Adaptive Shrinkage and False Discovery Rates by Laplace Approximation

Matthew Stephens

i-like, Warwick, May 2013

Inference from Summary Statistics

Genetic association studies aim to identify genetic variants that modify a phenotype.

Often they focus on clinically-relevant phenotypes (e.g. LDL cholesterol).

Other times they use molecular phenotypes (e.g. gene expression).

The idea is absurdly simple: measure genetic variants (usually SNPs), and phenotypes in randomly-sampled individuals, and identify which SNPs are correlated with phenotypes.

Genetic Association Studies

So a typical GWAS analysis involves fitting millions of simple regressions, and testing effects for significance.

Notation: y for phenotype, g for genotype, β for genetic effects:

$$y_i = \mu + \beta g_i + \epsilon_i \qquad (i = 1, \dots, n)$$

Genetic Association Studies and Heterogeneity

We have been developing statistical methods for association mapping in multiple subgroups, incorporating heterogeneity of effects. See also Lebrec et al (2010); Han and Eskin (2011,2012).

Motivating examples include:

1. The “Global Lipids Consortium” Genome-wide Association Study meta-analysis (Teslovich et al, 2010).
2. Gene expression analysis among multiple tissues (e.g. Dimas et al, 2009).

Example 1: Global Lipids meta-analysis

GWAS data from the Global Lipids consortium (Teslovich et al, 2010) on $> 100,000$ individuals from at least 25 separate studies.

Four phenotypes: total cholesterol (TC), low-density lipoprotein cholesterol (LDL-C), high-density lipoprotein cholesterol (HDL-C) and triglycerides (TG).

The original association analyses performed a *fixed-effects meta-analysis*. That is, assume the effects in each subgroup are all equal ($\beta_s = \beta \, \forall s$), and test $H_0 : \beta = 0$.

They reported a total of 95 SNPs as being associated with one or more phenotypes.

Example 1: Global Lipids meta-analysis

Question: Given that this study involved 25 separate subgroups, many with quite different recruitment criteria (e.g. some were recruited as cases for a particular disease; others were recruited as controls; etc), would an analysis that allows for heterogeneity across studies identify more associations?

Example 2: eQTL studies across multiple cell-types

Dimas et al (2009), measured expression data in 75 individuals, in 3 cell types: Fibroblasts, LCLs and T-cells.

A key goal was to identify genetic variants associated with expression that were shared among cell types, or were specific to some subset of cell-types. (Identifying eQTLs specific to individual cell types may shed insight into cell-type-specific regulation mechanisms.)

Original analysis performed association analysis separately in each cell type, and then looked at the overlap of the resulting associations. The overlap was small (14%), and they concluded that many eQTLs occur in only one cell type.

Example 2: eQTL studies across multiple cell-types

Question: Incomplete power may cause this analysis to underestimate sharing; does a joint analysis of all cell types come to the same conclusion?

Methods

Focus first on meta-analysis, where effect β may vary across subgroups s:

$$y_{si} = \mu_s + \beta_s g_i + \epsilon_{si} \text{ with } \epsilon_{si} \sim \mathcal{N}(0, \sigma_s^2).$$

Primary goal of the meta-analysis is to identify SNPs for which there is strong evidence against $H_0 : \beta_s \equiv 0 \, \forall s$.

Alternatives to H_0

To assess evidence against H_0 we introduce a set of alternative models, indexed by parameters ϕ, ω, to be compared with H_0.

$H_1(\phi, \omega)$: β_s normally distributed about common mean $\bar{\beta}$.

$$\beta_s | \bar{\beta} \sim N(\bar{\beta}, \phi^2); \qquad \bar{\beta} \sim N(0, \omega^2).$$

Note 1: $\phi = 0$ corresponds to the usual “fixed effects” alternative, $\beta_s = \bar{\beta} \, \forall s$.

