UC DAVIS
DEPARTMENT OF PLANT SCIENCES

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

RJ Cody Markelz[1], Sayan Mukherjee[2], Daniel Runcie[3]

[1]Department of Plant Biology and [3]Department of Plant Sciences, UC Davis. [2]Department of Statistical Science, Duke University

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:

- Biological systems tend to be modular in organization.
- Complex experimental or breeding designs and observational studies of natural populations with structure require flexible hierarchical models.

Building on our earlier genetic sparse factor model [3], we extend the specification to allow a full linear mixed effect model to capture complex experimental designs. Using an efficient Bayesian MCMC algorithm, the model 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

High-dimensional phenotypes (ex RNAseq, metabalomcs, hyperspectral imaging) promise to provide an unprecedented window into the complex functions of biological organisms. Settings we have in mind include:

- A set of genetically related individuals assigned to 2+ treatments to assess gene-environment interactions.
- Multi-factor experiments with split-plot or repeated measures designs that are common in agriculture settings.
- Case-control studies with individuals of varying ancestry.

In each case the experimental goals are to identify sets of traits associated with each experimental factor and to understand 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.
- Interactions (ex. Genotype x Treatment) greatly increase the number of parameters for each trait.
- 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, treatments) cause perturbations to the modules of the developmental system.
- The traits we observe are independent conditional on the state of the underlying developmental system.
- 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.

