Matrix Factorization with Binary Components with Application to Deconvolution of DNA Methylation Data

Matthias Hein
Joint work with: Martin Slawski, Pavlo Lutsik, Jörn Walter, David James and Felix Krahmer

Department of Mathematics and Computer Science
Saarland University, Saarbrücken, Germany

International Symposium: Big Data and Predictive Computational Modeling
Munich, 20.05.2015
Given: \( D \in \mathbb{R}^{m \times n} \), \( n \) number of data points, \( m \) number of features

Find matrices \( T \in \mathbb{R}^{m \times r} \), \( A \in \mathbb{R}^{r \times n} \) such that

\[
D = TA \quad \text{or} \quad \min_{T \in \mathbb{R}^{m \times r}, A \in \mathbb{R}^{r \times n}} \| D - TA \|_F^2,
\]

Exact case

Approximate case

where \( r \) is typically small

Globally optimal solution:

- Singular Value Decomposition (SVD)

\[
D = U \Sigma V^T \quad \implies \quad T = U \Sigma, \quad A = V^T.
\]

Best rank \( r \) approximation obtained by taking top \( r \) singular values

Problem: Factors often lack interpretation
Non-negative Matrix Factorization (NMF)

- **Non-negativity constraints on the factors:**
  find matrices \( T \in \mathbb{R}^{m \times r}_+ \), \( A \in \mathbb{R}^{r \times n}_+ \) such that

\[
D = TA \quad \text{or} \quad \min_{T \in \mathbb{R}^{m \times r}_+, A \in \mathbb{R}^{r \times n}_+} \| D - TA \|_F^2.
\]

(taken from Lee, Seung: Learning the parts of objects by NMF, Nature(1999))
Nonnegative Matrix Factorization (NMF)

- Non-negativity constraints on the factors: find matrices $T \in \mathbb{R}_+^{m \times r}$, $A \in \mathbb{R}_+^{r \times n}$ such that

$$D = TA \quad \text{or} \quad \min_{T \in \mathbb{R}_+^{m \times r}, A \in \mathbb{R}_+^{r \times n}} \| D - TA \|_F^2.$$

Prior work:

- used for finding latent factors/components $T$
- solved via alternating least squares but convergence can only proven to critical point $\Rightarrow$ no guarantee to find global optimum
- In 2012 Arora, Ge, Kanna, Moitra propose an algorithm for exact NMF with runtime $O((nm)r^2)$.
- In the case where $T$ is separable, algorithm runs in polynomial time (improved by Bittorf et al (2013))

**Goal:** extend conditions on NMF for which solution can be found efficiently
Matrix Factorization with Binary Components

Our model:

\[
\begin{bmatrix}
1 & 0 & 1 & 1 \\
0 & 1 & 1 & 0 \\
1 & 1 & 0 & 1 \\
0 & 1 & 0 & 0 \\
1 & 1 & 0 & 0 \\
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 1 \\
\end{bmatrix}
\]

\[D \in \mathbb{R}^{m \times n} \quad T \in \{0, 1\}^{m \times r} \quad A \in \mathbb{R}^{r \times n} \text{ and } 1^T A = 1^T\]

Motivation:

- DNA methylation data
  
  \((m \approx 480000 \text{ sites}, \ n < 1000 \text{ patients}, \ r < 20 \text{ celltypes})\)
  
  CpG sites are (most of the time) either methylated (1) or unmethylated (0)

  **Binary latent factor** \(T\) correspond to celltypes
Our model:

\[
D = \begin{bmatrix}
1 & 0 & 1 & 1 \\
0 & 1 & 1 & 0 \\
1 & 1 & 0 & 1 \\
0 & 1 & 0 & 0 \\
1 & 1 & 0 & 0 \\
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 1 \\
\end{bmatrix}
\]

\[T \in \{0, 1\}^{m \times r} \quad A \in \mathbb{R}^{r \times n} \text{ and } 1^T A = 1^T\]

Prior work (approximate case):

- Relaxation of \{0, 1\} into \([0, 1]\) (biconvex problem) - then alternating least squares (block coordinate descent)
- EM-type algorithms

**Exact case:** naive solution, check all \(2^{mr}\) possibilities for \(T\)
Main result

Assumptions:

