Principal Component of Explained Variance

High-Dimensional Estimation and Inference

Maxime Turgeon

December 11th, 2017

McGill University
Department of Epidemiology, Biostatistics, and Occupational Health

Introduction

- B.Sc. (UOttawa) et M.Sc. (McGill) en Mathématiques
    - Théorie des nombres, sous la direction de Jayce Getz
- Sur le point de terminer (i.e. Printemps 2018) mon doctorat en Biostatistique sous la direction de Celia Greenwood et Aurélie Labbe
- Biostatisticien sénior à l'Agence de Santé de la Saskatchewan

Saskatchewan Health Authority

- Supervise une équipe composée de deux biostatisticiens juniors
- Utilise des bases de données administratives pour répondre à des questions de recherche et d'amélioration qualitative
- Support analytique et consultation méthodologique
  - Saskatchewan Centre for Patient-Oriented Research
  - Clinical Quality Improvement Program
  - First Nation and Métis Health

Introduction

- In modern statistics, we often encounter multivariate variables of large dimension ($p > n$).
    - In biomedical sciences (e.g. neuroimaging, genomics), pattern recognition, text recognition, finance, etc.
- We are often faced with the following problem:
    - Given two sets of multivariate variables $\{W_1, \ldots, W_p\}$ and $\{Z_1, \ldots, Z_q\}$, **how do we test for global association, and how do we identify which variables drive the association?**

Introduction

- Regression: $E(W|Z) = \beta Z$.
  - The regression parameter β controls the global association **and** the contribution of each Z.
- Regularized regression can also be used to detect sparse signals.
- However, this framework can be cumbersome when W has dimension greater than one, especially when we have heterogeneous variable types (e.g. continuous and categorical).

Motivating Examples

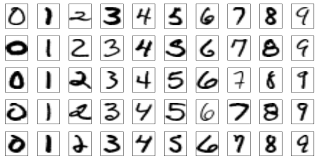
Motivating Examples

The next examples have the following in common:

We have a (possibly high-dimensional) multivariate vector $\mathbf{Y}$ and a set of covariates X.

We are interested in low dimensional representations of $\mathbf{Y}$ that **summarise** the relationship between $\mathbf{Y}$ and X.

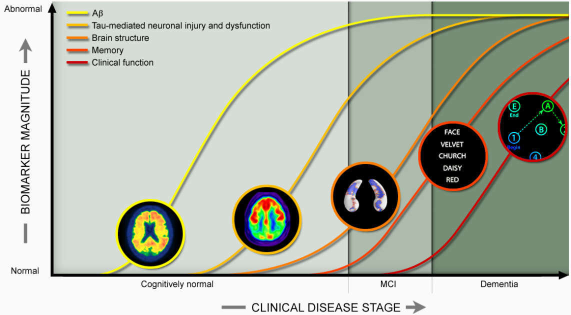
Motivating Example #1



- *Digit recognition*: A famous example in machine learning coming from Le Cun *et al.* (1990).
- Consists of 16×16 gray scale images of digits (i.e. 256 pixels), where the goal is to automatically identify the digit.
- $\mathbf{Y}$ is the set of gray scale values for each pixel, and X is the digit to which the image corresponds
- We would like to extract lower-dimensional features to use for prediction.

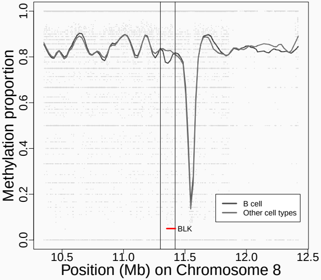
Motivating Example #2

- Data from 340 participants from the Alzheimer's Disease Neuroimaging Initiative (ADNI)
- Brain imaging was employed to assess amyloid-β protein load in 96 brain regions
- $\mathbf{Y}$ is the set of Aβ load values for each brain region, and X is the (binary) disease status.



Motivating Example #3

- The dataset consists of 40 blood samples, separated into different cell types (T cells, B cells, monocytes), and for which methylation levels were measured at 24,000 locations along the genome.
- **Y** is the set of DNA methylation values for all 24,000 locations, and X is the cell type.



