Principal Component of Explained Variance

An Efficient and Optimal Data Dimension Reduction Framework for Association Studies

Maxime Turgeon

May 30th, 2016

McGill University
Department of Epidemiology, Biostatistics, and Occupational Health

Introduction

Introduction

- In genetics and brain imaging studies, we are often interested in studying multivariate outcomes of large dimension ($p > n$).

Introduction

- In genetics and brain imaging studies, we are often interested in studying multivariate outcomes of large dimension ($p > n$).
- One popular method to analyse such datasets is to use *component-based dimension reduction methods*

Introduction

- In genetics and brain imaging studies, we are often interested in studying multivariate outcomes of large dimension ($p > n$).
- One popular method to analyse such datasets is to use *component-based dimension reduction methods*
    - The idea is to summarise a dataset into a single component based on a defined criterion

Introduction

- In genetics and brain imaging studies, we are often interested in studying multivariate outcomes of large dimension ($p > n$).
- One popular method to analyse such datasets is to use *component-based dimension reduction methods*
    - The idea is to summarise a dataset into a single component based on a defined criterion
    - E.g. Principal Component Analysis (PCA)

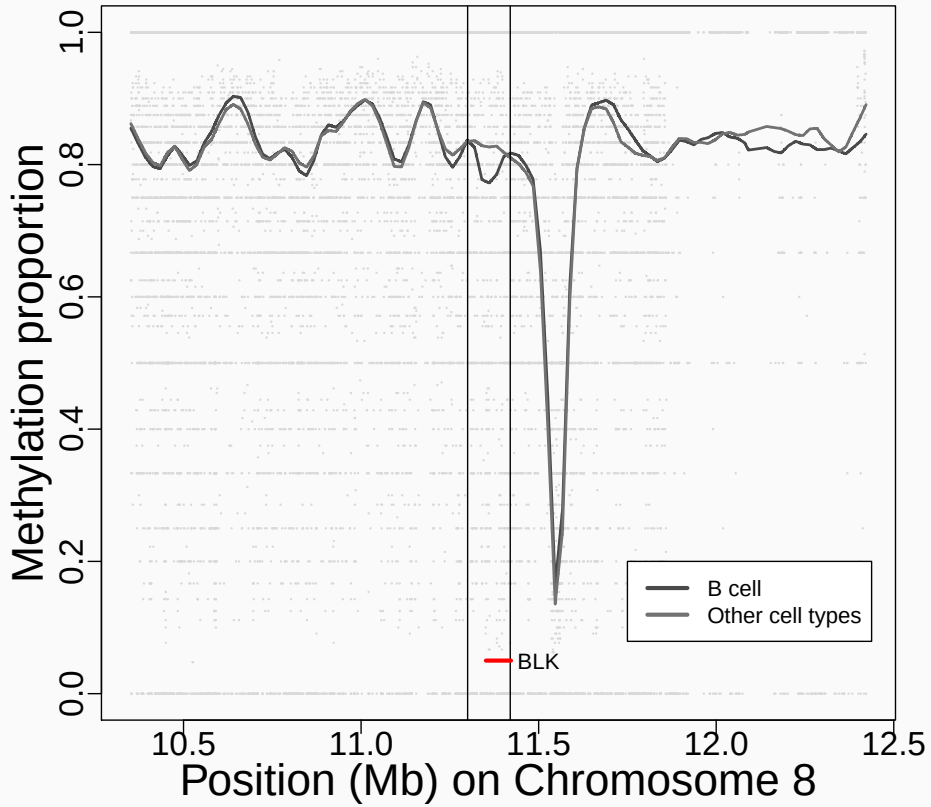
Introduction

- In genetics and brain imaging studies, we are often interested in studying multivariate outcomes of large dimension ($p > n$).
- One popular method to analyse such datasets is to use *component-based dimension reduction methods*
  - The idea is to summarise a dataset into a single component based on a defined criterion
  - E.g. Principal Component Analysis (PCA)
- There is also a need for **fast** computational methods which can handle **high-dimensional** outcomes

Motivating example

Motivating example

B-Lymphoid Tyrosine Kinase (BLK) gene is known to be differentially methylated with respect to blood cell types.



