A Tracy-Widom Empirical Estimator For Valid P-values With High-Dimensional Datasets

Maxime Turgeon

10 August 2019

University of Manitoba
Departments of Statistics and Computer Science

Motivating Example

Systemic Autoimmune Diseases

- Systemic Autoimmune diseases, e.g. Rheumatoid arthritis, Lupus, Scleroderma, impact many systems at once.
- We want to study the association between DNA methylation and these diseases
- To account for the complex biological architecture, we want to measure the association at the *genetic pathway level*
- **High-Dimensional Data**

How can we efficiently compute valid p-values?

High-dimensional inference

Double Wishart Problem

- Many multivariate methods involve maximising a Rayleigh quotient:
$$R^2(w) = \frac{w^T A w}{w^T (A + B) w}.$$
- This approach is equivalent to finding the largest root λ of a *double Wishart problem*:
$$\det\left(\mathbf{A} - \lambda(\mathbf{A} + \mathbf{B})\right) = 0.$$

Double Wishart Problem

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;

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

Contributions

The main contribution:

1. 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.

Two `R` packages implement this method: `pcev` and `covequal` (both available on CRAN)

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.
- The null distribution of λ is asymptotically equal to that of the largest root of a scaled Wishart (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).
    - The actual procedure is problem-specific.
2. For each permutation, compute the largest root statistic.
3. Fit a location-scale variant of the Tracy-Widom distribution.

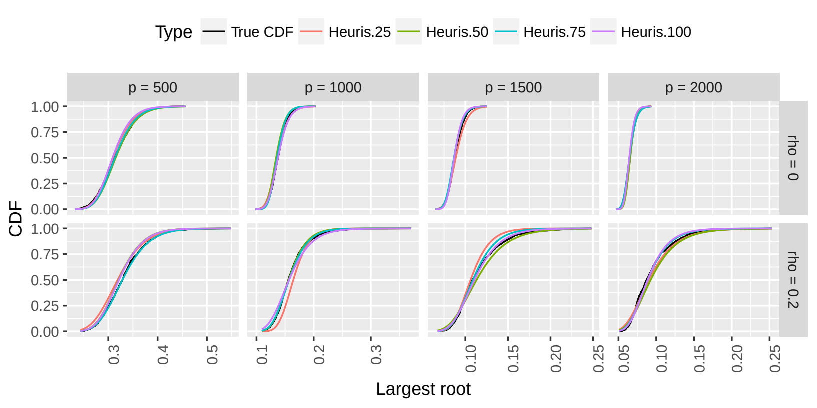
Numerical investigations support this approach for computing p-values. The main advantage over a traditional permutation strategy is the computation time.

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 exchangeable correlation structure with parameter $\rho = 0.2$.
- We looked at four different numbers of permutations for the empirical estimator: $K = 25, 50, 75, 100$.
- We compared graphically the CDF estimated from the empirical estimate with the true CDF

Distribution Estimation



P-value Comparison

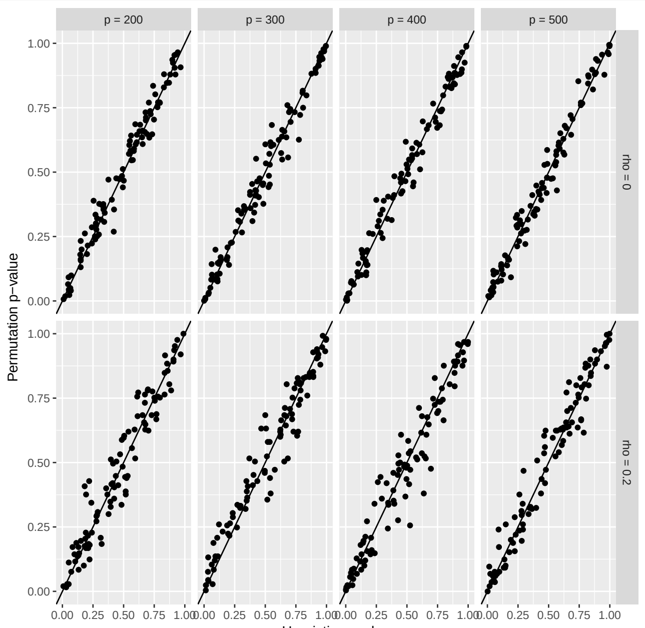
We looked at the following high-dimensional simulation scenario:

- We fixed $n = 100$.
- We generated $X \sim N_p(0, I_p)$ and $\mathbf{Y} \sim N_p(0, \Sigma)$, with $p = 200$, 300, $400, 500$.
- We selected an autocorrelation structure Σ:

$$\mathrm{Cov}(Y_i, Y_j) = \rho^{|i-j|}, \qquad \rho = 0, 0.2$$

- We compared the empirical estimate with a permutation procedure (250 permutations).
- Each simulation was repeated 100 times.

P-value Comparison



Data Analysis

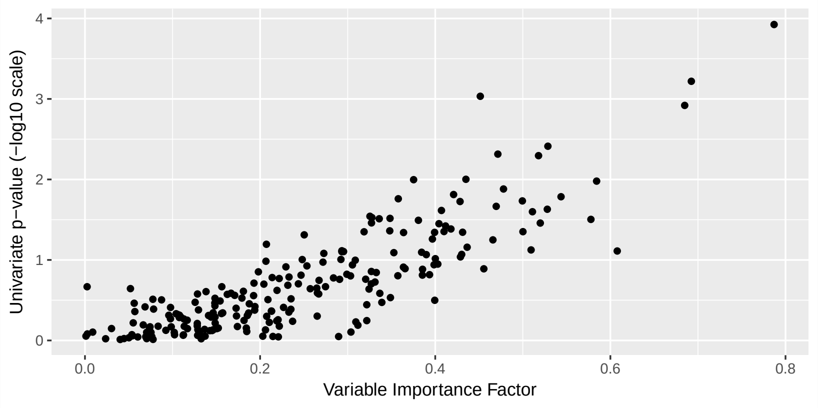
Data

- 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 $\Rightarrow$ 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.
- In a high-dimensional setting, **estimation** and **inference** are more challenging
  - Estimation: Truncated SVD
  - Inference: Fitted location-scale Tracy-Widom
- Our approach is computationally simple.
- Everything presented today has been implemented in two `R` packages.

Demo

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*

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:

$$R^2(w) = \frac{w^T V_M w}{w^T(V_M + V_R)w}.$$

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

Key observation: $R^2(w)$ measures the strength of the association

Acknowledgements

- 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.

CENTRE LUDMER CENTER
NEUROINFORMATIQUE & SANTÉ MENTALE | NEUROINFORMATICS & MENTAL HEALTH

Questions or comments?

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