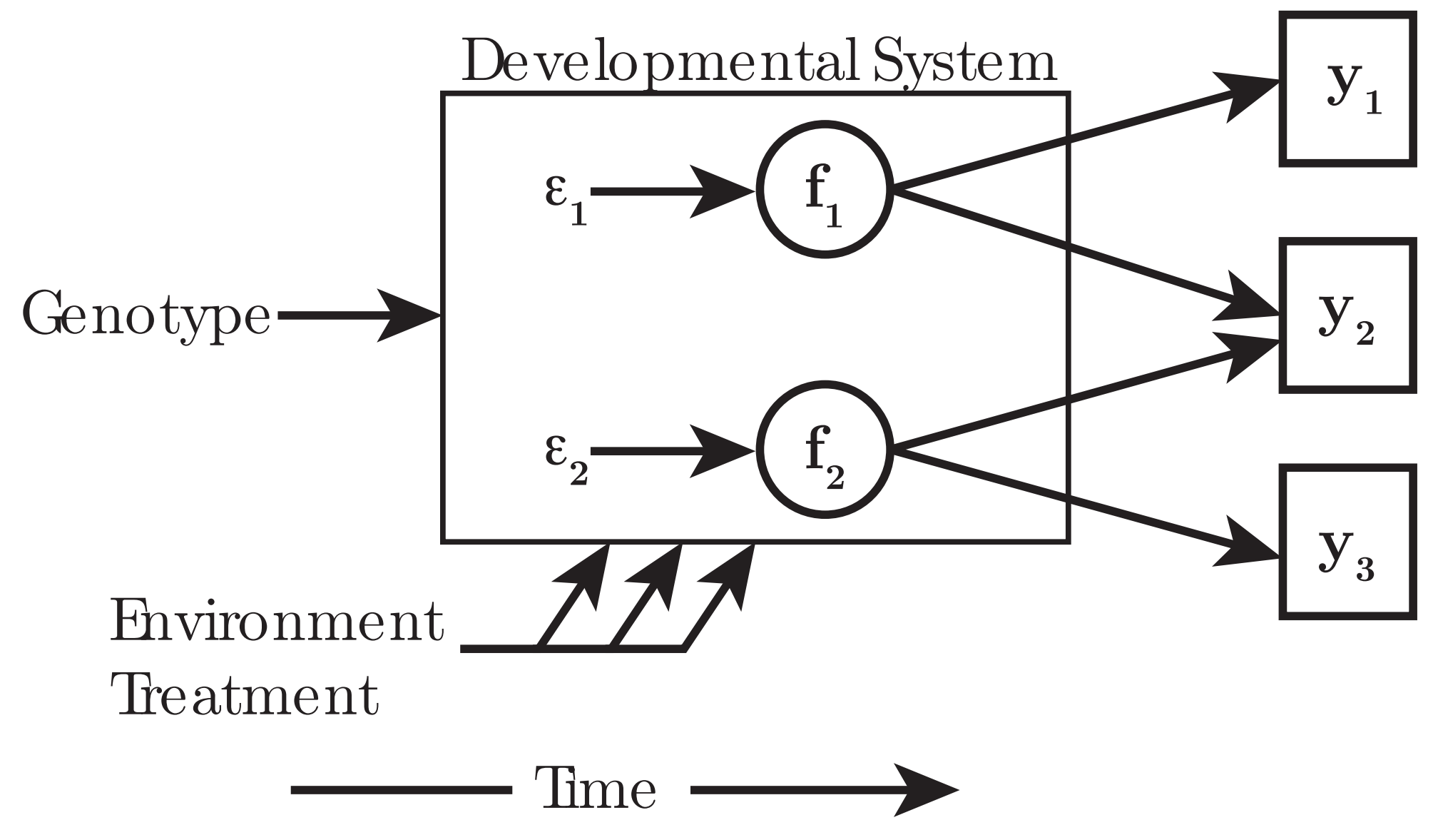

Figure 1. We assume that the influence of genotype and treatment on the observed traits $\mathbf{y}_j$ is mediated through developmental modules $\mathbf{f}_l$.