Principal Component of Explained Variance (PCEV)

- Provides an **optimal** strategy for selecting a low dimensional summary of $\mathbf{Y}$ that can be used to test for association with one or several covariates of interest.
- **Goal**: Find the linear combination (or component) that maximises the *proportion of variance explained by the covariates*

Contributions

1. Estimation strategies
2. Analytical framework for hypothesis testing
    - High-dimensional inference
3. An R package implementing this method (pcev available on CRAN)

Methods

PCEV: Statistical Model

Let $\mathbf{Y}$ be a multivariate outcome of dimension p and X, a vector of covariates.

We assume a linear relationship:

$$\mathbf{Y} = \beta^T X + \varepsilon.$$

The total variance of the outcome can then be decomposed as

$$\begin{aligned} \mathrm{Var}(\mathbf{Y}) &= \mathrm{Var}(\beta^T X) + \mathrm{Var}(\varepsilon) \\ &= V_M + V_R. \end{aligned}$$

PCEV: Statistical Model

Decompose the total variance of $\mathbf{Y}$ into:

1. Variance explained by the covariates;
2. Residual variance.

PCEV: Statistical Model

The PCEV framework seeks a linear combination $w^T\mathbf{Y}$ such that the proportion of variance explained by X is maximised; this proportion is defined as the following Rayleigh quotient:

$$h(w) = \frac{w^T V_M w}{w^T (V_M + V_R) w}.$$

A solution to this maximisation problem can be obtained through a combination of Lagrange multipliers and linear algebra.

Key observation: $h(w)$ measures the strength of the association

More Dimension Reduction

- **PCA**: Maximise total variance
- **CCA**: Maximise correlation
- **PLS**: Maximise covariance
- **RDA**: Maximise redundancy index
- **PCEV**: Maximise proportion of variance explained

All these methods (except PCA) have serious limitations with high-dimensional data.

Block-diagonal Estimator

We propose a **block approach** to the computation of PCEV in the presence of high-dimensional outcomes.

- Suppose the outcome variables can be divided in blocks of variables in such a way that
    - Variables **within** blocks are correlated
    - Variables **between** blocks are uncorrelated

$$\mathrm{Cov}(\mathbf{Y}) = \begin{pmatrix} * & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & * & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & * \end{pmatrix}$$

Block-diagonal Estimator

- We can perform PCEV on each of these blocks, resulting in a PCEV for each block.
- Treating all these “partial” PCEVs as a new, multivariate pseudo-outcome, we can perform PCEV again; the result is a linear combination of the original outcome variables.

With the above assumption, I showed that this is **mathematically equivalent** to performing PCEV in a single-step.

Finally, we can compute p-values using a permutation procedure.

Simulations

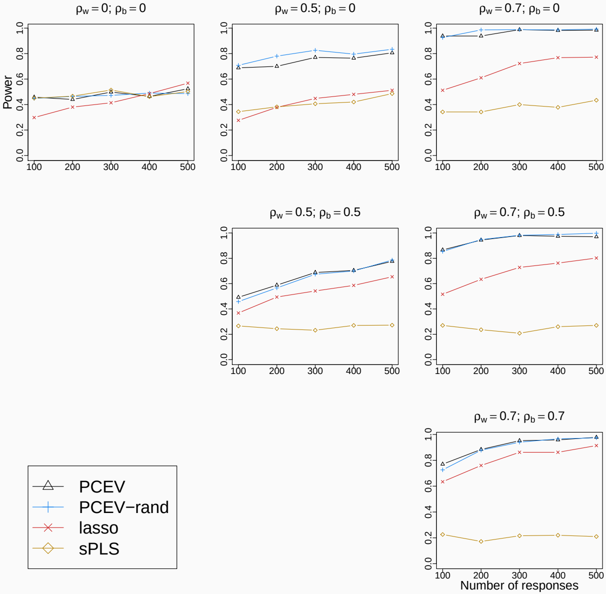
Simulation Setting

- We compared 4 different approaches:
    - PCEV-block, with blocks assumed known a priori
    - PCEV-block, with blocks selected randomly
    - Lasso
    - Sparse Partial Least Squares (sPLS)
- We fixed the sample size at $n = 100$ and simulated $p = 100, 200, 300, 400, 500$ outcomes; we distributed the outcome variables in 10 blocks.
- We also varied the correlation between (ρ_b) and within (ρ_w) blocks (0, 0.5, 0.7).
- We simulated a single continuous covariate from a standard normal distribution. 25% of the outcomes in each block are associated with X.

