Reduced-Rank Singular Value Decomposition for Dimension Reduction with High-Dimensional Data

Maxime Turgeon

June 12th, 2017

McGill University
Department of Epidemiology, Biostatistics, and Occupational Health

Acknowledgements

- Stepan Grinek (BC Cancer Agency)
- Celia Greenwood (McGill University)
- Aurélie Labbe (HEC Montréal)

Introduction

- Modern genomics bring an abundance of high-dimensional, correlated measurements $\mathbf{Y}$.
- We are interested in describing the relationship between such a $\mathbf{Y}$ and a set of covariates X.
- Our approach is to **summarise** this relationship using the largest root λ of a *double Wishart problem*:

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

Double Wishart Problem

There are many well-known examples:

- 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)**.

Main contribution

In this work:

1. We explain how to solve the double Wishart problem in a high-dimensional setting.
2. We provide a heuristic for assessing the significance of the largest root of the determinantal equation.

In what follows, we illustrate this approach using PCEV, but it is applicable to **any** double Wishart problem (e.g. CCA).

Methods

PCEV: Statistical model

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

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:

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

For the corresponding Wishart problem, we have

$$A = V_M, B = V_R.$$

We also have $\lambda = \max_w h(w)$.

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**.

- Johnstone: $(\log(\lambda) - \mu)/\sigma \to TW$ when $p < n$.
- **Turgeon et al.**: The null distribution of λ is asymptotically the same as the largest root of a scaled Wishart.
    - The null distribution of the largest root of a Wishart is also related to TW.
- More generally, random matrix theory suggests that the Tracy-widom distribution is key in central-limit-like theorem for random matrices.

Inference – Heuristic

Estimate the null distribution

1. Perform a small number of permutations (~ 25) 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.

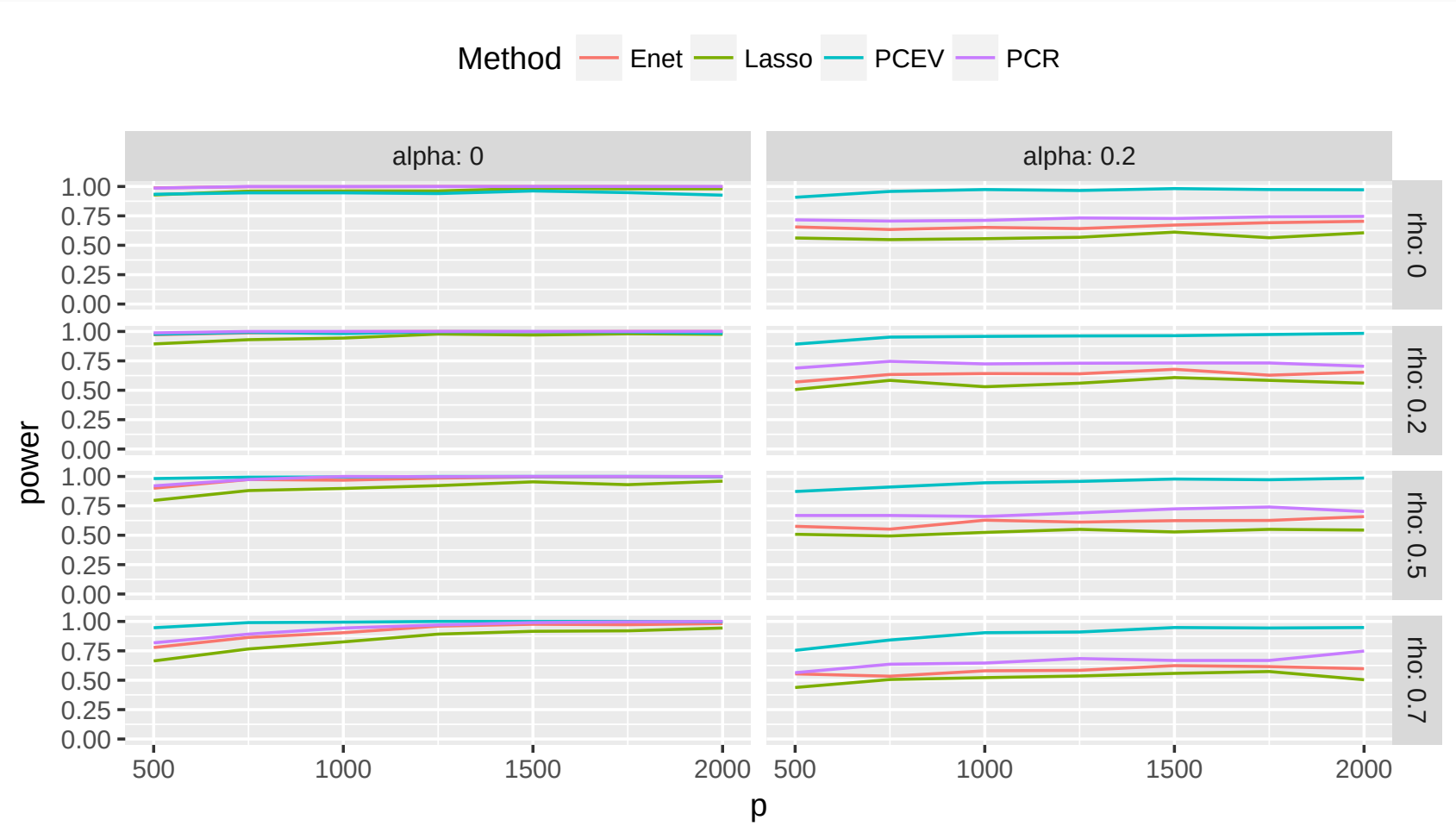
Numerical investigations support this approach for computing p-values. The main advantage over a traditional permutation strategy is the computation time.

