Dissecting High-Dimensional Phenotypes with Bayesian Sparse Factor Analysis of Genetic Covariance Matrices

Marco Lopez-Cruz

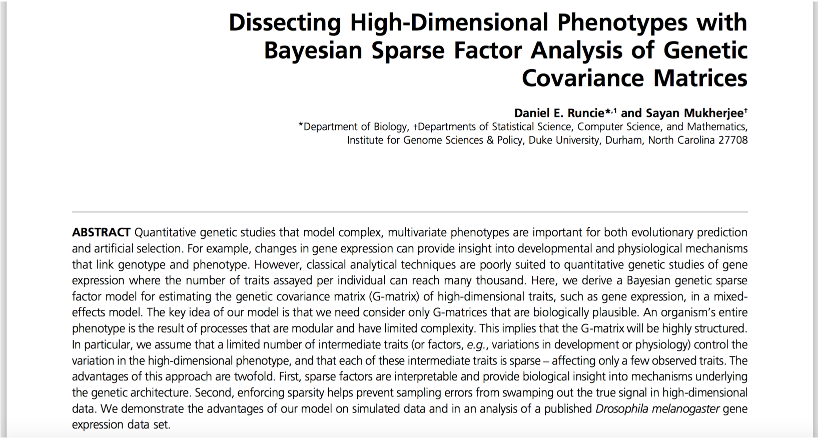
Michigan State University

lopezcru@msu.edu

September 24, 2015

Overview

1. Brief review of Factor Analysis
   - Latent Variable Model
   - Example
2. Introduction
3. Methods
   - Priors
   - Implementation
   - Simulation
   - Gene expression analysis
4. Results
   - Simulation
   - Gene expression analysis
5. Discusion and Conclusion



Dissecting High-Dimensional Phenotypes with Bayesian Sparse Factor Analysis of Genetic Covariance Matrices

Daniel E. Runcie*,[1] and Sayan Mukherjee†
*Department of Biology, †Departments of Statistical Science, Computer Science, and Mathematics,
Institute for Genome Sciences & Policy, Duke University, Durham, North Carolina 27708

ABSTRACT Quantitative genetic studies that model complex, multivariate phenotypes are important for both evolutionary prediction and artificial selection. For example, changes in gene expression can provide insight into developmental and physiological mechanisms that link genotype and phenotype. However, classical analytical techniques are poorly suited to quantitative genetic studies of gene expression where the number of traits assayed per individual can reach many thousand. Here, we derive a Bayesian genetic sparse factor model for estimating the genetic covariance matrix (G-matrix) of high-dimensional traits, such as gene expression, in a mixed-effects model. The key idea of our model is that we need consider only G-matrices that are biologically plausible. An organism's entire phenotype is the result of processes that are modular and have limited complexity. This implies that the G-matrix will be highly structured. In particular, we assume that a limited number of intermediate traits (or factors, *e.g.*, variations in development or physiology) control the variation in the high-dimensional phenotype, and that each of these intermediate traits is sparse – affecting only a few observed traits. The advantages of this approach are twofold. First, sparse factors are interpretable and provide biological insight into mechanisms underlying the genetic architecture. Second, enforcing sparsity helps prevent sampling errors from swamping out the true signal in high-dimensional data. We demonstrate the advantages of our model on simulated data and in an analysis of a published *Drosophila melanogaster* gene expression data set.

Latent Variable Model

Factor analysis is a statistical method used to describe variability among observed, correlated variables in terms of a potentially lower number of unobserved variables called **factors** or **latent** variables.

Latent Variable Model

Suppose we have a set of p observable random variables, $x_1, \ldots, x_p$ on individuals $1, 2, \ldots, n$.

Factor analysis assumes that, for individual i and condition j, x_{ij} is a combination of k ($k < p$) unobserved factors (f), i.e.

$$x_{ij} = \lambda_{j1} f_{i1} + \lambda_{j2} f_{i2} + \ldots + \lambda_{jk} f_{ik} + \delta_{ij}$$

where the λ terms are **factor loadings** to be estimated, and δ_{ij} is the measurement error in x_{ij} (or **idiosyncratic noise**).

Latent Variable Model

In matrix notation:

$$\boldsymbol{x}_i = \boldsymbol{\Lambda} \boldsymbol{f}_i + \boldsymbol{\delta}_i$$

where $\boldsymbol{x}_i$ is a $p \times 1$ vector of observed variables for individual i. $\boldsymbol{\Lambda}$ is a $p \times k$ matrix of factor loadings to be estimated, $\boldsymbol{f}_i$ is a $k \times 1$ vector of scores, and $\boldsymbol{\delta}_i$ is the $p \times 1$ vector of measurement errors.

