# Probabilistic Count Matrix Factorization for Single Cell Expression Data Analysis

Ghislain DURIF

July 4th 2019

Single-cell RNA-seq day, Toulouse

CNRS, IMAG, Université de Montpellier, France
`ghislain.durif@umontpellier.fr`
`https://gdurif.perso.math.cnrs.fr/`

# Outline

1. Introduction

2. Dimension reduction with matrix factorization

3. Matrix Factorization for count data

4. Experiments

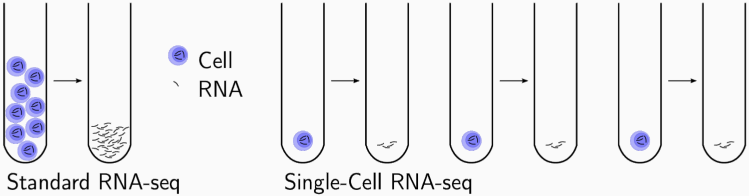
5. To conclude

# Introduction

# Single-cell RNA-seq

## RNA-seq

- Quantification of gene expression on a genomic scale



## Single-cell level (scRNA-seq)

- gene-to-gene variability: expression dynamics (low expression genes)
- cell-to-cell variability: diversity within a population of cells

# High-dimensional data

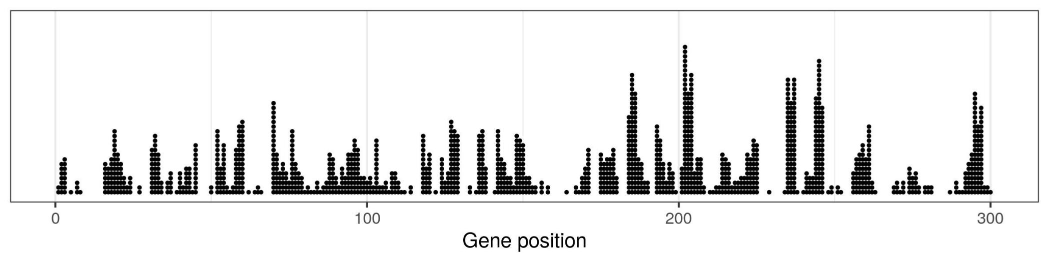
$x_{ij}$ = expression of gene $j$ in sample $i$

$$\mathbf{X}_{n\times p} = \begin{array}{l} \left[\begin{array}{ccccccc} \; & \; & \; & \; & \; & \; & \; \\ & & & x_{ij} & & & \\ & & & & & & \end{array}\right] \begin{array}{c} 1 \\ \vdots \\ n \end{array} \left.\vphantom{\begin{array}{c}1\\ \vdots\\ n\end{array}}\right\} \text{cells} \\ \underbrace{1 \quad \dots \quad \dots \quad \dots \quad \dots \quad p}_{\text{genes}} \end{array}$$

- **High dimension:** $n$ grows but $\ll p$

$\rightarrow$ $n \sim 100/1000$ and $p \sim 10000$

# RNA-seq = count data



$x_{ij}$ = height at position $j$

- Count data with drop-out events in single-cell RNA-seq (zero-inflation)

→ number of reads that map to a gene position

# Issues with high dimensional data ($p \gg n$)

*"The curse of high-dimensionality"* (Donoho, 2000)

- **Geometry:** counter-intuitive behavior of metrics (Aggarwal et al., 2001)

  → Representation: how to visualize thousands of variables?

- **Optimization:** numerical singularities due to complex dependencies (colinearity)

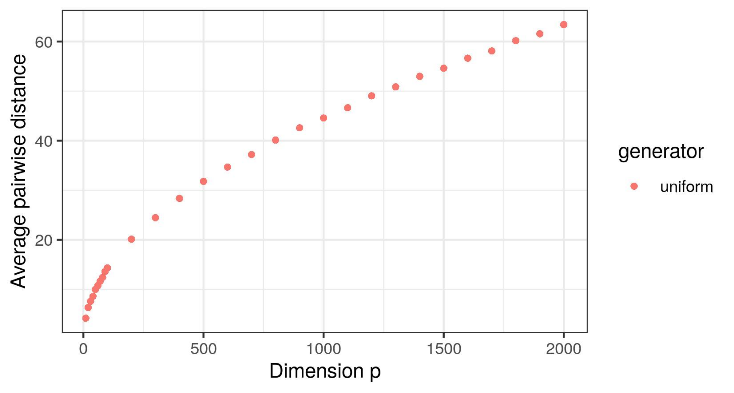
- **Computational efficiency and scalability**

# Issues with high dimensional data ($p \gg n$)

- **Geometry:** counter-intuitive behavior of metrics (Aggarwal et al., 2001)

$$p \mapsto \|\mathbf{x}_1 - \mathbf{x}_2\|_2$$

with $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^p$



# Issues with high dimensional data ($p \gg n$)

*"The curse of high-dimensionality"* (Donoho, 2000)

- **Geometry:** counter-intuitive behavior of metrics (Aggarwal et al., 2001)

  → Representation: how to visualize thousands of variables?

- **Optimization:** numerical singularities due to complex dependencies (colinearity)

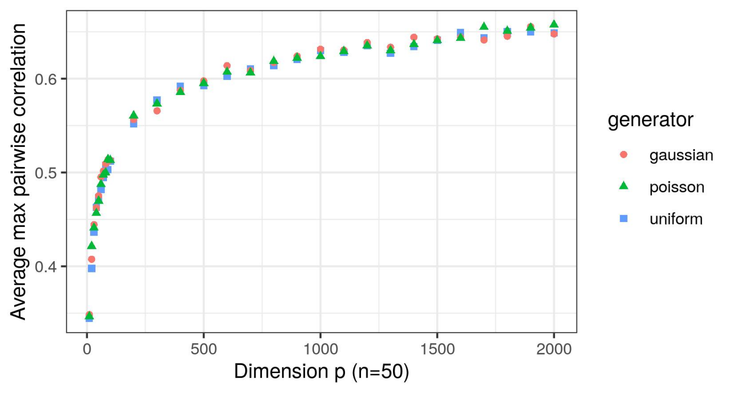
- **Computational efficiency and scalability**

# Issues with high dimensional data ($p \gg n$)

- **Optimization:** numerical singularities due to complex dependencies (colinearity)

$$\mathrm{X}_{n\times p} = \left[\ \mathrm{x}_1 \ \middle|\ \dots \ \middle|\ \mathrm{x}_p\ \right]$$

$$p \mapsto \max_{j\neq\ell\in\{1,\dots,p\}} \left|\mathrm{Corr}(\mathrm{x}_j, \mathrm{x}_\ell)\right|$$

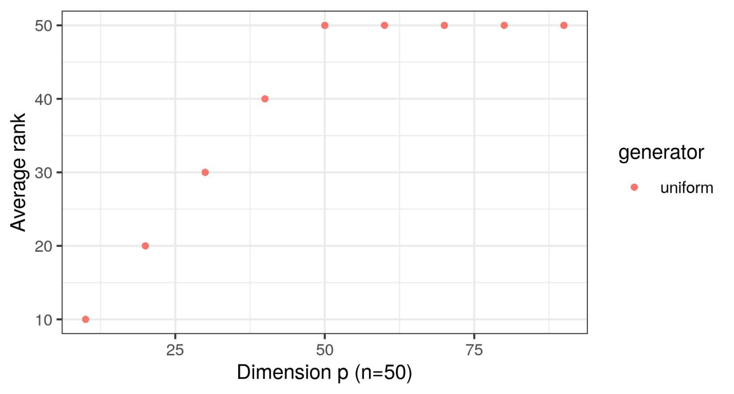


# Issues with high dimensional data ($p \gg n$)

- **Optimization:** numerical singularities due to complex dependencies (colinearity)

$p \mapsto \mathrm{rank}(\mathbf{X}^T\mathbf{X})$

**Note:** if $\mathrm{rank}(\mathbf{X}^T\mathbf{X}) < p$ then $\mathbf{X}^T\mathbf{X}$ is not invertible



# Issues with high dimensional data ($p \gg n$)

*"The curse of high-dimensionality"* (Donoho, 2000)

- **Geometry:** counter-intuitive behavior of metrics (Aggarwal et al., 2001)
  - → Representation: how to visualize thousands of variables?
- **Optimization:** numerical singularities due to complex dependencies (colinearity)
- **Computational efficiency and scalability**

# Issues with high dimensional data ($p \gg n$)

*"The curse of high-dimensionality"* (Donoho, 2000)

- **Geometry:** counter-intuitive behavior of metrics (Aggarwal et al., 2001)

  → Representation: how to visualize thousands of variables?

