Online nonnegative matrix factorization for Markovian data

Deanna Needell

Joint work with Hanbaek Lyu, Laura Balzano and Chris Strohmeier.
Partially supported by NSF DMS #2011140 and NSF BIGDATA #1740325.

deanna@math.ucla.edu

UCLA

Overview

Introduction

Algorithms for online NMF and their convergence

Applications of ONMF

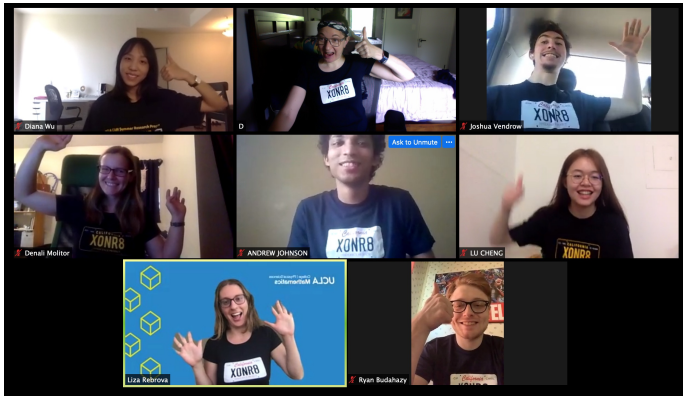
Collaborators

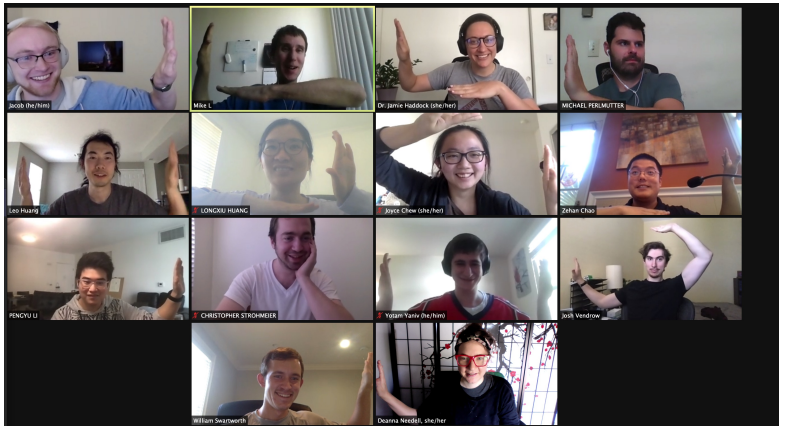
Dr. Hanbaek Lyu
(postdoc, UCLA
→ U. Wisconsin)

Prof. Laura Balzano
(Univ Michigan)

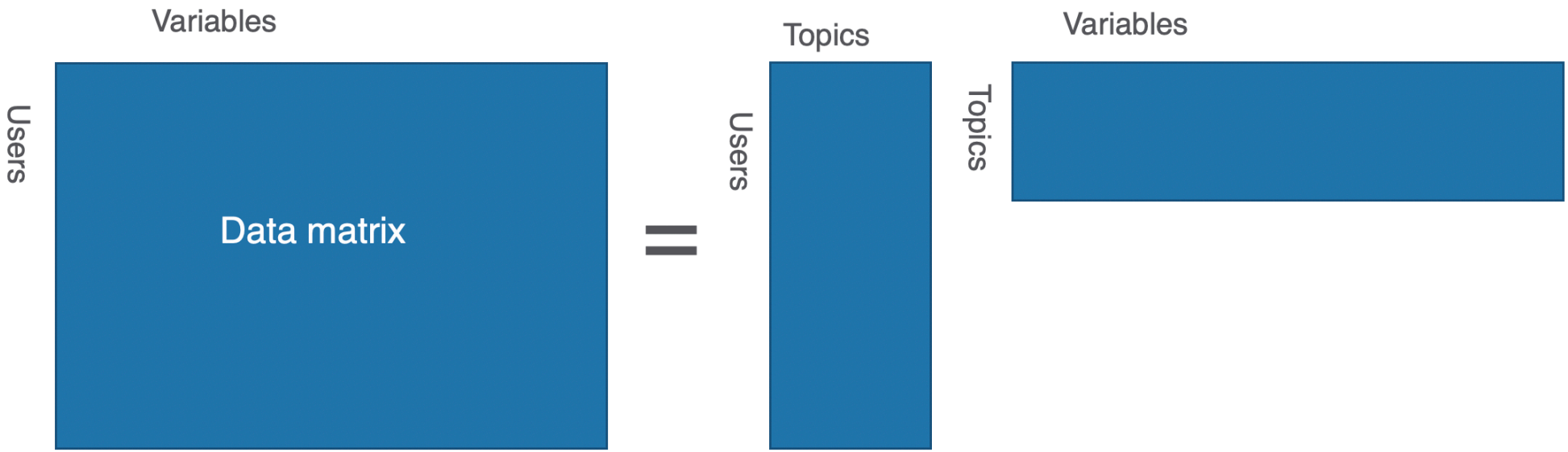
Chris Strohmeier
(PhD student, UCLA)

Joint work with

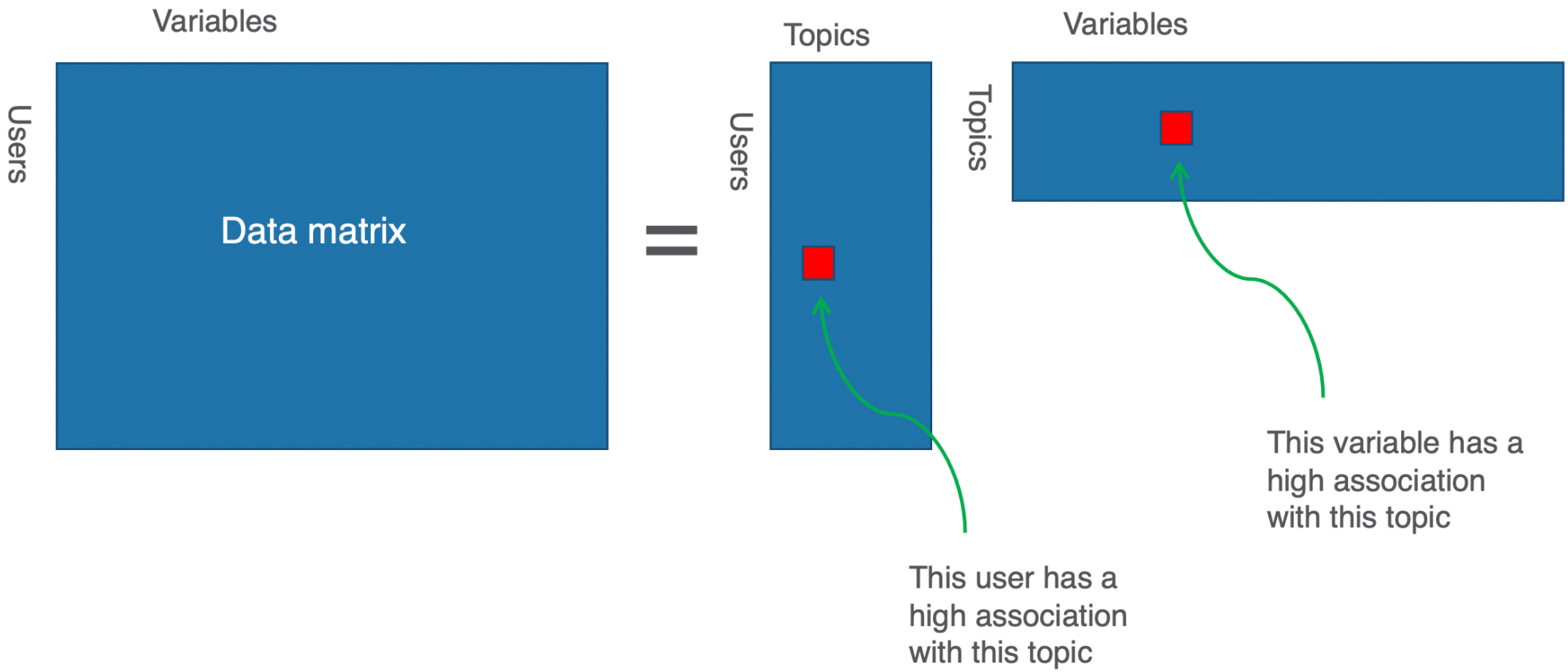




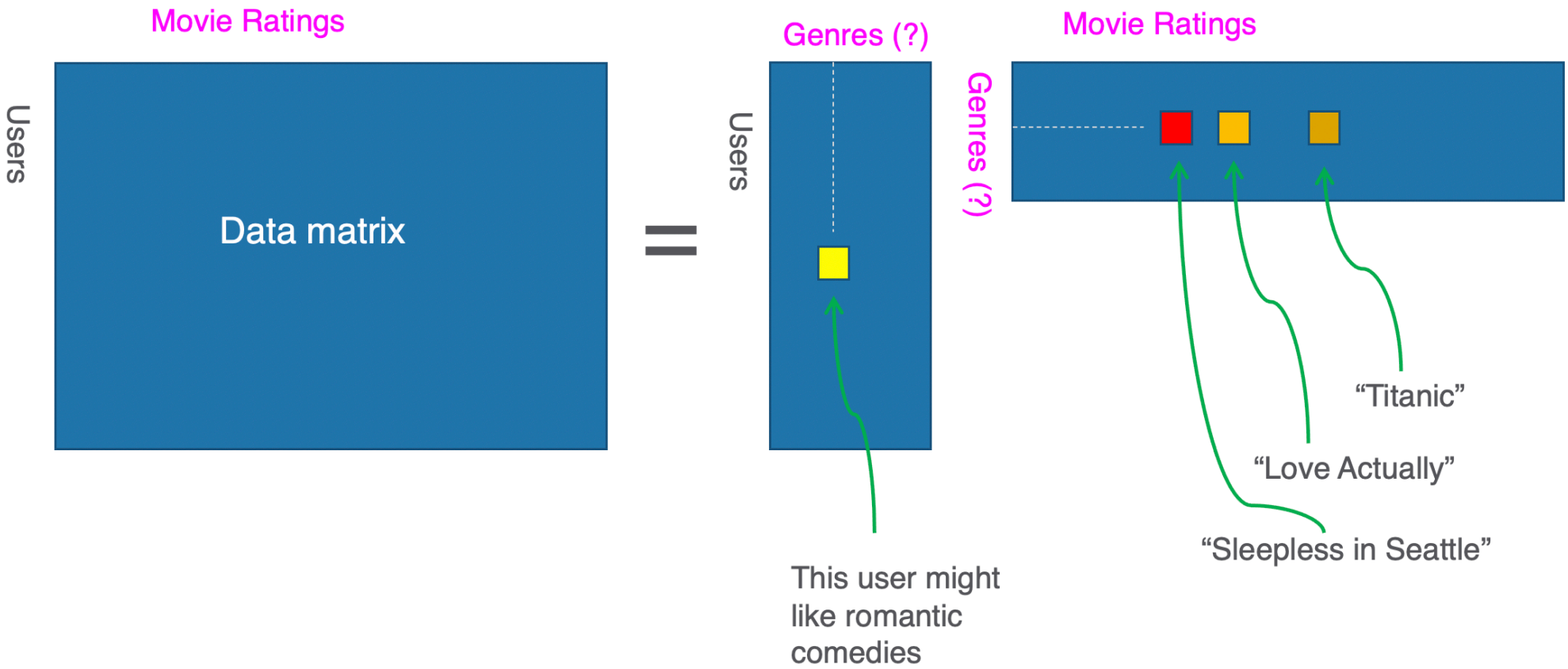
Non-negative matrix factorization



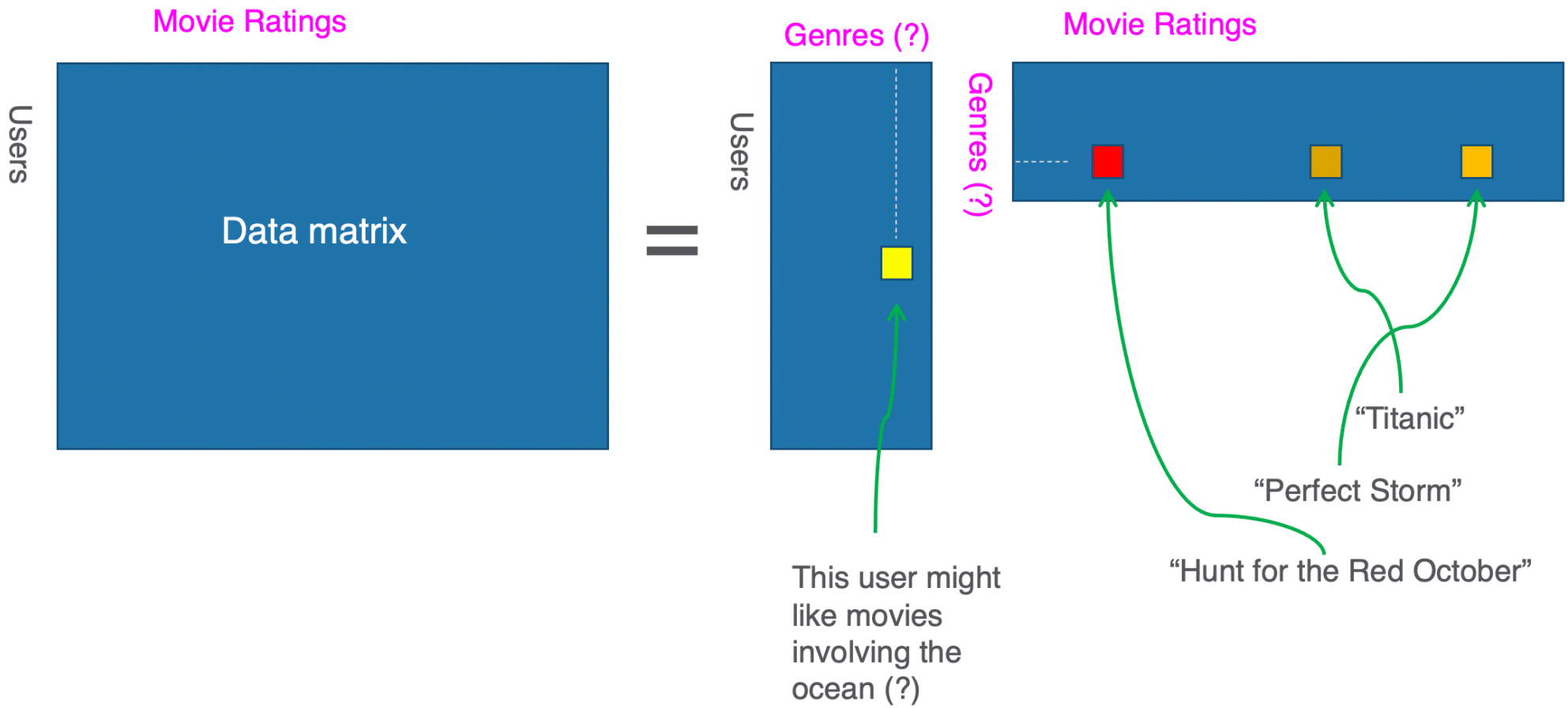
Non-negative matrix factorization



What is nonnegative matrix factorization?



What is nonnegative matrix factorization?