Motivating example

- The data consist of 40 cell-separated whole-blood samples (T cells, B cells, monocytes), for which methylation levels were measured at 24,000 CpG sites using bisulfite sequencing.

Motivating example

- The data consist of 40 cell-separated whole-blood samples (T cells, B cells, monocytes), for which methylation levels were measured at 24,000 CpG sites using bisulfite sequencing.
- The figure above was obtained using smoothing techniques: the methylation levels for a particular cell-type is smoothed across the 24,000 loci.

Principal Component of Explained Variance (PCEV)

Principal Component of Explained Variance (PCEV)

- Provides an **optimal** strategy for selecting components for association with one or several covariates of interest.

Principal Component of Explained Variance (PCEV)

- Provides an **optimal** strategy for selecting components for association with one or several covariates of interest.
- **Goal**: Find the component that maximises the proportion of variance explained by the covariates

Principal Component of Explained Variance (PCEV)

- Provides an **optimal** strategy for selecting components for association with one or several covariates of interest.
- **Goal**: Find the component that maximises the proportion of variance explained by the covariates
- In the literature, PCEV was formerly known as the **Principal Component of Heritability** (PCH).

Our Contributions

Our Contributions

1. An analytical framework for hypothesis testing.

Our Contributions

1. An analytical framework for hypothesis testing.
2. A high-dimensional approach that does not require any tuning parameter.

Our Contributions

1. An analytical framework for hypothesis testing.
2. **A high-dimensional approach that does not require any tuning parameter.**

Our Contributions

1. An analytical framework for hypothesis testing.
2. **A high-dimensional approach that does not require any tuning parameter.**

A manuscript describing our work is currently available on bioRxiv (search for "Principal Component of Explained Variance").

Methods

PCEV: Statistical model

PCEV: Statistical model

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

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

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

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

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:

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

PCEV

PCEV

- **Input**: a set of outcomes and a set of covariates

PCEV

- **Input**: a set of outcomes and a set of covariates
- **Output**:

PCEV

- **Input**: a set of outcomes and a set of covariates
- **Output**:
    - One or more components maximising the proportion of variance explained by the covariates

PCEV

- **Input**: a set of outcomes and a set of covariates
- **Output**:
  - One or more components maximising the proportion of variance explained by the covariates
  - A set of weights (also known as loadings): one for each combination of trait and component

PCEV

- **Input**: a set of outcomes and a set of covariates
- **Output**:
  - One or more components maximising the proportion of variance explained by the covariates
  - A set of weights (also known as loadings): one for each combination of trait and component
  - A measure of variable importance: one for each combination of trait and component. This is defined as the correlation between a single outcome and the component (in absolute value).

PCEV

- **Input**: a set of outcomes and a set of covariates
- **Output**:
  - One or more components maximising the proportion of variance explained by the covariates
  - A set of weights (also known as loadings): one for each combination of trait and component
  - A measure of variable importance: one for each combination of trait and component. This is defined as the correlation between a single outcome and the component (in absolute value).
  - A p-value for the association between the PCEV and the covariates

PCEV

- **Input**: a set of outcomes and a set of covariates
- **Output**:
  - One or more components maximising the proportion of variance explained by the covariates
  - A set of weights (also known as loadings): one for each combination of trait and component
  - A measure of variable importance: one for each combination of trait and component. This is defined as the correlation between a single outcome and the component (in absolute value).
  - A p-value for the association between the PCEV and the covariates

An `R` package called `pcev` is available on CRAN.

