Latent Variable Methods for the Analysis of Genomic Data

http://genomine.org/talks/
Data – $m$ variables and $n$ observations

- high-dimensional variables
  - observations
  - gene expression
  - SNP genotypes

- primary vars
  - observations
  - intercept + treatment
  - intercept + clinical outcome
  - intercept + trait
High-dimensional regression model

\[ Y = B X + E \]

Goal is to do inference on each row of \( B \), which connects variation in \( X \) to each row of \( Y \)
Example steps for simultaneous inference

1. Formulate model $Y = BX + E$, which has null version $Y = B_0X + E$

2. Fit models to obtain $\hat{B}$ and $\hat{B}_0$

3. Compare goodness of fits $\hat{B}$ and $\hat{B}_0$ for each genomic variable to obtain $m$ test statistics

4. Calculate a null distribution to obtain $m$ p-values

5. Perform some type of multiple testing correction (such as FDR or local FDR)
Incorporating latent variables

1. Formulate model $Y = BX + E$, which has null version $Y = B_0X + E$

2. Fit models to obtain $\hat{B}$ and $\hat{B}_0$

3. Compare goodness of fits $\hat{B}$ and $\hat{B}_0$ for each genomic variable to obtain $m$ test statistics

4. Calculate a null distribution to obtain $m$ p-values

5. Perform some type of multiple testing correction (such as FDR or local FDR)

We have been developing models, theory, and methods that replaces $Y = BX + E$ with $Y = BX + \Phi M + E$ where $M$ are low-dimensional latent variables.
Known variable + latent variable model

\[ g(\mathbb{E}[Y|X,M]) = B X + \Phi M \]

\( M \) accounts for dependence, unknown sources of systematic variation, structure, etc.
Gene expression studies

\[ g(E[Y|X, M]) = B X + \Phi M \]

- \( Y_{m \times n} \) gene expression data,
- \( X_{d \times n} \) study design matrix
- \( M \) batch effects, biological heterogeneity, etc.
- \( Y|X, M \sim \) Normal, Poisson, or NegBin
Genome-wide association studies

\[
g(E[X|Y,M]) = B Y + \Phi M
\]

\(X_{m \times n}\) SNP genotypes, \(Y_{1 \times n}\) trait variable
\(M\) population structure
\(X|Y,M \sim \text{Binomial}\)
Surrogate variable analysis (SVA)

Introduced in Leek and Storey (2007) *PLoS Genetics*

\[
E[Y \mid X, M] = B X + \Phi M
\]

\[Y_{m \times n} = (y_{ij})\] response variables where \(y_{ij} \in \mathbb{R}\)

\[X_{d \times n}\] study design matrix

\(Y \mid X, M \sim \text{Normal}\) (not necessary, but helpful)
Generalized surrogate variable analysis (gSVA)

\[ g(E[Y|X,M]) = BX + \Phi M \]

- \( Y_{m \times n} \) response variables
- \( X_{d \times n} \) study design matrix
- \( Y|X, M \sim \text{exponential family distribution} \)
- \( g(\cdot) \) link function


Data

\[ X_{m \times n} = (x_{ij}) \]
\[ x_{ij} \in \mathbb{R} \]

Gene expression

\[ Y_{d \times n} \]

Study design
Motivating study: “Inflammation and the Host Response to Injury”

Leukocyte mRNA Expression

Clinical Data
~400 clinical variables

Patient 1
Patient 2
....
Patient 168

MOF multiple organ failure score

Leukocyte mRNA Expression

Clinical Data
~400 clinical variables
Poor replication

Phase 1

Phase 3

Phase 2

Phase 4
Accounting for latent variables

**Phase 1**

**Phase 2**

**Phase 3**

**Phase 4**

**Surrogate Variable Analysis**
SVA model

\[
\mathbb{E}[Y|X, M] = B X + \Phi M
\]

\(Y_{m \times n}\) gene expression data
\(X_{d \times n}\) study design matrix
\(Y|X, M \sim \text{Normal}\)
Normal model

\[ Y = BX + E \]

\[ e^j \overset{iid}{\sim} N(0, \Sigma) \]
Leek and Storey (2007) introduce the SVA model, motivated as a way to capture unmodeled systemic variation

Leek and Storey (2008) show that the SVA model also captures general forms of dependence among variables

Desai and Storey (2012) study the SVA model under the Normal distribution and derive a penalized maximum likelihood approach for estimating it called “cross-dimensional inference” (CDI)

Storey, Desai, and Chen (2016) show that CDI applied to the SVA model yields the Bayes estimator under the multivariate Normal scenario with Normal prior on $B$

Storey, Desai, and Chen (2016) provide an empirical Bayes CDI approach for estimating the SVA model
Part 2. Genotype-conditional association test


Data

\[ X_{m \times n} = (x_{ij}) \]
\[ x_{ij} \in \{0, 1, 2\} \]
SNP genotypes

\[ y = (y_1, y_2, \ldots, y_n) \]
trait of interest
Trait models

Quantitative trait: \[ y_j = \alpha + \sum_{i=1}^{m} \beta_i x_{ij} + \lambda_j + \epsilon_j \]

Binary trait: \[ \log \left( \frac{\Pr(y_j=1)}{\Pr(y_j=0)} \right) = \alpha + \sum_{i=1}^{m} \beta_i x_{ij} + \lambda_j \]