(Semi)supervised NMF

- Incorporates label information to NMF
- $Y \in \mathbb{R}^{c \times n}$ = label matrix for c classes
- Find $A, S, B \in \mathbb{R}^{c \times k}$ by

$$\min_{A \geq 0, S \geq 0, B \geq 0} \underbrace{\|(X - AS)\|_F^2}_{\text{Reconstruction Error}} + \lambda \underbrace{\|(Y - BS)\|_F^2}_{\text{Classification Error}}$$

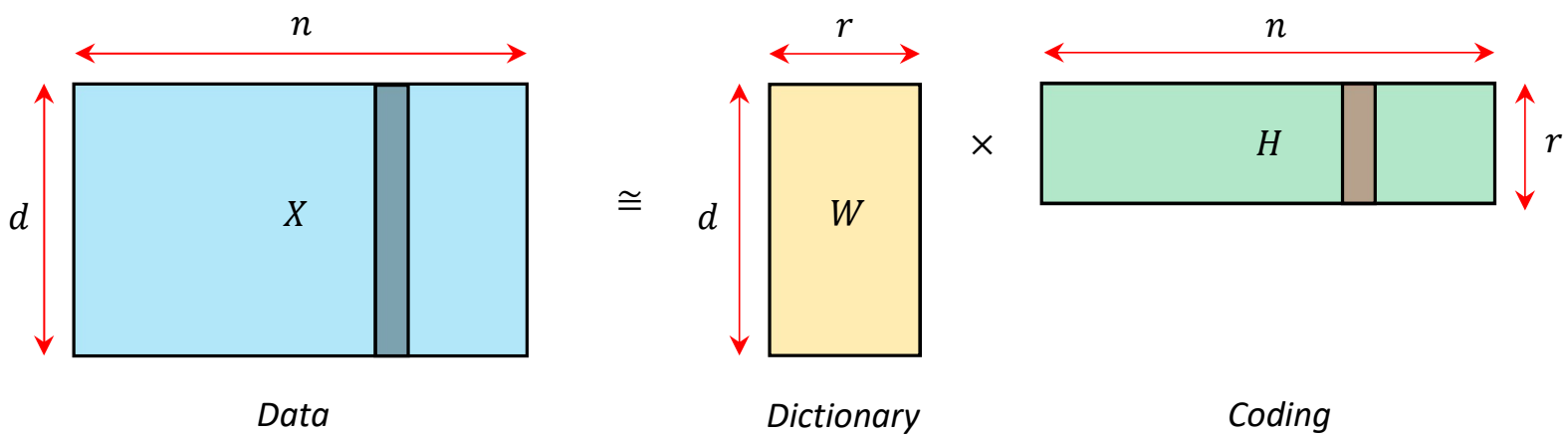
What is nonnegative matrix factorization?

- The goal of **nonnegative matrix factorization** (NMF) is to factorize a data matrix $X \in \mathbb{R}^{d \times n}_{\geq 0}$ into a pair of low-rank nonnegative matrices $W \in \mathbb{R}^{d \times r}$ and $H \in \mathbb{R}^{r \times n}$ by solving the following optimization problem

$$\inf_{W \in \mathbb{R}^{d \times r}_{\geq 0},\, H \in \mathbb{R}^{r \times n}_{\geq 0}} \|X - WH\|_F^2,$$

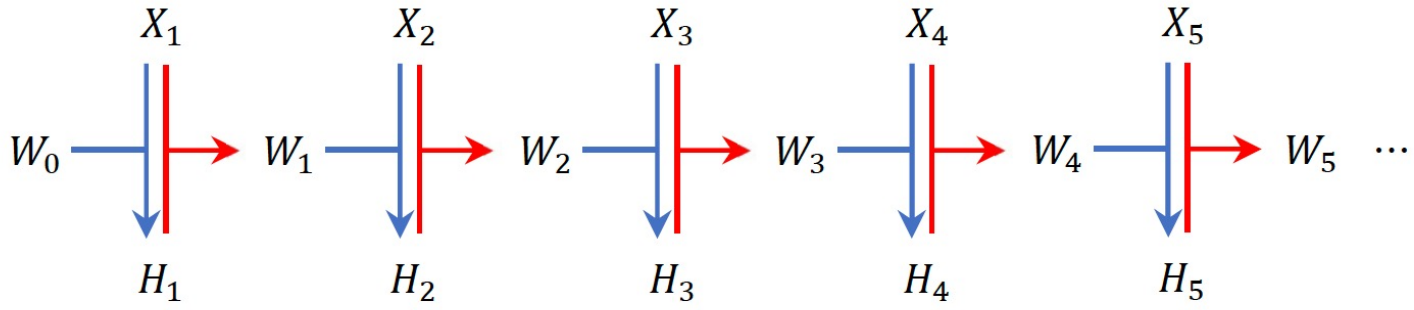
  where $\|A\|_F^2 = \sum_{i,j} A_{ij}^2$ denotes the matrix Frobenius norm.

- $\texttt{Data} \approx \texttt{Dictionary} \times \texttt{Coding}$



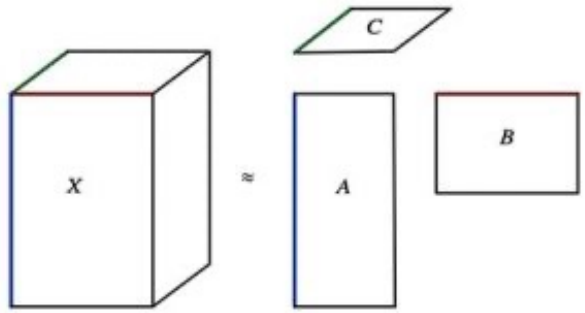
Online NMF

- Considers data that is streaming in over time
- *Learns a factorization that is best (in expectation)*
- Can be used for prediction in time series data
  - Uses "windows" across time to update factors and then predicts into a future window using one of the factors

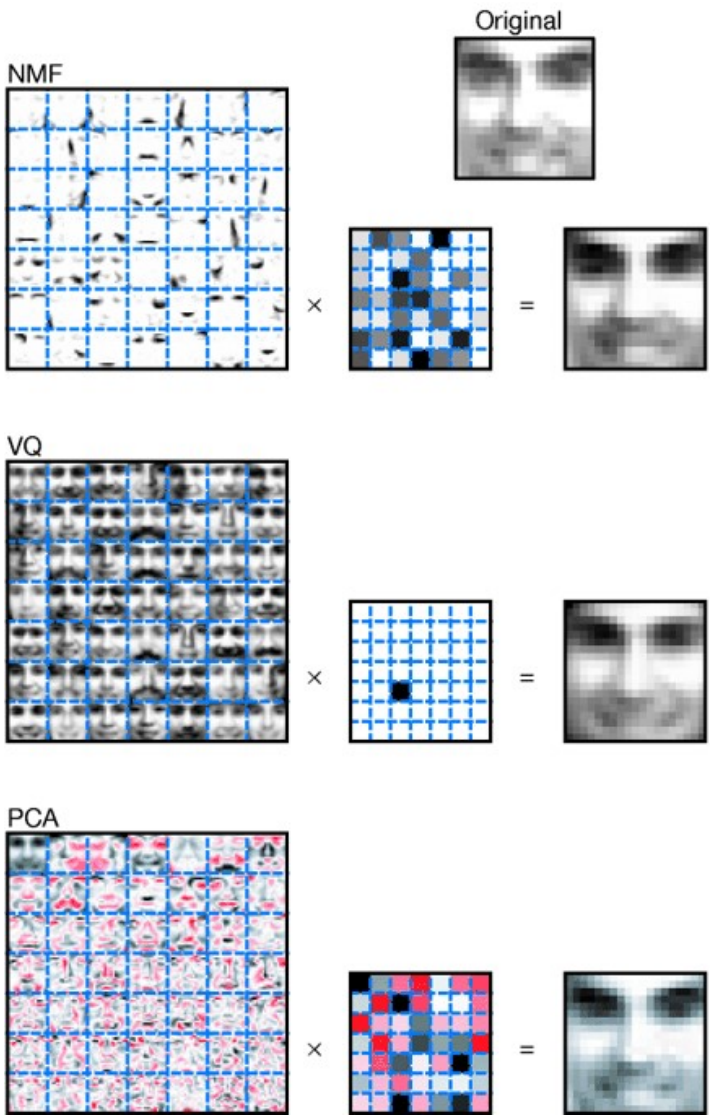


Non-negative Tensor Factorization (NTF)

- Can be extended to tensors in a (nontrivial but) analogous way

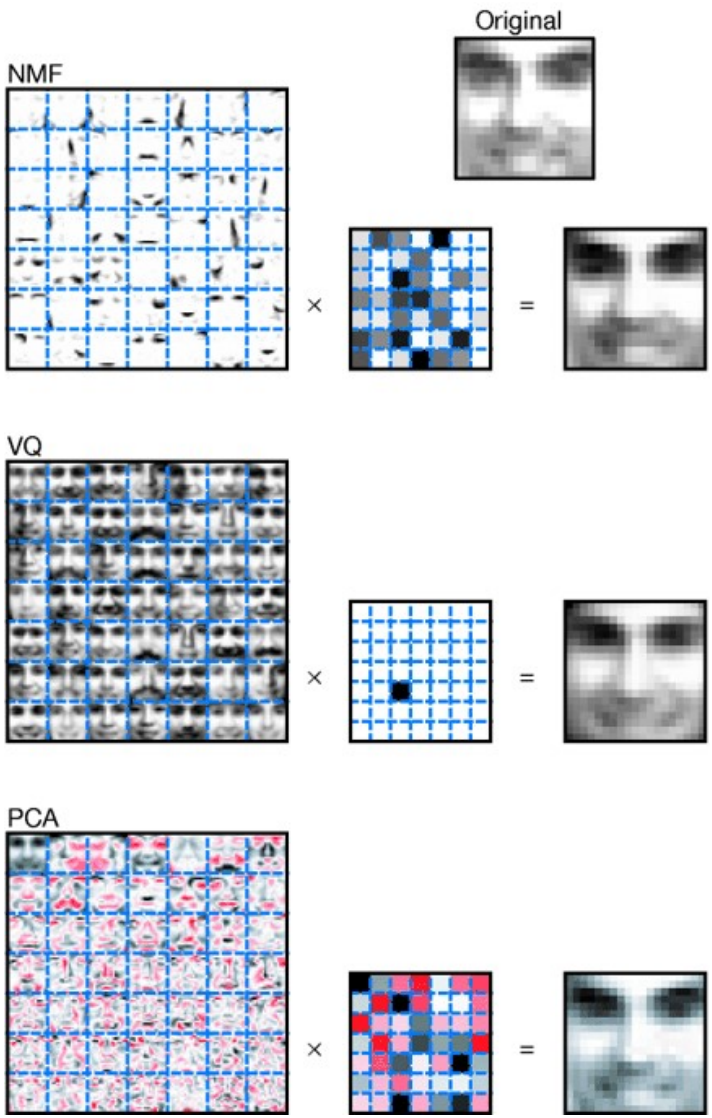


Static NMF algorithms



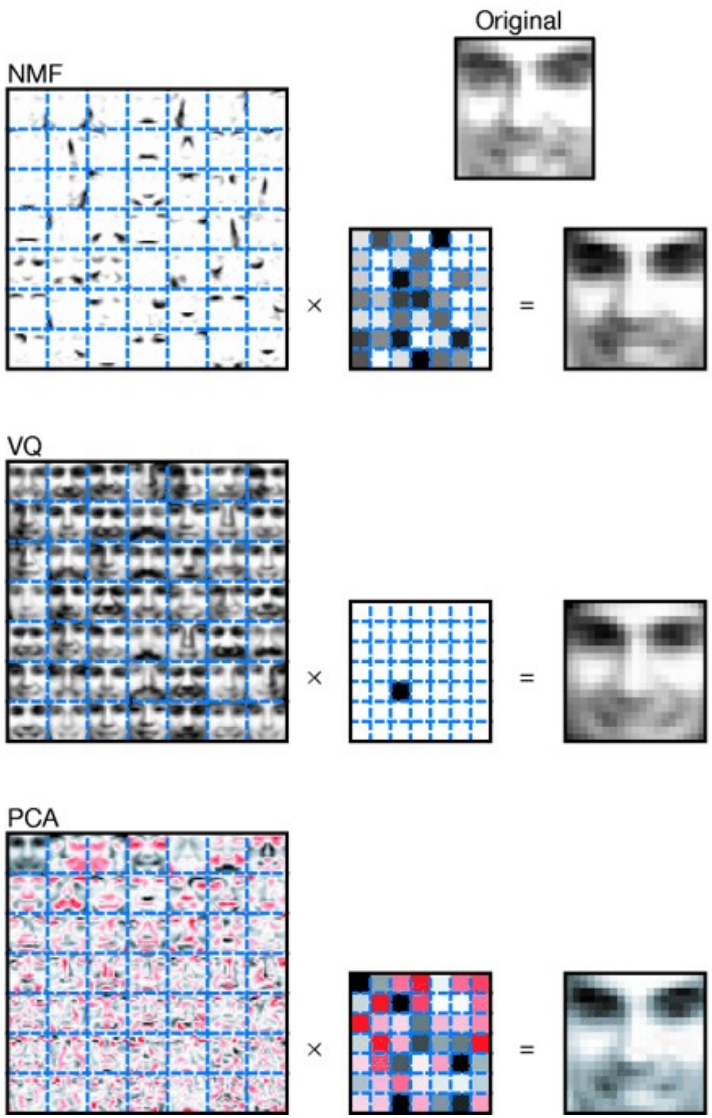
- In NMF, each column of the data matrix has to be represented as a non-negative linear combination of dictionaries

Static NMF algorithms



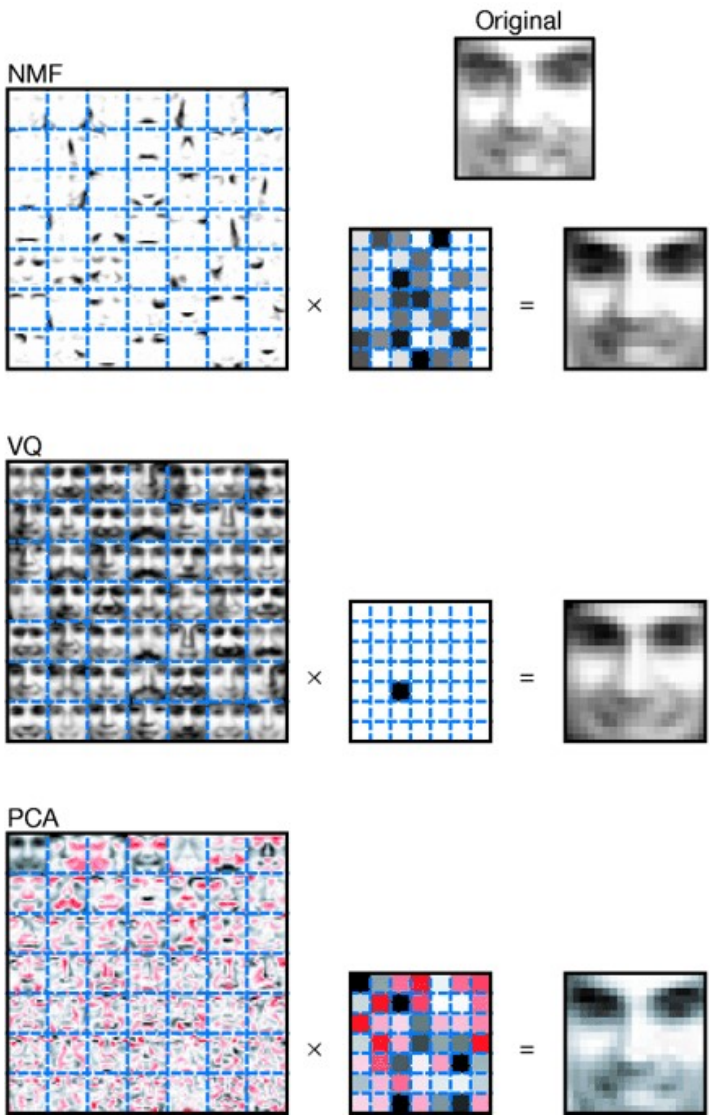
- In NMF, each column of the data matrix has to be represented as a non-negative linear combination of dictionaries
- Hence the dictionaries must be "positive parts" of the columns of the data matrix