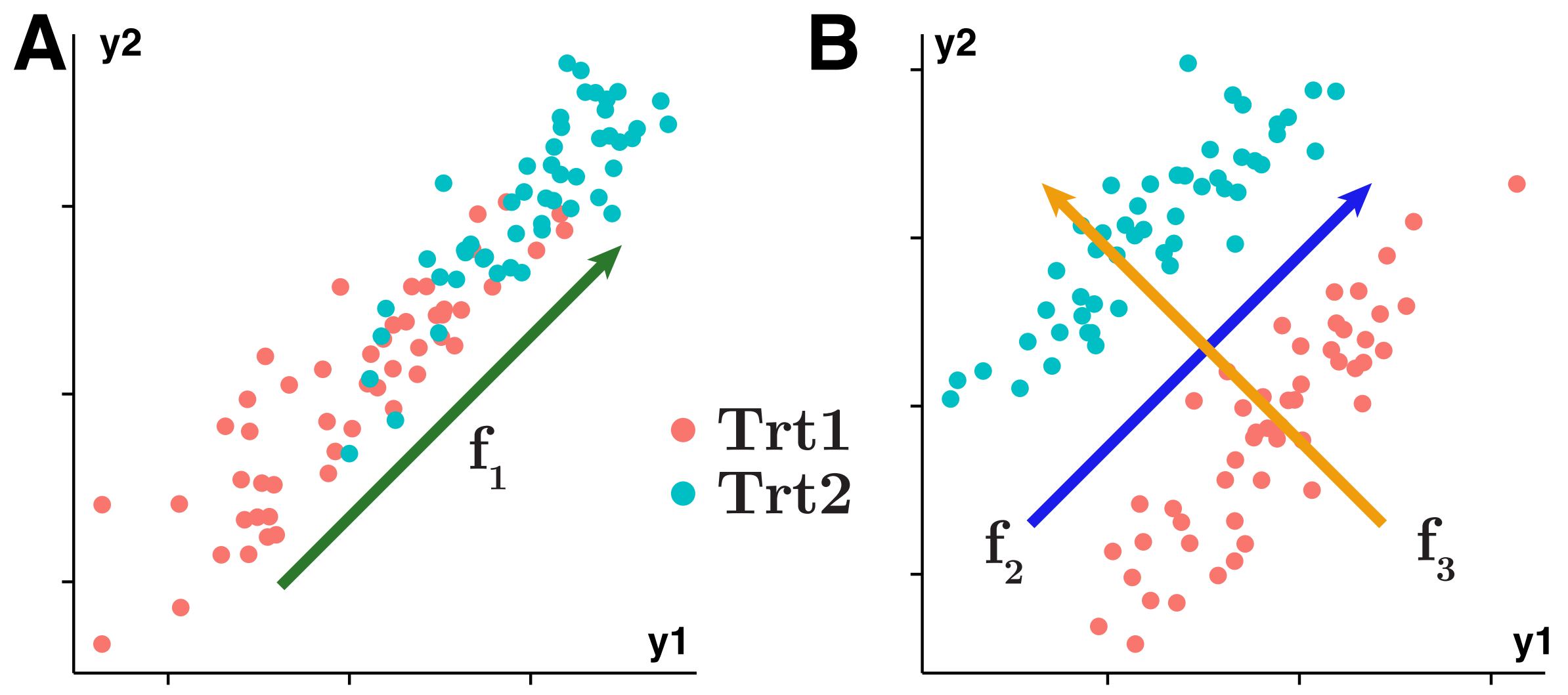

Figure 2. When treatment and residual correlations align (as in **A** but not **B**) the same factor can explain variation in both traits.

Model specification

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:

$$\mathbf{Y} = \mathbf{f}_1\boldsymbol{\lambda}_1' + \mathbf{f}_k\boldsymbol{\lambda}_k' + \dots \mathbf{f}_k\boldsymbol{\lambda}_k' + \mathbf{E} \tag{1}$$

Conditional on the factors, the residuals (rows of $\mathbf{E}$) are independent and assigned Gaussian priors. Sparsity in the factor loadings $\boldsymbol{\lambda}_l$ is induced with an "infinite factor"[1] prior structure, using local/global shrinkage with increasing shrinkage on higher-indexed factors. This induces a ranking of the factors and reduces the burden of pre-selecting the number of factors.

$$\begin{gathered} \lambda_{jl} \sim \mathrm{N}(0, \tau_k^{-1}\psi_{jl}^{-1}) \quad \psi_{jl} \sim \mathrm{Ga}(\nu/2, \nu/2) \\ \delta_1 \sim \mathrm{Ga}(\alpha_0, \beta_0) \quad \delta_l \sim \mathrm{Ga}(\alpha_1, 1) \quad \tau_l = \prod_{i=1}^{k} \delta_l \end{gathered} \tag{2}$$

Mixed Effect Model (MEM)

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

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

$$\beta_l \sim \mathrm{N}_b(\mathbf{0}, \sigma_{b_l}^2(\mathbf{X}'\mathbf{X})^{-1}) \qquad \text{Treatment effect}$$

$$\mathbf{a}_{1l} \sim \mathrm{N}_r(\mathbf{0}, \sigma_{a_{1l}}^2\mathbf{A}), \mathbf{a}_{2l} \sim \mathrm{N}_r(\mathbf{0}, \sigma_{a_{2l}}^2\mathbf{A}) \qquad \text{G and GxT effects}$$

$$\epsilon_l \sim \mathrm{N}_r(\mathbf{0}, \sigma_{e_l}^2\mathbf{I}) \qquad \text{Residual variation}$$

$$\sigma_{b_l}^2 + \sigma_{a_{1l}}^2 + \sigma_{a_{2l}}^2 + \sigma_{e_l}^2 = 1 \tag{4}$$

Summary

The key features of this specification are:

- Specifying the MEM for $\mathbf{f}_l$ instead of $\mathbf{y}_j$ is a massive reduction of complexity because the number of traits is small $k << p$ and the latent traits are assumed uncorrelated.
- While factors with shared genetic, treatment and residual components are preferred (Figure 2), purely residual factors are allowed.
- Given a small k, not all variation may be accounted for by the latent factors. We model 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.

Implementation

We use a Gibbs sampler implemented in R/Rcpp to estimate the posterior distribution of all model parameters and also coded the model in Stan[2].

Acknowledgements

We would like to thank Upendra Kumar Devisetty and Julin Maloof for sharing the gene expression data, NSF grants DMS-1418261 and IIS 1546331 to SM and XX to RJCM and the UC Davis Department of Plant Sciences for funding.

Data example

Background

Results

Our model identified five large factors that accounted for the majority of the covariance in the data.