Simulations

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

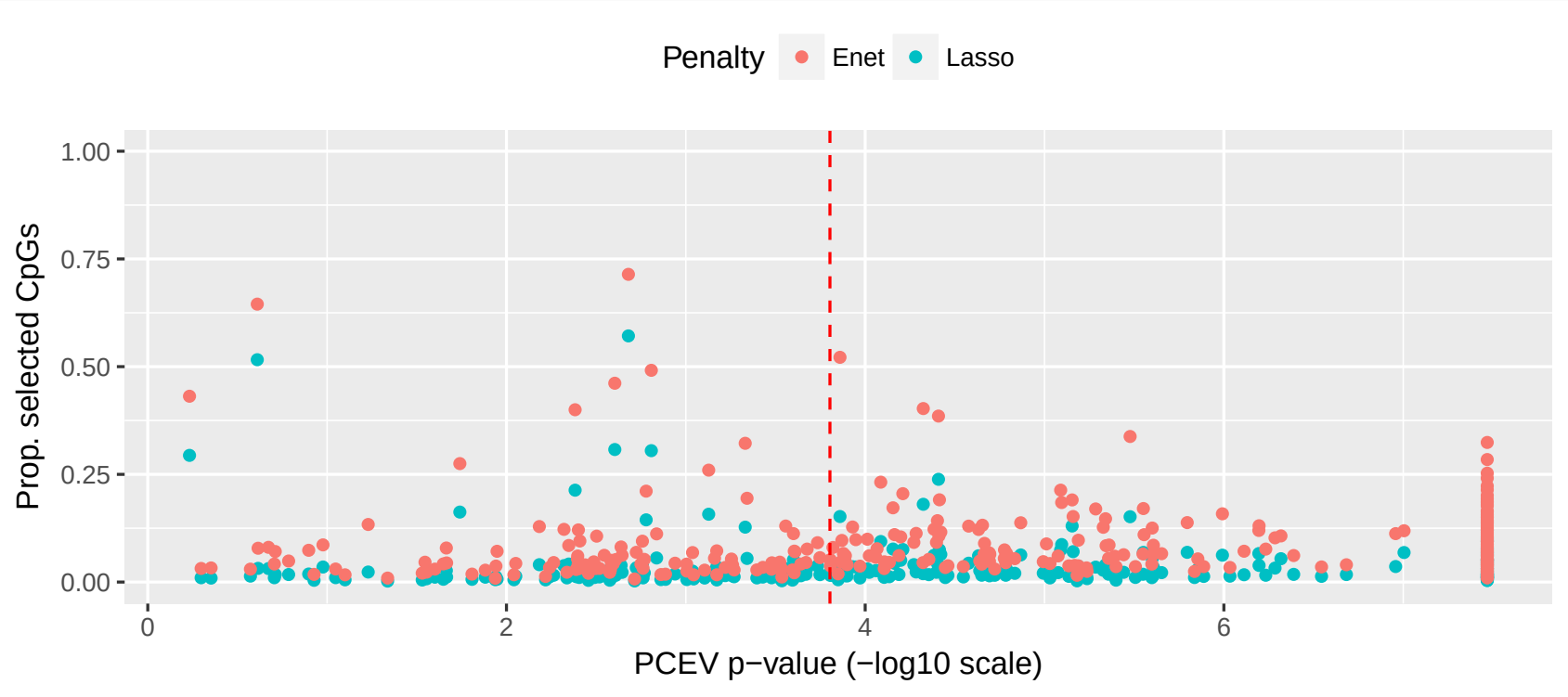


Data analysis

Data

- DNA methylation measured with Illumina 450k on 120 cell-separated samples
- We focus on Monocytes only.
- 18 controls; 35 Rheumatoid arthritis, 24 Lupus, 43 Scleroderma
- We group CpGs by KEGG pathways
    - On average about 1500 CpGs per pathway; max of 21,800.
- We compare PCEV to Lasso and Elastic-net.

Results


Penalty
Enet
Lasso
1.00
0.75
0.50
0.25
0.00
Prop. selected CpGs
0
2
4
6
PCEV p−value (−log10 scale)

Results

Pathway	PCEV pvalue	Lasso Prop.	Enet Prop.
Vitamin B6 metabolism	$< 3.4 \times 10^{-8}$	0.12	0.32
Primary bile acid biosynthesis	$< 3.4 \times 10^{-8}$	0.10	0.28
Fatty acid biosynthesis	$< 3.4 \times 10^{-8}$	0.07	0.25
Ascorbate and aldarate metabolism	$< 3.4 \times 10^{-8}$	0.10	0.24
Steroid biosynthesis	$< 3.4 \times 10^{-8}$	0.08	0.22
Glycosphingolipid biosynthesis	$< 3.4 \times 10^{-8}$	0.06	0.21
Histidine metabolism	$< 3.4 \times 10^{-8}$	0.07	0.20
Thiamine metabolism	$< 3.4 \times 10^{-8}$	0.10	0.19
Folate biosynthesis	$< 3.4 \times 10^{-8}$	0.10	0.19
Other types of O-glycan biosynthesis	$< 3.4 \times 10^{-8}$	0.09	0.19

Conclusion

- Data summary is an important feature in data analysis, and this can be achieved using dimension reduction techniques.
- In a high-dimensional setting, **estimation** and **inference** are more challenging
  - Estimation: Reduced-rank SVD;
  - Inference: Fitted location-scale Tracy-Widom.
- 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 Lasso and Elastic-net regression.

Questions or comments?

For more information and updates, visit

`maxturgeon.ca`.