Assumptions:
$\boldsymbol{\delta}_i \sim N(\mathbf{0}, \boldsymbol{\Psi})$, where $\boldsymbol{\Psi} = diag(\psi_1, \psi_2, \ldots, \psi_p)$.
$\boldsymbol{f}_i \sim N(\mathbf{0}, \boldsymbol{I}_k)$
$\boldsymbol{f}_i$ and $\boldsymbol{\delta}_i$ are independent.

Latent Variable Model

Suppose $\boldsymbol{x}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma}_p)$

Then we have $cov(\boldsymbol{x}_i) = \boldsymbol{\Sigma} = cov(\boldsymbol{\Lambda} \boldsymbol{f}_i + \boldsymbol{\delta}_i)$. This is:

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^{'} + \boldsymbol{\Psi}$$

Having the n individuals in the same equation, we have,

$$\boldsymbol{X} = \boldsymbol{F}\boldsymbol{\Lambda}^{'} + \boldsymbol{\Delta}$$

where $\boldsymbol{X} = [\boldsymbol{x}_1^{'} \ldots \boldsymbol{x}_n^{'}]^{'}$ is a $n \times p$ matrix of observed responses, $\boldsymbol{F} = [\boldsymbol{f}_1^{'} \ldots \boldsymbol{f}_n^{'}]^{'}$ is a $n \times k$ matrix of scores, and $\boldsymbol{\Delta}$ is a $n \times p$ matrix of measurement errors.

Example in R

```
data(mtcars)
(fm <- factanal(mtcars,factors=1))
(fm <- factanal(mtcars,factors=2))
(fm <- factanal(mtcars,factors=3))

Vhat <- fm$loadings%*%t(fm$loadings)+diag(fm$uniquenesses)
V <- cor(mtcars)
```

Introduction

- By scaling classic methods to hundreds or thousands of traits provoke the number of modeling parameters to grow exponentially.
- An unconstrained G-matrix for p traits requires $p(p+1)/2$ parameters.
- Large numbers of parameters can lead to instability in parameter estimates.
- It is expect the G-matrix to be sparse, generally G-matrices that are modular (traits varying together) and of low rank (few modules).

Motivation

Two assumptions are made:

1. A limited number of modules contribute to trait variation.
2. Each module affects a limited number of traits.

Based on this, the authors present a Bayesian sparse factor model for inferring G-matrices for hundreds or thousands of traits, called *Bayesian sparse factor analysis of genetic covariance matrices* or BSFG.

Model

For a single trait, the following linear mixed-effects model is commonly used to explain phenotypic variation (Henderson 1984),

$$\boldsymbol{y}_i = \boldsymbol{X}\boldsymbol{b}_i + \boldsymbol{Z}\boldsymbol{u}_i + \boldsymbol{e}_i$$

where $\boldsymbol{y}_i$ is the vector of observations of the i^{th} trait on n individuals; $\boldsymbol{b}_i$ is the vector of coefficients for fixed effects with design matrix $\boldsymbol{X}$; $\boldsymbol{u}_i \sim N(\mathbf{0}, \sigma^2_{G_i}\boldsymbol{A})$ is the random $r \times 1$ vector of additive genetic effects with incidence matrix $\boldsymbol{Z}$, and $\boldsymbol{e}_i \sim N(\mathbf{0}, \sigma^2_{R_i}\boldsymbol{I}_n)$ is the residual error.

Model

In going from one trait to p traits, we can align the vectors for each trait in the previous model to form the following multivariate model,

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{Z}\boldsymbol{U} + \boldsymbol{E}$$

where $\boldsymbol{Y} = [\boldsymbol{y}_1 \dots \boldsymbol{y}_p]$, $\boldsymbol{B} = [\boldsymbol{b}_1 \dots \boldsymbol{b}_p]$, $\boldsymbol{U} = [\boldsymbol{u}_1 \dots \boldsymbol{u}_p]$.
In addition,

$$\boldsymbol{U} \sim MN_{r,p}(\boldsymbol{0}; \boldsymbol{A}_r, \boldsymbol{G}) \qquad \boldsymbol{E} \sim MN_{n,p}(\boldsymbol{0}; \boldsymbol{I}_n, \boldsymbol{R})$$

where $\boldsymbol{A}$ and $\boldsymbol{I}_n$ specify the covariances of each trait among individuals, and $\boldsymbol{G}$ and $\boldsymbol{R}$ (both to be estimated) specify the additive genetic and residual covariances among traits.