Simulation Setting

- Whereas PCEV treats the multivariate, p-dimensional $\mathbf{Y}$ as the outcome variable and X as the covariate, we inverted these roles for both Lasso and sPLS, so that variable selection happens on $\mathbf{Y}$.
- The test statistics for Lasso and sPLS were as follows:
  - **Lasso**: Correlation between X and $\hat{\beta}_L\mathbf{Y}$
  - **sPLS**: Maximised covariance
- P-values were computed using a permutation procedure.

Simulation Results: Power analysis



Data analysis

Motivating Example #2

- Recall: Data on amyloid-β accumulation in 96 brain regions, measured on 340 subjects. We are interested in the association with Alzheimer's disease.
- We used this dataset to compare the block approach to the traditional approach
- We defined blocks using hierarchical clustering.

Results

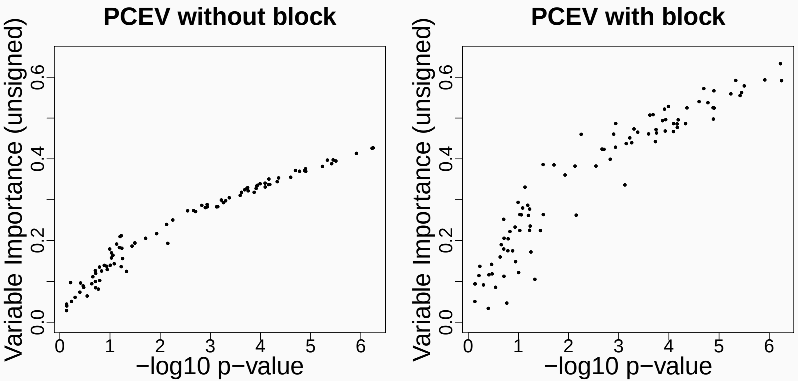
P-values for the joint association between amyloid-β accumulation and disease status. Permutation tests were performed using 100,000 permutations.

	PCEV	PCEV with blocks
Exact test	8.13×10^{-5}	—
Permutation test	2×10^{-5}	5×10^{-5}

Variable Importance Factor

- **VIF**: Correlation between a single variable Y_i in $\mathbf{Y}$ and the PCEV component.
- VIF allows us to decompose the global association into individual components; the higher the VIF, the stronger the contribution of an individual variable.

Variable Importance Factor

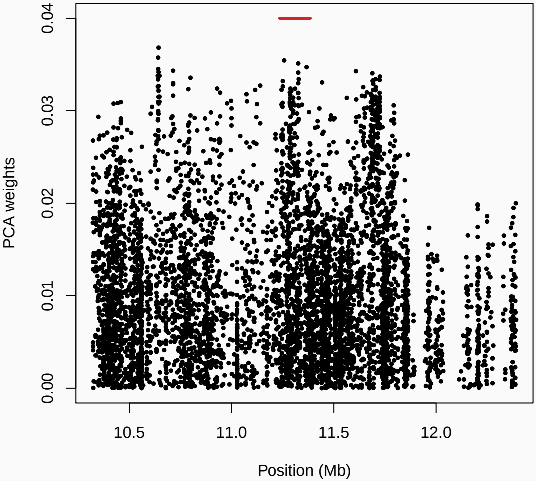


Motivating Example #3

- BLK gene, located on chromosome 8
- Data provided by Tomi Pastinen (McGill)
- 40 blood samples, from 3 different cell types
  - B cells (n=8)
  - T cells (n=19)
  - Monocytes (n=13)
- 24,068 locations on the DNA

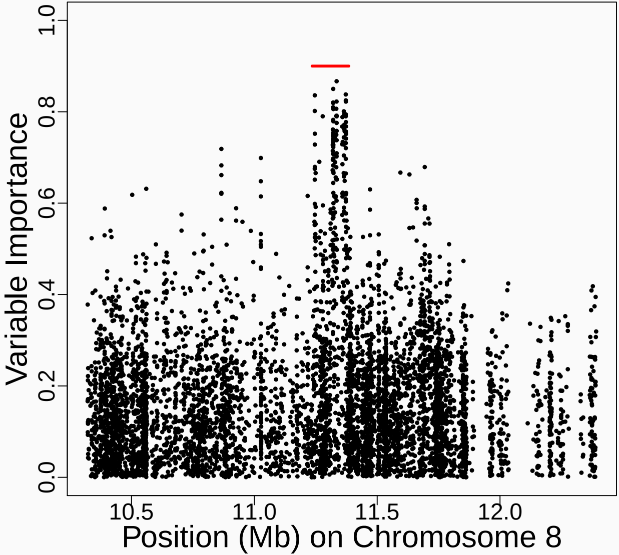
Goal: Investigate the association between methylation levels in the BLK region (outcomes) and cell type (covariate: B cell vs T cell and monocytes)