PCEV: High-dimensional outcomes

Our main contribution is an extension of PCEV to high-dimensional settings that is

- Simple

PCEV: High-dimensional outcomes

Our main contribution is an extension of PCEV to high-dimensional settings that is

- Simple
- Computationally very fast

PCEV: High-dimensional outcomes

Our main contribution is an extension of PCEV to high-dimensional settings that is

- Simple
- Computationally very fast
- Works with $p \gg n$

PCEV: High-dimensional outcomes

Our main contribution is an extension of PCEV to high-dimensional settings that is

- Simple
- Computationally very fast
- Works with $p \gg n$
- Free of tuning parameters

PCEV: High dimensional outcomes

PCEV: High dimensional outcomes

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

PCEV: High dimensional outcomes

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

- Suppose the outcome variables (e.g. methylation levels) can be divided in blocks of traits in such a way that

PCEV: High dimensional outcomes

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

- Suppose the outcome variables (e.g. methylation levels) can be divided in blocks of traits in such a way that
  - Traits **within** blocks are correlated

PCEV: High dimensional outcomes

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

- Suppose the outcome variables (e.g. methylation levels) can be divided in blocks of traits in such a way that
  - Traits **within** blocks are correlated
  - Traits **between** blocks are uncorrelated

PCEV: High dimensional outcomes

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

- Suppose the outcome variables (e.g. methylation levels) can be divided in blocks of traits in such a way that
  - Traits **within** blocks are correlated
  - Traits **between** blocks are uncorrelated
- If each block is small enough, we can perform PCEV on each of them, resulting in a PCEV for each block.

PCEV: High dimensional outcomes

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

- Suppose the outcome variables (e.g. methylation levels) can be divided in blocks of traits in such a way that
  - Traits **within** blocks are correlated
  - Traits **between** blocks are uncorrelated
- If each block is small enough, we can perform PCEV on each of them, 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.

PCEV: High dimensional outcomes

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

- Suppose the outcome variables (e.g. methylation levels) can be divided in blocks of traits in such a way that
  - Traits **within** blocks are correlated
  - Traits **between** blocks are uncorrelated
- If each block is small enough, we can perform PCEV on each of them, 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.

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

Simulations

Simulation setting

Simulation setting

- We compared 4 different approaches:

Simulation setting

- We compared 4 different approaches:
    - PCEV-block, with blocks assumed known a priori

Simulation setting

- We compared 4 different approaches:
  - PCEV-block, with blocks assumed known a priori
  - PCEV-block, with blocks selected randomly

Simulation setting

- We compared 4 different approaches:
  - PCEV-block, with blocks assumed known a priori
  - PCEV-block, with blocks selected randomly
  - Lasso

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)

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 simulated $p = 100, 200, 300, 400, 500$ outcomes,