Note 2: Can alternatively work with the “standardized” effect sizes, $b_s = \beta_s / \sigma_s$, which generally yields similar (but not identical) results.

Bayes Factors

The Bayes Factor

$$\mathrm{BF}(\phi, \omega) = \mathrm{p}(\mathrm{y}|\mathrm{g}, \mathrm{H}_1(\phi, \omega))/\mathrm{p}(\mathrm{y}|\mathrm{H}_0)$$

measures the support for $H_1(\phi, \omega)$ vs H_0, with large values indicating strong evidence against H_0.

Although $\mathrm{BF}(\phi, \omega)$ depends on priors for nuisance parameters (μ, σ_s^2), it is not very sensitive, and sensible default choices exist.

Hyperparameters ϕ and ω must be chosen to reflect expected effect sizes and levels of heterogeneity (but can average over several values to reflect uncertainty in choice of appropriate values).

Computation

$\mathrm{BF}(\phi, \omega)$ can be quickly and accurately approximated by Laplace approximation.

In the simplest cases these approximations depend only on the summary statistics in each study, $\hat{\beta}_s$ and $\mathrm{se}(\hat{\beta}_{\mathrm{s}})$. (Details: Wen and Stephens, 2011).

Bayes Factors and standard test statistics

This framework includes some commonly-used frequentist test statistics as special cases.

For example, if we allow ω to vary across SNPs according to the inverse of the standard error of $\bar{\beta}$ then $BF(\phi = 0, \omega)$ is monotonic with the weighted Z score

$$\mathcal{Z} = \frac{\sum_s w_s Z_s}{\sqrt{\sum_{s'} w_{s'}^2}} \tag{1}$$

where $Z_s = \hat{\beta}_s / \mathrm{se}(\hat{\beta}_s)$ and $w_s = \mathrm{se}(\hat{\beta}_s)^{-1}$. (Details: Wen and Stephens, 2011.)

In other words, we can see what implicit models for β are assumed by standard methods.

Example 1: Global Lipids meta-analysis

Question: Given that this study involved 25 separate subgroups, many with quite different recruitment criteria (e.g. some were recruited as cases for a particular disease; others were recruited as controls; etc), would an analysis that allows for heterogeneity across studies identify more associations?

Answer: Not much!

Example 1: Global Lipids meta-analysis

Question: Given that this study involved 25 separate subgroups, many with quite different recruitment criteria (e.g. some were recruited as cases for a particular disease; others were recruited as controls; etc), would an analysis that allows for heterogeneity across studies identify more associations?

Answer: Not much!

Results: Global Lipids GWAS

- Searched genome-wide for SNPs with strong signal when allowing for heterogeneity ($\mathrm{BF}_{\mathrm{het}} > 10^6$) but not when assuming no heterogeneity ($\mathrm{BF}_{\mathrm{no-het}} < 10^6$).
- 42 SNPs satisfied these criteria.
- But 36 of these were driven by apparently strong associations in a single study (Framingham heart Study), and seemed likely to be due to data processing errors.
- Two more showed similarly suspicious patterns (association in just one study, a subset of the WTCCC).

Results: Global Lipids GWAS

Phenotype	SNP	Gene	$\log_{10}(\mathrm{BF}_{\text{no-het}})$	$\log_{10}(\mathrm{BF}_{\text{het}})$
LDL	rs1800978	ABCA1	5.2	6.0
TG	rs1562398	nr KLF14	5.3	6.5
HDL	rs11229165	nr OR4A16	4.6	6.4
HDL	rs7108164	nr OR4A42P	4.2	6.3

Example 2: eQTL sharing across cell-types

In meta-analysis application the primary goal was to reject the global null, $H_0 : \beta_s = 0 \, \forall s$.

Mapping eQTLs among multiple cell-types (subgroups) differs in that we care more about *which* β_s are non-zero, and patterns of sharing among subgroups.