PCA



Results

- We used the block approach, where blocks were defined using physical distance: CpGs within 500kb are grouped together
  - 951 blocks were analysed
- Using PCEV, we obtained a single p-value, which is less than 6×10^{-5} (using 100,000 permutations)
- Hence, a single test for all variables, and no tuning parameter was required.


Variable Importance
0.0
0.2
0.4
0.6
0.8
1.0
10.5
11.0
11.5
12.0
Position (Mb) on Chromosome 8

Summary

- The block approach has good power compared to common high-dimensional methods
- Results are robust to how blocks are defined
    - P-values are similar
    - Power is similar
    - Variable Importance Factors are also similar

High-dimensional inference

Double Wishart Problem

- Recall that PCEV is maximising a Rayleigh quotient:

$$h(w) = \frac{w^T V_M w}{w^T (V_M + V_R) w}.$$

- This approach is equivalent to finding the largest root λ of a *double Wishart problem*:

$$\det\left(\mathbf{A} - \lambda(\mathbf{A} + \mathbf{B})\right) = 0,$$

where $A = V_M, B = V_R$.

Double Wishart Problem

There are many well-known examples of double Wishart problems:

- Multivariate Analysis of Variance (MANOVA);
- Canonical Correlation Analysis (CCA);
- Testing for independence of two multivariate samples;
- Testing for the equality of covariance matrices of two independent samples from multivariate normal distributions;
- **Principal Component of Explained Variance (PCEV)**.

In all the examples above, the largest root λ summarises the strength of the association.

Contributions

In what follows:

1. I will explain how to solve the double Wishart problem in a high-dimensional setting. (Already present in the pattern recognition literature)
2. I will provide an empirical estimate of the distribution of the largest root of the determinantal equation. This estimate can be used to compute valid p-values and perform high-dimensional inference. (**Original contribution**)

I illustrate this approach using PCEV, but it is applicable to **any** double Wishart problem (e.g. CCA and RDA).

Singular Value Decomposition

From the theory of SVD, we know there exists an orthogonal matrix T such that

$$D := T^T (V_R + V_M) T$$

is diagonal.

When $p > n$, the diagonal matrix D is **singular**, with rank $r < p$.

Solution: Focus only on the nonzero diagonal elements.

Reduced-Rank SVD

Let $\tilde{T} = T_{[r]} D_{[r]}^{-1/2}$. Therefore we get:

$$\tilde{T}^T \left(V_R + V_M\right) \tilde{T} = I_r.$$

Similarly, we can diagonalise $\tilde{T}^T V_M \tilde{T}$ via an orthogonal transformation S:

$$S^T \left(\tilde{T}^T V_M \tilde{T}\right) S = \Lambda.$$

The largest root λ of the double Wishart problem is the largest element on the diagonal of Λ.

Note: the vector w maximising the proportion of variance $h(w)$ is the column of $\tilde{T}S$ corresponding to the largest root.

Inference

There is evidence in the literature that the null distribution of the largest root λ should be related to the **Tracy-Widom distribution**.

Theorem

(Johnstone 2008) Assume $\mathbf{A} \sim W_p(\Sigma, m)$ *and* $\mathbf{B} \sim W_p(\Sigma, n)$ *are independent, with* Σ *positive-definite and* $\boldsymbol{n} \leq \boldsymbol{p}$*. As* $p, m, n \to \infty$*, we have*

$$\frac{\operatorname{logit} \lambda - \mu}{\sigma} \xrightarrow{\mathcal{D}} TW(1),$$

where $TW(1)$ *is the Tracy-Widom distribution of order 1, and* μ, σ *are explicit functions of* p, m, n*.*

Inference

- However, Johnstone's theorem requires an invertible matrix.
- **Grinek & Turgeon (In preparation)**: The null distribution of λ is asymptotically equal to that of the largest root of a scaled Wishart (based on results by Srivastava).
  - The null distribution of the largest root of a Wishart is also related to the Tracy-Widom distribution.
- More generally, random matrix theory suggests that the Tracy-widom distribution is key in central-limit-like theorems for random matrices.