Static NMF algorithms



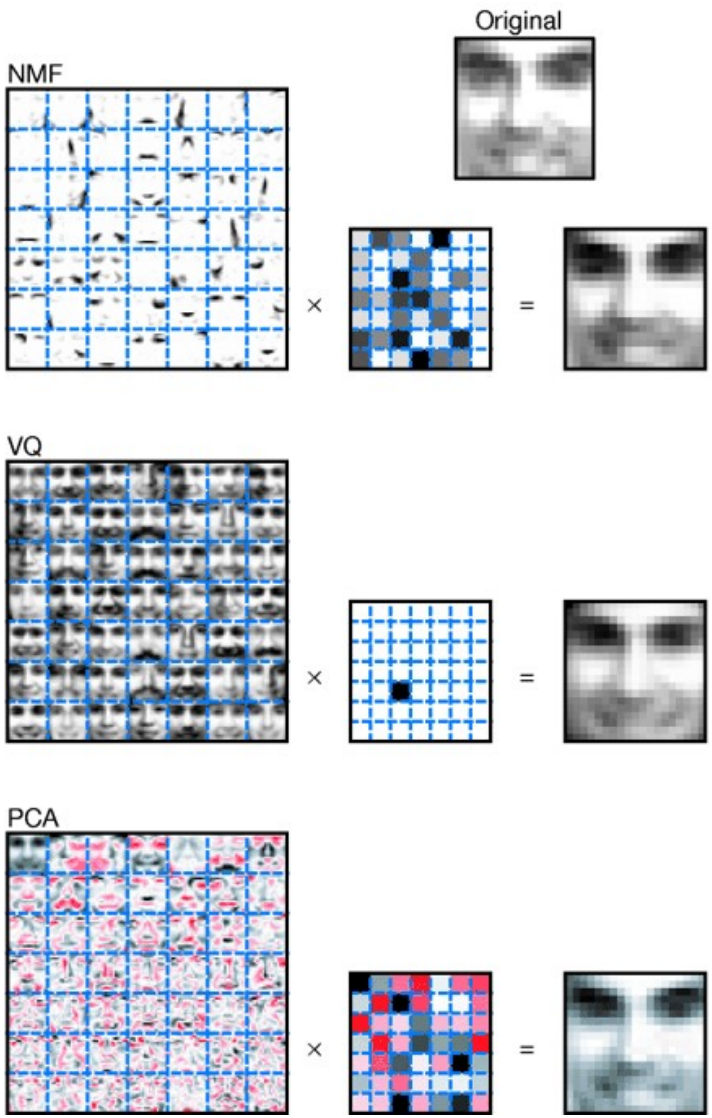
- In NMF, each column of the data matrix has to be represented as a non-negative linear combination of dictionaries
- Hence the dictionaries must be "positive parts" of the columns of the data matrix
- When each column consists of a human face image, NMF learns the parts of human face (e.g., eyes, nodes, mouth)

Static NMF algorithms



- In NMF, each column of the data matrix has to be represented as a non-negative linear combination of dictionaries
- Hence the dictionaries must be "positive parts" of the columns of the data matrix
- When each column consists of a human face image, NMF learns the parts of human face (e.g., eyes, nodes, mouth)
- This is in contrast to principal component analysis and vector quantization: Due to cancellation between eigenvectors, each 'eigenface' does not have to be parts of face

Static NMF algorithms



- In NMF, each column of the data matrix has to be represented as a non-negative linear combination of dictionaries
- Hence the dictionaries must be "positive parts" of the columns of the data matrix
- When each column consists of a human face image, NMF learns the parts of human face (e.g., eyes, nodes, mouth)
- This is in contrast to principal component analysis and vector quantization: Due to cancellation between eigenvectors, each 'eigenface' does not have to be parts of face
- NMF was popularized by Lee and Seung in their Nature paper in 1999

Applications: Learning parts from images

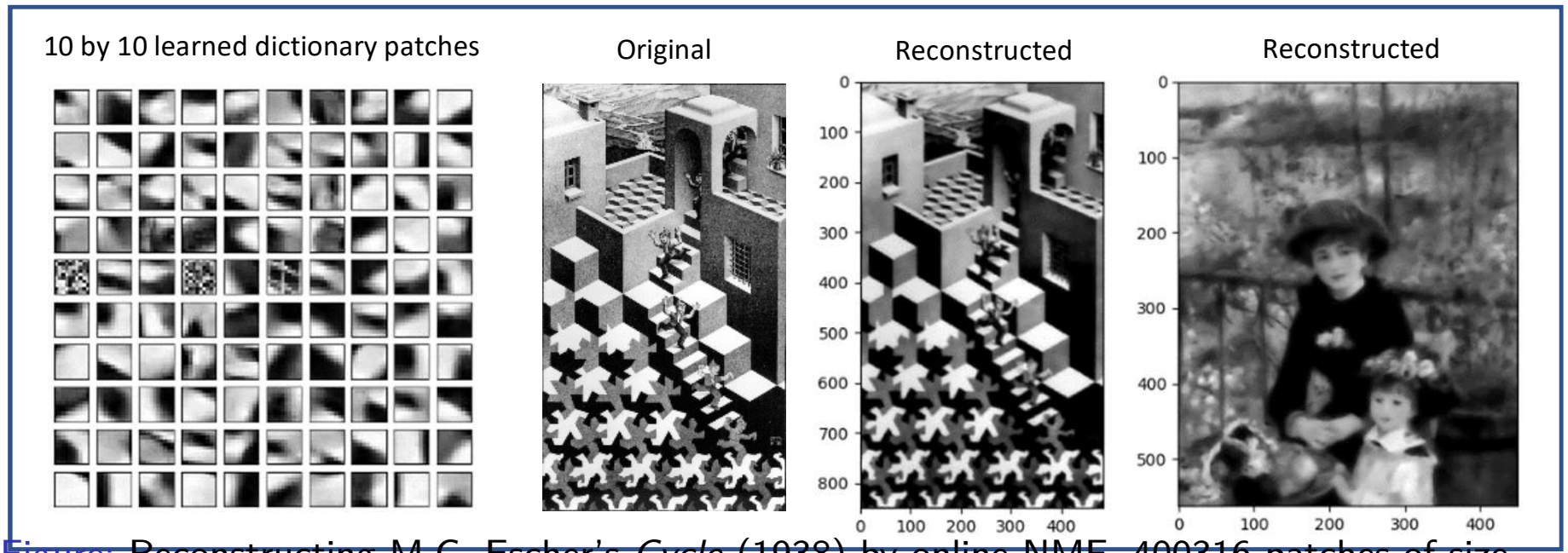
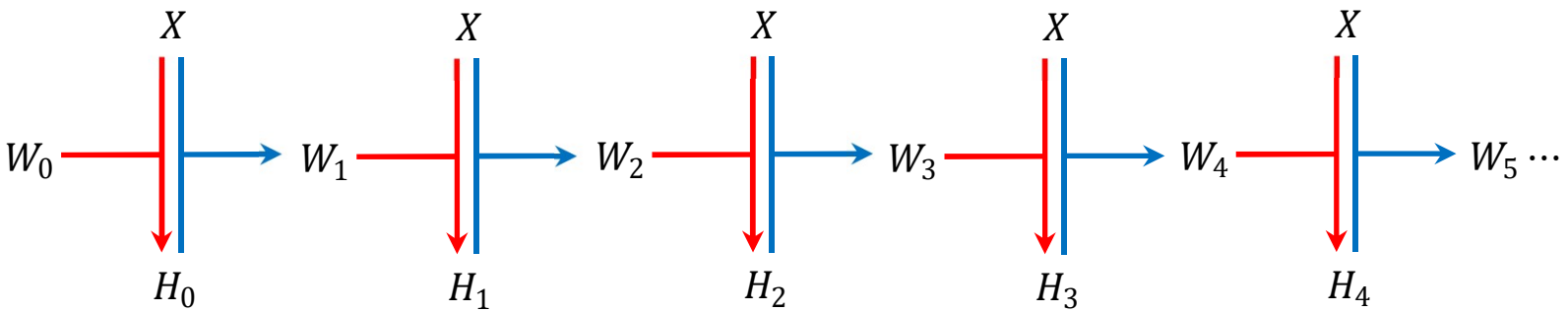


Figure: Reconstructing M.C. Escher's *Cycle* (1938) by online NMF. 400316 patches of size 10×10 are extracted from the original image, and 100 dictionary patches (left) are learned by NMF. Original and reconstructed image using the learned dictionaries shown in the middle. The last shows the reconstructed image of Pierre-Auguste Renoir's *Two Sisters* (1882) (original image omitted) using the dictionary patches learned from Escher's *Cycle* in the left.

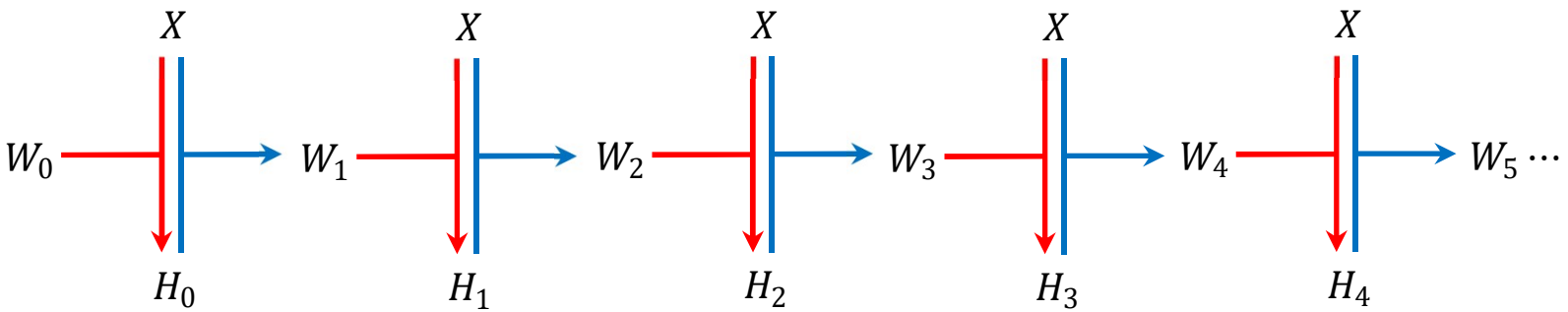
Algorithms for online NMF and their convergence

- In order to minimize $\|X - WH\|_F$, one can use block coordinate descent, by iteratively fixing W or H and minimizing the error w.r.t. the other factor



Static NMF algorithms

- In order to minimize $\|X - WH\|_F$, one can use block coordinate descent, by iteratively fixing W or H and minimizing the error w.r.t. the other factor

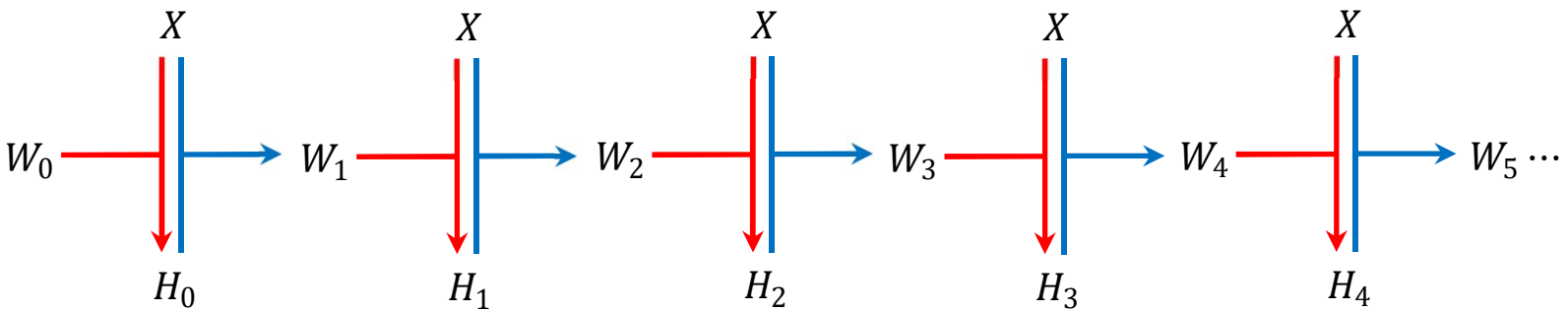


- One of the most popular static NMF algorithm is the **Multiplicative Update** by Lee and Seung: Update all entries of H and W alternatively using the following update

$$H_{ij} \leftarrow H_{ij} \frac{[W^T X]_{ij}}{[W^T W X]_{ij}}, \qquad W_{ij} \leftarrow W_{ij} \frac{[X H^T]_{ij}}{[X H H^T]_{ij}}.$$

Static NMF algorithms

- In order to minimize $\|X - WH\|_F$, one can use block coordinate descent, by iteratively fixing W or H and minimizing the error w.r.t. the other factor



- One of the most popular static NMF algorithm is the **Multiplicative Update** by Lee and Seung: Update all entries of H and W alternatively using the following update

$$H_{ij} \leftarrow H_{ij} \frac{[W^T X]_{ij}}{[W^T W X]_{ij}}, \qquad W_{ij} \leftarrow W_{ij} \frac{[X H^T]_{ij}}{[X H H^T]_{ij}}.$$

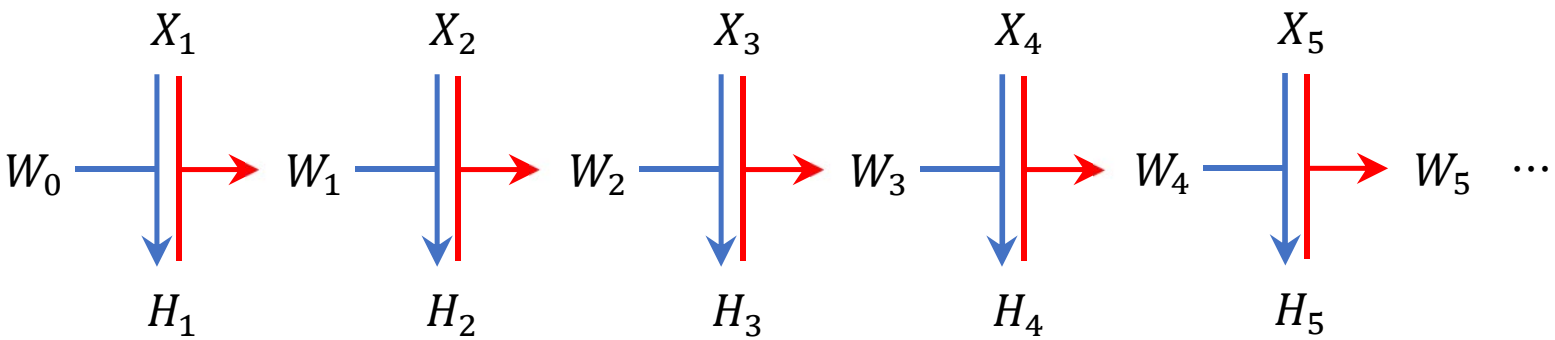
- It is known that the error $\|X - WH\|_F^2$ is non-increasing under the above update, but there is no guarantee to converge to a stationary point.

Online NMF

- If the data matrix X is randomly drawn from a sample space $\Omega \subseteq \mathbb{R}_{\geq 0}^{d \times n}$ according to a distribution π, can we still learn the 'best dictionaries' that describe X in law?

