Exploiting Adaptive Bayesian Regression Shrinkage to identify Exome Sequence Variants associated with Gene Expression.

Elizabeth Boggis[1], Marta Milo[2], Kevin Walters[1]

Bayesian Young Statisticians Meeting (BAYSM), Milan June, 5-6, 2013
Paper no. 5

[1] School of Mathematics and Statistics, University of Sheffield, Sheffield, UK
e.boggis@sheffield.ac.uk
k.walters@sheffield.ac.uk
[2] Department of Biomedical Science, University of Sheffield, Sheffield, UK
m.milo@sheffield.ac.uk

Abstract

Using Bayesian adaptive shrinkage in the form of the Normal-Gamma prior we show that causal DNA sequence variants associated with a change in gene expression can be successfully detected. Taking a fully Bayesian approach allows our model to be developed to include uncertainty in gene expression and SNP calls, and to include biological information from online databases.

Bayesian shrinkage; SNPs; linear model; sequencing

1 Introduction

Next-Generation exome sequencing identifies thousands of DNA sequence variants in each individual. Methods are needed that can effectively identify which of these variants are associated with changes in gene expression, a measure of the activity of the gene. As we expect only a few SNPs (single DNA base changes) to be causal, i.e. to cause disease, we need methods that induce sparse models. The Normal-Gamma prior has been shown to induce adaptive shrinkage within the Bayesian linear model framework (large effects are shrunk proportionally less than small effects) [1]. Using simulated data we assess the efficacy and limitations of this Bayesian shrinkage method in comparison to other published

methods in parsimoniously identifying such sequence variants. The model is then validated using publicly available human and yeast datasets. We further develop the model to include the uncertainty in gene expression; SNP functional information (information on the known biological effect of the single point mutation) obtained from online databases; and the uncertainty in the DNA base calls.

2 Modelling Using the Normal-Gamma Prior.

The Normal-Gamma hierarchical prior [1] is given by:

$$\begin{aligned}\pi(\beta_i|\psi_i) &\sim N(0,\psi_i)\\ \pi(\psi_i|\lambda,\gamma) &\sim Ga\left(\lambda,\frac{1}{2\gamma^2}\right)\end{aligned}$$

which has var $(\boldsymbol{\beta}|\lambda,\gamma) = 2\lambda\gamma^2$ which we assign an $IG(2,M)$ prior. Consider the standard linear model

$$y_{ij} = \sum_{k=1}^{p_j} \beta_{jk} x_{ijk} + \epsilon_{ij},$$

where i, j, k represent individual, gene and SNP respectively, and p_j represents the number of SNPs in the model for gene j. β_{jk} is the effect size of the k^{th} SNP in gene j.

2.1 Including Uncertainty in Gene Expression (y).

To account for the uncertainty in gene expression ($\mathbf{y}$) we use a more complex error structure. We propose to decompose the error variance into $\Sigma + \sigma^2\Omega$ that includes the weighted technical variance of the gene expression due to differences in non-biological aspects of the gene expression ($\sigma^2\Omega$) and a covariance matrix of errors (Σ). We define

$$\begin{aligned}\pi(\Sigma) &\sim Inv - Wishart\\ \pi(\sigma^2) &\propto 1\\ \Omega &= \text{diag(technical variance)},\end{aligned}$$

where the technical gene expression variance can be obtained from PUMA [2]. This variance decomposition estimates the variability due to technical effects and to other sources, e.g environmental, epigenetic (changes affecting gene expression that are not related to changes in the DNA) etc.

2.2 Including Uncertainty in SNP calls (X).

Exome sequence base calls have associated Phred-based quality scores (Q) which are a function of the base calling error probability (P), where $P = 10^{\frac{-Q}{10}}$ (high Q means high certainty in allele call). This probability is used in a Bernoulli sampler (0 represents wildtype, 1 represents SNP) to "update the matrix of SNPs at each iteration of the MCMC. This modification should improve detection of causal SNPs with poor quality scores that might otherwise be discarded if a quality score threshold is applied.

2.3 Including Functional Annotation Information (F).

Functional annotation information is increasingly widely available in online databases. Novel SNPs, SNPs that have not previously been found and recorded, are not annotated and the only information obtainable is whether the SNP is synonymous, have no change on the protein they code for, or non-synonymous, cause a change in the protein they code for. By combining a given set of biological parameters on annotated SNPs into one score [3], we can obtain a distribution of scores (ω) independently for synonymous and non-synonymous SNPs. We can use these empirical distributions to inform our hierarchical prior for $\boldsymbol{\beta}$, thus enforcing more shrinkage on parameter estimates of *a priori* less important SNPs with respect to association with disease.

