Enter the matrix: Interpreting unsupervised feature learning with matrix decomposition to discover hidden knowledge in high-throughput omics data

Genevieve L. Stein-O’Brien, Raman Arora, Aedin C. Culhane, Alexander Favorov, Casey Greene, Loyal A. Goff, Yifeng Li, Alioune Ngom, Michael F. Ochs, Yanxun Xu, Elana J. Fertig

IDIES Symposium
20 Oct 2017
Novel techniques for bulk to adapt to single cell


Defining tissue-specific peaks from chromatin

EVA analysis of tumor-specific chromatin genes

Bahman Afsari    Genevieve Stein-O’Brien, Wai-Shing Lee
Smooth-sparse non-negative matrix factorization in the R/Bioconductor package CoGAPS.

**A**

Data

Amplitude

Pattern

- **Data**
  - Rows of continuous or binary weights associate with samples (each sample is a column) capturing relationships including cell-types/lines, patients, or experimental conditions.

- **Amplitude**
  - Columns of genetic, epigenetic, or protein weights (each row is a unique molecule) associated with a given sample feature and often reflective of co-regulation.

- **Pattern**
  - Rows of continuous or binary weights associated with samples (each sample is a column) capturing relationships including cell-types/lines, patients, or experimental conditions.

**B**

- **Associated molecular signatures** (metagenes, modules, etc.)
- **Patterns in groups of samples**

Genevieve Stein-O’Brien, Wai-Shing Lee
Confirmatory Results

Enter the matrix: Interpreting unsupervised feature learning with matrix decomposition to discover hidden knowledge in high-throughput omics data

Genevieve L. Stein-O’Brien, Ramon Arella, Aedin C. Cullinan, Alexander Favorov, Casey Greene, Loyal A. Goff, Yifeng Li, Alouane Nguyen, Michael F. Ochoa, Yanxin Xu, Elana J. Fertig

doi: https://doi.org/10.1101/196915

Abstract

High-dimensional data is currently standard for biological inquiry. Biological systems are comprised of interrelated gene regulatory mechanisms, gene-gene interactions, and cellular interactions. These interactions induce low-dimensional structure within the high-dimensional data. Matrix factorization, also known as compressed sensing, learns low-dimensional mathematical representations from high-dimensional data. These factorization techniques can embed assumptions about pleiotropy, epistasis, inter-relationships between complex traits, and context-dependent interactions. They have been applied to uncover new biological knowledge in a breadth of topics ranging from pathway discovery to time course analysis. These techniques have been applied to data from diverse high-throughput omics technologies, including bulk and single-cell data. There are numerous computational techniques within the class of matrix factorization, each of which provides a unique interpretation of the processes in high-dimensional data. We review the visualization and applications of matrix factorization to systems-level analyses, which are diverse and require standardization to enable biological interpretation. Codifying the techniques to decipher biologically relevant features with matrix factorization enables their broad application to discovery beyond the limits of current biological
Pattern detection is critical in the genomics big data era.

Stephens et al. 2015 Big Data: Astronomical or Genomical? *PLoS Biology*

Sample sizes of single cell growing by $\sim 2$ orders of magnitude!
Pattern detection is critical in the genomics big data era

How do we learn the biological processes in each sample to answer unasked questions from big data?
Genomics pattern detection with matrix factorization

![Diagram of matrix factorization]

**A**
- **Data**: Rows of continuous or binary weights associate with samples (each sample is a column) capturing relationships including cell-types/lines, patients, or experimental conditions.

- **Amplitude**: Columns of genetic, epigenetic, or protein weights (each row is a unique molecule) associated with a given sample feature and often reflective of co-regulation.

- **Pattern**: Rows of continuous or binary weights associate with samples (each sample is a column) capturing relationships including cell-types/lines, patients, or experimental conditions.
Genomics pattern detection with matrix factorization

A. **Data**
- Rows of continuous or binary weights associate with samples (each sample is a column) capturing relationships including cell-types/lines, patients, or experimental conditions.
- Columns of genetic, epigenetics, or protein weights (each row is a unique molecule) associated with a given sample feature and often reflective of co-regulation.

B. **Amplitude**
- Associated molecular signatures (metagenes, modules, etc.)

**Pattern**
- Patterns in groups of samples
- Magnitude of rows of the "Pattern" matrix scaled from $\max(P)$

**E. J. Fertig (JHMI SKCCC)**

Enter the matrix 20 Oct 2017 8 / 21
Data-driven gene sets

Projection into new data

Amplitude Matrix

Enrichment analysis can be done directly or via thresholding

A priori functionally annotated sets

Unique “patternMarkers” can be used for biological validation
PCA  Columns of the amplitude matrix and rows of the pattern matrix are “orthogonal”, with each explaining increasing amounts of variance in the data.

ICA  Columns of the amplitude matrix and rows of the pattern matrix are statistically independent.