Empirical Estimate

We propose to obtain an empirical estimate as follows:

Estimate the null distribution

1. Perform a small number of permutations (~ 50) on the rows of $\mathbf{Y}$;
2. For each permutation, compute the largest root statistic.
3. Fit a location-scale variant of the Tracy-Widom distribution.

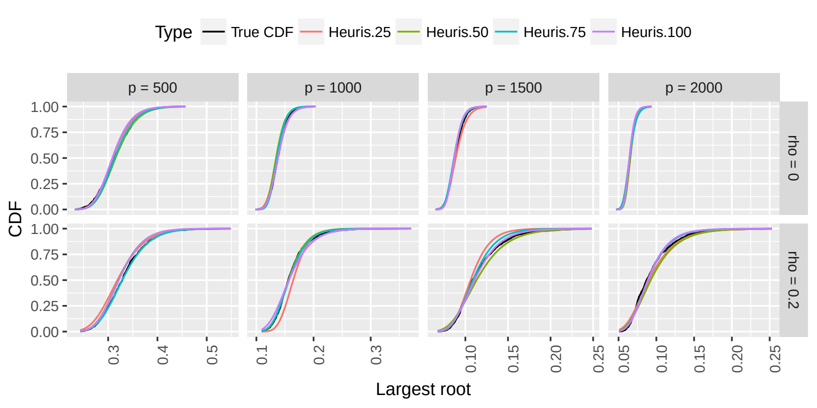
Numerical investigations support this approach for computing p-values. The main advantage over a traditional permutation strategy is the computation time.

Simulations

Distribution Estimation

- We generated 1000 pairs of Wishart variates $\mathbf{A} \sim W_p(\Sigma, m)$, $\mathbf{B} \sim W_p(\Sigma, n)$ with $m = 96$ and $n = 4$ fixed
  - MANOVA: this would correspond to four distinct populations and a total sample size of 100
- We varied $p = 500, 1000, 1500, 2000$
- We looked at two different covariance structures: $\Sigma = I_p$, and an exchangable correlation structure with parameter $\rho = 0.2$.
- We looked at four different numbers of permutations: $K = 25, 50, 75, 100$.
- We compared graphically the CDF estimated from the empirical estimate with the true CDF