- the columns of $T$ are affinely independent
  \[(\forall \lambda \in \mathbb{R}^r \text{ with } \lambda^T \mathbf{1}_r = 0 \text{ and } T \lambda = 0 \implies \lambda = 0)\]
- $\text{rank}(A) = r$

Theorem (Slawski, Hein, Lutsik (NIPS 2013))

*The minimal exact factorization can be computed in $O(rm2^r + mnr + r^2n)$ operations.*

Discussion:

- exponential runtime only in the (small) rank $r$ (in practice more like $O(r^2 2^r)$ instead of $O(rm2^r)$)
- intuition based on Littlewood-Offord Lemma allows us to solve problems up to $r = 80$ ($2^{80} \approx 10^{24}$)
Lemma

The affine hull of $T$ and $D$ agree, $\text{aff}(D) = \text{aff}(T)$.

Illustration for $m = 3$ - note that $\text{aff}(D) \cap \{0, 1\}^m = T$

(worst case: $|\text{aff}(D) \cap \{0, 1\}^m| = 2^{r-1}$)
Algorithm

Exact Matrix Factorization with Binary Components:

1. Obtain \( V = \text{aff}(D) \cap \{0, 1\}^m \)
2. Select \( r \) affinely independent elements of \( V \) as columns of \( T \).
3. Obtain \( A \) as solution of the linear system \( \begin{pmatrix} 1_r^\top \\ T \end{pmatrix} A = \begin{pmatrix} 1_n^\top \\ D \end{pmatrix} \).

How to obtain \( \text{aff}(D) \cap \{0, 1\}^m \)?

1. \( P \) denotes centered data matrix \( D \)
2. obtain \( r - 1 \) linearly independent columns \( C \) and rows \( R \) of \( P \)
   \[ \implies P_{R,C} \in \mathbb{R}^{r-1 \times r-1} \]
3. solve \( P_{R,C} \lambda = b \) for all \( b \in \{0, 1\}^{r-1} \)
4. compute \( Z = P_{:,C} \lambda = P_{:,C}(P_{R,C})^{-1}b \) and check which columns are in \( \{0, 1\}^m \)
**Exact Matrix Factorization with Binary Components:**

1. Obtain $V = \text{aff}(D) \cap \{0, 1\}^m$
2. Select $r$ affinely independent elements of $V$ as columns of $T$.
3. Obtain $A$ as solution of the linear system

$$
\begin{pmatrix}
1_r^	op & T
\end{pmatrix}
\begin{pmatrix}
A
\end{pmatrix}
= 
\begin{pmatrix}
1_n^	op \\
D
\end{pmatrix}.
$$

**How to obtain $\text{aff}(D) \cap \{0, 1\}^m$ ?**

1. $P$ denotes centered data matrix $D$
2. obtain $r-1$ linearly independent columns $C$ and rows $R$ of $P$
   \[ P_{R,C} \in \mathbb{R}^{r-1 \times r-1} \text{ (QR-Factorization - cost } O(mnr)) \]
3. solve $P_{R,C} \lambda = b$ for all $b \in \{0, 1\}^{r-1}$ \text{ (cost } O(r^3))
4. compute $Z = P_{:,C} \lambda = P_{:,C} (P_{R,C})^{-1} b$ and check which columns are in $\{0, 1\}^m$ \text{ (cost } O(rm2^r))
Extension to the NMF Setting

Our model:

\[
D \in \mathbb{R}^{m \times n} \quad T \in \{0, 1\}^{m \times r} \quad A \in \mathbb{R}_+^{r \times n} \quad \text{and} \quad 1^T A = 1^T
\]

Extension of the algorithm to the non-negative case:

- worst case: \( V = \text{aff}(D) \cap \{0, 1\}^m \) contains \( 2^{r-1} \) vertices of \( \{0, 1\}^m \).
- have to check all \( \binom{|V|}{r} \) possibilities to select \( r \) vertices.