Defining the factor model

$\boldsymbol{U}$ and $\boldsymbol{E}$ are specified via the following hierarchical factor model,

$$\begin{aligned} \boldsymbol{U} &= \boldsymbol{F}_a\boldsymbol{\Lambda}^{'} + \boldsymbol{E}_a, & \boldsymbol{E} &= \boldsymbol{F}_r\boldsymbol{\Lambda}^{'} + \boldsymbol{E}_r \\ \boldsymbol{F}_a &\sim MN_{r,k}(\mathbf{0}; \boldsymbol{A}, \boldsymbol{\Sigma}_a), & \boldsymbol{F}_r &\sim MN_{n,k}(\mathbf{0}; \boldsymbol{I}_n, \boldsymbol{\Sigma}_r) \\ \boldsymbol{E}_a &\sim MN_{r,p}(\mathbf{0}; \boldsymbol{A}, \boldsymbol{\Psi}_a), & \boldsymbol{E}_r &\sim MN_{n,p}(\mathbf{0}; \boldsymbol{I}_n, \boldsymbol{\Psi}_r) \\ \boldsymbol{\Lambda} &\sim \pi(\theta) \end{aligned}$$

Matrices $\boldsymbol{F}_a$ and $\boldsymbol{F}_r$ partition the among-individual variation in the latent factors. $\boldsymbol{\Sigma}_a$ and $\boldsymbol{\Sigma}_r$ model the among-factor (within-individual) covariances of $\boldsymbol{F}_a$ and $\boldsymbol{F}_r$, which are assumed to be diagonal ($\boldsymbol{\Sigma}_a = \text{diag}(\sigma^2_{a_j})$, $\boldsymbol{\Sigma}_r = \text{diag}(\sigma^2_{r_j})$). $\boldsymbol{\Psi}_a$ and $\boldsymbol{\Psi}_r$ are the idiosyncratic (trait-specific) variances.

Defining the factor model

Restrictions: $\boldsymbol{\Sigma}_a + \boldsymbol{\Sigma}_r = \boldsymbol{I}_k$ which means $\boldsymbol{\Sigma}_{h^2} = \boldsymbol{\Sigma}_a = \boldsymbol{I}_k - \boldsymbol{\Sigma}_r$. Note that the diagonal elements of this matrix is the narrow-sense heritability $(h_j^2 = \sigma^2_{a_j}/(\sigma^2_{a_j} + \sigma^2_{r_j}) = \sigma^2_{a_j})$ of latent trait j.

$\boldsymbol{G}$ and $\boldsymbol{R}$ can be written as:

$$\boldsymbol{G} = \boldsymbol{\Lambda}\boldsymbol{\Sigma}_{h^2}\boldsymbol{\Lambda}^{'} + \boldsymbol{\Psi}_a$$
$$\boldsymbol{R} = \boldsymbol{\Lambda}(\boldsymbol{I}_k - \boldsymbol{\Sigma}_{h^2})\boldsymbol{\Lambda}^{'} + \boldsymbol{\Psi}_r$$

Therefore, the model for the total phenotypic covariance $\boldsymbol{P} = \boldsymbol{G} + \boldsymbol{R}$ is:

$$\boldsymbol{P} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^{'} + \boldsymbol{\Psi}_a + \boldsymbol{\Psi}_r$$

Priors

To impose sparsity (shrinkage toward zero) and low effective rank, we assume a prior on $\boldsymbol{\Lambda}$ specified as a hierarchical distribution, commonly used to impose sparsity (Neal 1996, Tipping 2001), on each element λ_{ij}:

$$
\begin{aligned}
&\lambda_{ij}|\phi_{ij} \sim N(0, \phi_{ij}^{-1}\tau_j^{-1}), \qquad i = 1, \ldots, p, \quad j = 1, \ldots, k \\
&\phi_{ij} \sim \text{Ga}(v/2, v/2) \\
&\tau_j = \prod_{l=1}^{m} \delta_l \\
&\delta_1 \sim \text{Ga}(a_1, b_1), \quad \delta_l \sim \text{Ga}(a_2, b_2) \quad \text{for } l = 2, \ldots, k
\end{aligned}
$$

where τ_j is a parameter that controls the overall variance explained by factor j.

Priors

The prior on the heritability of each of latent factor trait is a discrete set of values in the unit interval, which gives equal weight to $h_j^2 = 0$ and $h_j^2 > 0$.

Also, inverse gamma priors with parameters a_a, b_a and a_r, b_r are imposed on each diagonal element of $\boldsymbol{\Psi}_a$ and $\boldsymbol{\Psi}_r$, respectively.

Priors on each element of $\boldsymbol{B}$ are normal distributions with very large (> 106) variances.