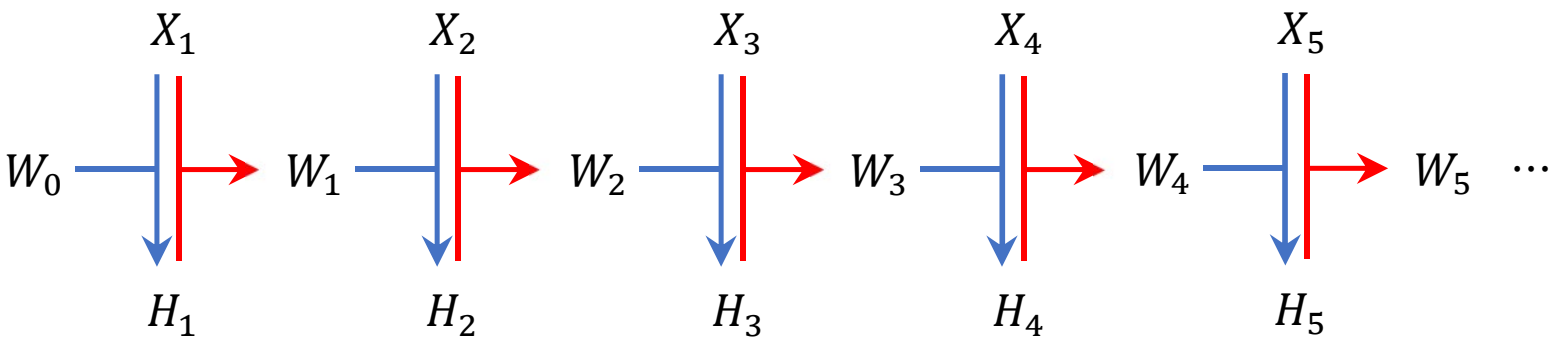
Online NMF

- If the data matrix X is randomly drawn from a sample space $\Omega \subseteq \mathbb{R}_{\geq 0}^{d \times n}$ according to a distribution π, can we still learn the 'best dictionaries' that describe X in law?
- The **online Non-negative Matrix Factorization** (ONMF) problem concerns a similar matrix factorization problem for a sequence of input matrices $(x_T)_{t \geq 0}$.



Online NMF

- If the data matrix X is randomly drawn from a sample space $\Omega \subseteq \mathbb{R}_{\geq 0}^{d \times n}$ according to a distribution π, can we still learn the 'best dictionaries' that describe X in law?
- The **online Non-negative Matrix Factorization** (ONMF) problem concerns a similar matrix factorization problem for a sequence of input matrices $(x_T)_{t \geq 0}$.



- Suppose $(X_t)_{t \geq 1}$ is an irreducible Markov chain on a sample space Ω with unique stationary measure π. The goal of ONMF problem is to construct a sequence $(W_t, H_t)_{t \geq 1}$ of dictionary $W_t \in \mathbb{R}^{r \times d}$ and a coding $H_t \in \mathbb{R}_{\geq 0}^{r \times n}$ such that (almost surely)

$$\|X_t - W_{t-1} H_t\|_F^2 \longrightarrow \inf_{W \in \mathbb{R}^{d \times r},\, H \in \mathbb{R}^{r \times n}} \mathbb{E}_{X \sim \pi} \left[\|X - WH\|_F^2 \right]$$

Naive ONMF algorithm minimizing empirical loss

- Fix $\lambda > 0$ and define the following the **quadratic loss function**

$$\ell(X, W) = \inf_{H \in \mathbb{R}^{r \times n}} \|X - WH\|_F^2 + \lambda \|H\|_1,$$

Define the **expected loss** and **empirical loss** functions

$$f(W) = \mathbb{E}_{X \sim \pi}[\ell(X, W)], \qquad f_t(W) = \frac{1}{t} \sum_{s=1}^{t} \ell(X_s, W)$$

Naive ONMF algorithm minimizing empirical loss

- Fix $\lambda > 0$ and define the following the **quadratic loss function**

$$\ell(X, W) = \inf_{H \in \mathbb{R}^{r \times n}} \|X - WH\|_F^2 + \lambda \|H\|_1,$$

Define the **expected loss** and **empirical loss** functions

$$f(W) = \mathbb{E}_{X \sim \pi}[\ell(X, W)], \qquad f_t(W) = \frac{1}{t} \sum_{s=1}^{t} \ell(X_s, W)$$

- By Markov chain ergodic theorem, for each dictionary W, the empirical loss $f_t(W)$ converges almost surely to the expected loss $f(W)$:

$$\lim_{t \to \infty} f_t(W) = f(W) \quad \text{a.s.}$$

Naive ONMF algorithm minimizing empirical loss

- Fix $\lambda > 0$ and define the following the **quadratic loss function**

$$\ell(X, W) = \inf_{H \in \mathbb{R}^{r \times n}} \|X - WH\|_F^2 + \lambda \|H\|_1,$$

Define the **expected loss** and **empirical loss** functions

$$f(W) = \mathbb{E}_{X \sim \pi}[\ell(X, W)], \qquad f_t(W) = \frac{1}{t} \sum_{s=1}^{t} \ell(X_s, W)$$

- By Markov chain ergodic theorem, for each dictionary W, the empirical loss $f_t(W)$ converges almost surely to the expected loss $f(W)$:

$$\lim_{t \to \infty} f_t(W) = f(W) \quad \text{a.s.}$$

- A naive solution to ONMF based on block optimization scheme:

$$\text{Upon arrival of } X_t: \quad \begin{cases} H_t = \operatorname{argmin}_{H \in \mathbb{R}_{\geq 0}^{r \times n}} \|X_t - W_{t-1} H\|_F^2 + \lambda \|H\|_1 \\ W_t = \operatorname{argmin}_{W \in \mathcal{C}} f_t(W), \end{cases}$$

where $\mathcal{C} \subseteq \mathbb{R}^{d \times r}$ is the set of admissible dictionaries.

Naive ONMF algorithm minimizing empirical loss

- Fix $\lambda > 0$ and define the following the **quadratic loss function**

$$\ell(X, W) = \inf_{H \in \mathbb{R}^{r \times n}} \|X - WH\|_F^2 + \lambda \|H\|_1,$$

Define the **expected loss** and **empirical loss** functions

$$f(W) = \mathbb{E}_{X \sim \pi}[\ell(X, W)], \qquad f_t(W) = \frac{1}{t} \sum_{s=1}^{t} \ell(X_s, W)$$

- By Markov chain ergodic theorem, for each dictionary W, the empirical loss $f_t(W)$ converges almost surely to the expected loss $f(W)$:

$$\lim_{t \to \infty} f_t(W) = f(W) \quad \text{a.s.}$$

- A naive solution to ONMF based on block optimization scheme:

$$\text{Upon arrival of } X_t: \qquad \begin{cases} H_t = \operatorname{argmin}_{H \in \mathbb{R}_{\geq 0}^{r \times n}} \|X_t - W_{t-1} H\|_F^2 + \lambda \|H\|_1 \\ W_t = \operatorname{argmin}_{W \in \mathcal{C}} f_t(W), \end{cases}$$

where $\mathcal{C} \subseteq \mathbb{R}^{d \times r}$ is the set of admissible dictionaries.

- This requires to store all previous matrices and solve many optimization instances.

Asymptotic solution minimizing surrogate loss function

- Mairal, Bach, Ponce, and Sapiro gave an influential solution to the ONMF problem with a rigorous derivation of almost sure convergence of the empirical loss over time for i.i.d. data matrices.

Asymptotic solution minimizing surrogate loss function

- Mairal, Bach, Ponce, and Sapiro gave an influential solution to the ONMF problem with a rigorous derivation of almost sure convergence of the empirical loss over time for i.i.d. data matrices.
- The idea is to solve the following approximate problem

$$\textit{Upon arrival of } X_t\textit{:} \qquad \begin{cases} H_t = \operatorname{argmin}_{H \in \mathbb{R}^{r \times n}_{\geq 0}} \|X_t - W_{t-1}H\|_F^2 + \lambda \|H\|_1 \\ W_t = \operatorname{argmin}_{W \in C} \hat{f}_t(W), \end{cases}$$

where $\hat{f}_t(W)$ is a convex upper bounding **surrogate** for $f_t(W)$ defined by

$$\hat{f}_t(W) = \frac{1}{t} \sum_{s=1}^{t} (\|X_s - WH_s\|_F^2 + \lambda \|H_s\|_1).$$

Asymptotic solution minimizing surrogate loss function

- Mairal, Bach, Ponce, and Sapiro gave an influential solution to the ONMF problem with a rigorous derivation of almost sure convergence of the empirical loss over time for i.i.d. data matrices.
- The idea is to solve the following approximate problem

$$\textit{Upon arrival of } X_t\textit{:} \qquad \begin{cases} H_t = \operatorname{argmin}_{H \in \mathbb{R}^{r\times n}_{\geq 0}} \|X_t - W_{t-1}H\|_F^2 + \lambda \|H\|_1 \\ W_t = \operatorname{argmin}_{W \in \mathcal{C}} \hat{f}_t(W), \end{cases}$$

where $\hat{f}_t(W)$ is a convex upper bounding **surrogate** for $f_t(W)$ defined by

$$\hat{f}_t(W) = \frac{1}{t}\sum_{s=1}^{t} (\|X_s - WH_s\|_F^2 + \lambda \|H_s\|_1).$$

- Namely, we recycle the previously found coding $H_1, \cdots, H_t$ and use them as approximate solutions of the sub-problems. Hence, there is only a single optimization for W_t in the above relaxed problem

Asymptotic solution minimizing surrogate loss function

- Mairal, Bach, Ponce, and Sapiro gave an influential solution to the ONMF problem with a rigorous derivation of almost sure convergence of the empirical loss over time for i.i.d. data matrices.
- The idea is to solve the following approximate problem

$$\textit{Upon arrival of } X_t\textit{:} \qquad \begin{cases} H_t = \operatorname{argmin}_{H \in \mathbb{R}^{r\times n}_{\geq 0}} \|X_t - W_{t-1}H\|_F^2 + \lambda\|H\|_1 \\ W_t = \operatorname{argmin}_{W\in\mathcal{C}} \hat{f}_t(W), \end{cases}$$

where $\hat{f}_t(W)$ is a convex upper bounding **surrogate** for $f_t(W)$ defined by

$$\hat{f}_t(W) = \frac{1}{t}\sum_{s=1}^{t}(\|X_s - WH_s\|_F^2 + \lambda\|H_s\|_1).$$

- Namely, we recycle the previously found coding $H_1, \cdots, H_t$ and use them as approximate solutions of the sub-problems. Hence, there is only a single optimization for W_t in the above relaxed problem
- But we still need to store the entire history $X_1, \cdots, X_t$ and $H_1, \cdots, H_t$. Do we?

Asymptotic solution minimizing surrogate loss function

- In fact, the approximate ONMF problem is equivalent to

$$\textit{Upon arrival of } X_t\textit{:} \quad \begin{cases} H_t = \operatorname{argmin}_{H \in \mathbb{R}^{r \times n}_{\geq 0}} \|X_t - W_{t-1}H\|_F^2 + \lambda \|H\|_1 \\ A_t = t^{-1}((t-1)A_{t-1} + H_t H_t^T) \\ B_t = t^{-1}((t-1)B_{t-1} + H_t X_t^T) \\ W_t = \operatorname{argmin}_{W \in \mathcal{C} \subseteq \mathbb{R}^{d \times r}_{\geq 0}} \left(\operatorname{tr}(WA_tW^T) - 2\operatorname{tr}(WB_t)\right), \end{cases}$$

where A_0 and B_0 are zero matrices of size $r \times r$ and $r \times d$, respectively.

Asymptotic solution minimizing surrogate loss function

- In fact, the approximate ONMF problem is equivalent to

$$\textit{Upon arrival of } X_t\textit{:} \quad \begin{cases} H_t = \operatorname{argmin}_{H \in \mathbb{R}^{r\times n}_{\geq 0}} \|X_t - W_{t-1}H\|_F^2 + \lambda \|H\|_1 \\ A_t = t^{-1}((t-1)A_{t-1} + H_t H_t^T) \\ B_t = t^{-1}((t-1)B_{t-1} + H_t X_t^T) \\ W_t = \operatorname{argmin}_{W \in \mathcal{C} \subseteq \mathbb{R}^{d\times r}_{\geq 0}} \left(\operatorname{tr}(WA_tW^T) - 2\operatorname{tr}(WB_t)\right), \end{cases}$$

  where A_0 and B_0 are zero matrices of size $r \times r$ and $r \times d$, respectively.
- So we only need to store two summary matrices $A_t \in \mathbb{R}^{r\times r}_{\geq 0}$ and $B_t \in \mathbb{R}^{r\times d}$.

Asymptotic solution minimizing surrogate loss function

- In fact, the approximate ONMF problem is equivalent to

$$\textit{Upon arrival of } X_t\textit{:} \quad \begin{cases} H_t = \operatorname{argmin}_{H \in \mathbb{R}^{r\times n}_{\geq 0}} \|X_t - W_{t-1}H\|_F^2 + \lambda \|H\|_1 \\ A_t = t^{-1}((t-1)A_{t-1} + H_t H_t^T) \\ B_t = t^{-1}((t-1)B_{t-1} + H_t X_t^T) \\ W_t = \operatorname{argmin}_{W \in \mathcal{C} \subseteq \mathbb{R}^{d\times r}_{\geq 0}} \left(\operatorname{tr}(WA_tW^T) - 2\operatorname{tr}(WB_t)\right), \end{cases}$$

  where A_0 and B_0 are zero matrices of size $r \times r$ and $r \times d$, respectively.
- So we only need to store two summary matrices $A_t \in \mathbb{R}^{r\times r}_{\geq 0}$ and $B_t \in \mathbb{R}^{r\times d}$.
- Computing W_t also requires solving only a single optimization instance

Convergence under i.i.d. assumption

$$\textit{Upon arrival of } X_t\textit{:} \quad \begin{cases} H_t = \operatorname{argmin}_{H \in \mathbb{R}^{r \times n}_{\geq 0}} \|X_t - W_{t-1}H\|_F^2 + \lambda \|H\|_1 \\ A_t = t^{-1}((t-1)A_{t-1} + H_t H_t^T) \\ B_t = t^{-1}((t-1)B_{t-1} + H_t X_t^T) \\ W_t = \operatorname{argmin}_{W \in \mathcal{C} \subseteq \mathbb{R}^{d \times r}_{\geq 0}} \left(\operatorname{tr}(WA_tW^T) - 2\operatorname{tr}(WB_t)\right), \end{cases}$$

$f_t =$ empirical loss, $\quad \hat{f}_t =$ surrogate loss, $\quad f =$ expected loss

Theorem (Mairal, Bach, Ponce, and Sapiro '10)

Suppose $(X_t)_{t \geq 0}$ are i.i.d. with common distribution π. Let $(W_{t-1}, H_t)_{t \geq 1}$ be the optimal solution to the above ONMF algorithm.

Convergence under i.i.d. assumption

$$\textit{Upon arrival of } X_t\textit{:} \quad \begin{cases} H_t = \operatorname{argmin}_{H \in \mathbb{R}^{r \times n}_{\geq 0}} \|X_t - W_{t-1}H\|_F^2 + \lambda \|H\|_1 \\ A_t = t^{-1}((t-1)A_{t-1} + H_t H_t^T) \\ B_t = t^{-1}((t-1)B_{t-1} + H_t X_t^T) \\ W_t = \operatorname{argmin}_{W \in \mathcal{C} \subseteq \mathbb{R}^{d \times r}_{\geq 0}} \left(\operatorname{tr}(WA_tW^T) - 2\operatorname{tr}(WB_t)\right), \end{cases}$$

$f_t = \text{empirical loss}, \qquad \hat{f}_t = \text{surrogate loss}, \quad f = \text{expected loss}$

Theorem (Mairal, Bach, Ponce, and Sapiro '10)

Suppose $(X_t)_{t\geq 0}$ are i.i.d. with common distribution π. Let $(W_{t-1}, H_t)_{t\geq 1}$ be the optimal solution to the above ONMF algorithm.

(i) *$(f_t(W_t))_{t\geq 1}$ and $(\hat{f}_t(W_t))_{t\geq 1}$ converge to the same constant almost surely.*

Convergence under i.i.d. assumption

$$\textit{Upon arrival of } X_t\textit{:} \quad \begin{cases} H_t = \operatorname{argmin}_{H \in \mathbb{R}^{r \times n}_{\geq 0}} \|X_t - W_{t-1}H\|_F^2 + \lambda \|H\|_1 \\ A_t = t^{-1}((t-1)A_{t-1} + H_t H_t^T) \\ B_t = t^{-1}((t-1)B_{t-1} + H_t X_t^T) \\ W_t = \operatorname{argmin}_{W \in \mathcal{C} \subseteq \mathbb{R}^{d \times r}_{\geq 0}} \left(\operatorname{tr}(WA_tW^T) - 2\operatorname{tr}(WB_t)\right), \end{cases}$$

$f_t =$ empirical loss, $\quad \hat{f}_t =$ surrogate loss, $\quad f =$ expected loss

Theorem (Mairal, Bach, Ponce, and Sapiro '10)

Suppose $(X_t)_{t\geq 0}$ are i.i.d. with common distribution π. Let $(W_{t-1}, H_t)_{t\geq 1}$ be the optimal solution to the above ONMF algorithm.