Infeasible in practice if \( |V| \gg r \)

but

if \( |V| = r \) then the factorization is unique and extension trivial!
Uniqueness of the Factorization

$$D = \begin{bmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 \end{bmatrix}, \quad T = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad T' = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

uniqueness is crucial for the interpretability of the factors
uniqueness does not hold in general even under the additional positivity constraint on $A$

$m = 2$ - the data lies in the convex hull of the blue and red vertices

Matthias Hein (Saarland University) Matrix Factorization with Binary Components
Uniqueness - General result

General factorization model:
\[ T \in \{0, 1\}^{m \times r}, \ A \in \mathbb{R}^{r \times n}, \ 1^T A = 1^T. \]

Lemma

The factorization \( D = TA \) is unique (modulo permutation of columns) if and only if
\[ \text{aff}(T) \cap \{0, 1\}^m = \{T_1, \ldots, T_r\}. \]

Nonnegative factorization model:
\[ T \in \{0, 1\}^{m \times r}, \ A \in \mathbb{R}_+^{r \times n}, \ 1^T A = 1^T. \]

\[ \text{aff}(T) \cap \{0, 1\}^m = \{T_1, \ldots, T_r\}. \]

is sufficient but not necessary.
Uniqueness - $T$ is separable

**Prior Work:** Donoho/Stodden(2003) show that NMF is unique if $T$ is separable.

**Proposition**

If $T$ is separable that is there exists a permutation $\Pi$ such that

$$\Pi T = \begin{pmatrix} M \\ I_r \end{pmatrix},$$

where $M \in \{0,1\}^{m-r \times r}$, then the factorization is unique.

**Proof:** $\text{aff}(T) \ni b \in \{0,1\}^m$ iff there exists $\lambda \in \mathbb{R}^r$, $\lambda^\top 1_r = 1$ such that

$$T \lambda = b \iff \Pi T \lambda = \Pi b \iff \begin{pmatrix} M \\ I_r \end{pmatrix} \lambda = \Pi b.$$

Since $\Pi b \in \{0,1\}^m$, we get $\lambda \in \{0,1\}^r$. Then $\lambda^\top 1_r = 1$ implies that $\lambda$ must be one of the $r$ canonical basis vectors of $\mathbb{R}^r$. □
Conjecture (James, Krahmer, Hein (2015))

Let $T$ be a random $m \times r$ matrix whose entries are drawn i.i.d. from \{0, 1\} with probability $p$ ($t_{ij} = 1$) with $p \in (0, 1)$. If there is an $\varepsilon > 0$ such that $r < m(1 - \varepsilon)$, then as $m \rightarrow \infty$,

$$
P\left( \text{aff}(T) \cap \{0, 1\}^m = \{T_1, \ldots, T_r\} \right) \geq 1 - 4 \binom{r}{3} (1 - p(1 - p))^m + o((1 - p(1 - p))^m)
$$

- the leading term is

$$
4 \binom{r}{3} (1 - p + p^2)^m \text{ reduces to } 4 \binom{r}{3} \left( \frac{3}{4} \right)^m \text{ for } p = \frac{1}{2}
$$

the probability that three columns of $T$ yield again a \{0, 1\}-vector (Odlyzko (1988)).

- independence of sufficiently many rows is sufficient
What is DNA methylation?

DNA methylation:
- (most) important epigenetic modification (no change of DNA but its interpretation e.g. gene expression)
- influence of the environment leads to changes in the DNA methylation (changes in phenotype)
- certain alterations in DNA methylation seem to be related to cancer

DNA methylation data:
- Measurements: $m \approx 480000$ sites (approx 2% of all sites: $\approx 26 \cdot 10^6$) CpG sites in a single cell are either methylated (1) or unmethylated (0)

but

In mixtures of cells of the same cell type the (average) methylation of a CpG site can be non-binary (even though most of the time it is still close to 0 or 1)
Deconvolution of DNA methylation data

**Problem:** Given measurement of DNA methylation of a certain tissue from different patients infer the contained cell types and their proportions (blind deconvolution).

**Relevance?**
- A disease can manifest both in a change of proportions of cell types (e.g. blood) or in the change of the DNA methylation pattern.
- A proper study of a disease needs to separate these two effects otherwise there is the danger of false discoveries.