- **Optimization:** numerical singularities due to complex dependencies (colinearity)

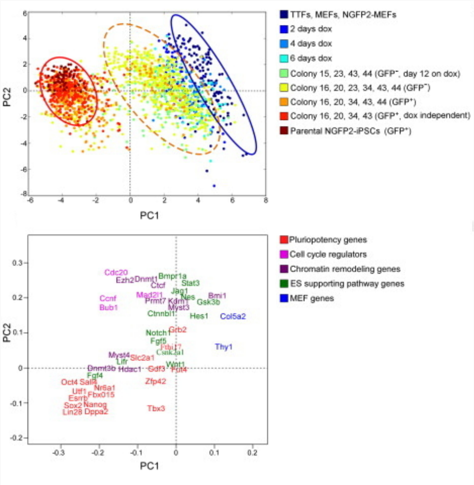
- **Computational efficiency and scalability**

→ Dimension reduction approaches

# Dimension reduction

## Statistical challenges

- Data exploration
- Visualization / clustering

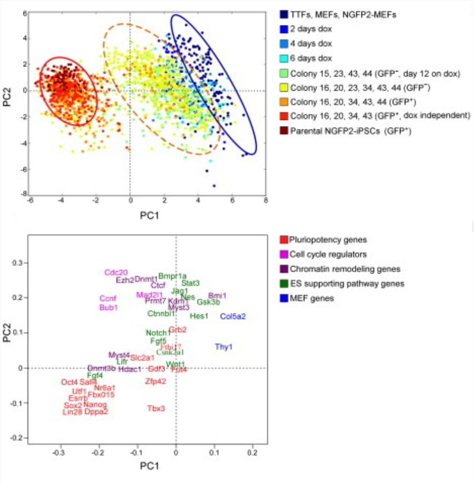


Buganim et al. (2012)

# Dimension reduction

- Representation of the data in a **lower dimensional subspace**
- **Consider sparsity**

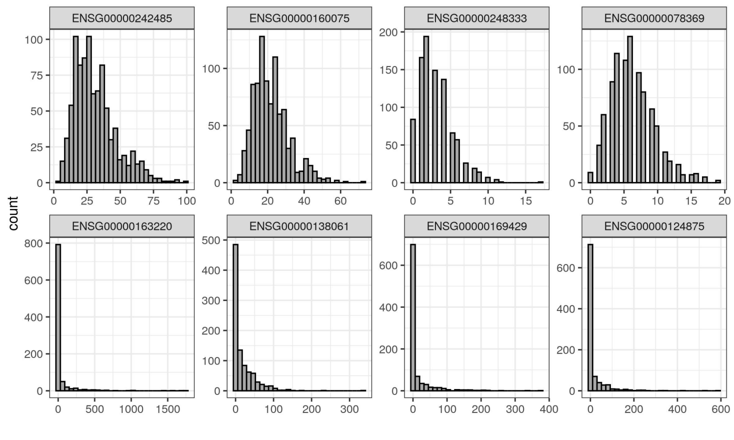
  → select the variables (genes) that contribute to this representation



Buganim et al. (2012)

# Zero-inflation and drop-out events

- No gene expression
- Transcription is **bursty** (cells are not synchronized)
- Failure of the sequencing (**dropout events** = loss of the information)



Gene expression distribution (Freytag et al., 2018, “goldstandard” dataset)

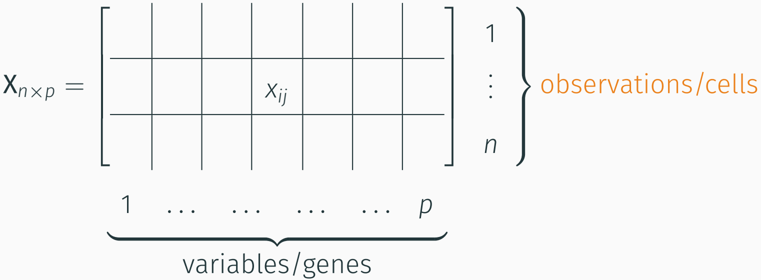
# Dimension reduction with matrix factorization

# Data dimension

$$\mathsf{X}_{n\times p} = \begin{array}{c} \left[\begin{array}{c|c|c|c|c|c|c} \quad & \quad & \quad & \quad & \quad & \quad & \quad \\ \hline & & & x_{ij} & & & \\ \hline & & & & & & \end{array}\right] \begin{array}{c} 1 \\ \vdots \\ n \end{array} \left.\vphantom{\begin{array}{c} 1 \\ \vdots \\ n \end{array}}\right\} \text{observations/cells} \\ \underbrace{\begin{array}{cccccc} 1 & \dots & \dots & \dots & \dots & p \end{array}}_{\text{variables/genes}} \end{array}$$

- $n$ individuals (cells) with $p$ recordings
- $p$ variables (genes) with $n$ observations

# Data dimension



$X_{n \times p} =$ $x_{ij}$

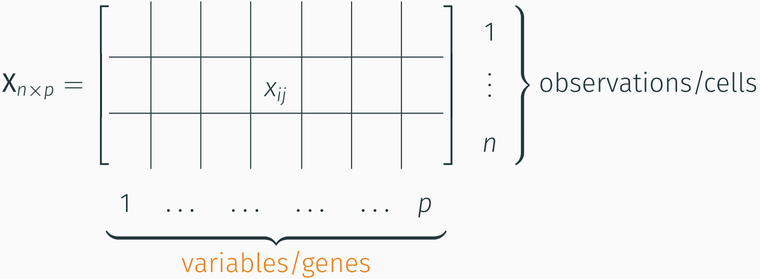
$1$ $\vdots$ $n$ observations/cells

$1$ $\dots$ $\dots$ $\dots$ $\dots$ $p$

variables/genes

- $n$ individuals (cells) with $p$ recordings
- $p$ variables (genes) with $n$ observations

# Data dimension



$$X_{n\times p} =$$

$x_{ij}$

1 ⋮ $n$

observations/cells

1 … … … … $p$

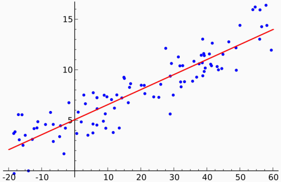
variables/genes

- $n$ individuals (cells) with $p$ recordings
- $p$ variables (genes) with $n$ observations

# Data visualization

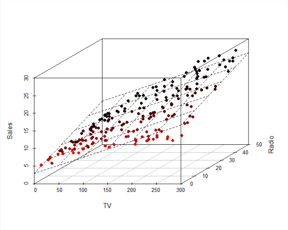
## Regression

### 2-d



Source: `wikipedia.org`

### 3-d



Source: `stackoverflow.com`

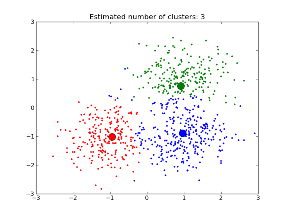
### Dimension > 3

?

# Data visualization

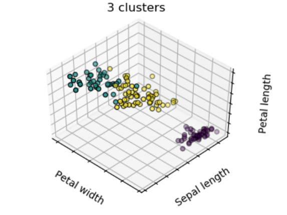
## Clustering

### 2-d



Source: `stackoverflow.com`

### 3-d



Source: `scikit-learn.org`

### Dimension > 3

?

# Low dimensional representation

$$\mathbf{U}_{n\times K} = \begin{array}{cc} \left(\begin{array}{c|c|c} & & \\ \hline & u_{ik} & \\ \hline & & \end{array}\right) & \begin{array}{c} 1 \\ \vdots \\ n \end{array} \\ \mathbf{u}_1 \ \ldots \ \mathbf{u}_K & \end{array} \quad \text{and} \quad \mathbf{V}_{p\times K} = \begin{array}{cc} \begin{array}{c} 1 \\ \vdots \\ \vdots \\ p \end{array} & \left(\begin{array}{c|c|c} & & \\ \hline & v_{jk} & \\ \hline & & \\ \hline & & \end{array}\right) \\ & \mathbf{v}_1 \ \ldots \ \mathbf{v}_K \end{array}$$

## Visualization/clustering

- Represention of individuals/cells (columns of **U**) in dimension $K < p$
- Contribution of variables/genes (columns of **V**) in dimension $K < n$

# Low dimensional representation

$$\mathbf{U}_{n\times K} = \begin{array}{cc} \begin{pmatrix} & & \\ & u_{ik} & \\ & & \end{pmatrix} & \begin{array}{c} 1 \\ \vdots \\ n \end{array} \\ \mathbf{u}_1 \;\ldots\; \mathbf{u}_K & \end{array} \qquad \text{and} \qquad \mathbf{V}_{p\times K} = \begin{array}{cc} \begin{array}{c} 1 \\ \vdots \\ \vdots \\ p \end{array} & \begin{pmatrix} & & \\ & v_{jk} & \\ & & \\ & & \end{pmatrix} \\ & \mathbf{v}_1 \;\ldots\; \mathbf{v}_K \end{array}$$

## Visualization/clustering

- Represention of individuals/cells (columns of **U**) in dimension $K < p$
- Contribution of variables/genes (columns of **V**) in dimension $K < n$

# Low dimensional representation

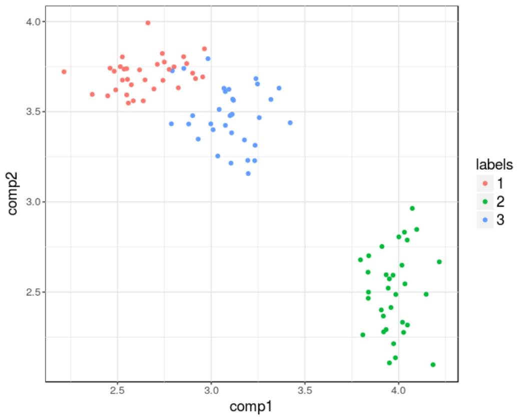
$$\mathbf{U}_{n\times K} = \begin{array}{cc} \left(\begin{array}{c|c|c} & & \\ \hline & u_{ik} & \\ \hline & & \end{array}\right) & \begin{array}{c} 1 \\ \vdots \\ n \end{array} \\ \mathbf{u}_1 \; \ldots \; \mathbf{u}_K & \end{array} \quad \text{and} \quad \mathbf{V}_{p\times K} = \begin{array}{cc} \begin{array}{c} 1 \\ \vdots \\ \vdots \\ p \end{array} & \left(\begin{array}{c|c|c} & & \\ \hline & v_{jk} & \\ \hline & & \\ \hline & & \end{array}\right) \\ & \mathbf{v}_1 \; \ldots \; \mathbf{v}_K \end{array}$$

## Visualization/clustering

- Represention of individuals/cells (columns of **U**) in dimension $K < p$
- Contribution of variables/genes (columns of **V**) in dimension $K < n$

# Low dimensional representation

Data visualization with $K = 2$

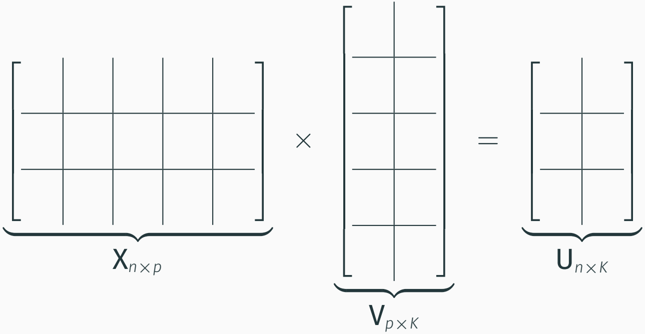


- scatter plot of $(u_{i1}, u_{i2})_{i=1:n}$

- Expose hidden/latent structures

- Principal Component Analysis (PCA)?