Implementation

Inference in the model uses an adaptive Gibbs sampler.

The code has been implemented in Matlab and can be found at the website `http://www.stat.duke.edu/~sayan/bfgr/index.html`.

The model BSFG was tested on simulated and real data coming from a published study.

Case 1: Simulation

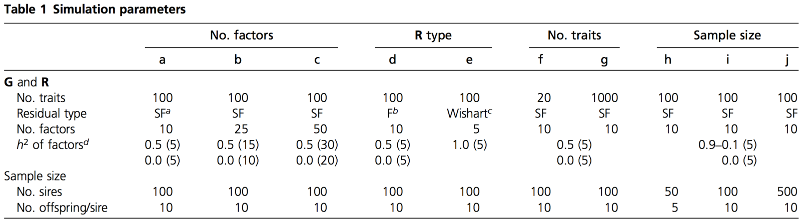


Table 1 Simulation parameters

	No. factors			R type		No. traits		Sample size		
	a	b	c	d	e	f	g	h	i	j
G and **R**										
No. traits	100	100	100	100	100	20	1000	100	100	100
Residual type	SF[a]	SF	SF	F[b]	Wishart[c]	SF	SF	SF	SF	SF
No. factors	10	25	50	10	5	10	10	10	10	10
h^2 of factors[d]	0.5 (5) 0.0 (5)	0.5 (15) 0.0 (10)	0.5 (30) 0.0 (20)	0.5 (5) 0.0 (5)	1.0 (5)	0.5 (5) 0.0 (5)		0.9–0.1 (5) 0.0 (5)		
Sample size										
No. sires	100	100	100	100	100	100	100	50	100	500
No. offspring/sire	10	10	10	10	10	10	10	5	10	10

Scenarios

- a–c test covariance matrices ($\boldsymbol{G}$ and $\boldsymbol{R}$) composed of different numbers of factors.
- d–e test residual covariance matrices $\boldsymbol{R}$ that are not sparse.
- f–g test different numbers of traits and latent traits heritabilities.
- h–j test different sample sizes.

Simulation

All simulations followed a paternal half-sib breeding design. Each simulation was run 10 times.

To fit the simulated data, the hyperparameters were set to: $v = 3$, $a_1 = 2$, $b_1 = 1/20$, $a_2 = 3$, $b_2 = 1$.

Gibbs sampler was run for $12,000$ iterations, discarded the first $10,000$ samples as burn-in, and collected $1,000$ posterior samples with a thinning rate of two.

Simulation

Accuracy of the method was compared to a method of moments estimate of $\boldsymbol{G}$ calculated as $\boldsymbol{G}_m = 4(\boldsymbol{B} - \boldsymbol{W})/n$, where $\boldsymbol{B}$ and $\boldsymbol{W}$ are the between- and within-sire matrices of mean squares and cross products and n is the number of offspring per sire.

The accuracy of the moments estimator $\boldsymbol{G}_m$ to the posterior mean $\hat{\boldsymbol{G}}$ from BSFG model by calculating the Frobenius norm of the errors: $|\boldsymbol{G}_m - \boldsymbol{G}|$ and $|\hat{\boldsymbol{G}} - \boldsymbol{G}|$.

Case 2: Gene expression analysis

Data from 414 genes, plus competitive fitness measured on flies from 40 lines of *Drosophila melanogaster* from ArrayExpress (`http://dgrp.gnets.ncsu.edu/`).

Dataset was normalized to correspond to the analyses of Ayroles *et al*. (2009). They included fixed effect of *sex* and independent random effects of the *sex:line* interaction for each gene.

Gibbs sampler was run for 40,000 iterations (same hyperparameters as before), discarded the first 20,000 as a burn-in, and collected 1000 posterior samples (thinning=20).

Simulation

The BSFG model′s estimates of genetic covariances were considerably more accurate than estimates based on unbiased methods of moments estimators.

Simulation

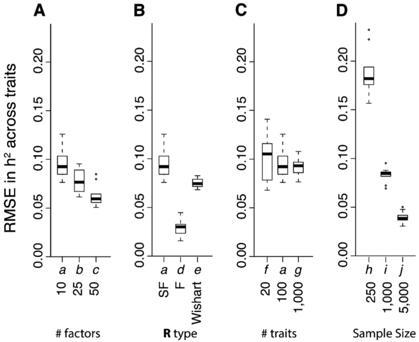


Figure: The heritability of each individual trait was calculated as $h_i^2 = \boldsymbol{G}_{ii}/\boldsymbol{P}_{ii}$. (A) Increasing numbers of simulated factors. (B) Different types of ***R*** matrices. (C) Different numbers of traits. (D) Different numbers of sampled individuals.