(i) *$(f_t(W_t))_{t\geq 1}$ and $(\hat{f}_t(W_t))_{t\geq 1}$ converge to the same constant almost surely.*

(ii) $\limsup_{t\to\infty} \|\nabla f(W_t)\|_{\text{op}} = 0$ *almost surely.*

Convergence under Markovian dependence

$$\textit{Upon arrival of } X_t\textit{:} \quad \begin{cases} H_t = \operatorname{argmin}_{H \in \mathbb{R}^{r\times n}_{\geq 0}} \|X_t - W_{t-1}H\|_F^2 + \lambda \|H\|_1 \\ A_t = t^{-1}((t-1)A_{t-1} + H_t H_t^T) \\ B_t = t^{-1}((t-1)B_{t-1} + H_t X_t^T) \\ W_t = \operatorname{argmin}_{W \in \mathcal{C} \subseteq \mathbb{R}^{d\times r}_{\geq 0}} \left(\operatorname{tr}(WA_tW^T) - 2\operatorname{tr}(WB_t)\right), \end{cases}$$

$f_t =$ empirical loss, $\quad \hat{f}_t =$ surrogate loss, $\quad f =$ expected loss

Theorem (Balzano, Lyu, Needell '19+)

Suppose $(X_t)_{t\geq 0}$ is an irreducible MC on a finite state space with unique stationary distribution π. Let $(W_{t-1}, H_t)_{t\geq 1}$ be a solution to the above ONMF algorithm. Then the following hold.

(i) $\lim_{t\to\infty} \mathbb{E}[f_t(W_t)] = \lim_{t\to\infty} \mathbb{E}[\hat{f}_t(W_t)] < \infty.$

Convergence under Markovian dependence

$$\textit{Upon arrival of } X_t\textit{:} \quad \begin{cases} H_t = \operatorname{argmin}_{H \in \mathbb{R}^{r\times n}_{\geq 0}} \|X_t - W_{t-1}H\|_F^2 + \lambda \|H\|_1 \\ A_t = t^{-1}((t-1)A_{t-1} + H_t H_t^T) \\ B_t = t^{-1}((t-1)B_{t-1} + H_t X_t^T) \\ W_t = \operatorname{argmin}_{W \in \mathcal{C} \subseteq \mathbb{R}^{d\times r}_{\geq 0}} \left(\operatorname{tr}(WA_tW^T) - 2\operatorname{tr}(WB_t)\right), \end{cases}$$

$f_t = \text{empirical loss}, \qquad \hat{f}_t = \text{surrogate loss}, \quad f = \text{expected loss}$

Theorem (Balzano, Lyu, Needell '19+)

Suppose $(X_t)_{t\geq 0}$ is an irreducible MC on a finite state space with unique stationary distribution π. Let $(W_{t-1}, H_t)_{t\geq 1}$ be a solution to the above ONMF algorithm. Then the following hold.

(i) $\lim_{t\to\infty} \mathbb{E}[f_t(W_t)] = \lim_{t\to\infty} \mathbb{E}[\hat{f}_t(W_t)] < \infty.$

(ii) $f_t(W_t) - \hat{f}_t(W_t) \to 0$ *as* $t \to \infty$ *almost surely.*

Convergence under Markovian dependence

$$\textit{Upon arrival of } X_t\textit{:} \quad \begin{cases} H_t = \operatorname{argmin}_{H \in \mathbb{R}^{r\times n}_{\geq 0}} \|X_t - W_{t-1}H\|_F^2 + \lambda \|H\|_1 \\ A_t = t^{-1}((t-1)A_{t-1} + H_t H_t^T) \\ B_t = t^{-1}((t-1)B_{t-1} + H_t X_t^T) \\ W_t = \operatorname{argmin}_{W \in \mathcal{C} \subseteq \mathbb{R}^{d\times r}_{\geq 0}} \left(\operatorname{tr}(WA_tW^T) - 2\operatorname{tr}(WB_t)\right), \end{cases}$$

$f_t = \text{empirical loss}, \qquad \hat{f}_t = \text{surrogate loss}, \quad f = \text{expected loss}$

Theorem (Balzano, Lyu, Needell '19+)

Suppose $(X_t)_{t\geq 0}$ is an irreducible MC on a finite state space with unique stationary distribution π. Let $(W_{t-1}, H_t)_{t\geq 1}$ be a solution to the above ONMF algorithm. Then the following hold.

(i) $\lim_{t\to\infty} \mathbb{E}[f_t(W_t)] = \lim_{t\to\infty} \mathbb{E}[\hat{f}_t(W_t)] < \infty$.

(ii) $f_t(W_t) - \hat{f}_t(W_t) \to 0$ *as* $t \to \infty$ *almost surely.*

(iii) $\limsup_{t\to\infty} \|\nabla f(W_t)\|_{\text{op}} = 0$ *almost surely.*

Applications of ONMF

MyLymeData

- Lyme disease a vector-borne disease typically transmitted by tick or insect bite or blood-blood contact
  - Symptoms often mimic those of others, e.g. MS / ALS / Parkinsons / FMA … and can become chronic
- CDC estimates 300,000 new diagnoses each year
  - Likely a grandiose underestimate
- Poorly understood, poorly funded, poorly diagnosed, poorly treated

MyLymeData
A PROJECT OF LYMEDISEASE.ORG

MyLymeData

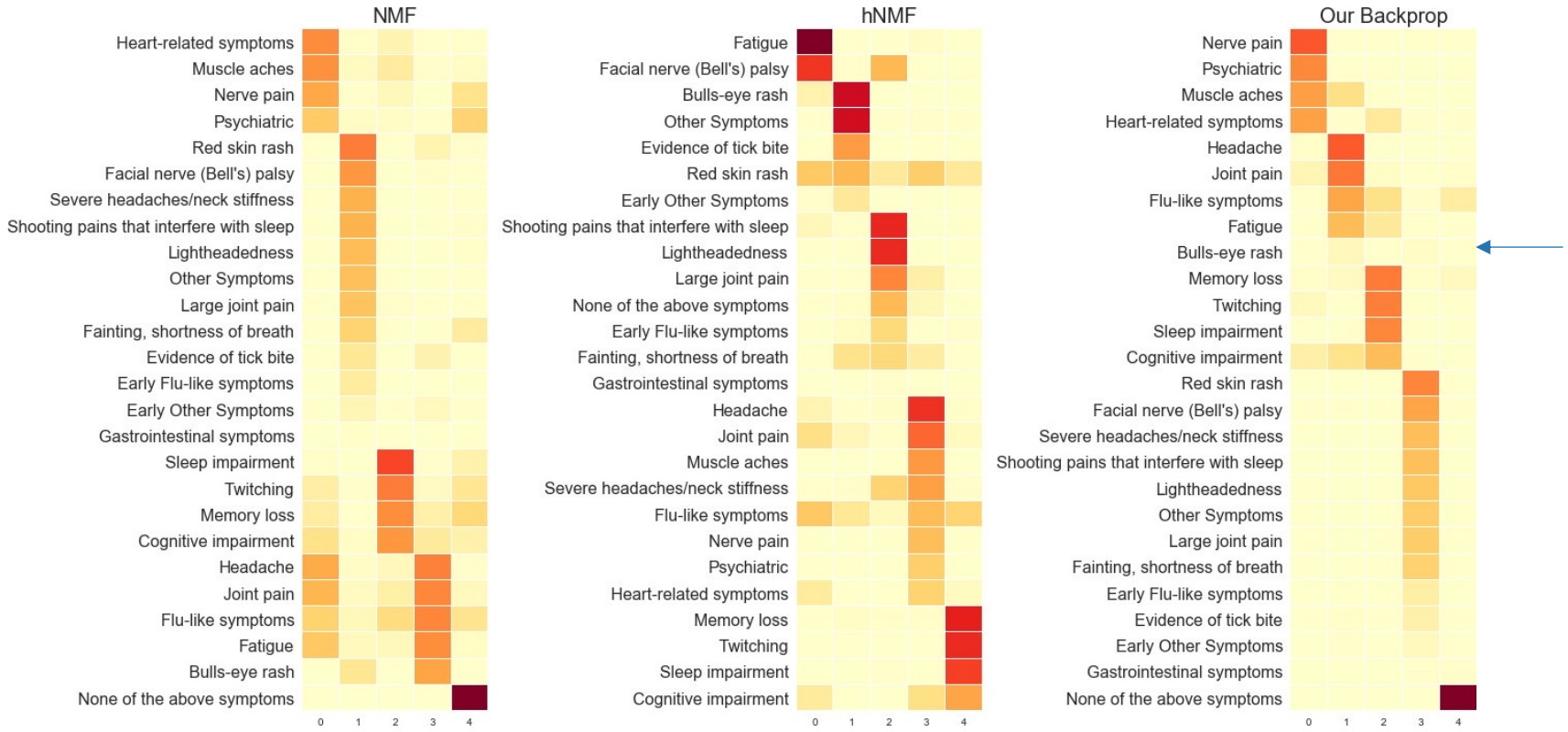
- MyLymeData launched Nov 2015 → ~14,000 patients registered

Type of data	# questions	Examples
Demographic	23	sex, age, birth country
Tick bite info	19	presence of EM rash, knowledge of bite, location, tick test results
Diagnostic info	34	# mis-diagnoses, time until diagnosis, # specialists
Early symptoms	11	rash, flu, neurologic, joint issues
Basic symptoms	13	(similar), with severity
Extended symptoms	54	(similar), with severity
Lab tests	21	CDC 2-tier, Western Blot, various blood markers
Co-infections	8	Babesia, Bartonella, Mycoplasma, Rickettsia
Treatment	129	effectiveness, type/style of antibiotics, duration, other treatments
Quality of life	54	sleep, ability to work, physical activity, improvements

Table 1: Patient supplied survey data description

MyLymeData
A PROJECT OF LYMEDISEASE.ORG

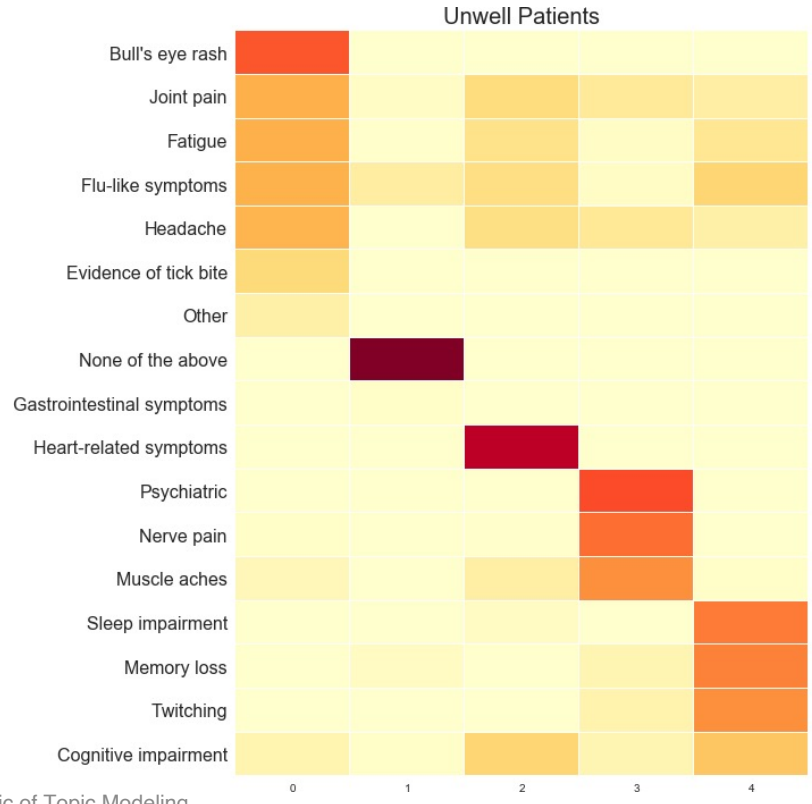
Comparisons on Lyme data

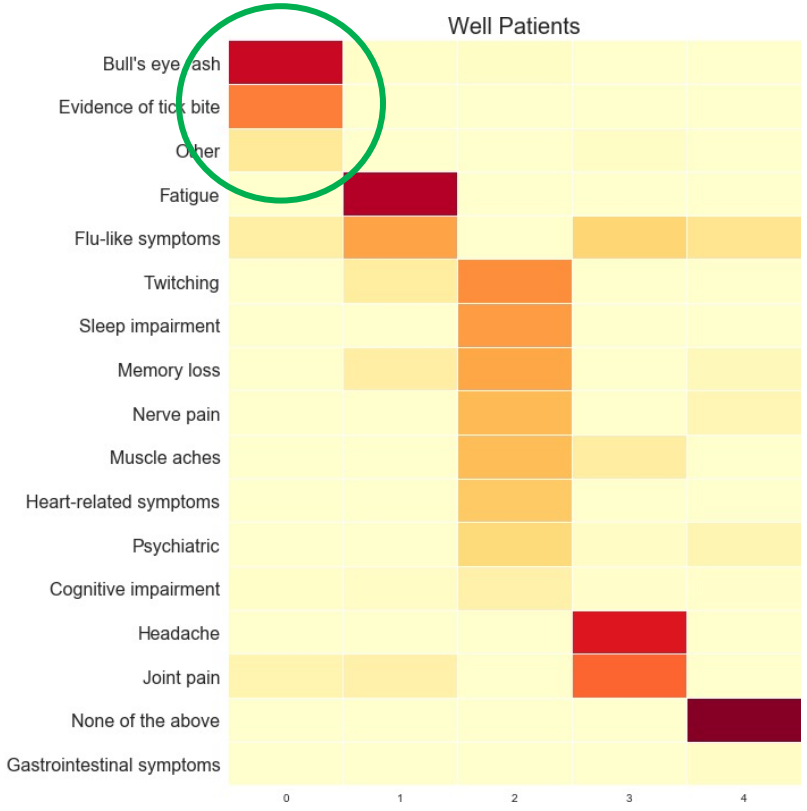


The hidden topics here may provide insight on how symptoms manifest themselves

More Lyme data results

- Run our backpropagation separately on unwell and well datasets, then compare
- Notice the topics are **very** different!





