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

Sarah Urbut

July 24, 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

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:

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

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

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

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:

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

When we integrate over $\bar{\boldsymbol{b}}$, we could have just done the integral over $\boldsymbol{b_s}$ rather than $\boldsymbol{b_s}\bar{\boldsymbol{b}}$ which would be $\boldsymbol{b_s} \sim (\mathbf{0},\mathbf{W})$, where W is the prior matrix of covariance effects, with $\phi+\omega$ on the diagonal and ω on the off-diagonal. This is akin to what is done in the biometric paper, where:

$$\begin{aligned}p(\boldsymbol{b_s}|\boldsymbol{X_s}\tau) \propto \exp(\frac{-1}{2}[(\boldsymbol{b_s}-\hat{\boldsymbol{\beta}}_{\boldsymbol{s}})^t\boldsymbol{V^{-1}}(\boldsymbol{b_s}-\hat{\boldsymbol{\beta}}_{\boldsymbol{s}})+(\boldsymbol{b_s}^t\boldsymbol{W^{-1}}\boldsymbol{b_s})]\\ \propto \exp(\frac{-1}{2}[(\boldsymbol{b_s}^t(\boldsymbol{V^{-1}}+\boldsymbol{W^{-1}})\boldsymbol{b_s})]-(\boldsymbol{b_s}^t(\boldsymbol{V^{-1}}\hat{\boldsymbol{\beta}}_{\boldsymbol{s}})+(\hat{\boldsymbol{\beta}}_{\boldsymbol{s}}{}^t\boldsymbol{V^{-1}})\boldsymbol{b_s}))\end{aligned} \tag{9}$$

Analogous to the previous situation, we have omitted an additional term contained in the likelihood of $\Pr(\hat{\boldsymbol{\beta}}_{\boldsymbol{s}}|\boldsymbol{X}_{\boldsymbol{s}}, \mathbf{E})$ because it didn't contain $\boldsymbol{b}_{\boldsymbol{s}}$, $\hat{\boldsymbol{\beta}}_{\boldsymbol{s}}{}^t \boldsymbol{V}^{-\mathbf{1}} \hat{\boldsymbol{\beta}}_{\boldsymbol{s}}$ and we have added an additional term:

$$\boldsymbol{b}_{\boldsymbol{s}}{}^t (\boldsymbol{V}^{-\mathbf{1}} \hat{\boldsymbol{\beta}}_{\boldsymbol{s}})(\boldsymbol{V}^{-\mathbf{1}} + \boldsymbol{W}^{-\mathbf{1}})(\hat{\boldsymbol{\beta}}_{\boldsymbol{s}}{}^t \boldsymbol{V}^{-\mathbf{1}})\boldsymbol{b}_{\boldsymbol{s}} \tag{10}$$

Crucially, in computing the Bayes Factor

$$\frac{\Pr(\hat{\boldsymbol{\beta}}_{\boldsymbol{s}}|\boldsymbol{X}_{\boldsymbol{s}}, \mathbf{E}, M1)}{\Pr(\hat{\boldsymbol{\beta}}_{\boldsymbol{s}}|\boldsymbol{X}_{\boldsymbol{s}}, \mathbf{E}, M0)} \tag{11}$$

Only the added term is present in the numerator and not the denominator, and so we can see that the Bayes Factor will be proportional to: $\exp(\boldsymbol{b}_{\boldsymbol{s}}{}^t (\boldsymbol{V}^{-\mathbf{1}} \hat{\boldsymbol{\beta}}_{\boldsymbol{s}})(\boldsymbol{V}^{-\mathbf{1}} + \boldsymbol{W}^{-\mathbf{1}})(\hat{\boldsymbol{\beta}}_{\boldsymbol{s}}{}^t \boldsymbol{V}^{-\mathbf{1}})\boldsymbol{b}_{\boldsymbol{s}})$

which, intuitively, augments the likelihood of the data by the prior covariance matrix of the effects, incorporating prior belief about shared effect and thus increasing our power to detect such homogeneity.

- 1. In the simplest case, assuming the τ is known, if we consider as the product of all marginal likelihoods, then how are we really exploiting the covariance in effects? I still don't see it – the Y vectors are n x 1, and so the covariance in effects in the offi diagonal terms is not in the ABF_{random}

  This is only because we integrate over $\boldsymbol{b}_{\boldsymbol{s}}$ first - if we had integrated over $\bar{\boldsymbol{b}}$ first, it would not have been so. We would have brought everything back into 1, and then it becomes factored again because the residuals are uncorrelated.

- 2. We can use a product because we assume the vectors of residual errors are independent across populations, but the effects are not (in any case by the max H case). So is the product because of the conditional exchangeability?

  Yes!

- 3. I see that we are pooling the Bayes Factors by the product, which I guess increases our sample size (by sheer more terms in our product) but even if the residuals are independent, we miss the covariance in effects inherent in the data

- 4. Perhaps the product refers to conditional on the mean $\bar{\boldsymbol{b}}$, they are exchangeable (i.e., independent) but then we are replying only on the BF(fixed) to reinforce our covariance in effects.

From William's paper:

> *The off diagonal of matrix $\mathbf{W_g}$ defines context-dependent prior correlation between non-zero regression coefficients. Incorporating they information enables borrowing strength across correlated components in β_g, thereby improving the efficiency of model selection.*

But if we use a product over all $\boldsymbol{b}_{\boldsymbol{s}}$, then how are we incorporating the off-diagonal elements? Is it just through the fixed effect term on the left, or does the term on the right incorporate the greater probability that $\boldsymbol{b}_{\boldsymbol{s}}$arises from a shared effects mode? I don't think it can because it doesn't pay any attention to other data. Maybe if I could see where the prior on shared effect came into the Bayes factor ...

Also: Kass LaPlace Approximation: I see that the majority of the Mass of the BF will fall at the MLE, but don't we still need to weight by the prior probability of that parameter?

3 Answers!

2. We can use a product because we assume the vectors of residual errors are independent across populations, but the effects are not (in any case by the max H case). So is the product because of the conditional exchangeability?

Yes!