3 Preliminary Results.

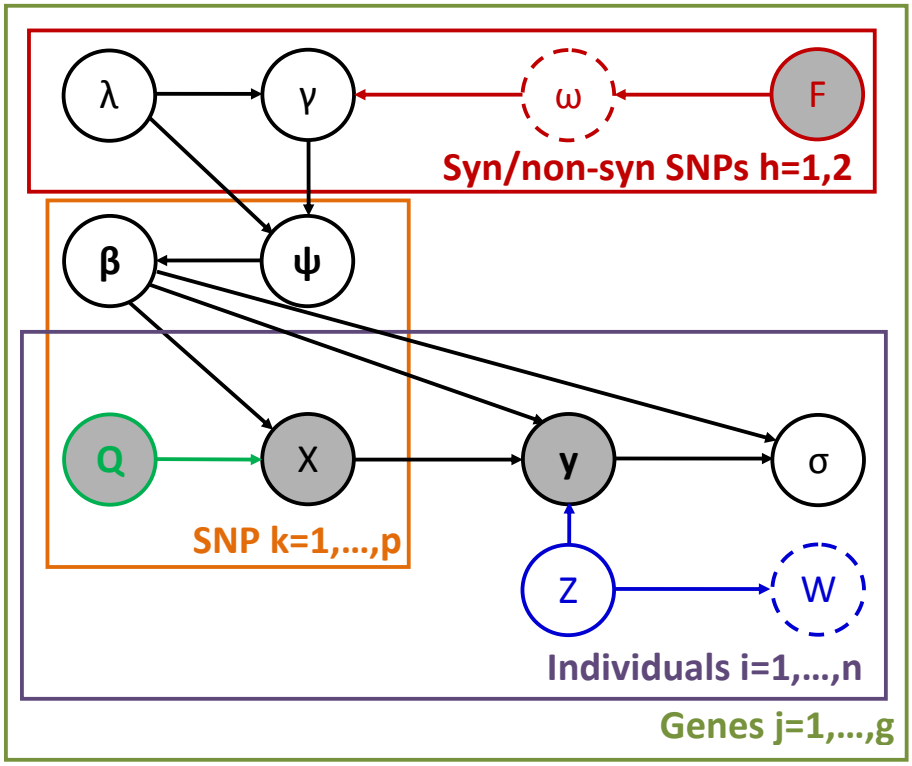
Figure 1 includes the weights (W) of confounding factors (Z) such as age, gender, population structure etc. To avoid having to incorporate confounding in our model, we use PANAMA [4] to deconfound the gene expression signal ($\mathbf{y}$).

In the initial simulation study 8 causal β (ranging from 2 to 0.4 in magnitude) are fixed and non-causal β are simulated to have an effect sampled from a $N(0, 0.01)$ with gene expression a linear sum of the weighted SNPs plus a $N(0, 1)$ error.

In comparison with the least squares estimates, see Figure 1, the Normal-Gamma [1] prior detects all truly causal SNPs at a lower false positive rate. This is due to comparatively less differential shrinkage across all β_{jk}. Comparing with the HyperLasso [5] which enforces similar shrinkage, and piMASS [6] which uses Bayesian selection, the Normal-Gamma model has similar performance (Figure 1).

4 Conclusion

Our developments to the Normal-Gamma prior provide a suitable framework, that has been shown via simulation, to successfully identify causal DNA sequence variants (SNPs) affecting the gene expression level. Taking a fully Bayesian



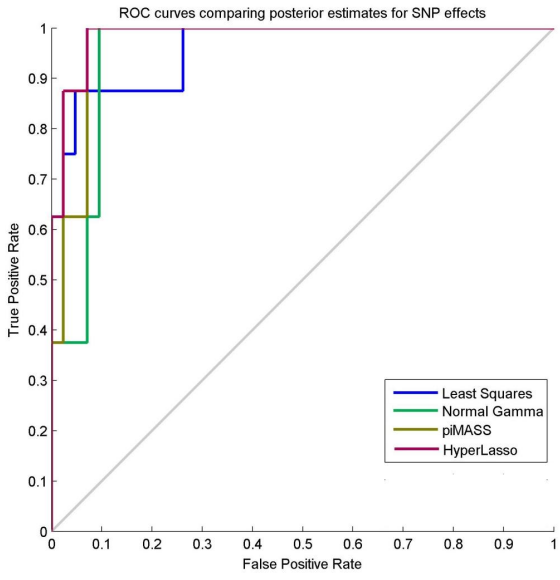


Figure 1: TOP: A graphical model of the relationships between the parameters in our extended Normal-Gamma model. BOTTOM: ROC curves generated from the Least Squares, Normal-Gamma [1], piMASS [6], and HyperLasso [5] on simulated data.

approach, permitted by the Normal-Gamma prior, allows for the various sources of uncertainty to be incorporated in a coherent manner.

References

[1] J.E. Griffin, P.J. Brown. **Inference with normal-gamma prior distributions in regression problems**. *Bayesian Analysis*; 2010; 5(1); pp. 171-188.

[2] X. Liu, M. Milo, N.D. Lawrence, M. Rattray. **A tractable probabilistic model for Affymetrix probe-level analysis across multiple chips**. *Bioinformatics*; 2005; 21(18); pp. 3637-3644.

[3] P.H. Lee, H. Shatkay. **An integrative scoring system for ranking SNPs by their potential deleterious effects**. *Bioinformatics*; 2009; 25(8); pp. 1048-1055.

[4] N. Fusi, O. Stegle, N.D. Lawrence. **Joint Modelling of Confounding Factors and Prominent Genetic Regulators Provides Increased Accuracy in Genetical Genomics Studies**. *PLoS Computational Biology*; 2012; 8(1); pp. e1002330.

[5] C.J. Hoggart, J.C. Whittaker, M. De Iorio, D.J. Balding. **Simultaneous Analysis of All SNPs in Genome-Wide and Re-Sequencing Association Studies**. *PLoS Genet*; 2008; 4(7); pp. e1000130.

[6] Y. Guan, M. Stephens. **Bayesian Variable Selection Regression for Genome-Wide Association Studies and Other Large-Scale Problems**. *The Annals of Applied Statistics*; 2011; 5(3); pp. 1780-1815