Dimension Reduction and High-Dimensional Data

Estimation and Inference with Application to Genomics and Neuroimaging

Maxime Turgeon

April 9, 2019

McGill University
Department of Epidemiology, Biostatistics, and Occupational Health

Introduction

- Data revolution fueled by technological developments, era of "big data".
- In genomics and neuroimaging, high-throughput technologies lead to *high-dimensional data*.
  - High costs lead to small-to-moderate samples size.
  - More **features** than **samples** (large p, small n)

Omnibus Hypotheses and Dimension Reduction

- Traditionally, analysis performed *one feature at a time*.
  - Large computational burden
  - Conservative tests and low power
  - Ignore correlation between features
- From a biological standpoint, there are natural groupings of measurements
- **Key**: Summarise group-wise information using *latent* features
  - Dimension Reduction

High-dimensional data–Estimation

- Several approaches use regularization
  - Zou *et al.* (2006) Sparse PCA
  - Witten *et al.* (2009) Penalized Matrix Decomposition
- Other approaches use structured estimators
  - Bickel & Levina (2008) Banded and thresholded covariance estimators
- All of these approaches require tuning parameters, which increases computational burden

High-dimensional data–Inference

- Double Wishart problem and largest root
- Distribution of largest root is difficult to compute
  - Several approximation strategies presented
  - Chiani found simple recursive equations, but computationally unstable
- Result of Johnstone gives an excellent good approximation
  - Does not work with high-dimensional data

Contribution of the thesis

In this thesis, I address the limitations outlined above.

- Block-independence leads to simple approach **free of tuning parameters**
- Empirical estimator that extends Johnstone's theorem to **high-dimensional data**
- **Application** of these ideas to sequencing study of DNA methylation and ACPA levels.

First Manuscript–Estimation

Principal Component of Explained Variance

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

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

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:

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

Maximisation using a combination of Lagrange multipliers and linear algebra.

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

Block-diagonal Estimator

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

- Suppose the outcome variables $\mathbf{Y}$ can be divided in blocks of variables in such a way that
  - Variables **within** blocks are correlated
  - Variables **between** blocks are uncorrelated

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

Block-diagonal Estimator

- We can perform PCEV on each of these blocks, resulting in a component 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.
    - Mathematically equivalent to performing PCEV in a single-step (under assumption)
- Extensive simulation study shows *good power* and *robustness of inference* to violations of assumption.
- Presented application to genomics and neuroimaging data.

Second Manuscript–Inference

Double Wishart Problem

- Recall that PCEV is maximising a Rayleigh quotient:

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

- Equivalent to finding largest root λ of a *double Wishart problem*:

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

where $A = V_M, B = V_R$.

Inference

- Evidence in the literature that the null distribution of the largest root λ should be related to the **Tracy-Widom distribution**.
- Result of Johnstone (2008) gives an excellent approximation to the distribution using an explicit location-scale family of the TW(1).

Inference

- However, Johnstone's theorem requires a rank condition on the matrices (rarely satisfied in high dimensions).
- 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

I proposed 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**.

Third Manuscript–Application

Data

- Anti-citrullinated Protein Antibody (ACPA) levels were measured in 129 levels without any symptom of Rheumatoid Arthritis (RA).
- DNA methylation levels were measured from whole-blood samples using a targeted sequencing technique
    - CpG dinucleotides were grouped in regions of interest before the sequencing
- We have 23,350 regions to analyze individually, corresponding to multivariate datasets $Y_k, k = 1, \ldots, 23,350$.

Method

- PCEV was performed independently on all regions.
  - Significant amount of missing data; complete-case analysis.
- Analysis was adjusted for age, sex, and smoking status.
- ACPA levels are dichotomized into high and low.
- For the 2519 regions with more CpGs than observations, we used the Tracy-Widom empirical estimator to obtain p-values.

Results

- There were 1062 statistically significant regions at the $\alpha = 0.05$ level.
- Univariate analysis of 175,300 CpG dinucleotides yielded 42 significant results
    - These 42 CpG dinucleotides were in 5 distinct regions.

Discussion

Summary

- This thesis described specific approaches to dimension reduction with high-dimensional datasets.
- *Manuscript 1*: Block-independence assumption leads to convenient estimation strategy that is free of tuning parameters.
- *Manuscript 2*: Empirical estimator provides valid p-values for high-dimensional data by leveraging Johnstone's theorem.
- *Manuscript 3*: Application of this thesis' ideas to a study of the association between aCPA levels and DNA methylation.
- All methods from Manuscripts 1 & 2 are part of the `R` package `pcev`.

Limitations

- *Inference* for PCEV-block is robust to block-independence violations, but *not* estimation
    - Could have impact on downstream analyses.
- Empirical estimator does not address limitations due to power
    - But combining with shrinkage estimator should improve power.
- Missing data and multivariate analysis

Future Work

- Estimate effective number of independent tests in region-based analyses
- Multiple imputation and PCEV
- Nonlinear dimension reduction

Thank you

The slides can be found at

`maxturgeon.ca/talks`.