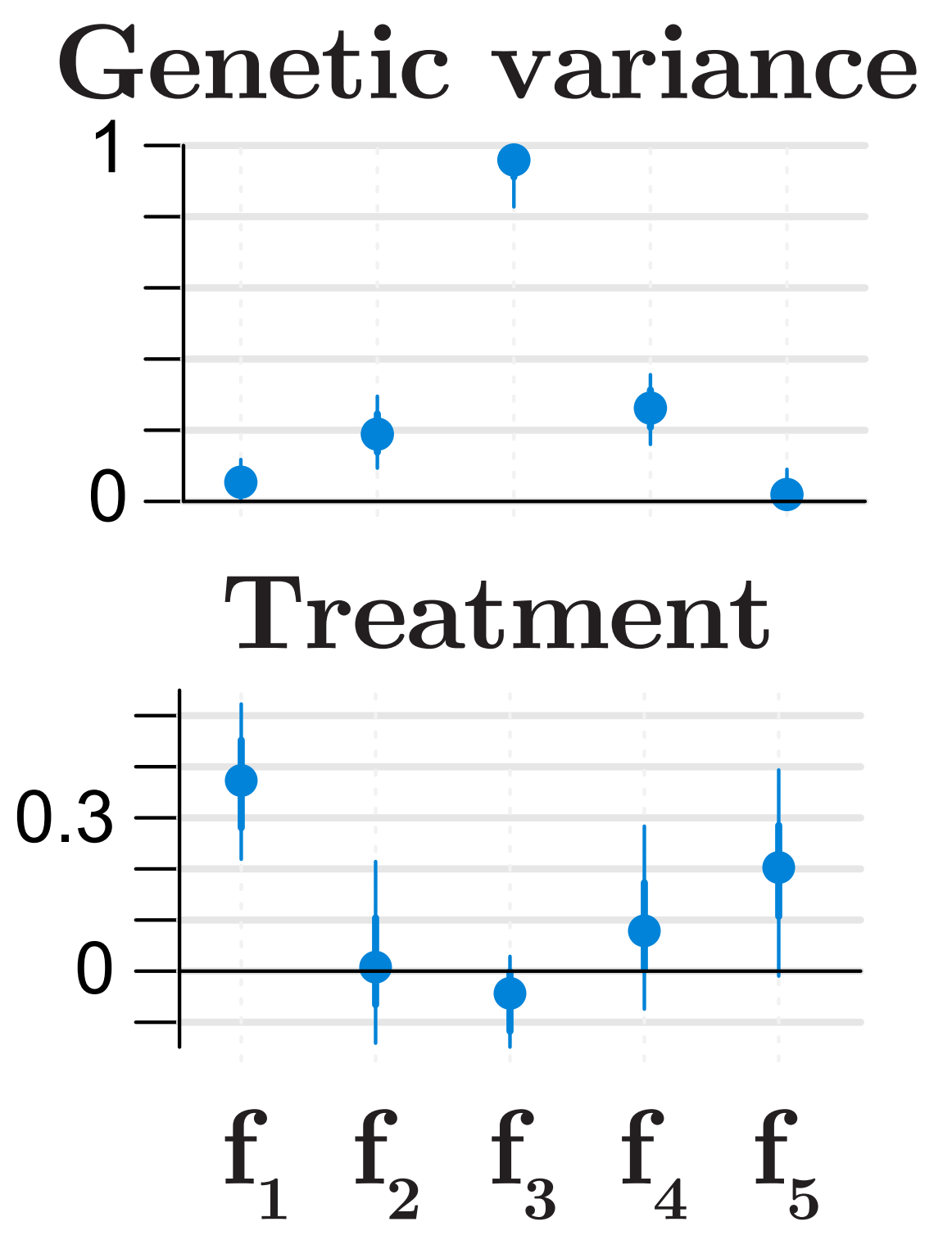

Figure 4. Among-line genetic variance and response to crowded conditions for each of the five major factors.

References

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

[2] Bob Carpenter, Andrew Gelman, Matt Hoffman, Daniel Lee, Ben Goodrich, Michael betancourt, marcus A Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A Probabilistic Programming Language. *J Stat Softw*, 2015.

[3] 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.