E.g. Dimas et al (2009), identified eQTLs separately in 3 cell types, and asked which eQTLs are shared among cell types (subgroups).

Example 2: eQTL sharing across cell-types

To address this we expand our alternative models $H_1(\phi, \omega)$ to allow that effects may be zero in some subgroups.

Introduce a configuration γ indicating which subgroups have non-zero effect.

- E.g. $\gamma = [110]$ corresponds to non-zero effect in the first two subgroups.

See also Han & Eskin (PloS Genetics, 2012).

Bayesian Model Averaging and hierarchical modeling

The support in the data for configuration γ can be measured by the Bayes Factor

$$\mathrm{BF}_\gamma(\phi,\omega) = \frac{\mathrm{p}(\mathrm{y}|\mathrm{H}_1(\gamma,\phi,\omega))}{\mathrm{p}(\mathrm{y}|\mathrm{H}_0)}$$

.

Overall evidence against H_0 can be measured by averaging over γ, ϕ, ω: BMA $= \sum_{\gamma,\phi,\omega} \eta_{\gamma,\phi,\omega} \mathsf{BF}_\gamma(\phi,\omega)$

Estimate proportions $\eta_{\gamma,\phi,\omega}$ using a hierarchical model to combine information across genes.

Example 2: eQTL studies across multiple cell-types

Dimas et al (2009), measured expression data in 75 individuals, in 3 cell types: Fibroblasts, LCLs and T-cells.

They identified eQTLs separately in each cell-type, and found small overlap of results (14%).

Question: Incomplete power may cause this analysis to underestimate sharing; does a joint analysis of all cell types come to the same conclusion?

Joint Analysis Increases Power

Gain in power from the joint analysis

Joint analysis suggests much more sharing of eQTLs

Wrong tissue-specific call by the tissue-by-tissue analysis

Example of gene ENSG00000106153 and SNP rs4948093 (MAF=0.23).
See also Ding *et al.* (2010, AJHG).

The next challenge - more subgroups!

- This “configuration-based” framework can deal satisfactorily with, perhaps, 6-10 subgroups.
- The NIH GTEX project is currently collecting data on upwards of 20 tissues.
- More generally, in genomics, one might have hundreds of “observations” on each unit...
- ... and, potentially, relevant covariates.

The next challenge - a general framework for data integration?

- Summarize the "data" on each SNP in each subgroup (or experimental condition) as $(\hat{\beta}, se(\hat{\beta}))$.
- Arrange these in a big p by S matrix.
- Goal: identify the elements that correspond to non-zero (or "large") β, exploiting combined structure across rows of the matrix.

Challenge: exploit the many available tools – clustering, PCA, factor analysis,etc – to do this in a flexible and powerful way.

Acknowledgments

- Xiaoquan Wen, Timothée Flutre, Jonathan Pritchard.
- Global Lipids Consortium for making data available
- Manoulis Dermitzakis for expression data.
- Funding: NHGRI and NIH GTEX consortium.

Selected References:

- Wen & Stephens (2011, arXiv), Wen (2012, arXiv), Flutre et al (2013, PloS Genetics).
- Han & Eskin (2011, AJHG; 2012, PloS Genetics).
- Ding *et al.* (2010, AJHG).
- Lebrec *et al.* (2010, SAGMB).

`http://stephenslab.uchicago.edu/publications.html`

Gain in power from the joint analysis

Gain in power from the joint analysis

Where next?

- Larger-scale problems (e.g. GTEx collecting data on 30 tissues)
- Multi-SNP multi-phenotype? (e.g. Verzilli et al (2005); Banerjee et al (2008)).
- Dealing with non-normality; Outliers; Binary outcome with intermediate quantitative phenotypes.
- “Response” phenotypes. (Maranville et al, PloS Genetics, 2011).

Some eQTLs may be shared; others tissue-specific

Some eQTLs may be shared; others tissue-specific

Some eQTLs may be shared; others tissue-specific