# Latent space projection

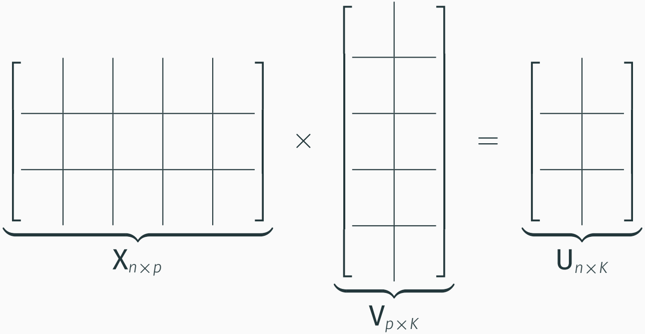

$\mathbf{X}_{n \times p}$
$\times$
$\mathbf{V}_{p \times K}$
$=$
$\mathbf{U}_{n \times K}$

- Linear **projection** of X onto a lower dimensional space (of dim. $K$)

$$u_{ik} = \sum_{j=1}^{p} x_{ij}\, v_{jk}$$

$$\mathbf{u}_k = \mathbf{X}\,\mathbf{v}_k \;\text{ with }\; \begin{cases} \mathbf{u}_k \in \mathbb{R}^n \\ \mathbf{v}_k \in \mathbb{R}^p \end{cases}$$

# Principal Component Analysis (PCA; Pearson, 1901; Hotelling, 1933)



$$\mathbf{v}_k = \underset{\mathbf{v}\in\mathbb{R}^p}{\operatorname{argmax}}\ \operatorname{Var}(\mathbf{X}\mathbf{v})$$

- Additional constraints: $\|v\|_2 = 1$ and $\mathbf{t}_k$ orthogonal to $\mathbf{u}_1, \ldots, \mathbf{u}_{k-1}$

# Variable selection

$$\underbrace{\begin{bmatrix} & & & & \\ & & & & \\ & & & & \end{bmatrix}}_{X_{n\times p}} \times \underbrace{\begin{bmatrix} 0 & \\ 0 & 0 \\ & \\ & 0 \\ 0 & 0 \end{bmatrix}}_{V_{p\times K}} = \underbrace{\begin{bmatrix} & \\ & \\ & \end{bmatrix}}_{U_{n\times K}}$$

- **Enforce sparsity:** only a few variables contribute to the model
- **Objective:** drop non relevant variables from the model

# Sparse PCA (Zou et al., 2006)

$$\underbrace{\begin{bmatrix} & & & & \\ & & & & \\ & & & & \end{bmatrix}}_{\mathbf{X}_{n\times p}} \times \underbrace{\begin{bmatrix} 0 & \\ 0 & 0 \\ & \\ & 0 \\ 0 & 0 \end{bmatrix}}_{\mathbf{V}_{p\times K}} = \underbrace{\begin{bmatrix} & \\ & \\ & \end{bmatrix}}_{\mathbf{U}_{n\times K}}$$

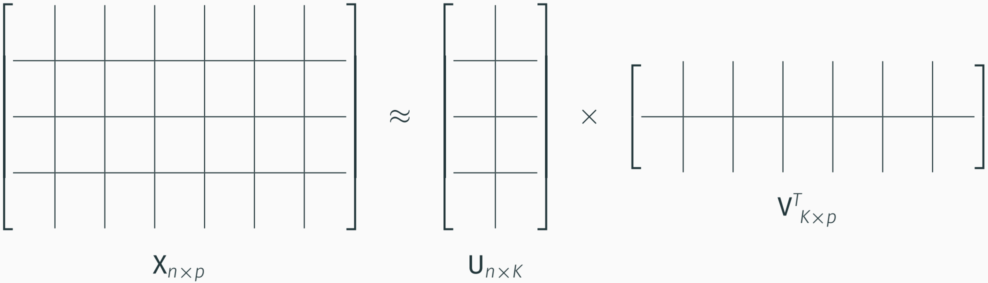
$$\underset{\mathbf{v}\in\mathbb{R}^p}{\operatorname{argmax}} \operatorname{Var}(\mathbf{X}\mathbf{v}) + \lambda \sum_{j=1}^{p} |v_j|$$

- Lasso principle Tibshirani (1996): $\ell_1$ penalty on the weights $\mathbf{v}$

# Why matrix factorization?

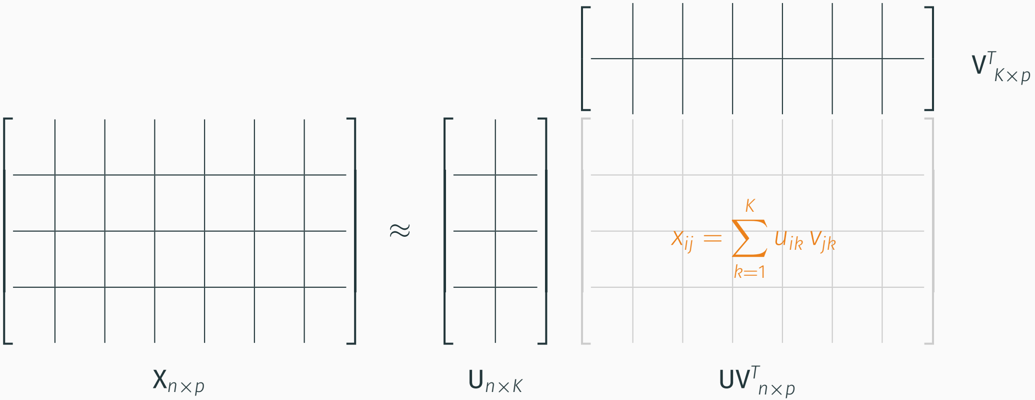
→ generalization of PCA principle

# Matrix factorization $X \approx UV^T$



$$x_{ij} = \sum_{k=1}^{K} u_{ik}\, v_{jk}$$

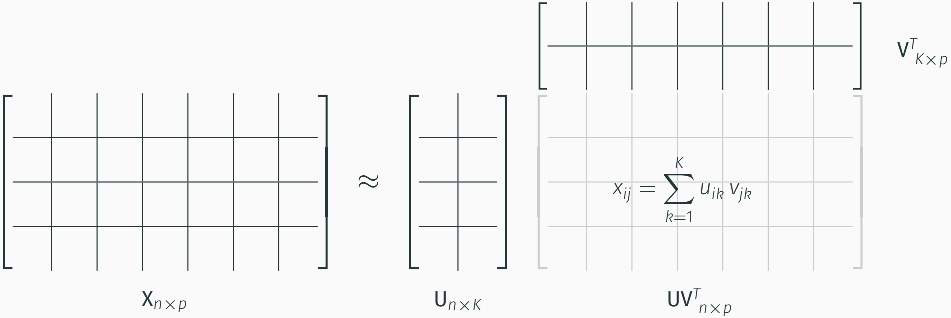
# Matrix factorization $X \approx UV^T$



$$x_{ij} = \sum_{k=1}^{K} u_{ik} \, v_{jk}$$

# Matrix factorization $\mathbf{X} \approx \mathbf{U}\mathbf{V}^T$

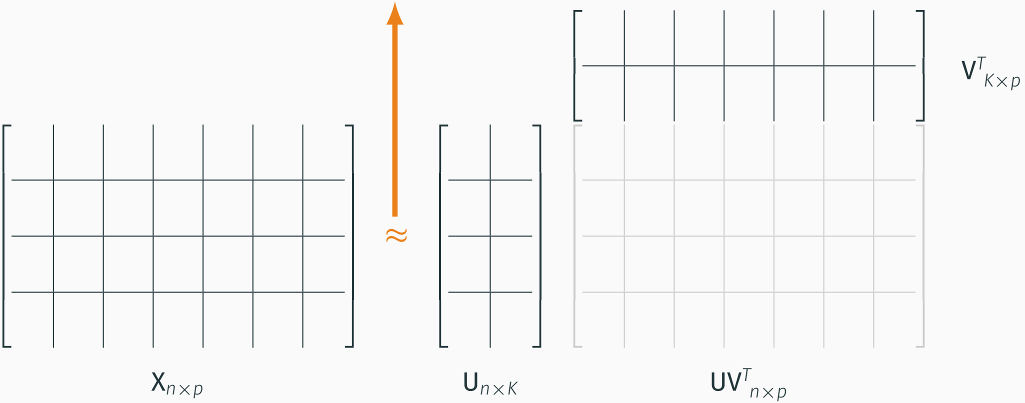
- Individuals (cells): $\mathbf{U} \in \mathbb{R}^{n \times K}$
- Variables (genes): $\mathbf{V} \in \mathbb{R}^{p \times K}$


$\mathbf{V}^T_{K \times p}$
$\mathbf{X}_{n \times p}$
$\approx$
$\mathbf{U}_{n \times K}$
$x_{ij} = \sum_{k=1}^{K} u_{ik} \, v_{jk}$
$\mathbf{U}\mathbf{V}^T_{n \times p}$

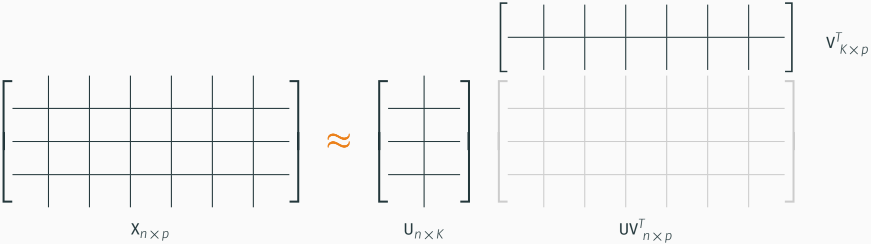
- $K$ = latent dimension (hopefully small)
- $\mathbf{U}\mathbf{V}^T$ = low-rank representation of $\mathbf{X}$

# Approximation $X \approx UV^T$?

Sense of the approximation?