Learning features from MCMC trajectories - Network Dictionary Learning

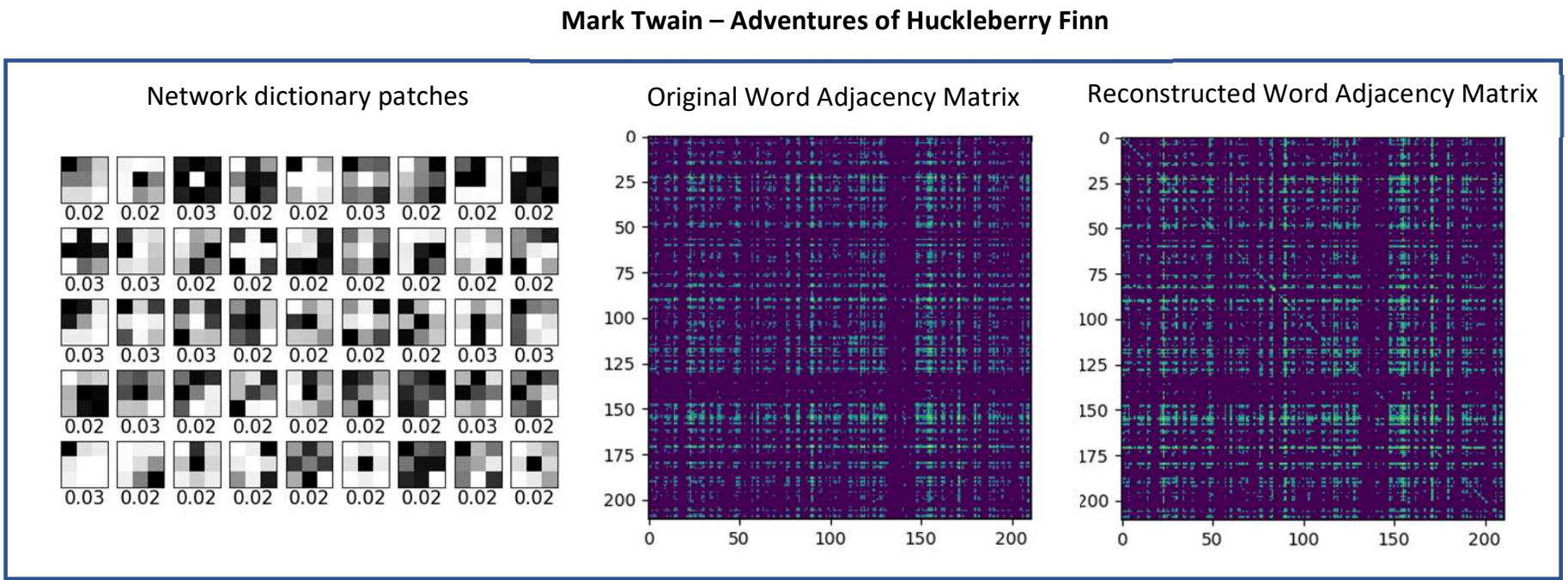


Figure: (Left) 45 learned 3 by 3 network dictionary patches from Glauber chain sampling from the Word Adjacency Matrix of "Mark Twain - Adventures of Huckleberry Finn". Black=1 and white = 0 with gray scale. (Middle) Heat map of the original Word Adjacency Matrix where blue = 0 and yellow = 1 (Right) the reconsrtuction.

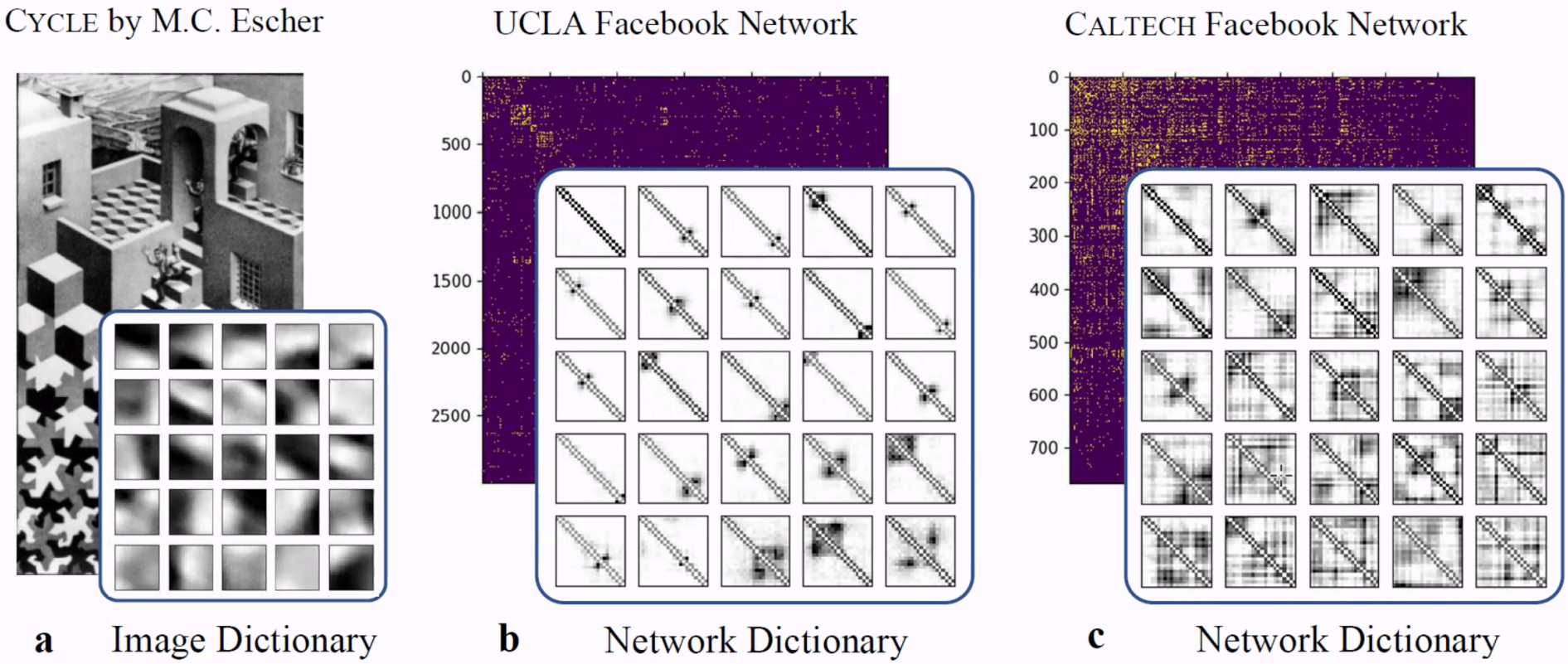


Figure: [6] (a) 25 latent shapes learned from an image by NMF. (b,c) 25 latent motifs for $k = 21$-node connected subgraphs learned from UCLA and CALTECH.

ONMF for image reconstruction

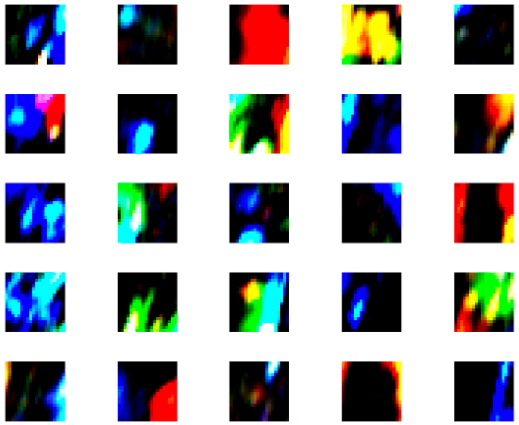
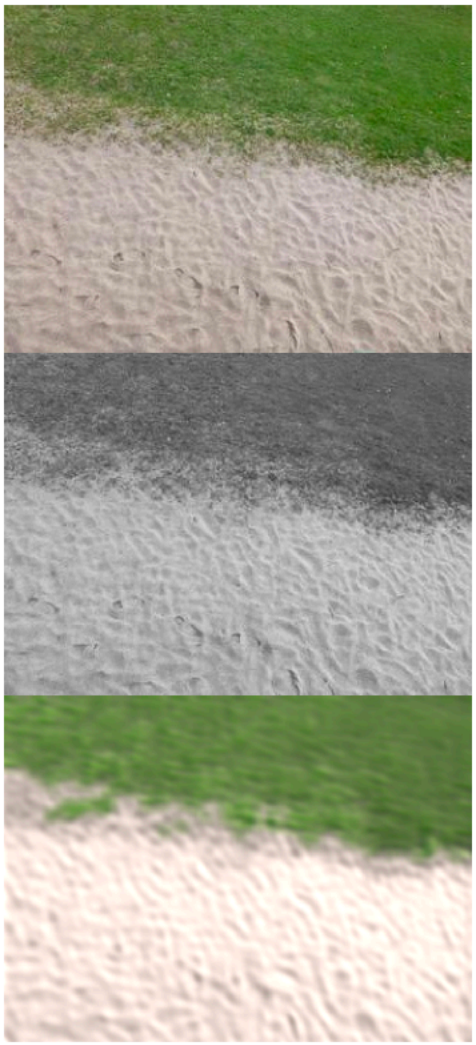


Fig. 7: Image Compression Via ONMF. (Top) uncompressed image of Leonid Afremov's famous painting "Rain's Rustle." (Middle) 25 of the 100 learned dictionary elements, reshaped from their vectorized form to color image patch form. (Bottom): Painting compressed using a dictionary of 100 vectorized 20×20 color image patches obtained from 30 data samples of ONMF, each consisting of 1000 randomly selected sample patches. We used an overlap length of 15 in the patch averaging for the construction of the compressed image.

ONMF for color restoration



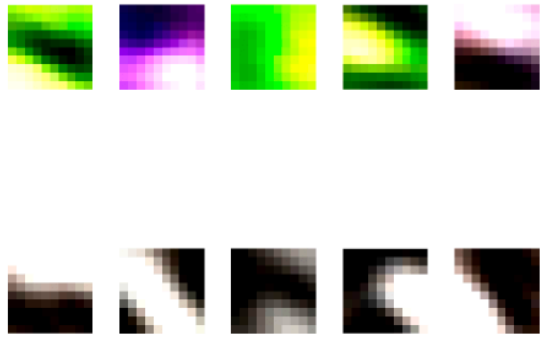
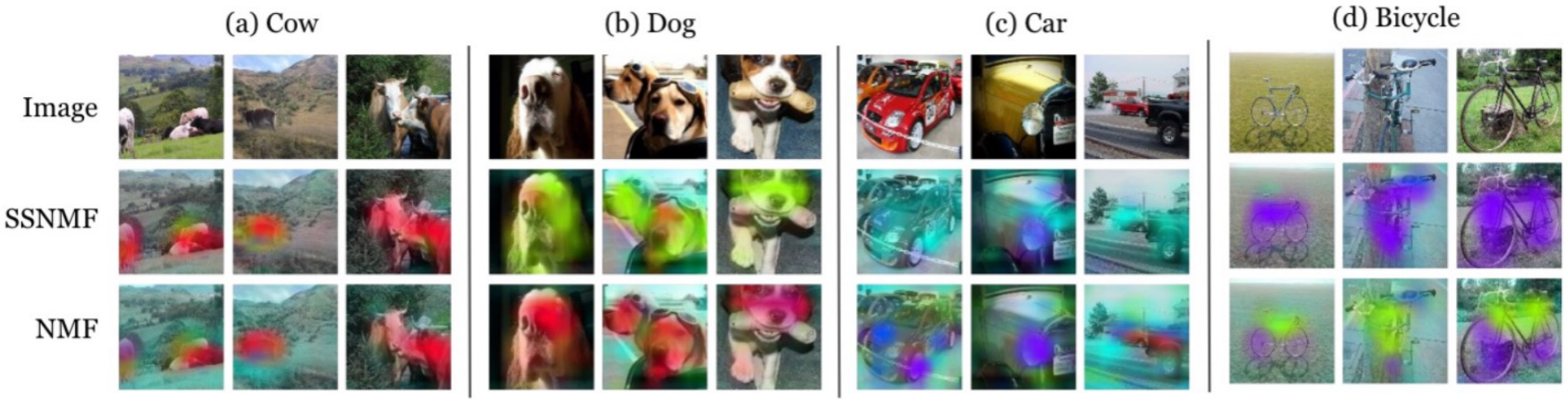


Fig. 8: Color Restoration. (First) Original image of grass and sand. (Second) Conversion of first image using matlab's default (linear) color to grayscale conversion function, rgb2gray. (Third) Restored color image, obtained by applying our method to the second (grayscale) image using the dictionary below. (Fourth) The first row displays the grass dictionary, the second row displays the sand dictionary. Each dictionary contains five 10 x 10 color image patches. These were trained on our ONMF with 20 batches of 1000 randomly selected sample patches each. We used maximal patch overlap in the restoration process.

More applications

(O)NMF for image co-segmentation



ONMF for video reconstruction

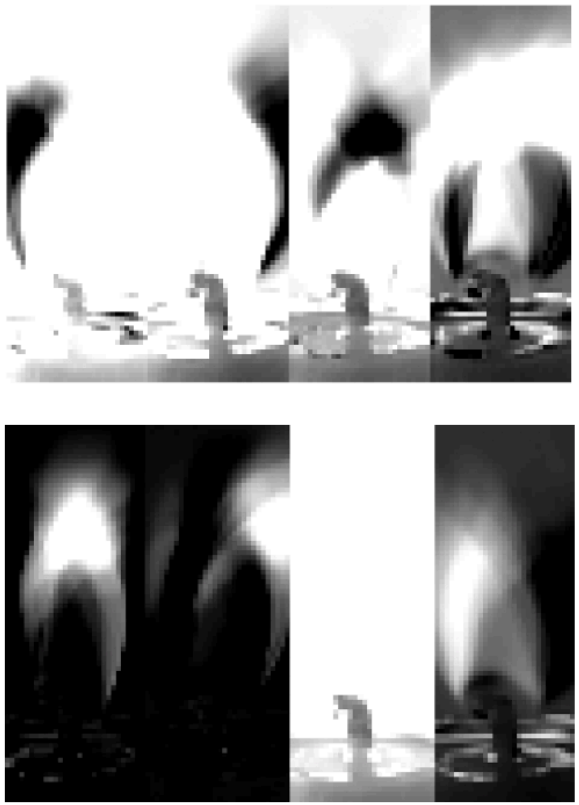


Fig. 4: Candle Video Dictionaries. The first dictionary consists of four elements and was trained by an alternating least squares-based, offline NMF, the second dictionary below was trained using ONMF, where each time frame of the video represented a new data point.

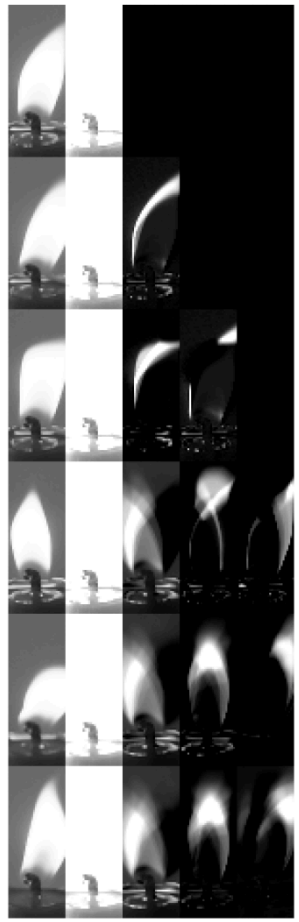


Fig. 5: Candle video and learned dictionary at various time frames (time goes from top to bottom). The left column corresponds to the actual video frame. The remaining four columns each correspond to a particular dictionary element. The six correspond to different time frames, 1, 5, 7, 15, 35, and 75

ONMF for video denoising

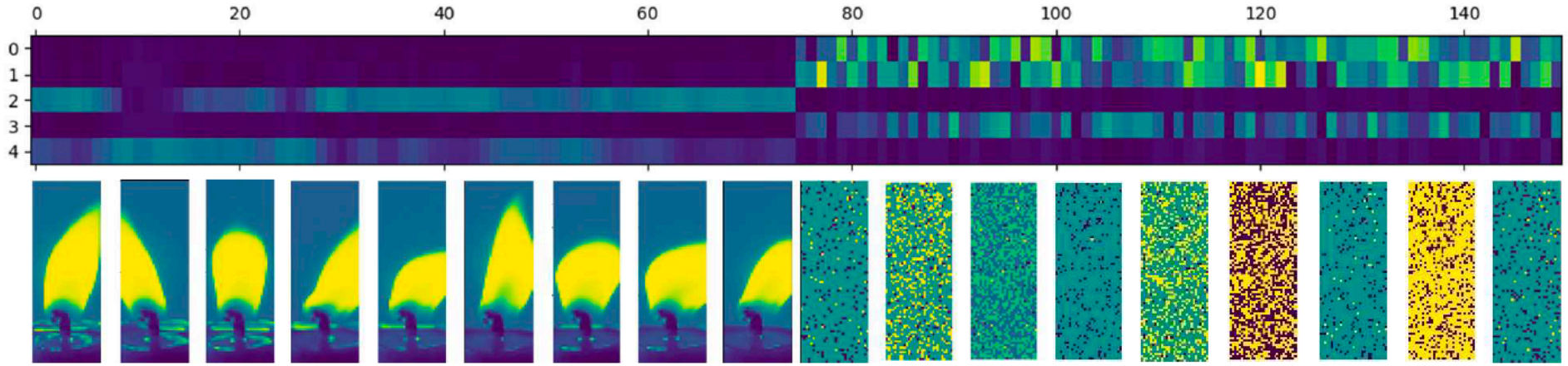


Fig. 6: Learning time evolution dictionary from a video frame using NMF. The first and last 75 frames of the video are from a candle video and white noise, respectively, as shown below. By an approximate factorization of shape [time x space] = [time x 5] [5 x space], we learn 150 x 5 dictionary matrix, whose columns give an approximate basis for the time evolution of each pixel of the video frame. The learned time evolution dictionaries detect the planeted 'phase transition' between frames 75 and 76.