NMF  Elements of the amplitude and pattern matrices are non-negative, resulting in a non-orthogonal and non-independent factorization.
Tailor method selection to analysis goal

Complex biological process (CBP) captured in sample features, e.g.:

Data Matrix

CBPs: A B C

Pattern Matrix (PCA)

PC1 (54.90% of the variance)

PC2 (27.31% of the variance)

PC3

Pattern Matrix (NMF)

PC1

PC2

PC3

Molecules of interest, i.e. genes

Samples

PC2 (27.31% of the variance)

PC1 (54.90% of the variance)

CBPs: A B C

Complex biological process (CBP) captured in sample features, e.g.:

E. J. Fertig (JHMI SKCCC)

Enter the matrix

20 Oct 2017

11 / 21

PCA learns components that separate groups but mix CBPs, whereas NMF learns components shared between groups and unique to CBPs.
Tailor method selection to analysis goal

**PCA**  Columns of the amplitude matrix and rows of the pattern matrix are “orthogonal”, with each explaining increasing amounts of variance in the data.

Optimal for clustering.

**ICA**  Columns of the amplitude matrix and rows of the pattern matrix are statistically independent.

Optimal for equally valid, but independent outcomes; e.g., branching in cell fate decisions.

**NMF**  Elements of the amplitude and pattern matrices are non-negative, resulting in a non-orthogonal and non-independent factorization.

Optimal for modeling gene reuse and concurrent processes; e.g., cell cycle.
Input datasets impact inferred patterns (RNA-seq for healthy human tissues in GTEx)

Dey et al. 2017 Visualizing the structure of RNA-seq expression data using grade of membership models *PLoS Genetics*
Multiple views of the data depending on the factorization dimensionality (expression arrays of head and neck tumors).

Fertig et al. 2013 Preferential Activation of the Hedgehog Pathway by Epigenetic Modulations in HPV Negative HNSCC Identified with Meta-Pathway Analysis *PLoS One*
Multiple data views raise critical unanswered questions

- What is the correct dimensionality of the data?
  - Is smaller always better?
  - What is the balance between uncovering structure in the data and overfitting noise?

- How can you quantify performance of the algorithm?
  - What is the best “norm” to quantify “fit” of inferred matrices to data?
  - How can accurate reflection of biology be encoded?
  - What views of the data are revealed by different algorithms?
Time course analysis with matrix factorization

**A**

Pattern Matrix (Time Course)

CBPs: A B C

**Time:** d1 ... d6 d1 ... d6 d1 ... d6

- P1
- P2
- P3

**Days**

- P1
  - Days 1 to 6

- P2
  - Days 1 to 6

- P3
  - Days 1 to 6
Bulk data (Illumina 450K DNA methylation) of acquired therapeutic resistance

A Clustering of DNA methylation data

B DNA methylation of PatternMarker genes for CoGAPS methylation patterns

C CoGAPS DNA methylation patterns (sample weights v. time)

Genevieve Stein-O’Brien, Luciane Kagohara, Lucy Li, Manjusha Thakar, Christine Chung
CoGAPS of single cell (smart-seq) retina development data

A. T-SNE of raw scRNAseq data colored by # genes expressed

B. T-SNE of CoGAPS patterns for scRNAseq colored by # genes expressed

C. T-SNE of CoGAPS patterns colored by Crx expression
   (Crx is a marker of photoreceptors)

D. T-SNE of CoGAPS patterns colored by Ccnd expression
   (Ccnd is a marker of cell cycle)

Genevieve Stein-O’Brien, Brian Clark, Gabriel Cannon, Loyal Goff, Seth Blackshaw
Single cell data poses new challenges

- How to visualize the dynamics uncovered in large numbers of patterns?
- How to account for missing data?
  - Does imputation induce unwanted correlation structures in the data?
  - Can it be incorporated directly in the model for data fit quality?
- Improving efficiency for larger datasets.
  - Will different sub-structures be apparent in factorizations of datasets of different sizes?
  - How can algorithms converge for $\sim 10^K$ molecular measurements AND $\sim 10^K$ samples?
  - Computational efficiency with local optima (gradient-based methods) vs inefficiency with robust, global optima (Bayesian methods)
Summary

- Matrix factorization has a long history in genomics (since 1990s), and is still critical to interpreting today’s large datasets.
- PCA is robust for sample clustering, ICA for learning independent components, and NMF at accounting for co-regulation of genes in concurrent biological processes.
- New metrics to quantify fit quality and assess dimensionality are needed to assess whether one of multiple truths are learned in a low-dimensional data reduction.
- Matrix factorization techniques are applicable to single cell datasets, and requires new methods for efficiency and visualization.
Acknowledgements

• Fertig Lab
  • Genevieve Stein-O’Brien
  • Luciane Kagohara
  • Sijia “Lucy” Li (UW)
  • Jacob Carey
  • Thomas Sherman
  • Michael Considine

• JHU collaborators
  • Raman Arora
  • Seth Blackshaw
  • Gabriel Cannon
  • Leslie Cope
  • Alexander V. Favorov
  • Daria A. Gaykalova
  • Loyal Goff
  • Manjusha Thakar
  • Yanxun Xu

• Harvard / Dana Farber
  • Aedin Culhane
  • Giovanni Parmigiani

• Other
  • Joseph Califano (UCSD)
  • Christine Chung (Moffitt)
  • Casey Greene (U Penn)
  • Yifeng Li (NRCC)
  • Alioune Ngom (U Windsor)
  • Michael Ochs (TCNJ)