**Why is that non-trivial?**
- For some cell types (e.g. blood) methylation profiles are available $\implies$ estimation of proportions is a simple regression problem.
- However for other tissues (e.g. brain) cell type profiles are not available and even if they are available it might be problematic to use them as DNA methylation changes e.g. with age/disease.
How to deal with noisy DNA methylation data?

Problems in the transition to noisy, non-binary data:

- Binary cell types are first proxy but the data suggests that methylation is in $[0, 1]$.
- Relaxation from $\{0, 1\}$ to $[0, 1]$ leads to non-unique factorization $\Rightarrow$ need to use of regularizer to re-establish uniqueness.

$m = 2$ - All Factorizations achieve zero reconstruction error $\|D - TA\|_F$.

Old Optimization Problem:

$$\min_{T \in \{0, 1\}^{m \times r}, A \in \mathbb{R}^{r \times n}} \| D - TA \|_F^2.$$
How to deal with noisy DNA methylation data?

Problems in the transition to noisy, non-binary data:

- Binary cell types are first proxy but the data suggests that methylation is in $[0, 1]$.
- Relaxation from $\{0, 1\}$ to $[0, 1]$ leads to non-unique factorization $\implies$ need to use of regularizer to re-establish uniqueness.

$m = 2$ - All Factorizations achieve zero reconstruction error $\|D - TA\|_F$.

New Optimization Problem (QuadHC):

$$
\min_{T \in [0,1]^{m \times r}, A \in \mathbb{R}_+^{r \times n}, 1_T^T A = 1_n} \|D - TA\|_F^2 + \lambda \sum_{i=1}^{m} \sum_{j=1}^{r} T_{ij}(1 - T_{ij}).
$$
Brain tissue
Artificial mixture

We generated an artificial mixture of isolated NeuN+ and NeuN- (proportion of NeuN+: 0.1 to 0.9 in steps of 0.1)

Results for $\lambda = 0.01$

Pearson correlation: 0.959

Pearson correlation: 0.997

Pearson correlation: 0.997

Pearson correlation: 0.956

Matthias Hein (Saarland University) Matrix Factorization with Binary Components 17 / 20
Results: Left, CV error suggests at least 3 components, Right, Identification of NeuN+ and split of NeuN- into 2 components.
The unknown component 1 is hypomethylated in gene PAX6 which is involved in the differentiation of brain cells in the human brain.

Plot shows that data cannot be well explained using NeuN+ and NeuN-.
Conclusion and Outlook

Conclusion:

- algorithm for exact matrix factorization with binary component $T$
- extension to NMF setting possible under uniqueness
- uniqueness of the factorization is shown under separability and random design of the binary factor
- relaxed version with regularization enforcing $\{0, 1\}$ for deconvolution of DNA methylation data

Outlook and open questions:

- extension of uniqueness results to more than two values in $T$ and general Bernoulli matrices
- use of factorization for exploratory data analysis - interpretation of third component for brain tissue data

**What are generic conditions on (non-convex) learning problems which allow their efficient solution?**
Choice of the number of Components $k$: 

![Graph showing the choice of $k$ via CV Error](image)

- **CV error**
- **$k$**
- **lambda values**: $1e-04$, $0.001$, $0.01$, $0.1$, $1$
Extensions of the model


\[ D \in \mathbb{R}^{m \times n} \]
\[ T \in \{0, 1\}^{m \times r}, \ A \in \{0, 1\}^{n \times r}, \ \text{and} \ W \in \mathbb{R}^{r \times r} \]
\[ \text{such that} \ D = TWA^T. \]

Algorithm for exact factorization:

- apply our algorithm to \( D \) (no affine constraint)
  yields \( D = TB \) with \( T \in \{0, 1\}^{m \times r} \) and \( B \in \mathbb{R}^{r \times n} \)
- apply our algorithm to \( B^T \) (no affine constraint)
  yields \( B^T = AW^T \) with \( A \in \{0, 1\}^{n \times r} \) and \( W \in \mathbb{R}^{r \times r} \)
  \[ D = TB = TWA^T \]
Theorem (Erdös (1951))

Let $x_1, \ldots, x_s \in \mathbb{R} \setminus \{0\}$ and $y \in \mathbb{R}$.