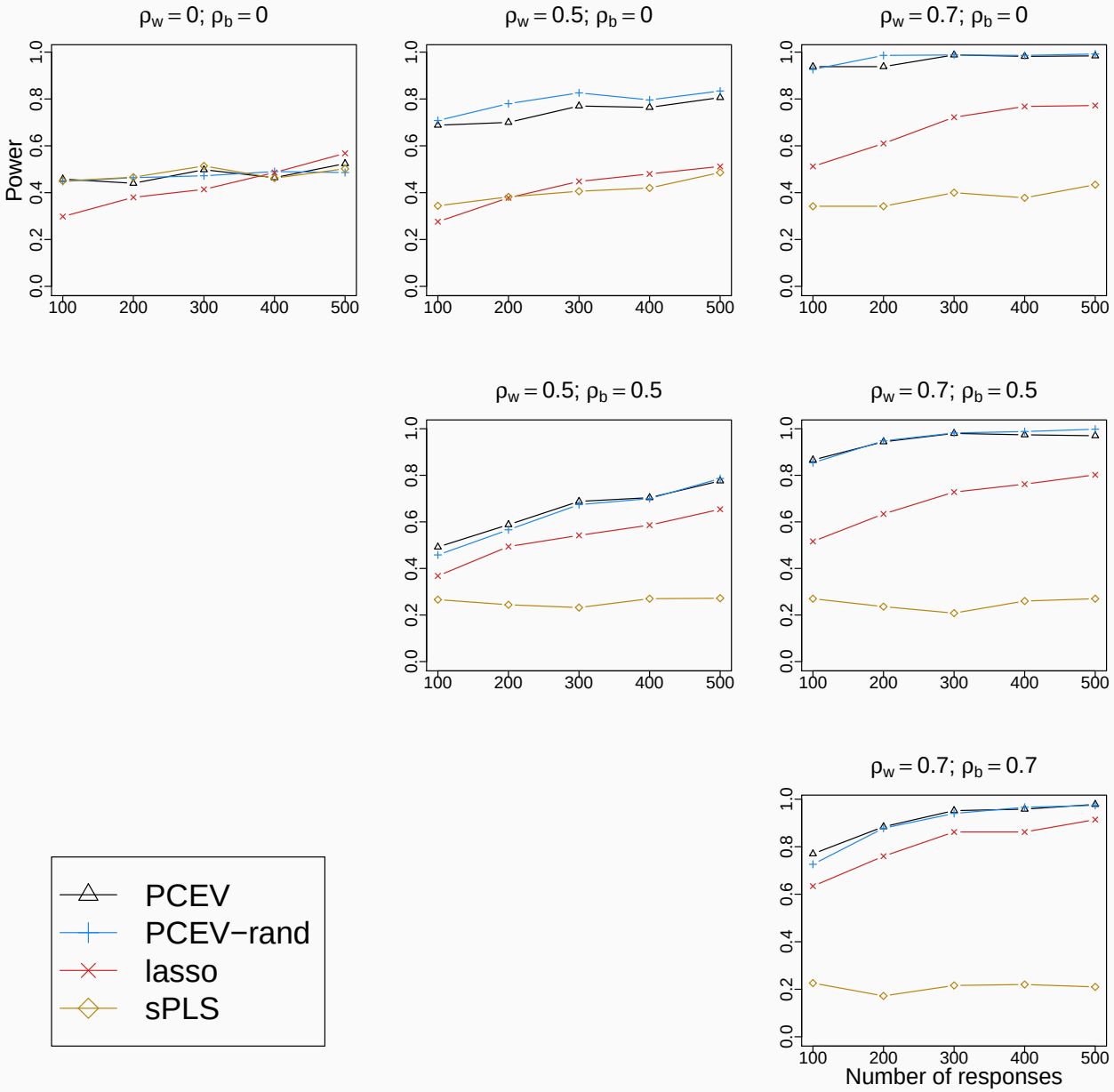
Simulation setting

- 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 simulated $p = 100, 200, 300, 400, 500$ outcomes,
- The parameters we varied are: number of outcomes (from 100 to 500), correlation between and within blocks (0, 0.5, 0.7).

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 simulated $p = 100, 200, 300, 400, 500$ outcomes,
- The parameters we varied are: number of outcomes (from 100 to 500), correlation between and within blocks (0, 0.5, 0.7).
- We fixed the sample size at $n = 100$ and simulated a single continuous covariate from a standard normal distribution. We distributed the outcome variables in 10 blocks. 25% of the outcomes in each block are associated with X.

Simulation results: Power analysis



Data analysis

BLK gene – Motivating example

BLK gene – Motivating example

- BLK gene, located on chromosome 8

BLK gene – Motivating example

- BLK gene, located on chromosome 8
- Data provided by Tomi Pastinen (McGill)

BLK gene – Motivating example

- BLK gene, located on chromosome 8
- Data provided by Tomi Pastinen (McGill)
- DNA methylation levels derived from bisulfite sequencing

BLK gene – Motivating example

- BLK gene, located on chromosome 8
- Data provided by Tomi Pastinen (McGill)
- DNA methylation levels derived from bisulfite sequencing
- 40 cell-separated samples, from 3 different cell types