\( \beta_i \) is the genetic effect of SNP \( i \) on the trait
\( \lambda_j \) is the random non-genetic effect
\( \epsilon_j \sim N(0, \sigma_j^2) \) is the random noise variation
Confounding in GWAS

\[ \pi_1(z) \rightarrow \text{SNP } x_1 \]
\[ \pi_2(z) \rightarrow \text{SNP } x_2 \]
\[ \cdots \]
\[ \pi_i(z) \rightarrow \text{SNP } x_i \]
\[ \cdots \]
\[ \pi_m(z) \rightarrow \text{SNP } x_m \]

\[ \text{Structure, Lifestyle, Environment } z \]

\[ \text{"non-genetic" } \lambda(z) \]

\[ \text{Trait } y \]
Confounding in GWAS

**Quantitative trait:** \( y_j = \alpha + \sum_{i=1}^{m} \beta_i x_{ij} + \lambda_j + \epsilon_j \)

**Binary trait:** \( \log\left( \frac{\Pr(y_j=1)}{\Pr(y_j=0)} \right) = \alpha + \sum_{i=1}^{m} \beta_i x_{ij} + \lambda_j \)

\( x^j = (x_{1j}, \ldots, x_{mj})^T \), \( \lambda_j \), and \( \sigma_j^2 \) [from \( \epsilon_j \sim N(0, \sigma_j^2) \)] may all be functions of \( z_j \)
Perform a hypothesis test of $b_i = 0$ vs. $b_i \neq 0$ in the following model:

$$\logit \left( \frac{E[x_{ij}|y_j, z_j]}{2} \right) = a_i + b_i y_j + \logit(\pi_{ij})$$

1. formulate and estimate a linear basis for $\logit(\pi_{ij})$, called logistic factors
2. fit a logistic regression of SNP $i \sim \text{trait} + \text{logistic factors}$
3. perform a generalized likelihood ratio test of the trait’s coefficient
Theorem

**Quantitative trait:** \[ y_j = \alpha + \sum_{i=1}^{m} \beta_i x_{ij} + \lambda_j + \epsilon_j \]

**Binary trait:** \[ \log \left( \frac{\Pr(y_j=1)}{\Pr(y_j=0)} \right) = \alpha + \sum_{i=1}^{m} \beta_i x_{ij} + \lambda_j \]

**GC model:** \[ \logit \left( \frac{\mathbb{E}[x_{ij}|y_j,z_j]}{2} \right) = a_i + b_i y_j + \logit(\pi_{ij}) \]

Suppose that trait values \( y_j \) are distributed according to either of the above trait models and SNPs \( x_{ij} | \pi_{ij} \sim \text{Binomial}(2, \pi_{ij}) \) as described earlier. Then \( \beta_i = 0 \) in either trait model implies that \( b_i = 0 \) in the GC model.
Latent variable model

\[ \logit(\mathbb{E}[X|Y, M]/2) = B Y + \Phi M \]

\[ X|Y, M \sim \text{Binomial} \]
Part 3. Jackstraw method

Focus on latent variable part of the model

\[ Y = \Phi M + E \]
How to do inference on $\Phi$?

$Y = \Phi M + E$
Estimating $M$

Let $Y$ be row-wise mean-centered and $Y = \Phi M + E$. SVD of $Y = UDV^T$, where $V_r^T$ is first $r$ rows of $V^T$.

Leek (2010) *Biometrics* shows that under reasonable assumptions:

$$\left\| [V_r^T V_r - \hat{\sigma}_{\text{avg}}^2 I] - MM^T \right\|_{\text{rowspace}} \to 0$$

as number of variables (rows of $Y$) goes to $\infty$ for fixed sample size (columns of $Y$).

In other words, $V_r^T$ is a reasonable estimate of $M$. 
Y = ΦM + E
= ΓV_T + E'

We perform inference on the rows of Γ: γ_1, ..., γ_m.

Suppose we want to test γ_i = 0 vs. γ_i ≠ 0 for each i = 1, ..., m.
Preserve structure and create synthetic nulls

- $Y$: $n$ observations and $m$ variables
- $Y^*$: Permute $s$ rows
- $V^T_r$: F-statistics
- $V'^T_r$: F-statistics
Preserve structure and create synthetic nulls

$n$ observations

$Y$

$m$ variables

$V_r^T$

Permute $s$ rows

$Y^*$

$V_r^{*T}$

Compare to get $p$-values
Yeast cell cycle analysis

(a) Principal Component Values

(b) Percent Variance Explained

(c) Minutes

0 100 200 300 400
−0.4 0.0 0.2 0.4

Minutes Post–Elucidation

1st PC
2nd PC

0 5 10 15 20 25

Percent Variance Explained

2 4 6 8 10 12

Principal Component
Consistently estimating $M$

$\mathbf{Y}_{m \times n} = (y_{ij})$ distributed according to a single parameter exponential family distribution

We show how to estimate with probability 1 the row-space of $M$ and consistently estimate its dimension
Acknowledgements

Funding:
National Institutes of Health
Office of Naval Research

Co-authors:
Xiongzhi “Chee” Chen
Neo Chung
Keyur Desai
Wei Hao
Jeff Leek
Minsun Song

Lab members:
Andrew Bass
Irineo Cabreros
Chee Chen
Sean Hackett
Wei Hao
Emily Nelson
Alejandro Ochoa