# Approximation $X \approx UV^T$?



Least Square Approximation?

$$\underset{\substack{U\in\mathbb{R}^{n\times K}\\ V\in\mathbb{R}^{p\times K}}}{\operatorname{argmin}} \|X - UV^T\|_F^2$$

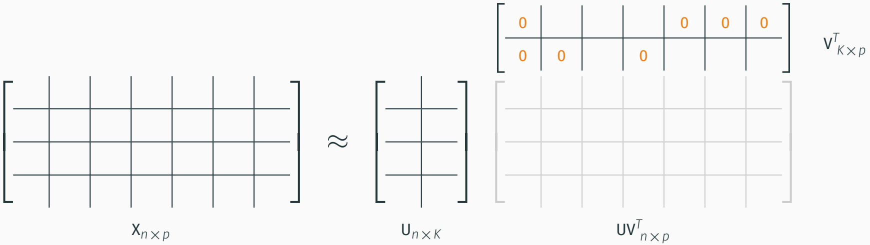
# Approximation $\mathbf{X} \approx \mathbf{U}\mathbf{V}^T$?

Least Square Approximation?

$$\underset{\substack{\mathbf{U}\in\mathbb{R}^{n\times K}\\ \mathbf{V}\in\mathbb{R}^{p\times K}}}{\operatorname{argmin}} \left\|\mathbf{X} - \mathbf{U}\mathbf{V}^T\right\|_F^2$$

- Solution given by Singular Value Decomposition (SVD Eckart and Young, 1936)
- PCA = SVD of $\widetilde{\mathbf{X}}$ where $\widetilde{x}_{ij} = x_{ij} - \bar{\mathbf{x}}_j$

# Sparse matrix factorization



Sparse SVD (Shen and Huang, 2008; Witten et al., 2009)

$$\underset{\substack{\mathbf{u}\in\mathbb{R}^n\\ \mathbf{v}\in\mathbb{R}^p}}{\operatorname{argmin}} \left\{ \left\| \mathbf{X} - \mathbf{u}\mathbf{v}^T \right\|_F^2 + \lambda \sum_{j=1}^{p} |v_j| \right\}$$

- $\ell_1$ penalty shrinks contributions of non pertinent variables to zero

# RNA-seq data = counts

Relation between geometry and underlying model

$\|\cdot\|_2 \leftrightarrow$ Gaussian distribution

Gaussian SVD?

- $\mathrm{X} \sim \mathcal{N}\left(\mathrm{UV}^T\,,\ \boldsymbol{\Sigma}^2\right)$ i.e. $X_{ij} \sim \mathcal{N}\left(\sum_k u_{ik}\, v_{jk}\,,\ \sigma_{ij}^2\right)$
- $\log \mathcal{L}(\mathrm{U}, \mathrm{V}) = \left\|\mathrm{X} - \mathrm{UV}^T\right\|_F^2$

# RNA-seq data = counts

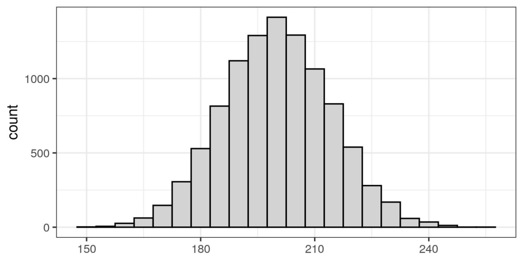
## Relation between geometry and underlying model

$\| \cdot \|_2 \leftrightarrow$ Gaussian distribution

- Count = not Gaussian
- First idea: $X_{ij} \sim \mathscr{P}(\lambda)$
- Highly expressed genes

  $\rightarrow$ large $\lambda$

  $\rightarrow$ Gaussian approximation



Empirical distribution, counts drawn from $\mathscr{P}(200)$

# RNA-seq data = counts

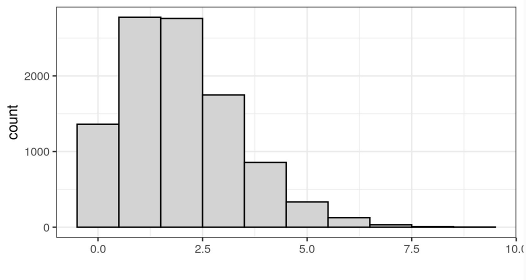
## Relation between geometry and underlying model

$\|\cdot\|_2 \leftrightarrow$ Gaussian distribution

- Lowly expressed genes in single-cell RNA-seq?
- Poisson assumption?

  → Why not Negative Binomial?

  → $\mathcal{NB}$ suitable for RNA-seq (Anders and Huber, 2010) and for scRNA-seq (Chen et al., 2016)



Empirical distribution, counts drawn from $\mathscr{P}(2)$

# RNA-seq data = counts

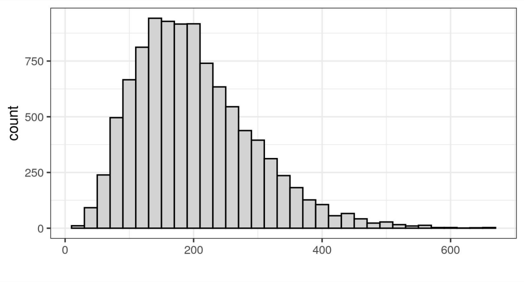
## Relation between geometry and underlying model

$\| \cdot \|_2 \leftrightarrow$ Gaussian distribution

- Lowly expressed genes in single-cell RNA-seq?
- Poisson assumption?

  → Why not Negative Binomial?

  → $\mathcal{NB}$ suitable for RNA-seq (Anders and Huber, 2010) and for scRNA-seq (Chen et al., 2016)



Empirical distribution, counts drawn from $\mathcal{NB}(n = 5, p = 2.5E-3)$

# Zero-inflated over-dispersed counts

1) Interest for lowly expressed genes in single-cell

2) Over-dispersion in RNA-seq data

   $\rightarrow \mathrm{Var}(X_{ij}) > \mathbb{E}[X_{ij}]$

3) Single-cell data: zero-inflation

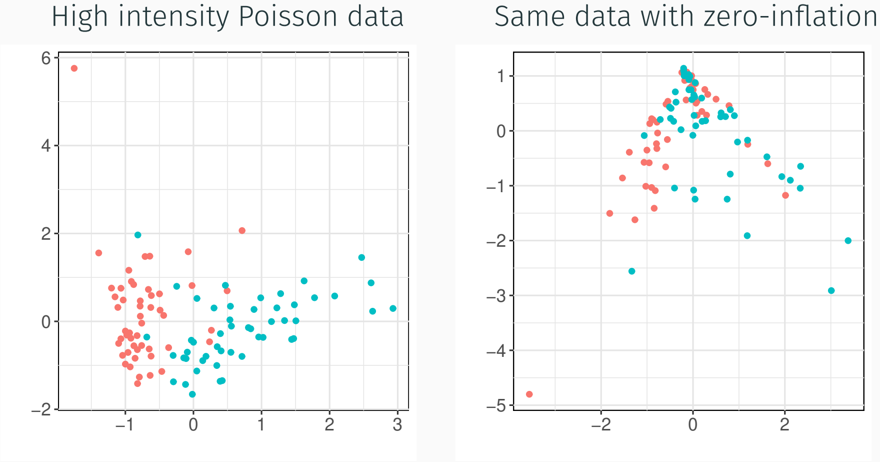
   $\rightarrow \mathbb{P}(X_{ij} = 0) > e^{-\lambda}$

# Zero-inflated over-dispersed counts

| Dataset | $n$ | $p$[1] | prop. 0 |
|---|---|---|---|
| Baron et al. (2016) | 1886 | 6080 | 80.9% |
| Freytag et al. (2018)<br>goldstandard | 925 | 8580 | 39.5% |
| Freytag et al. (2018)<br>silverstandard 5 | 8352 | 4547 | 86.3% |
| Llorens-Bobadilla et al. (2015) | 141 | 13826 | 64.8% |

[1] after pre-filtering

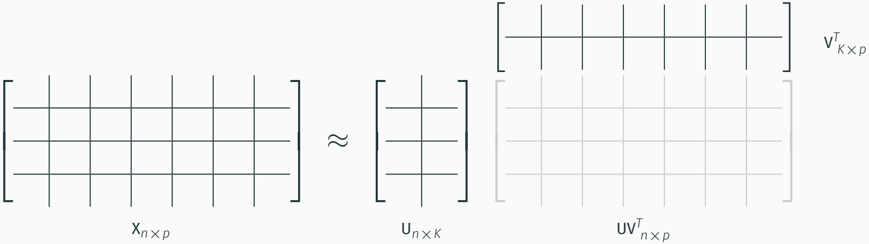
# PCA on zero-inflated count data



Observations scores over first two principal components

# Matrix Factorization for count data

# Non-negative Matrix Factorization (NMF, Lee and Seung, 1999)



Least Square Approximation with non-negativity constraints

$$\underset{\substack{U\in\mathbb{R}^{n\times K}\\ V\in\mathbb{R}^{p\times K}}}{\operatorname{argmin}} \left\|X - UV^T\right\|_F^2 \qquad \text{where} \qquad u_{ik}, v_{jk} \geq 0 \text{ for any } i, j, k$$

