Y_s}, \boldsymbol{X_s}, \boldsymbol{\Phi_s^{-1}})$ as:

$$\Pr(\boldsymbol{b_s}|\boldsymbol{Y_s}, \boldsymbol{X_s}, \boldsymbol{\Phi_s^{-1}}) \propto (\boldsymbol{b_s} - \boldsymbol{\mu_s})^t \boldsymbol{\Omega_s^{-1}} (\boldsymbol{b_s} - \boldsymbol{\mu_s}) \tag{6}$$

So that when we compute the marginal probability of Y:

$$\Pr(\boldsymbol{Y_s}|\boldsymbol{X_s}) = \frac{\Pr(\boldsymbol{b_s}|\boldsymbol{Y_s}, \boldsymbol{X_s}, \tau) \Pr(\boldsymbol{b_s}|\tau, \bar{\boldsymbol{b}})}{\Pr(\boldsymbol{b_s}|\boldsymbol{Y_s}, \boldsymbol{X_s}, \tau)} \tag{7}$$

It is obvious that the numerator will contain the two terms we neglected (because they didn't contain $\boldsymbol{b_s}$) in (3) and the denominator will contain the term we added in (4), $(\boldsymbol{\Phi_s^{-1}} \bar{\boldsymbol{b}} + \boldsymbol{X_s}^t \boldsymbol{Y_s})^t (\boldsymbol{\Phi_s^{-1}} - \boldsymbol{X_s}^t \boldsymbol{X_s}) (\boldsymbol{\Phi_s^{-1}} \bar{\boldsymbol{b}} + \boldsymbol{X_s}^t \boldsymbol{Y_s})$.

Thus the marginal likelihood is:

$$\begin{gathered}\Pr(\boldsymbol{Y_s}|\boldsymbol{X_s}) = \frac{2\pi}{\tau_s}^{-n_s/2} \left|\boldsymbol{\Phi_s^{-1}}\right|^{\frac{-1}{2}} \left|(\boldsymbol{\Phi_s^{-1}} + \boldsymbol{X_s}^t \boldsymbol{X_s})\right|^{\frac{-1}{2}} * \\ \exp(\frac{1}{2}(\boldsymbol{Y_s}^t \boldsymbol{Y_s} \text{ - } (\boldsymbol{\Phi_s^{-1}} \bar{\boldsymbol{b}} + \boldsymbol{X_s}^t \boldsymbol{Y_s})^t (\boldsymbol{\Phi_s^{-1}} - \boldsymbol{X_s}^t \boldsymbol{X_s})^{-1} (\boldsymbol{\Phi_s^{-1}} \bar{\boldsymbol{b}} + \boldsymbol{X_s}^t \boldsymbol{Y_s}) - \bar{\boldsymbol{b}}^t \boldsymbol{\Phi_s^{-1}} \bar{\boldsymbol{b}}))\end{gathered} \tag{8}$$