Gene expression analysis

Estimate of the G-matrix was qualitatively similar to the original estimate.

Estimates of the broad-sense heritability of each gene were also similar ($r = 0.74$).

Using the modulated modularity clustering (MMC), Ayroles et al. (2009) identified 20 modules of genetically correlated transcripts post hoc. BSFG model identified 27 latent factors, of which 13 were large factors.

Many factors were similar to the modules identified by MMC.

Simulation

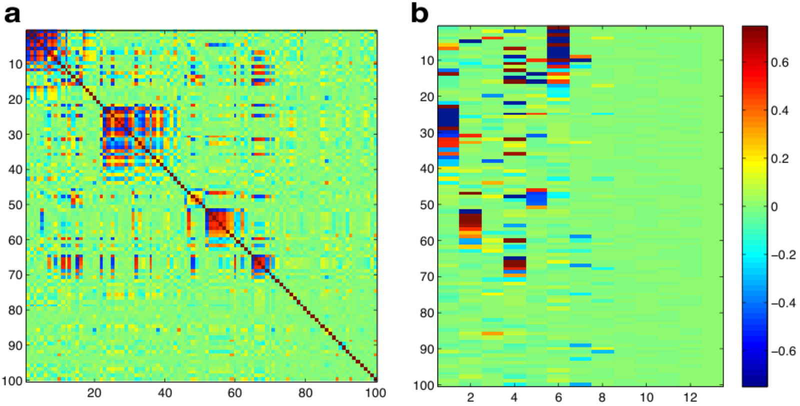


Figure: a) Posterior mean genetic covariance matrix. b) Posterior mean factor loadings.

Gene expression

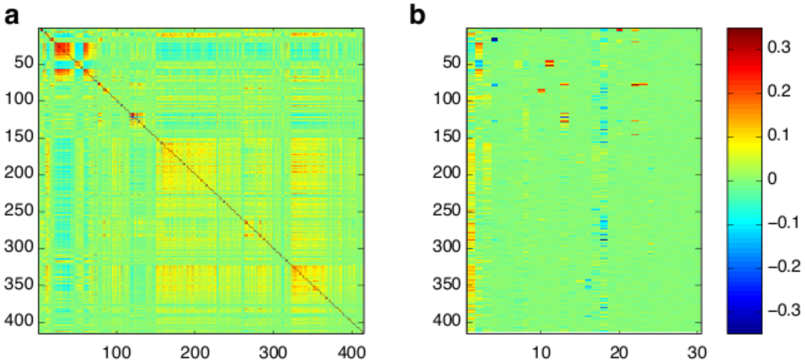


Figure: a) Posterior mean genetic covariance matrix. b) Posterior mean factor loadings.

Gene expression

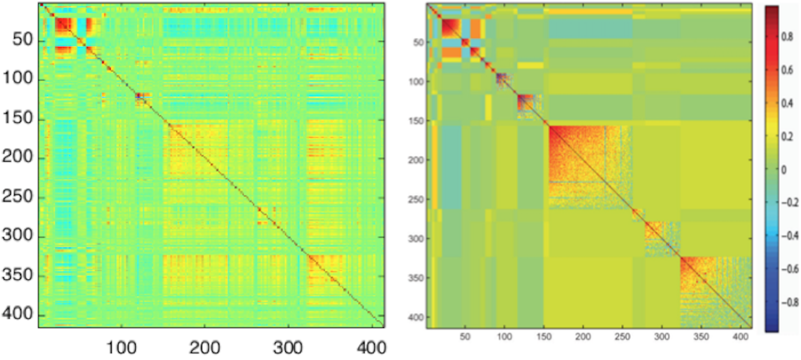


Figure: Estimated covariance matrix (left). Original estimated (right).

Discusion and Conclusion

The BSFG model performs well on both simulated and real data and opens the possibility of incorporating high-dimensional traits into evolutionary genetic studies and breeding programs.

Technologies for high-dimensional phenotyping are becoming widely available in evolutionary biology so methods for modeling such traits are needed.

Gene expression traits in particular provide a way to measure underappreciated molecular and developmental traits that may be important.

References

Ayroles, J. F., M. A. Carbone, E. A. Stone, K. W. Jordan, R. F. Lyman et al (2009)
Systems genetics of complex traits in Drosophila melanogaster
Nat Genet 41(3), 299 – 307.

Runcie D, Mukherjee S (2013)
Dissecting High-Dimensional Phenotypes with Bayesian Sparse Factor Analysis of Genetic Covariance Matrices
Genetics 194, 753 – 767.

The End