Distribution Estimation

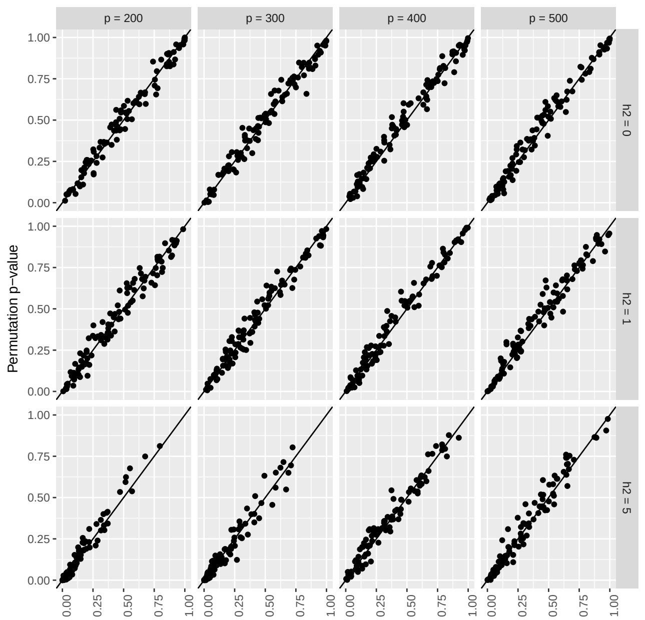


P-value Comparison

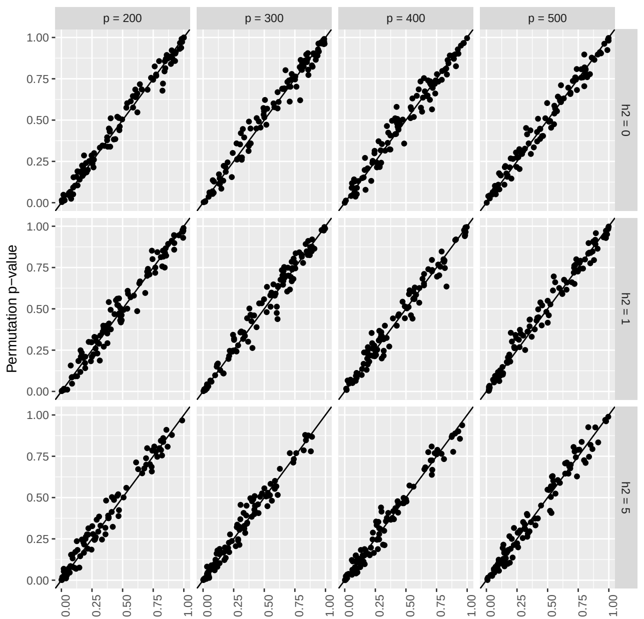
We looked at the following high-dimensional simulation scenario:

- We fixed $n = 100$ and a balanced binary covariate X.
- We varied the number of response variables $p = 200$, 300, $400, 500$ and the association between X and the first 50 response variables in $\mathbf{Y}$.
- We assumed a block structure for the covariance:
  - The p observations are grouped in 10 independent blocks of equal size and the correlation between two variables located in the same block varied with $\rho = 0, 0.5$.
- We compared the empirical estimate with a permutation procedure (250 permutations).
- Each simulation was repeated 100 times.

P-value Comparison: No correlation



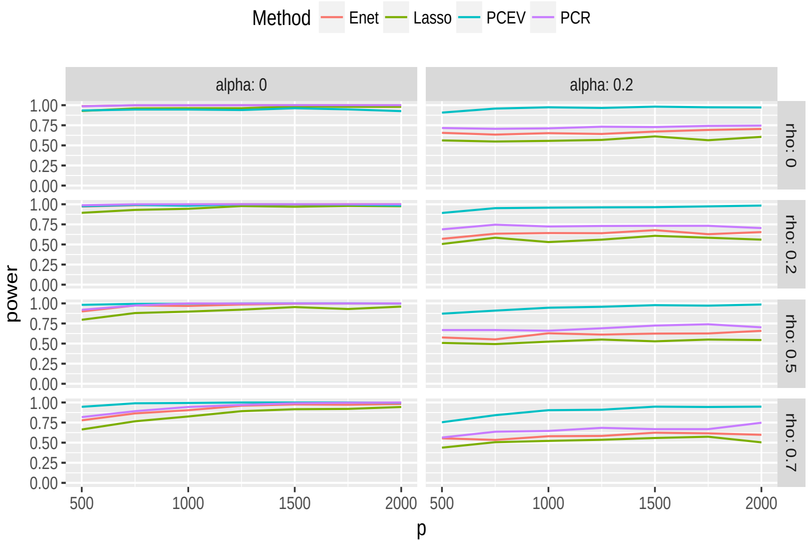
P-value Comparison: Mild correlation



Simulation setting

- We compared 4 different approaches:
  - PCEV with reduced-rank SVD
  - Lasso
  - Elastic net
  - Principal Component Regression
- We simulated $p = 500, 750, \ldots, 2000$ outcomes, 100 observations, one binary covariate.
- Covariance structure is block-diagonal:
  - 10 uncorrelated blocks of equal size
  - Within block is autoregressive (with parameter ρ) with baseline correlation α
- 25% of the outcomes in each block are associated with the covariate, with a fix effect size of 0.333.