# Generalized PCA (Collins et al., 2001)

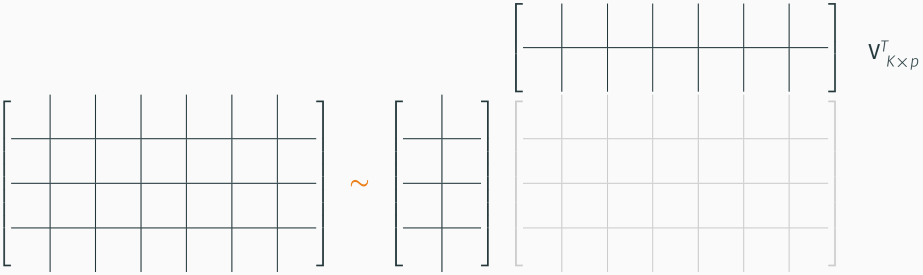
Embed PCA with a **probabilistic model**

- Replace $\|\cdot\|_2$ approximation by likelihood-based approaches
- $X_{ij} \sim$ probability distribution in the exponential family

→ Factorization of $\mathbb{E}[\mathbf{X}]$ rather than $\mathbf{X}$

$$\mathbb{E}[X_{ij}] = \sum_{k=1}^{K} u_{ik} v_{jk}$$

# Poisson-NMF (Lee and Seung, 1999)



$\sim$

$\mathbf{V}^T{}_{K\times p}$

$\mathbf{X}_{n\times p} \sim \mathscr{P}(\boldsymbol{\Lambda})$

$\mathbf{U}_{n\times K}$

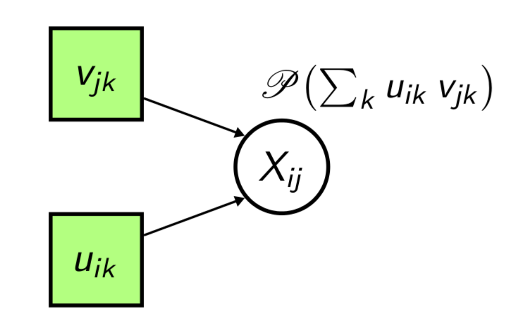
$\boldsymbol{\Lambda} \approx \mathbf{U}\mathbf{V}^T{}_{n\times p}$

- $X_{ij} \sim \mathscr{P}(\lambda_{ij})$ with the Poisson rate matrix $\boldsymbol{\Lambda} = [\lambda_{ij}]_{n\times p}$

- Factorization: $\mathbb{E}[\mathbf{X}] = \boldsymbol{\Lambda} \approx \mathbf{U}\mathbf{V}^T \quad \leftrightarrow \quad \lambda_{ij} \approx \sum_k u_{ik}\, v_{jk}$

# Poisson-NMF (Lee and Seung, 1999)

- Maximum Likelihood Estimation under non-negativity constraint over **U** and **V**



- **U** and **V** are parameters
- Optimization computationally expensive
- No account for over-dispersion or zero-inflation

# Geometry vs model

Bregman divergence between X and $\mathrm{UV}^T$

$$D\left(\mathrm{X} \mid \mathrm{UV}^T\right)$$

- Based on the parametrization in the exponential family

$$\text{Poisson:} \quad D(x_{ij} \mid \lambda_{ij}) = x_{ij} \log \frac{x_{ij}}{\lambda_{ij}} - x_{ij} + \lambda_{ij}$$

→ Connected to the likelihood

→ Choice of the geometry driven by the model

# Matrix factorization for over-dispersed zero-inflated count data?

- Probabilistic matrix factorization
- Hierarchical model: prior on factors $\mathbf{U}$ and $\mathbf{V}$
- Model inference: likelihood optimization → variational inference
- Impose sparsity on $\mathbf{V}$: how to select variables?
  → Probabilistic selection

# Using a Negative Binomial NMF?

- Negative Binomial = standard distribution for over-dispersed count
- $X_{ij} \sim \mathcal{NB}(r_{ij}, \pi_{ij}) \rightarrow$ complex optimization of the likelihood

# Using a Negative Binomial NMF?

- Negative Binomial = standard distribution for over-dispersed count

- $X_{ij} \sim \mathcal{NB}(r_{ij}, \pi_{ij}) \rightarrow$ complex optimization of the likelihood

- Gamma-Poisson hierarchical model

$$\lambda_{ij} \sim \Gamma(\alpha_1, \alpha_2)$$

$$X_{ij} \mid \lambda \sim \mathscr{P}(\lambda_{ij})$$

$\rightarrow$ Marginal distribution of $X_{ij}$ is a Negative Binomial distribution:

$$X_{ij} \sim \mathcal{NB}\left(\alpha_1, \frac{\alpha_2}{\alpha_2 + 1}\right)$$

# Gamma-Poisson factor model (Cemgil, 2009)

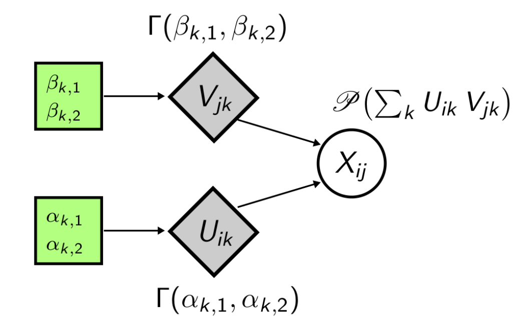
- Independent Gamma prior distributions over factors **U** and **V**:

$$U_{ik} \sim \Gamma(\alpha_{k,1}, \alpha_{k,2}) \text{ and } V_{jk} \sim \Gamma(\beta_{k,1}, \beta_{k,2})$$

- Conditional Poisson distribution over the data **X**:

$$X_{ij} \mid (U_{ik}, V_{jk})_{k=1:K} \sim \mathscr{P}(\textstyle\sum_k U_{ik} V_{jk})$$

# Gamma-Poisson factor model (Cemgil, 2009)



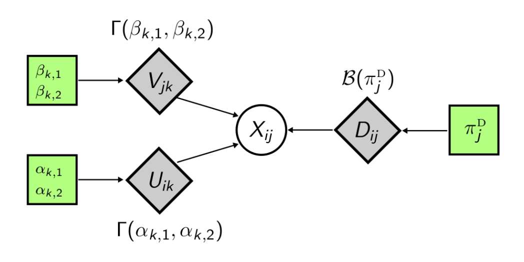
$\Gamma(\beta_{k,1}, \beta_{k,2})$
$\beta_{k,1}$
$\beta_{k,2}$
$V_{jk}$
$\mathscr{P}(\sum_k U_{ik} V_{jk})$
$X_{ij}$
$\alpha_{k,1}$
$\alpha_{k,2}$
$U_{ik}$
$\Gamma(\alpha_{k,1}, \alpha_{k,2})$

- Factors = latent variables
- Recover the posterior $\widehat{\mathsf{U}} = \mathbb{E}[\mathsf{U} \mid \mathsf{X}]$ and $\widehat{\mathsf{V}} = \mathbb{E}[\mathsf{V} \mid \mathsf{X}]$
- Marginal distribution is over-dispersed $\mathrm{Var}(X_{ij}) > \mathbb{E}[X_{ij}]$

# Probabilistic Count Matrix Factorization (pCMF, Durif et al., 2019)

1. “Zero-inflated” Gamma-Poisson factor model

- **Poisson-Dirac mixture:** $X_{ij} \mid (U_{ik}, V_{jk})_{k=1:K} \sim (1 - \pi_j^{\mathrm{D}}) \times \delta_0 + \pi_j^{\mathrm{D}} \times \mathscr{P}(\lambda_{ij})$
- $1 - \pi_j^{\mathrm{D}} \in [0, 1]$ is the dropout rate for gene $j$



- $D_{ij}$ = drop-out event indicator
- $\mathbb{P}(X_{ij} = 0 \mid \mathbf{U}, \mathbf{V}) > e^{-\lambda_{ij}}$

# Sparsity on V?

- Variable $j$ contributes to factor $k$ if $V_{jk} \neq 0$
- Objective: force the $V_{jk}$'s to be null for non pertinent genes

If **V** was a parameter → $\ell_1$ penalty

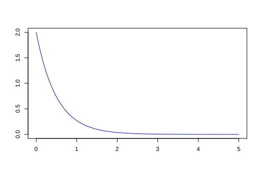
$$\underset{\substack{\mathbf{u}\in\mathbb{R}^n \\ \mathbf{v}\in\mathbb{R}^p}}{\text{argmin}} \left\{ \left\| \mathbf{X} - \mathbf{u}\mathbf{v}^T \right\|_F^2 + \lambda \sum_{j=1}^{p} |v_j| \right\}$$

# Probabilistic variable selection

- $V_{jk}$ is a random variable → necessary to use sparsity-inducing priors

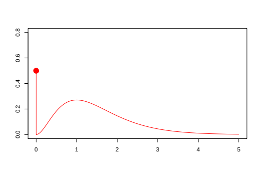
## Spike and slab model:

- Continuous one-group prior: shrinkage to small value near zero (ex: Bayesian Lasso with Laplace prior)