ONMF as a pattern detection and prediction tool – COVID19

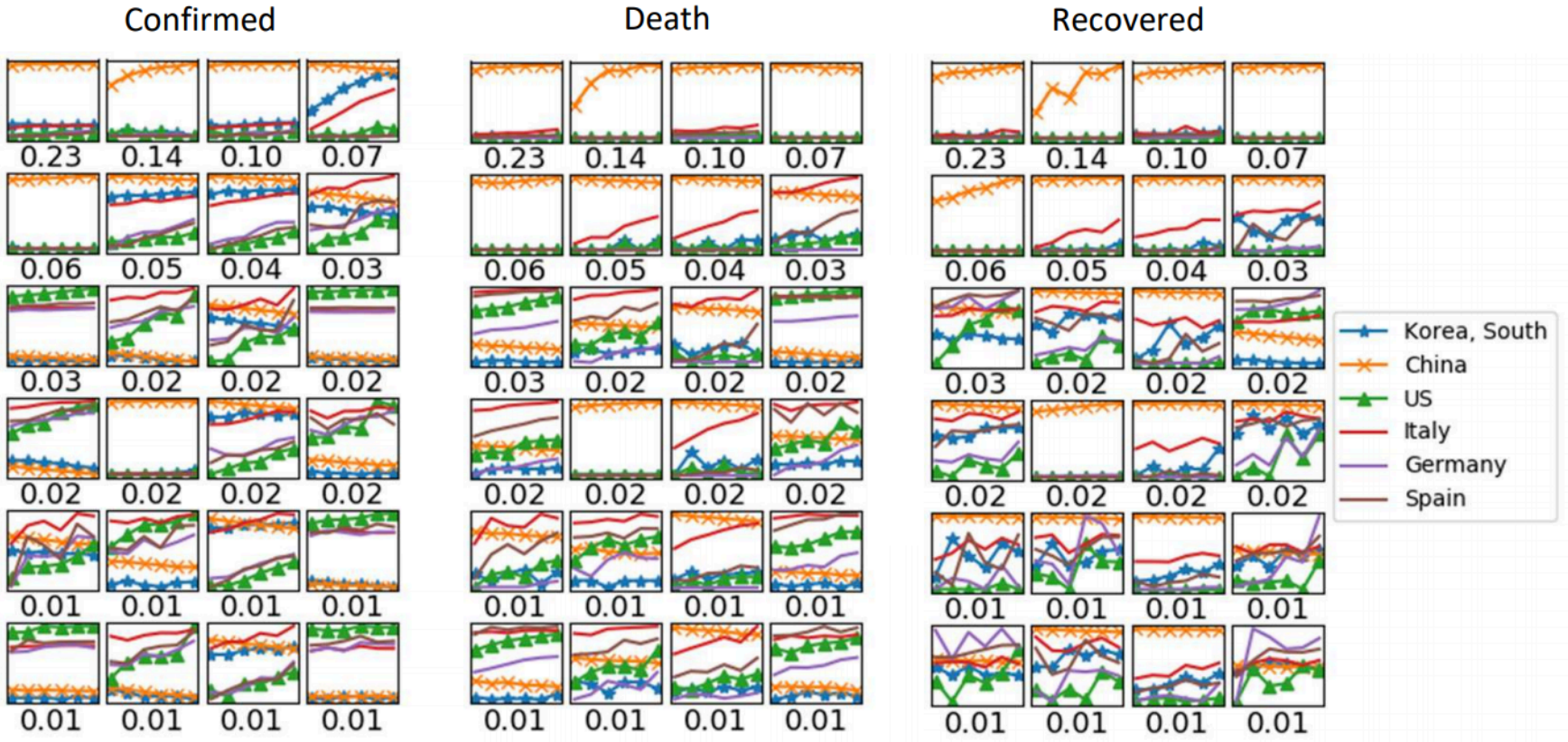


Fig. 2. 24 Joint dictionary atoms of 6-day evolution patterns of new daily cases (confirmed/death/recovered) in six countries (S. Korea, China, US, Italy, Germany, and France). Each dictionary atom is a $6 * 6 * 3 = 108$ dimensional vector corresponding to `time * country * case type`. The corresponding importance metric is shown below each atom. 50 atoms are learned and the figure shows top 24 with the highest importance metric.

ONMF as a pattern detection and prediction tool – COVID19

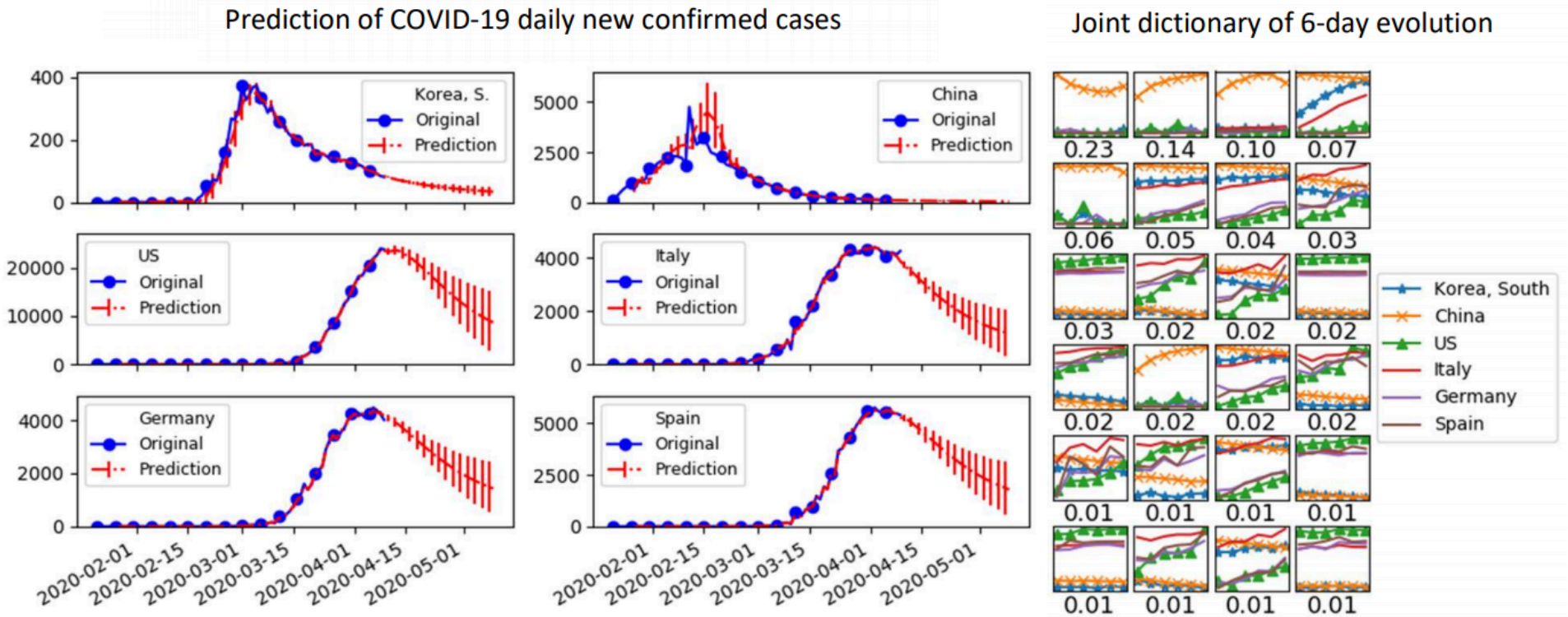


Fig. 3. Joint dictionary learning and prediction for the time-series of new daily cases (confirmed/death/recovered) in six countries (S. Korea, China, US, Italy, Germany, and France). After joint dictionary atoms are learned by minibatch learning, they are further adapted to the time-series data by concurrent online learning and predictions. (Right) Joint dictionary atoms of 6-day evolution patterns of new confirmed cases. The corresponding importance metric is shown below each atom. (Left) Plot of the original and predicted daily new confirmed cases of the six countries. The errorbar in the red plot shows standard deviation of 1000 trials.

ONMF as a pattern detection and prediction tool – COVID19

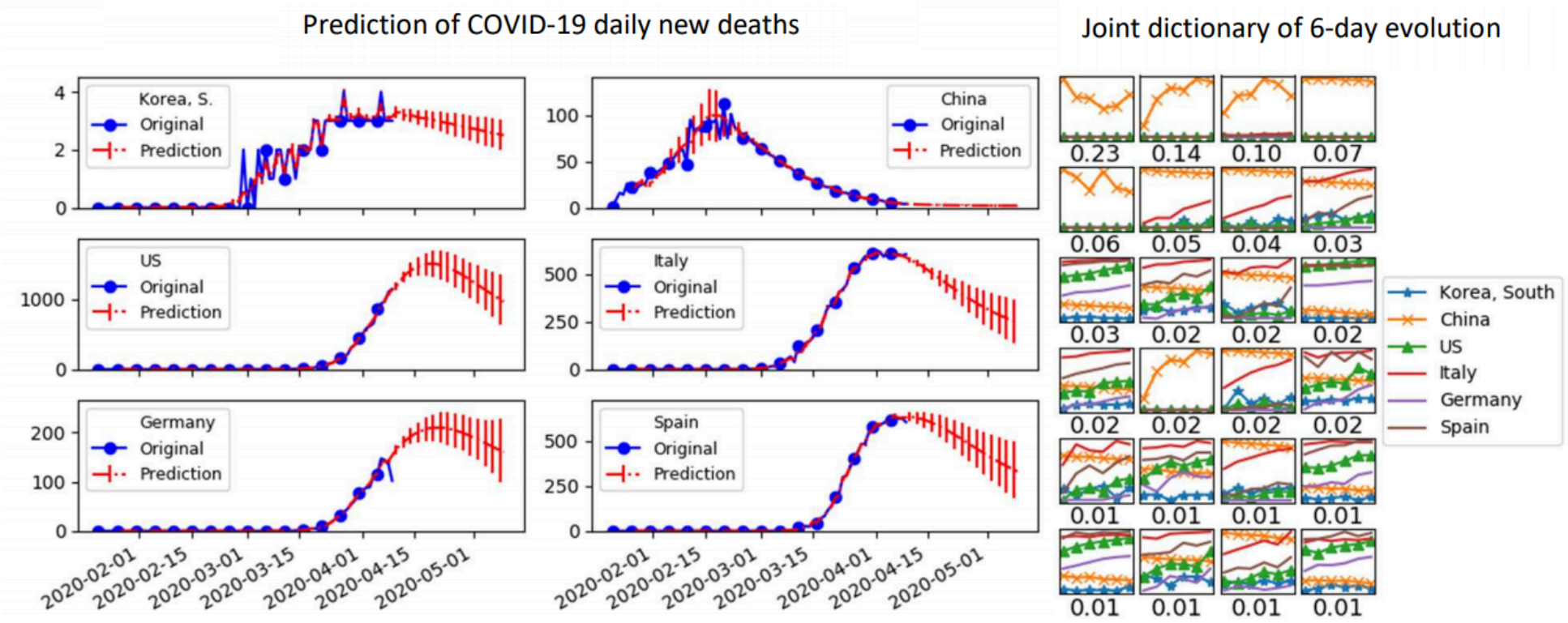


Fig. 4. Joint dictionary learning and prediction for the time-series of new daily cases (confirmed/death/recovered) in six countries (S. Korea, China, US, Italy, Germany, and France). After dictionary atoms representing fundamental joint time-series patterns are obtained by minibatch learning, they are further adapted to the time-series data by online learning while making predictions. (Right) Joint dictionary atoms of 6-day evolution patterns of new death cases. The corresponding importance metric is shown below each atom. (Left) The plot of the original and predicted daily new death cases of the six countries. The error bar in the red plot shows the standard deviation of 1000 trials.

ONMF as a pattern detection and prediction tool – COVID19

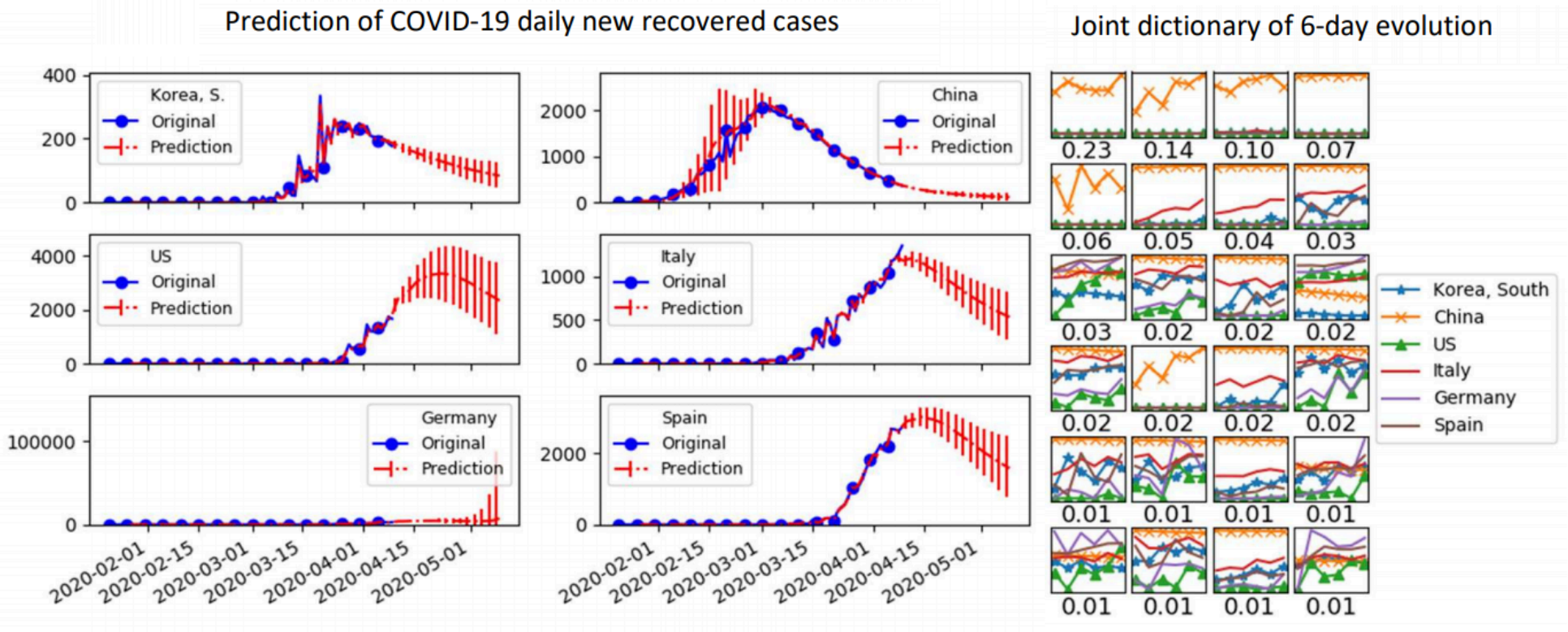


Fig. 5. Joint dictionary learning and prediction for the time-series of new daily cases (confirmed/death/recovered) in six countries (S. Korea, China, US, Italy, Germany, and France). After joint dictionary atoms are learned by minibatch learning, they are further adapted to the time-series by online learning while making predictions. (Right) Joint dictionary atoms of 6-day evolution patterns of new recovered cases. The corresponding importance metric is shown below each atom. (Left) The plot of the original and predicted daily new recovered cases of the six countries. The errorbar in the red plot shows the standard deviation of 1000 trials.