BLK gene – Motivating example

- BLK gene, located on chromosome 8
- Data provided by Tomi Pastinen (McGill)
- DNA methylation levels derived from bisulfite sequencing
- 40 cell-separated samples, from 3 different cell types
    - B cells (n=8)

BLK gene – Motivating example

- BLK gene, located on chromosome 8
- Data provided by Tomi Pastinen (McGill)
- DNA methylation levels derived from bisulfite sequencing
- 40 cell-separated samples, from 3 different cell types
  - B cells (n=8)
  - T cells (n=19)

BLK gene – Motivating example

- BLK gene, located on chromosome 8
- Data provided by Tomi Pastinen (McGill)
- DNA methylation levels derived from bisulfite sequencing
- 40 cell-separated samples, from 3 different cell types
  - B cells (n=8)
  - T cells (n=19)
  - Monocytes (n=13)

BLK gene – Motivating example

- BLK gene, located on chromosome 8
- Data provided by Tomi Pastinen (McGill)
- DNA methylation levels derived from bisulfite sequencing
- 40 cell-separated samples, from 3 different cell types
  - B cells (n=8)
  - T cells (n=19)
  - Monocytes (n=13)
- 24,068 CpG sites

BLK gene – Motivating example

- BLK gene, located on chromosome 8
- Data provided by Tomi Pastinen (McGill)
- DNA methylation levels derived from bisulfite sequencing
- 40 cell-separated samples, from 3 different cell types
  - B cells (n=8)
  - T cells (n=19)
  - Monocytes (n=13)
- 24,068 CpG sites

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

Results

Results

- Blocks are defined using physical distance: CpGs within 500kb are grouped together

Results

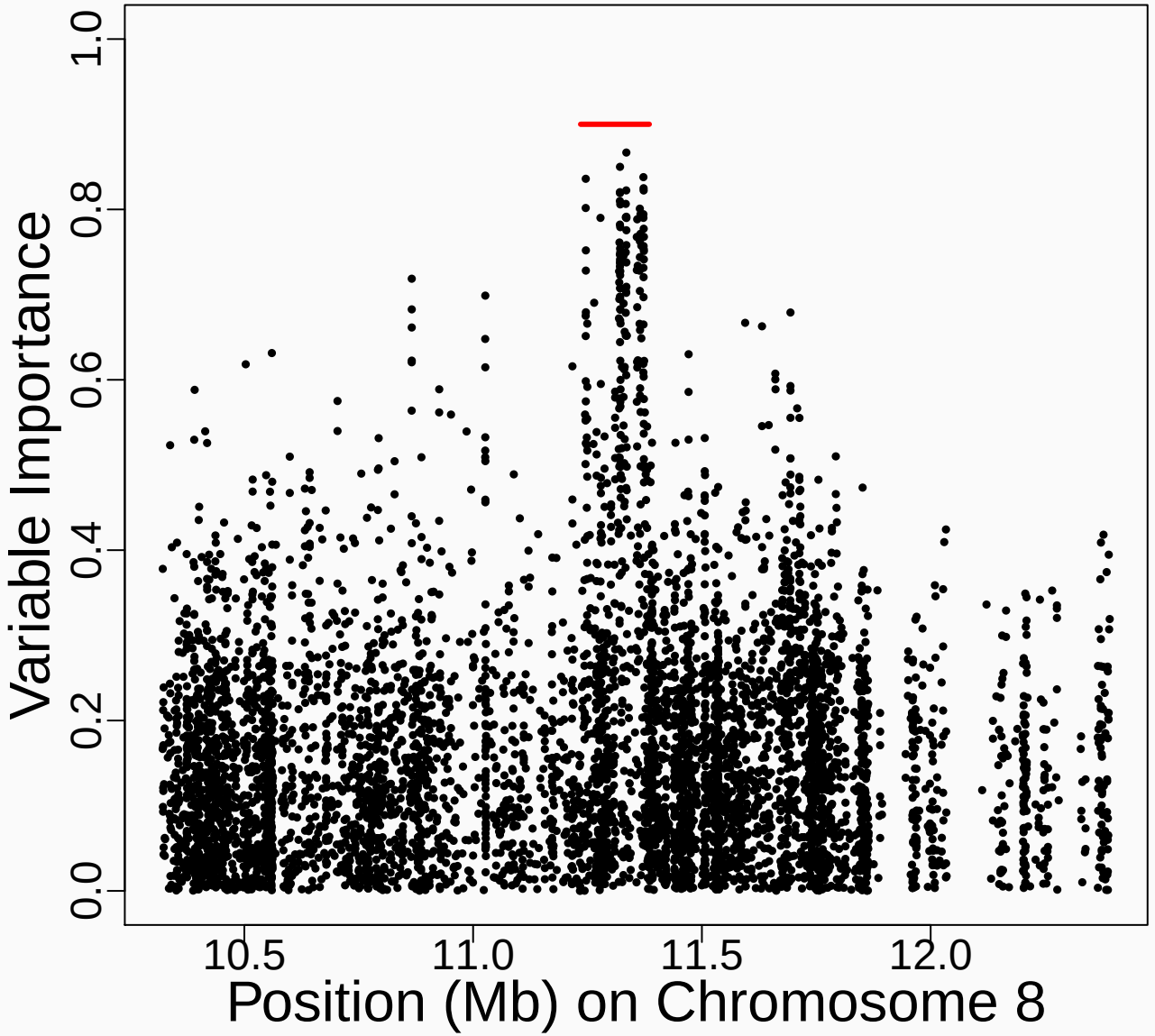
- Blocks are defined using physical distance: CpGs within 500kb are grouped together
    - 951 blocks were analysed