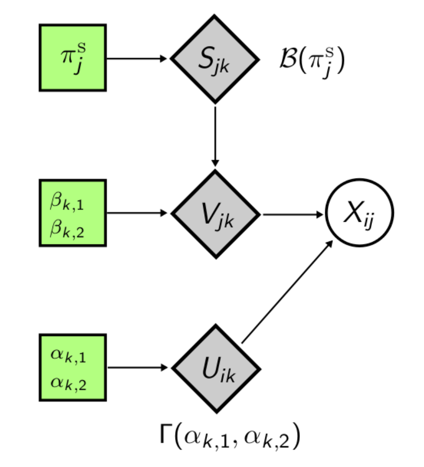
- **Two-group prior:** mixture between a Dirac and a continuous distribution, true mass at zero

  → to induce a "sparse" posterior with a mass at zero



# Probabilistic Count Matrix Factorization (pCMF, Durif et al., 2019)

2. Sparse Gamma-Poisson model



- Probabilistic variable selection
- **Gamma-Dirac mixture**
  $V_{jk} \sim (1 - \pi_j^S)\,\delta_0 + \pi_j^S\,\Gamma(\beta_{k,1}, \beta_{k,2})$
- $\pi_j^S \in [0, 1]$ probability that gene $j$ contributes to the model
- $S_{jk}$ = sparsity indicator

# Model inference in pCMF

- Objective: estimation of the factors $\mathbf{U}$ and $\mathbf{V}$

- $\mathscr{L}(\mathbf{U} \,|\, \mathbf{X})$ and $\mathscr{L}(\mathbf{V} \,|\, \mathbf{X})$ are not explicit

  $\rightarrow$ **Cannot use** Maximum a Posteriori (MAP) or Expectation-Maximization (EM)

  $\rightarrow$ Inference of the posterior of latent variables

- Markov Chain Monte Carlo (MCMC) are computationally expensive

- Variational inference[2]: approximation of the posterior

[2] See supplementary slides at the end for more details

# Model inference in pCMF

- Objective: estimation of the factors $\mathbf{U}$ and $\mathbf{V}$

- $\mathscr{L}(\mathbf{U} \,|\, \mathbf{X})$ and $\mathscr{L}(\mathbf{V} \,|\, \mathbf{X})$ are not explicit

  $\rightarrow$ **Cannot use** Maximum a Posteriori (MAP) or Expectation-Maximization (EM)

  $\rightarrow$ Inference of the posterior of latent variables

- Markov Chain Monte Carlo (MCMC) are computationally expensive

- Variational inference[2]: approximation of the posterior

[2] See supplementary slides at the end for more details

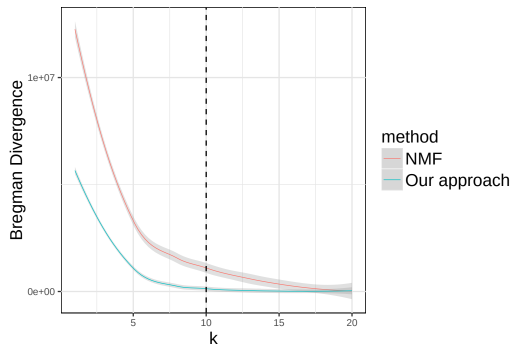
# Choice of $K$?

- Concerns all low rank methods
- No consensus procedure
- Explained variance

  $\rightarrow$ $\ell_2$ metric criterion
- Bregman divergence:

$$k \mapsto D\left(\mathbf{X} \mid \widehat{\mathbf{U}}_{1:k}(\widehat{\mathbf{V}}_{1:k})^T\right)$$

- Visualization: $K = 2$



# Quality of the model

## Percentage of explained deviance

$$\%\mathrm{dev} = \frac{\log p(\mathrm{X} \mid \boldsymbol{\Lambda} = \widehat{\mathrm{U}}\widehat{\mathrm{V}}^{T}) - \log p(\mathrm{X} \mid \boldsymbol{\Lambda} = \bar{\mathrm{X}})}{\log p(\mathrm{X} \mid \boldsymbol{\Lambda} = \mathrm{X}) - \log p(\mathrm{X} \mid \boldsymbol{\Lambda} = \bar{\mathrm{X}})}$$

- $\log p(\mathrm{X} \mid \boldsymbol{\Lambda})$: conditional distribution in the model
- $\boldsymbol{\Lambda} = \mathrm{X}$: saturated model
- $\boldsymbol{\Lambda} = \bar{\mathrm{X}}$: moment estimator (column-wise average)

# Quality of the model

## Percentage of explained deviance

$$\%\text{dev} = \frac{\log p(\mathrm{X} \mid \boldsymbol{\Lambda} = \widehat{\mathrm{U}}\widehat{\mathrm{V}}^T) - \log p(\mathrm{X} \mid \boldsymbol{\Lambda} = \bar{\mathrm{X}})}{\log p(\mathrm{X} \mid \boldsymbol{\Lambda} = \mathrm{X}) - \log p(\mathrm{X} \mid \boldsymbol{\Lambda} = \bar{\mathrm{X}})}$$

- $\log p(\mathrm{X} \mid \boldsymbol{\Lambda})$: conditional distribution in the model
- $\boldsymbol{\Lambda} = \mathrm{X}$: saturated model
- $\boldsymbol{\Lambda} = \bar{\mathrm{X}}$: moment estimator (column-wise average)

## Example in the Gaussian case

$$\%\text{dev} = \frac{\sum_{k=1}^{K} \sigma_k^2}{\sum_{\ell=1}^{\text{rk}(\mathrm{X})} \sigma_\ell^2} = \text{\% explained variance from } \textbf{PCA}$$

$\sigma_1 > \ldots > \sigma_{\text{rk}(\mathrm{X})}$ = singular values of **X**

# Quality of the model

t-SNE[3] (van der Maaten and Hinton, 2008)

- No measure of the quality of the representation
- How to choose the “perplexity” parameter?

[3] C.f. later

# Experiments

# t-SNE (van der Maaten and Hinton, 2008)

t-Stochastic Neighbourhood Embedding

**Dimension** $p$

- Observations: $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$
- Probabilistic distribution $\mathcal{P}$ on pairwise distance $d(\mathbf{x}_i, \mathbf{x}_{i'})$

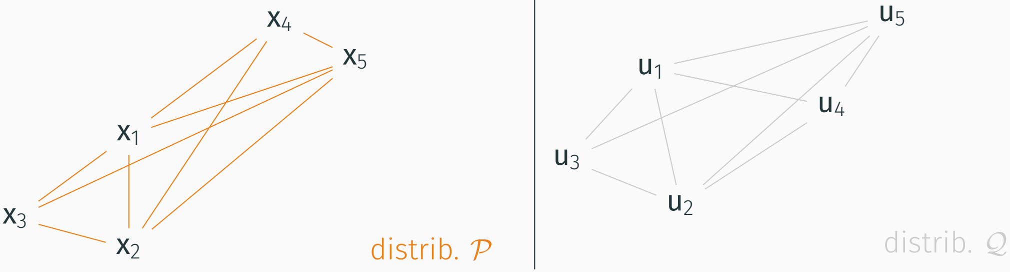
**Dimension** 2

- Low-dimensional obervations: $\mathbf{u}_1, \ldots, \mathbf{u}_n \in \mathbb{R}^2$
- Probabilistic distribution $\mathcal{Q}$ on pairwise distance $d(\mathbf{u}_i, \mathbf{u}_{i'})$

# t-SNE (van der Maaten and Hinton, 2008)

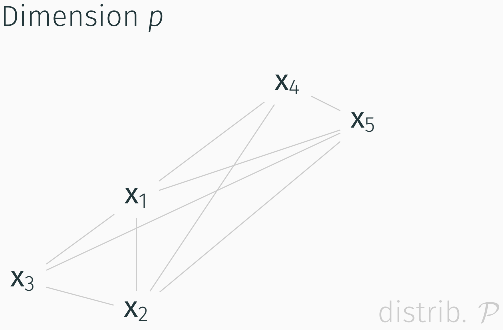
Dimension $p$

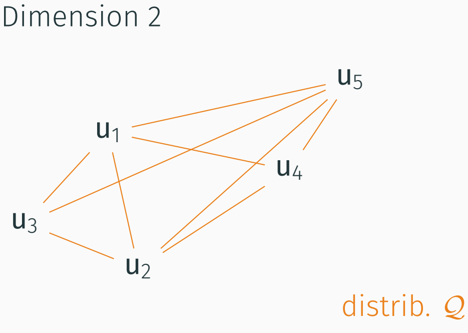
Dimension 2



$$\underset{U\in\mathbb{R}^{n\times 2}}{\operatorname{argmin}}\ \mathrm{KL}\left(\mathcal{P} \mid \mathcal{Q}\right)$$

# t-SNE (van der Maaten and Hinton, 2008)





$$\underset{\mathbf{U} \in \mathbb{R}^{n \times 2}}{\operatorname{argmin}} \operatorname{KL}\left(\mathcal{P} \mid \mathcal{Q}\right)$$

# t-SNE (van der Maaten and Hinton, 2008)

Perplexity parameter?