(i) 

$$\left| \{ t \in \{0, 1\}^s : \sum_{i=1}^{s} x_i t_i = y \} \right| \leq \binom{s}{\lfloor s/2 \rfloor} \leq \frac{2^s}{\sqrt{s}}.$$ 

(ii) If $|x_i| \geq 1$, $i = 1, \ldots, s$,

$$\left| \{ t \in \{0, 1\}^s : \sum_{i=1}^{s} x_i t_i \in (y, y + 1) \} \right| \leq \binom{s}{\lfloor s/2 \rfloor}.$$ 

Probabilistic formulation: $t \in \mathbb{R}^s$ with $P(t_i = 0) = P(t_i = 1) = \frac{1}{2}$

$$\sup_{y \in \mathbb{R}} P \left( \sum_{i=1}^{s} x_i t_i = y \right) \leq \frac{1}{\sqrt{s}}.$$
Algorithm runtime

Algorithm:

1. $P$ denotes centered data matrix $D$
2. obtain $r - 1$ linearly independent columns $C$ and rows $R$ of $P$
   \[ P_{R,C} \in \mathbb{R}^{r^{-1} \times r^{-1}} \] (QR-Factorization - cost $O(mnr)$)
3. solve $P_{R,C}\lambda = b$ for all $b \in \{0, 1\}^{r^{-1}}$ (cost $O(r^3)$)
4. compute $Z = P_{:,C}\lambda = P_{:,C}(P_{R,C})^{-1}b$ and check which columns are in
   \[ \{0, 1\}^m \] (cost $O(rm2^r)$)

Intuition from proof: after computation of $k$ rows of $Z$ approximately

$$2^r P_{r,p}^k = 2^r \left( \frac{1}{4p(1-p)r} \right)^k$$

vectors $b \in \{0, 1\}^{r^{-1}}$ are still feasible.

Runtime: $\frac{2r^2 2^r}{\log(4p(1-p)r)}$ instead of $O(mr2^r)$. 
Cardinality of the remaining potential solutions after componentwise check
Use of Integer Linear Programming

The continuous version of the LO-Lemma suggests that we can check

\[
\text{find } b \in \{0, 1\}^{r-1} \text{ subject to } 0 \leq P_{:,C}(P_{R,C})^{-1}b \leq 1, \tag{1}
\]

We solve this ILP with CPLEX.
Extension to the noisy case

Modification of the algorithm

- use truncated SVD to get approximate affine hull of $D$
- take $r - 1$ linearly independent columns of the left factor $U$
- determine the candidates for columns of $T$
- use the candidates $\hat{t}$ which are closest to their thresholded version (Euclidean distance)
- repeat the column selection several times in order to minimize

$$\| D - TA \|_F^2.$$ 

No guarantee for the inexact case!

Arora et al (2012), Bittorf et al (2013) can provide guarantees for the distance to the optimal solution in the near separable case.
Experiments I

**Setup:** \( D = T^* A^* + \alpha E \), \((m = 1000, r = 10 \text{ and } n = 20)\).
- \(T^*\) is chosen uniformly at random
- columns of \(A^*\) are sampled uniformly from the simplex
- \(E\) has i.i.d. standard Gaussian entries

Comparison to locally convergent methods and an oracle method (optimal \(A^*\) is given)
**Experiments II**

**Setup:** \( D = T^* A^* + \alpha E \), \((m=100, r=10 \text{ and } n=20)\)
- \( T^* = [M; I_r] \) and \( M \) is chosen uniformly at random
- \( A^* \) and \( E \) are generated as before.

![Graphs showing comparison to HotTopixx of Bittorf et al. (2013)](image-url)

**Comparison to HotTopixx of Bittorf et al. (2013)**
**Dataset:** $m = 500$ CpG sites and $n = 12$ samples of blood cells composed of four major types (B-/T-cells, granulocytes, monocytes), i.e. $r = 4$. Ground truth is partially available.

**Left:** mixture proportions of the ground truth.  
**Middle:** mixture proportions as estimated by our method.  
**Right:** RMSEs $\| D - \overline{T} \overline{A} \|_F / (mn)^{1/2}$ in dependency of $r$.  

Matthias Hein (Saarland University)  
Matrix Factorization with Binary Components