A Bayesian hierarchical sparse factor model for complex experiments in genetical genomics

Daniel Runcie, R Cody Markelz, Sayan Mukherjee

Contribution

We present a modeling framework aimed at studying the contributions of genes and the environment to phenotypic variation in high dimensions. Specifically, we are interested in leveraging high-dimensional phenotypes to study the combined effects of multiple factors on traits like yield in crops, fitness in natural populations, or health in patients. Our approach combines two well established principles:

- That biological systems tend to be modular in organization.
- That appropriately designed and analyzed experiments are necessary to address complex experimental questions.

We begin with the classic linear mixed effects model that forms the basis for the analysis of a wide range of experimental and observational studies. But whereas a traditional multivariate mixed model becomes intractable and highly sensitive with more than a handful of traits, our model uses a sparse factor structure to focus only on the strongest modules in the data. This builds upon our earlier genetic sparse factor model [2]. We have built an efficient Bayesian MCMC algorithm to fit the model that scales well both computationally and numerically to dimensions relevant to modern phenotype-centric datasets, and generates descriptions of modules that are interpretable with respect to the environmental or genetic factors.

Motivation

Technological advances are leading to a revolution in the collection of phenotype data - from the various *seq technologies (ex RNAseq, ChIPseq, DNase-Seq), to proteomics, metabalomcs, and 3D and hyperspectral imaging. These high-dimensional phenotypes promise to provide an unprecedented window into the complex functions of biological organisms.

0.1 Examples:

- A set of genetically related individuals are assigned to 2+ treatments and then a large set of traits (such as gene expression) are measured on each individual simultaneously.
- A multi-factor experiment with a split-plot or repeated measures design that are common in agriculture settings.
- A case-control studies with individuals of varying ancestry.

In all cases, the experimental goals are to understand which factors are important for sets of the traits, and to identify the biological bases of these effects.

Analytical challenges

In these examples, the analytical challenges are a combination of those that affect univariate quantitative genetic analyses and those that affect the study of high dimensional data:

- Individuals within (and across) treatments are correlated due to shared genetic history.
- Traits are correlated and may be functionally related
- The number of traits (and between-trait covariances) may be larger than the number of individuals ($n << p$)

Approach

We view the high-dimensional phenotype as a readout of a modular and time-ordered developmental process (Figure 1).
This model has the following implications:

- Experimental factors (genotypes, environments) cause perturbations to the modules of the developmental system.
- The traits we observe are independent conditional on the state of the underlying developmental system.
- We cannot directly observe the modules, but can indirectly identify them based on a) the correlation they induce among traits in each individual, and b) the correlation of modules across individuals induced by the experimental design and the genetic structure in the population.
- The same module might explain responses to multiple experimental factors (Figure 2).

We therefore focus on identifying the most important modules, which we assume will account for the majority of the effect of the experimental factors on the downstream high-dimensional phenotype.

Model specification

0.2 Sparse factor model

Motivated by the conceptual model in Figure 1, we model the $n \times p$ phenotype matrix $\mathbf{Y}$ with a sparse factor model, with k factors representing latent developmental modules:

$$\mathbf{Y} = \mathbf{F}\mathbf{\Lambda}' + \mathbf{E} \tag{1}$$

Conditional on the factors, the residuals (rows of $\mathbf{E}$) are independent and assigned Gaussian priors. An "infinite factor"[1] prior structure uses local/global shrinkage to induce sparsity on $\mathbf{\Lambda}$, with increasing shrinkage on higher-indexed factors. This induces a ranking of the factors.

$$\begin{aligned} \lambda_{jl} &\sim \mathrm{N}(0, \tau_k^{-1}\psi_{jl}^{-1}) \\ \psi_{jl} &\sim \mathrm{Ga}(\nu/2, \nu/2) \\ \delta_1 &\sim \mathrm{Ga}(\alpha_0, \beta_0)\ \delta_l \sim \mathrm{Ga}(\alpha_1, 1) \text{ for } l \in 2 \dots k \\ \tau_l &= \prod_{i=1}^{k} \delta_l \end{aligned} \tag{2}$$

0.3 Mixed Effect Model (MEM)

Columns of $\mathbf{F}$ represent trait values for each of the k latent modules across the n individuals. We assume that each module is independent and model its variation with a linear mixed effect model:

$$\mathbf{f}_{\cdot l} = \mathbf{X}\beta_l + \mathbf{Z}_1\mathbf{a}_{1l} + \mathbf{Z}_2\mathbf{a}_{2l} + \epsilon_l \tag{3}$$

$\beta_l \sim \mathrm{N}_b(\mathbf{0}, \sigma^2_{b_l}(\mathbf{X}'\mathbf{X})^{-1})$	Treatment effect
$\mathbf{a}_{1l} \sim \mathrm{N}_r(\mathbf{0}, \sigma^2_{a_{1l}}\mathbf{A}), \mathbf{a}_{2l} \sim \mathrm{N}_r(\mathbf{0}, \sigma^2_{a_{2l}}\mathbf{A})$	Genetic and GxT effects
$\epsilon_l \sim \mathrm{N}_r(\mathbf{0}, \sigma^2_{e_l}\mathbf{I})$	Residual variation

0.4 Summary

The key features of this specification are:

- Specifying the MEM for $\mathbf{F}$ instead of $\mathbf{Y}$ is a massive reduction of complexity because the number of traits is small $k << p$ and the latent traits are assumed uncorrelated.
- We place a simplex prior on the balance among the variance components as a simplex to constrain the total variance of each column (Figure XX). While factors with shared genetic, treatment and residual components are preferred, purely residual factors are allowed.
- Given a small k, not all variation may be accounted for by the latent factors. We each trait's residual vector (row of $\mathbf{E}$) with a parallel MEM, and place an additional prior π_j on the proportion of variation in that trait accounted for by the k factors.

0.5 Implementation

We have derived a Gibbs sampler to estimate the posterior distribution of all model parameters which we have implemented in R/Rcpp, and also coded the model in Stan.

Data example

0.6 Background

0.7 Results

0.8 Conclusions

Acknowledgements

References

[1] A Bhattacharya and D B Dunson. Sparse Bayesian infinite factor models. *Biometrika*, 98(2):291–306, May 2011.

[2] Daniel Runcie and S Mukherjee. Dissecting High-Dimensional Phenotypes with Bayesian Sparse Factor Analysis of Genetic Covariance Matrices. *Genetics*, 194(3):753–767, July 2013.