- “You see what you want”
- What happens if you don’t know what you are looking for?
- See `https://distill.pub/2016/misread-tsne/`

# Zero-Inflated Factor Analysis (ZIFA, Pierson and Yau, 2015)

Gaussian PCA with zero-inflated compartment

$$X_{ij} \;\sim\; (1-\pi_j)\delta_0 \;+\; \pi_j \mathcal{N}\left(\sum_k U_{ik} v_{jk}\,,\;\sigma_{ij}^2\right)$$

Latent factors:

- $U_{ik} \sim \mathcal{N}(\cdot, \cdot)$
- $v_{jk} =$ parameter

# Zero-Inflated Factor Analysis (ZIFA, Pierson and Yau, 2015)

Gaussian PCA with zero-inflated compartment

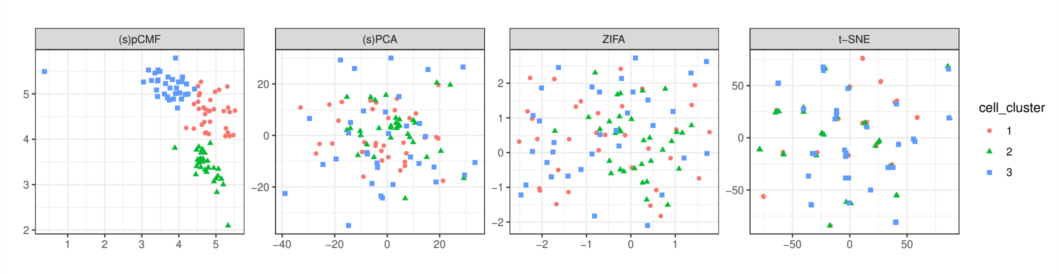
$$X_{ij} \;\sim\; (1-\pi_j)\delta_0 \;+\; \pi_j \mathcal{N}\left(\textstyle\sum_k U_{ik} v_{jk} \,,\; \sigma_{ij}^2\right)$$

Latent factors:

- $U_{ik} \sim \mathcal{N}(\cdot, \cdot)$
- $v_{jk} =$ parameter

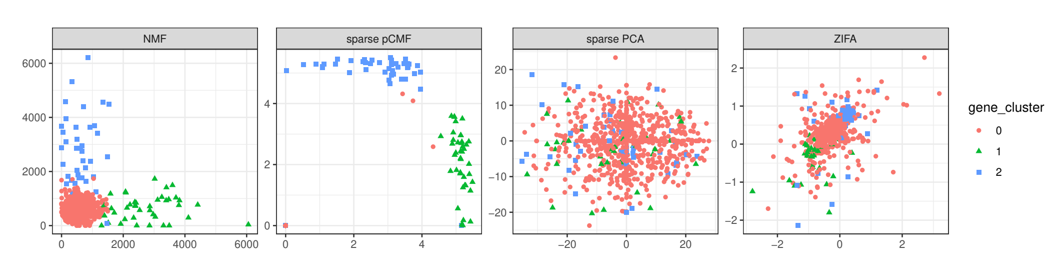
# Simulation: data visualization

## $\mathbf{u}_1$ vs $\mathbf{u}_2$ (individual representation)

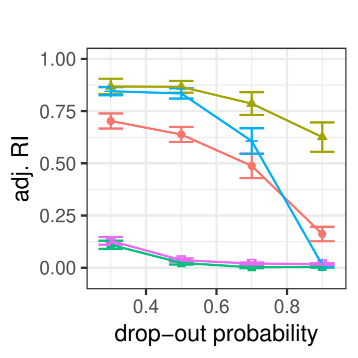


# Simulation: data visualization

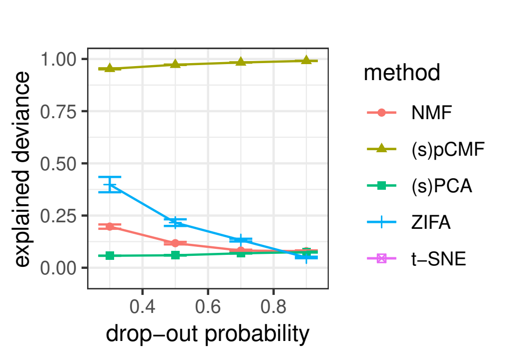
$\mathbf{v}_1$ vs $\mathbf{v}_2$ (variable representation)



# Simulation: quantitative results

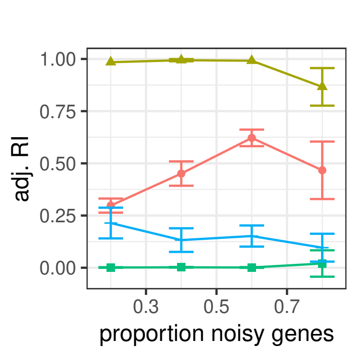


Clustering of individuals

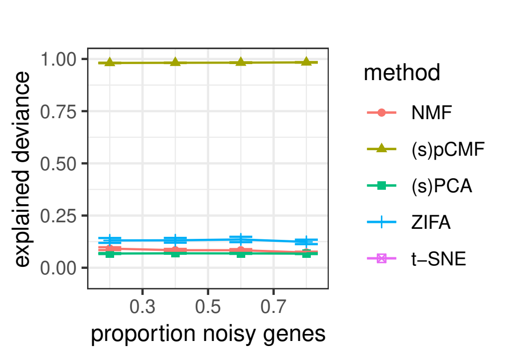


Quality of the model

# Simulation: quantitative results



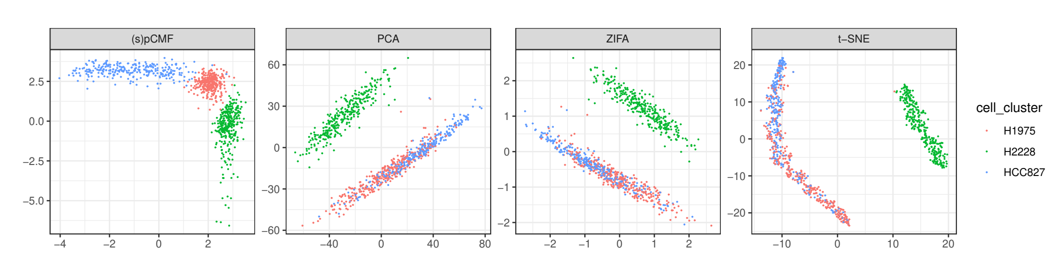
Clustering of variables



Quality of the model

# scRNA-seq: data visualization

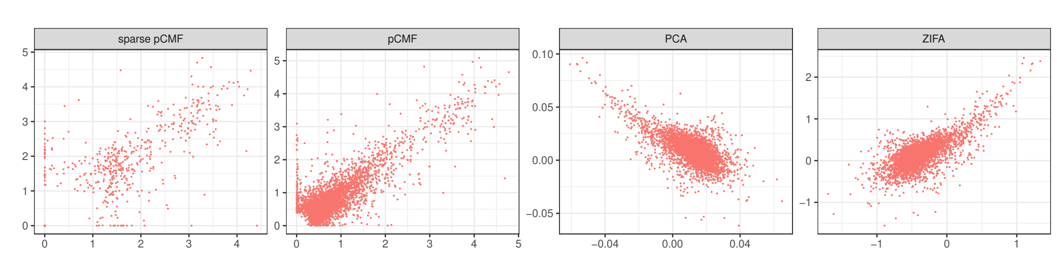
$\mathbf{u}_1$ vs $\mathbf{u}_2$ (individual representation)



Freytag et al. (2018) goldstandard dataset

# scRNA-seq: data visualization

$\mathbf{v}_1$ vs $\mathbf{v}_2$ (variable representation)



Freytag et al. (2018) goldstandard dataset

# scRNA-seq: quantitative results

| | prop. 0 | ngroup | | (s)pCMF | PCA | ZIFA | t-SNE |
|---|---|---|---|---|---|---|---|
| Baron et al. (2016) | 80.9% | 13 | adj. RI | 21.2% | 14.3% | 15.4% | 14.2% |
| | | | %dev | 73.2% | 41.6% | 53.5% | / |
| Freytag et al. (2018) | 39.5% | 3 | adj. RI | 81.3% | 60.1% | 56.8% | 60.5% |
| goldstandard | | | %dev | 55.7% | 65.6% | 48.6% | / |
| Freytag et al. (2018) | 86.3% | 11 | adj. RI | 24.2% | 16.2% | 19.8% | 24.8% |
| silverstandard 5 | | | %dev | 70.0% | 55.1% | / | / |
| Llorens-Bobadilla et al. (2015) | 64.8% | 6 | adj. RI | 40.1% | 25.3% | 38.3% | 29.8% |
| | | | %dev | 64.4% | 34.8% | 42.6% | / |

# To conclude

# What I left out regarding pCMF

- Optimization algorithm → variational inference
- Algorithm initialization

  → especially sparse compartment (gene pre-filtering strategy)

# What I left out regarding pCMF

- Optimization algorithm → variational inference

- Algorithm initialization

  → especially sparse compartment (gene pre-filtering strategy)

# Take-home message

- Data visualization
  - PCA

    → hidden PCA?

    → representation with low of explained variance ($< 10\%$)?
  - t-SNE: clustering vs visualization a posteriori
- Dimension reduction (unsupervised)
- probabilistic Count Matrix Factorization

# Take-home message

- Data visualization
  - PCA
    - → hidden PCA?
    - → representation with low of explained variance ($< 10\%$)?
  - t-SNE: clustering vs visualization a posteriori
- Dimension reduction (unsupervised)
- probabilistic Count Matrix Factorization

## Take-home message

- Data visualization
    - PCA
        - → hidden PCA?
        - → representation with low of explained variance ($< 10\%$)?
    - t-SNE: clustering vs visualization a posteriori
- Dimension reduction (unsupervised)
- probabilistic Count Matrix Factorization

# Take-home message

- Data visualization
- Dimension reduction (unsupervised)

  → Latent space projection

  → Variable selection (sparsity)
- probabilistic Count Matrix Factorization

# Take-home message

- Data visualization

- Dimension reduction (unsupervised)

  → Latent space projection

  → Variable selection (sparsity)

- probabilistic Count Matrix Factorization

  → Model-based

  → Data-driven (count, over-dispersion, zero-inflation)

# Perspectives

## Count matrix Factorization

- Model selection criterion (choice of $K$)
- Calibration of the sparsity hyper-parameter
- Stochastic procedure to improve the optimization
- Extension to account for covariates in the model

# Thanks for your attention

Collaborators

- Franck Picard (CNRS, LBBE)
- Sophie Lambert-Lacroix (Université Grenoble Alpes, TIMC)
- Laurent Modolo (CNRS, LBMC)
- Jeff Mold (Karolinska Institutet, Stockholm)

Institutions

- LBBE, Lyon 1 University
- Inria Grenoble, Thoth team

# Thanks for your attention

Funding

*ANR "Investissement d'Avenir"* project **ABS4NGS**
*ANR* project **MACARON**
ERC **SOLARIS**

Paper

`https://doi.org/10.1093/bioinformatics/btz177`
`https://arxiv.org/abs/1710.11028`

Software (R)

`https://github.com/gdurif/pCMF`
`https://github.com/gdurif/pCMF_experiments`

# References

Aggarwal, C. C., A. Hinneburg, and D. A. Keim (2001). On the surprising behavior of distance metrics in high dimensional space. In *International Conference on Database Theory*, pp. 420–434. Springer.