Simulation results: Power analysis



Data Analysis

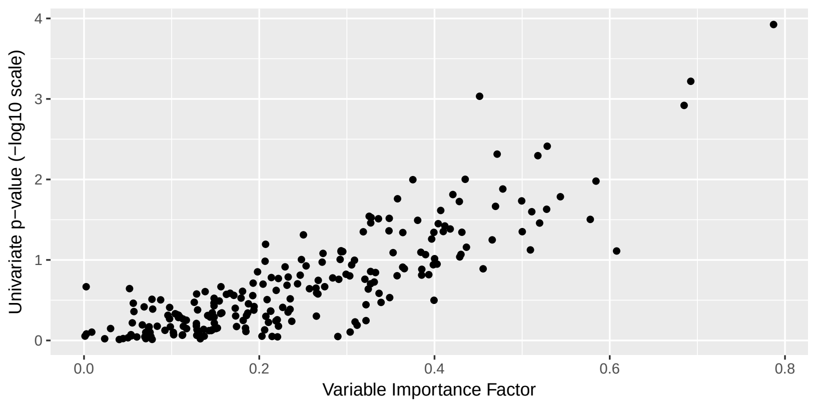
Data

- DNA methylation measured with Illumina 450k on 28 cell-separated samples
- We focus on Monocytes only.
- 18 patients suffering from Rheumatoid arthritis, Lupus, Scleroderma
- We group locations by biological KEGG pathways
    - The number of genomic locations per pathway ranged from 39 to 21,640, with an average around 2000 dinucleotides.
    - 134,941 CpG dinucleotides were successfully matched to one of 320 KEGG pathways
    - On average, each locations appears in 4.5 pathways ⇒ effectively 70 independent hypothesis tests

Results

Description	P-value	P-value (permutation)
Glutamatergic synapse	1.91×10^{-4}	7.00×10^{-4}
Ras signaling pathway	1.33×10^{-3}	1.40×10^{-3}
Circadian rhythm	1.52×10^{-3}	1.00×10^{-4}
Histidine metabolism	1.59×10^{-3}	3.00×10^{-4}
Pathogenic E. coli infection	1.65×10^{-3}	5.20×10^{-3}

Results



path:hsa00120—Glutamatergic synapse: Comparison of VIF and univariate p-values for the most significant pathway.

Conclusion

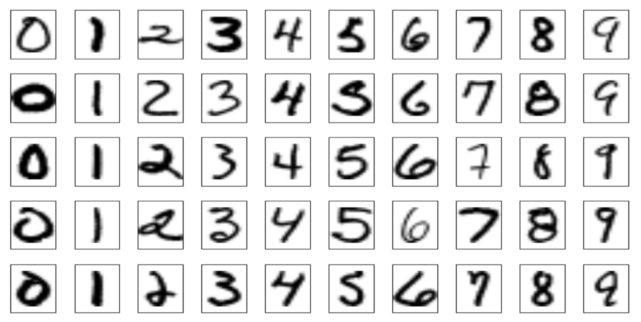
- Data summary is an important feature in data analysis, and this is the objective of dimension reduction techniques.
- Principal Component of Explained Variance is an interesting alternative to PCA
    - It is optimal in capturing the association with covariates
- In a high-dimensional setting, **estimation** and **inference** are more challenging
    - Estimation: Reduced-rank SVD, or block-diagonal estimator
    - Inference: Fitted location-scale Tracy-Widom, or permutation strategy.

Conclusion

- Our approach is computationally simple and provides good power.
- Simulations and data analyses confirm its advantage over a more traditional approach using PCA, as well as other high-dimensional approaches such as regularized regression and sparse PLS.
- The empirical estimate of the distribution of λ has already been succesfully applied to another double Wishart problem (test of covariance equality).
- Everything presented today has been implemented in an `R` package called `pcev` (available on CRAN).

Motivating Example #1

- PCEV could be used to extract features from data and possibly increase predictive accuracy.
- However, there is evidence in the literature that linear features have limited predictive power in pattern recognition.
- We would therefore need a nonlinear variant of PCEV



Future Work

- Investigate the data mining capabilities of PCEV
    - To some extent already in process, with genomic data
- Look into nonlinear alternatives to PCEV
- Extend results on empirical estimate to different variance estimators
    - Preliminary results with Ledoit-Wolf linear shrinkage estimator are promising
- Study the correlation of *pairs* of largest root statistics
    - Obtain less conservative Bonferroni corrections

Acknowledgements

- Karim Oualkacha (UQAM)
- Antonio Ciampi (McGill University)
- Stepan Grinek (BC Cancer Agency)
- Celia Greenwood (McGill University)
- Aurélie Labbe (HEC Montréal)

Funding for this project was provided by CIHR, FQR-NT, and the Ludmer Centre for Neuroinformatics and Mental Health.

Questions or comments?

For more information and updates, visit `maxturgeon.ca`.