Results

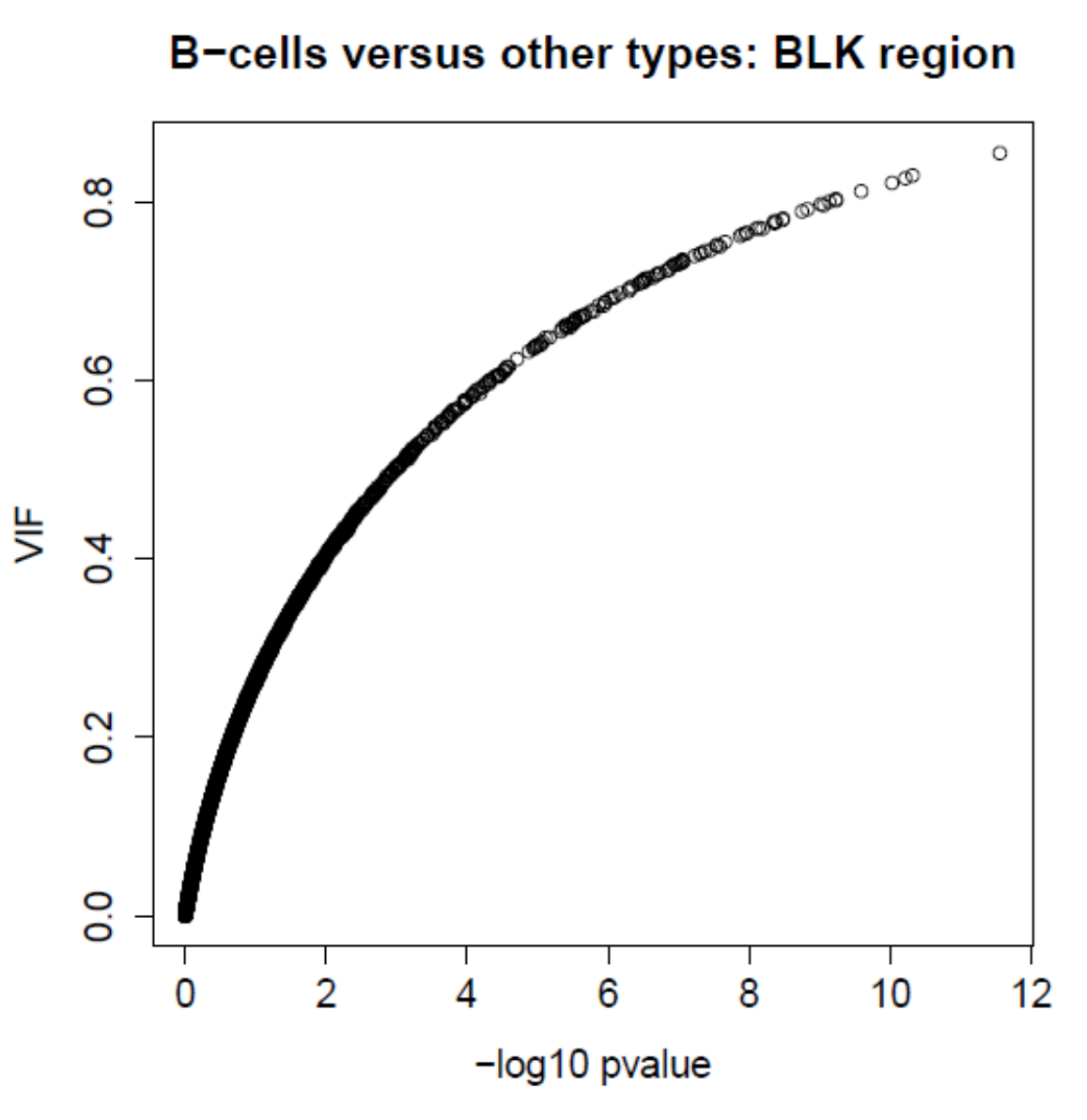
- Blocks are defined using physical distance: CpGs within 500kb are grouped together
  - 951 blocks were analysed
- Using PCEV, we obtained a single p-value, which is less than 6×10^{-5} (using 100,000 permutations)