Anders, S. and W. Huber (2010). Differential expression analysis for sequence count data. *Genome Biology 11*(10), R106.

Baron, M., A. Veres, S. L. Wolock, A. L. Faust, R. Gaujoux, A. Vetere, J. H. Ryu, B. K. Wagner, S. S. Shen-Orr, A. M. Klein, D. A. Melton, and I. Yanai (2016, October). A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cell systems 3*(4), 346–360.e4.

Beal, M. J. and Z. Ghahramani (2003). The variational Bayesian EM algorithm for incomplete data: With application to scoring graphical model structures. *Bayesian statistics 7*, 453–464.

Buganim, Y., D. A. Faddah, A. W. Cheng, E. Itskovich, S. Markoulaki, K. Ganz, S. L. Klemm, A. van Oudenaarden, and R. Jaenisch (2012, September). Single-Cell

Expression Analyses during Cellular Reprogramming Reveal an Early Stochastic and a Late Hierarchic Phase. *Cell 150*(6), 1209–1222.

Cemgil, A. T. (2009, January). Bayesian Inference for Nonnegative Matrix Factorisation Models. *Intell. Neuroscience 2009*, 4:1–4:17.

Chen, H.-I. H., Y. Jin, Y. Huang, and Y. Chen (2016, August). Detection of high variability in gene expression from single-cell RNA-seq profiling. *BMC Genomics 17*(Suppl 7).

Collins, M., S. Dasgupta, and R. E. Schapire (2001). A generalization of principal components analysis to the exponential family. In *Advances in Neural Information Processing Systems*, pp. 617–624.

Donoho, D. L. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. *AMS Math Challenges Lecture*, 1–32.

Durif, G., L. Modolo, J. E. Mold, S. Lambert-Lacroix, and F. Picard (2019).

Probabilistic Count Matrix Factorization for Single Cell Expression Data Analysis. *Bioinformatics In press*.

Eckart, C. and G. Young (1936). The approximation of one matrix by another of lower rank. *Psychometrika 1*(3), 211–218.

Freytag, S., L. Tian, I. Lönnstedt, M. Ng, and M. Bahlo (2018, August). Comparison of clustering tools in R for medium-sized 10x Genomics single-cell RNA-sequencing data. *F1000Research 7*.

Hoffman, M. D., D. M. Blei, C. Wang, and J. Paisley (2013, May). Stochastic Variational Inference. *J. Mach. Learn. Res. 14*(1), 1303–1347.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology 24*(6), 417.

Lee, D. D. and H. S. Seung (1999, October). Learning the parts of objects by non-negative matrix factorization. *Nature 401*(6755), 788–791.

Llorens-Bobadilla, E., S. Zhao, A. Baser, G. Saiz-Castro, K. Zwadlo, and A. Martin-Villalba (2015, September). Single-Cell Transcriptomics Reveals a Population of Dormant Neural Stem Cells that Become Activated upon Brain Injury. *Cell Stem Cell 17*(3), 329–340.

Pearson, K. (1901, November). LIII. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine Series 6 2*(11), 559–572.

Pierson, E. and C. Yau (2015). ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biology 16*, 241.

Shen, H. and J. Z. Huang (2008). Sparse principal component analysis via regularized low rank matrix approximation. *Journal of multivariate analysis 99*(6), 1015–1034.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological) 58*(1), 267–288.

van der Maaten, L. and G. Hinton (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research 9*(Nov), 2579–2605.

Witten, D. M., R. Tibshirani, and T. Hastie (2009, July). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics 10*(3), 515–534.

Zou, H., T. Hastie, and R. Tibshirani (2006). Sparse principal component analysis. *Journal of computational and graphical statistics 15*(2), 265–286.

# Variational inference (see Hoffman et al., 2013)

- $p(\mathrm{U}, \mathrm{V} \,|\, \mathrm{X})$ approximated by the **variational distribution** $q(\mathrm{U}, \mathrm{V})$

- Regarding the **Kullback-Leibler divergence**
  → quantify the "proximity" between two distributions of probability

$$q(\mathrm{U}, \mathrm{V}) \;=\; \underset{\text{distribution } \widetilde{q}}{\operatorname{argmin}} \; \mathrm{KL}\Big( \widetilde{q}(\mathrm{U}, \mathrm{V}) \,\Big|\, p(\mathrm{U}, \mathrm{V} \,|\, \mathrm{X}) \Big)$$

- **Constraints on $q$:**
  → $q$ is factorizable: independence between the factors

$$q(\mathrm{U}, \mathrm{V}) = \prod_{i,k} q(u_{ik} \,;\, \mathbf{a}_{ik}) \times \prod_{j,k} q(v_{jk} \,;\, \mathbf{b}_{jk})$$

  → $q$ respects the Gamma-Poisson conjugacy in the exponential family

# The Evidence Lower Bound (ELBO)

**Objective:** $J(q) = \mathbb{E}_q[\log p(\mathsf{X}, \mathsf{U}, \mathsf{V})] - \mathbb{E}_q[\log q(\mathsf{U}, \mathsf{V})]$

- A **lower bound** on the marginal log-likelihood: $\log p(\mathsf{X}) \geq J(q)$ (by Jensen's inequality)
- **Maximizing** $J(q)$ equivalent to **minimizing** $\mathrm{KL}\Big( q(\mathsf{U}, \mathsf{V}) \,|\, p(\mathsf{U}, \mathsf{V} \,|\, \mathsf{X}) \Big)$

  → because $J(q) = \log p(\mathsf{X}) - \mathrm{KL}\Big( q(\mathsf{U}, \mathsf{V}) \,|\, p(\mathsf{U}, \mathsf{V} \,|\, \mathsf{X}) \Big)$
- $J(q)$ is optimized regarding the variational parameters $\mathbf{a}_{ik}$ and $\mathbf{b}_{jk}$

# Optimization of the ELBO

- Gradient of $J(q)$ regarding the variational parameters:

$$\left.\begin{array}{l}\nabla_{\mathbf{a}_{ik}}\\ \nabla_{\mathbf{b}_{jk}}\\ \nabla_{(rijk)_k}\end{array}\right| J(q)$$

- Expression of the ELBO regarding the variational parameters:

$$\widetilde{J}(\mathbf{a}_{ik}) = \mathbb{E}_q[\log p(u_{ik} \,|\, -)] - \mathbb{E}_q[\log q(u_{ik}\,;\, \mathbf{a}_{ik})] + \mathrm{cst}$$

$$\widetilde{J}(\mathbf{b}_{jk}) = \mathbb{E}_q[\log p(v_{jk} \,|\, -)] - \mathbb{E}_q[\log q(v_{jk}\,;\, \mathbf{b}_{jk})] + \mathrm{cst}$$

$$\widetilde{J}\big((r_{ijk})_k\big) = \mathbb{E}_q\big[\log p\big((z_{ijk})_k \,|\, -\big)\big] - \mathbb{E}_q\big[\log q\big((z_{ijk})_k\,;\, (r_{ijk})_k\big)\big] + \mathrm{cst}$$

→ Explicit coordinates of the point that sets the gradient to zero

→ **Iterative optimization through a fixed-point algorithm**

# Variational EM algorithm (Beal and Ghahramani, 2003)

1) Variational E-step:

   → Estimation of the variational parameters $\mathbf{a}$ and $\mathbf{b}$

2) M-step:

   → Estimation of the prior parameters (Gamma parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$)

$$\text{EM:} \quad \underset{\boldsymbol{\alpha},\boldsymbol{\beta}}{\operatorname{argmax}} \, \mathbb{E}_{\mathsf{U},\mathsf{V}\,|\,\mathsf{X}}[\log \mathcal{L}(\mathsf{X}, \mathsf{U}, \mathsf{V}\,;\, \boldsymbol{\alpha}, \boldsymbol{\beta})]$$

Output: estimation of the factors by the variational expectation

$$\widehat{\mathsf{U}} = \mathbb{E}_q[\mathsf{U}] \quad \text{and} \quad \widehat{\mathsf{V}} = \mathbb{E}_q[\mathsf{V}]$$

# Variational EM algorithm (Beal and Ghahramani, 2003)

1) Variational E-step:

   → Estimation of the variational parameters **a** and **b**

2) M-step:

   → Estimation of the prior parameters (Gamma parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$)

$$\text{vEM: } \underset{\boldsymbol{\alpha},\boldsymbol{\beta}}{\operatorname{argmax}} \, \mathbb{E}_q[\log \mathcal{L}(\mathsf{X}, \mathsf{U}, \mathsf{V} \, ; \, \boldsymbol{\alpha}, \boldsymbol{\beta})]$$

Output: estimation of the factors by the variational expectation

$$\widehat{\mathsf{U}} = \mathbb{E}_q[\mathsf{U}] \text{ and } \widehat{\mathsf{V}} = \mathbb{E}_q[\mathsf{V}]$$