Results

- Blocks are 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

Variable importance



Conclusion

Conclusion

- Data summary is an important feature in data analysis, and this can be achieved using dimension reduction techniques.

Conclusion

- Data summary is an important feature in data analysis, and this can be achieved using dimension reduction techniques.
- Principal Component of Explained Variance is an interesting alternative to PCA

Conclusion

- Data summary is an important feature in data analysis, and this can be achieved using dimension reduction techniques.
- Principal Component of Explained Variance is an interesting alternative to PCA
    - It is optimal in capturing the association with covariates

Conclusion

- Data summary is an important feature in data analysis, and this can be achieved using dimension reduction techniques.
- Principal Component of Explained Variance is an interesting alternative to PCA
    - It is optimal in capturing the association with covariates
- Our block approach is a simple, computationally fast way of handling high-dimensional outcomes.

Conclusion

- Data summary is an important feature in data analysis, and this can be achieved using dimension reduction techniques.
- Principal Component of Explained Variance is an interesting alternative to PCA
    - It is optimal in capturing the association with covariates
- Our block approach is a simple, computationally fast way of handling high-dimensional outcomes.
    - It does not require any tuning parameter.

Conclusion

- Data summary is an important feature in data analysis, and this can be achieved using dimension reduction techniques.
- Principal Component of Explained Variance is an interesting alternative to PCA
    - It is optimal in capturing the association with covariates
- Our block approach is a simple, computationally fast way of handling high-dimensional outcomes.
    - It does not require any tuning parameter.
- Simulations and data analyses confirm its advantage over a more traditional approach using PCA (not shown), as well as other high-dimensional approaches such as Lasso and sPLS.

Acknowledgements

- Karim Oualkacha (UQAM)
- Antonio Ciampi (McGill University)
- **Aurélie Labbe** (McGill University)
- **Celia Greenwood** (McGill University)

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