Extension to Online NN *Tensor* Factorization (ONTF)

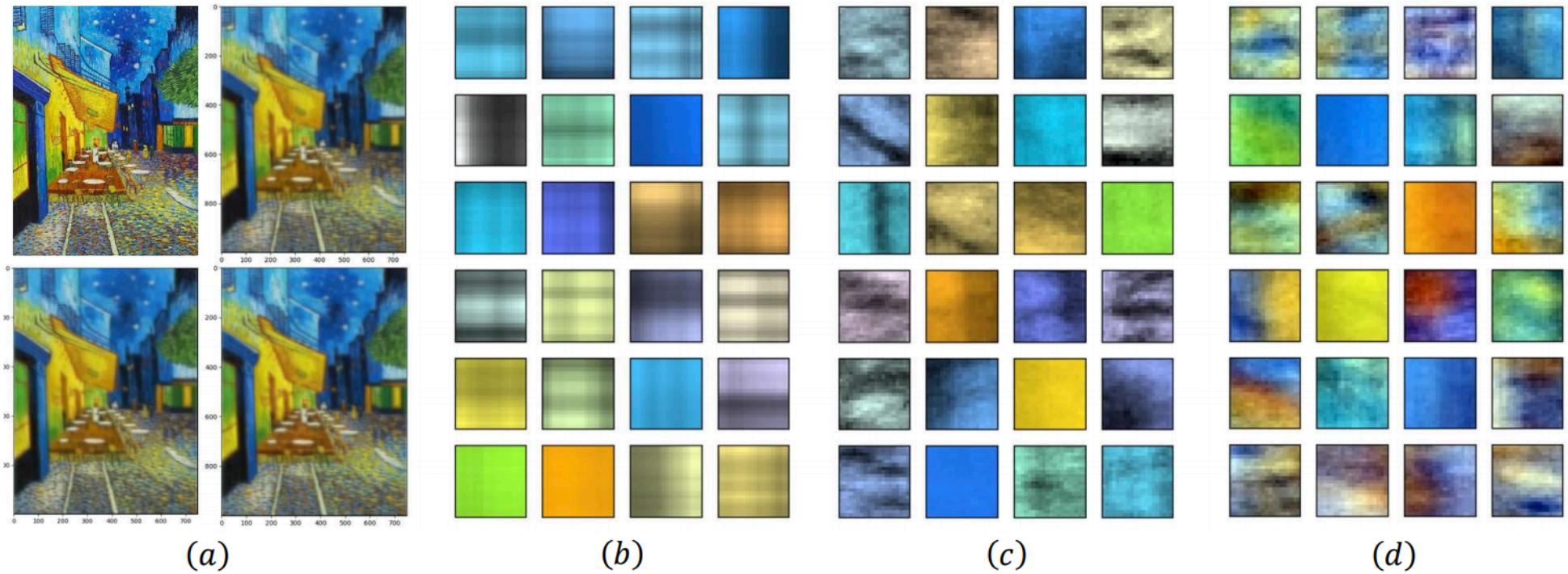
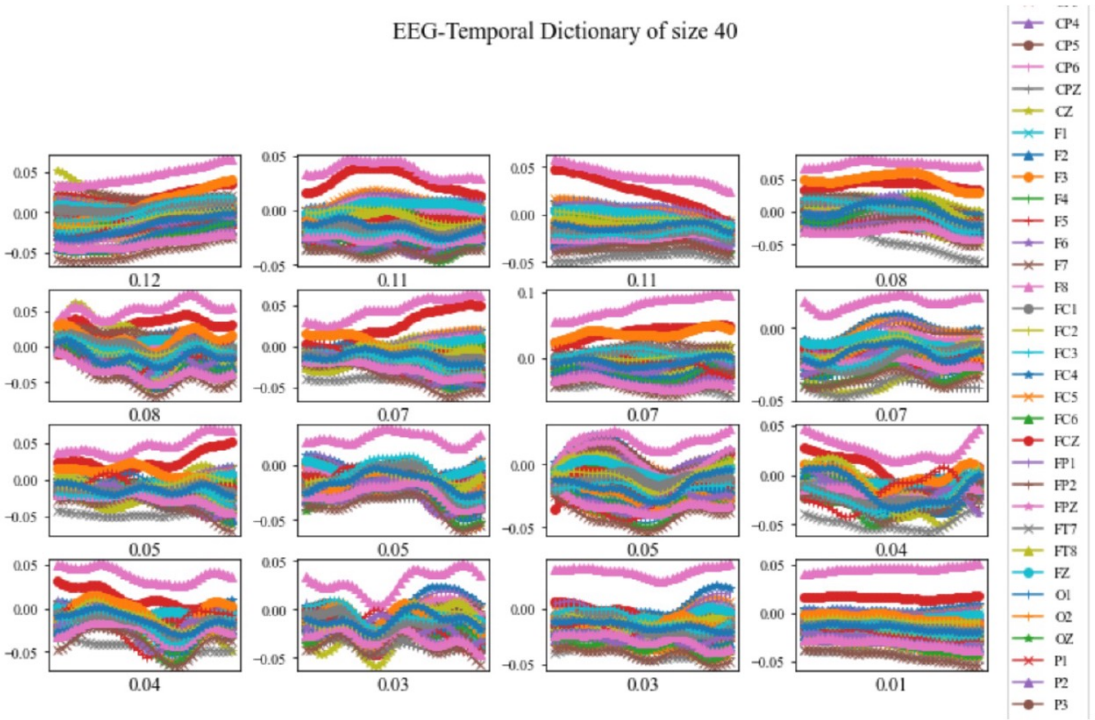


FIGURE 3. Learning 24 CP-dictionary patches by Online CPDL: (a) original (top left) and reconstructed images (top right from (b), bottom left from (c), and bottom right from (d)), (b) dictionary learned by Online CPDL, (c) dictionary learned by vectorizing the spatial modes and applying Online CPDL to resulting tensor then reshaping, (d) dictionary learned using online CPDL on fully vectorized image patch data.

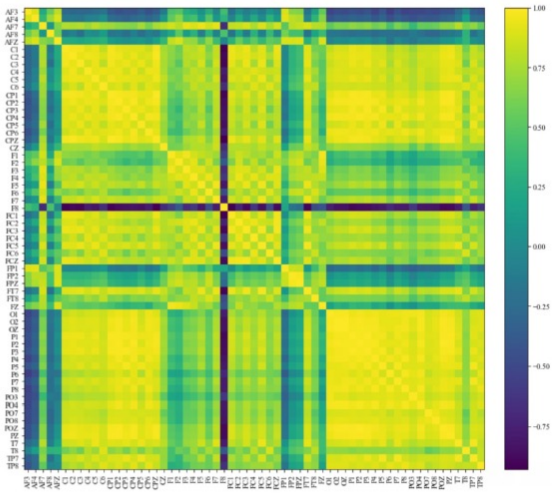
More applications

ONMF on EEG node correlations (UCI EEG Alcoholism data, 64 electrodes)

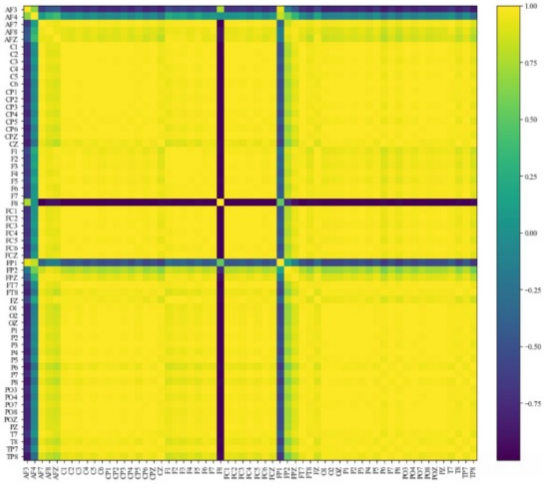


More applications

ONMF on EEG data (node correlations)



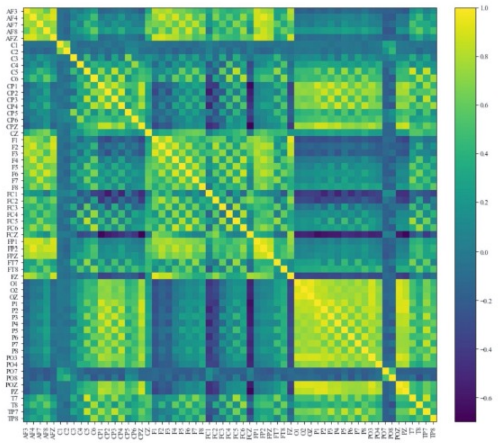
(Pearson)



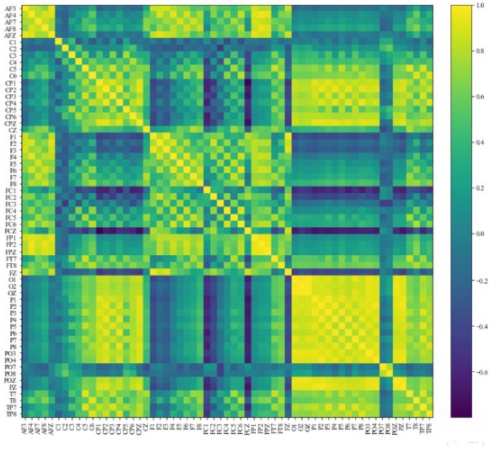
(ONMF)

More applications

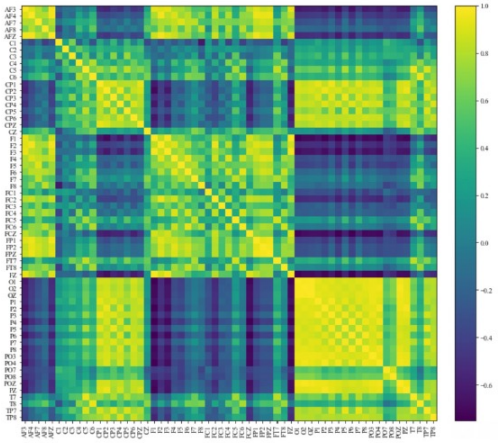
ONMF on EEG data (node correlations)



(Pearson w/ gradient)



(ONMF w/ gradient)



(ONMF w/o gradient, r=16)

ONTF to learn activation patterns in mouse cortex

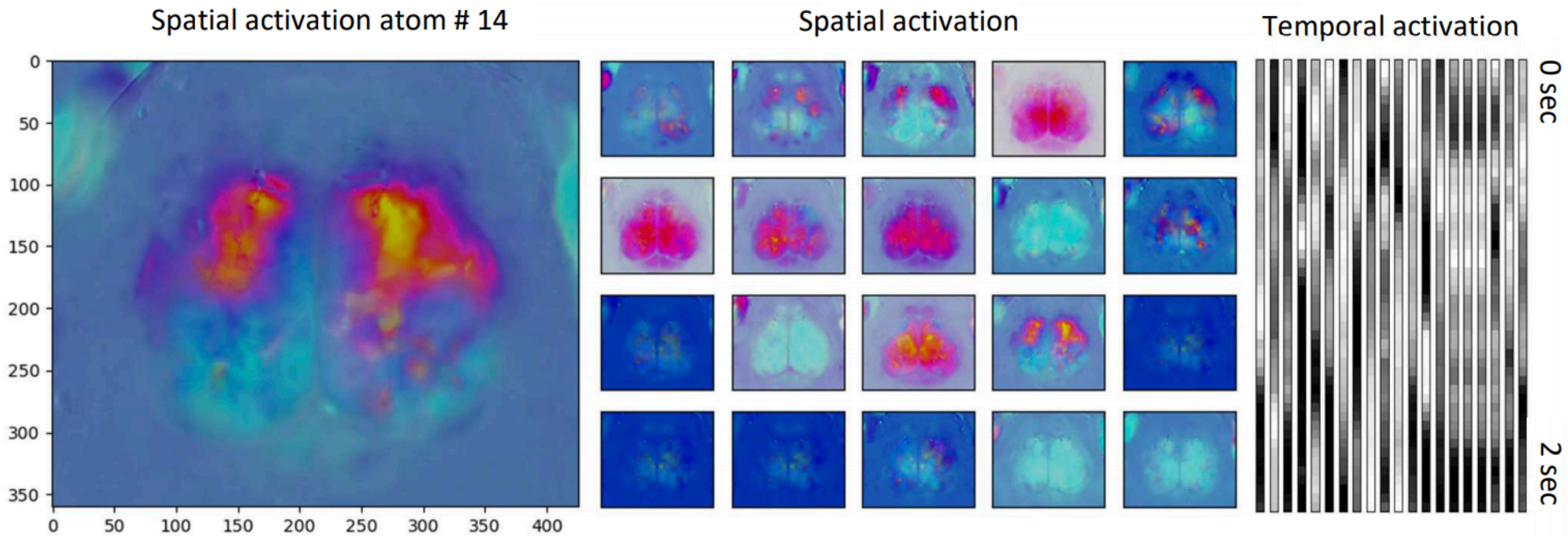


FIGURE 4. Learning 20 CP-dictionary patches from video frames on brain activity across the mouse cortex.

ONTF to learn weather patterns

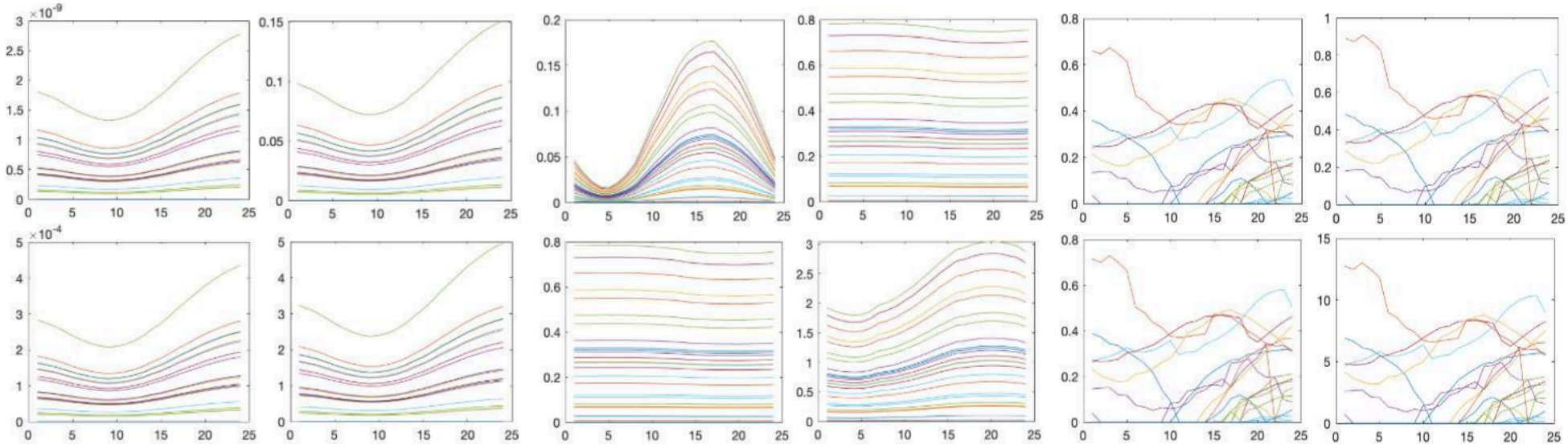


FIGURE 5. Display of one atom from three different dictionaries of 25 atoms which were obtained from Online CPDL on weather data: (a) no reshaping, (b) data which was reshaped to $36 \times (24 \times 4)$, and (c) data which was reshaped to $(36 \times 24) \times 4$.

Learning features from MCMC trajectories - Ising model

- The two dimensional Ising model (1920) is one of the most well-known spin systems in the physics literature, which models ferromagnetism.

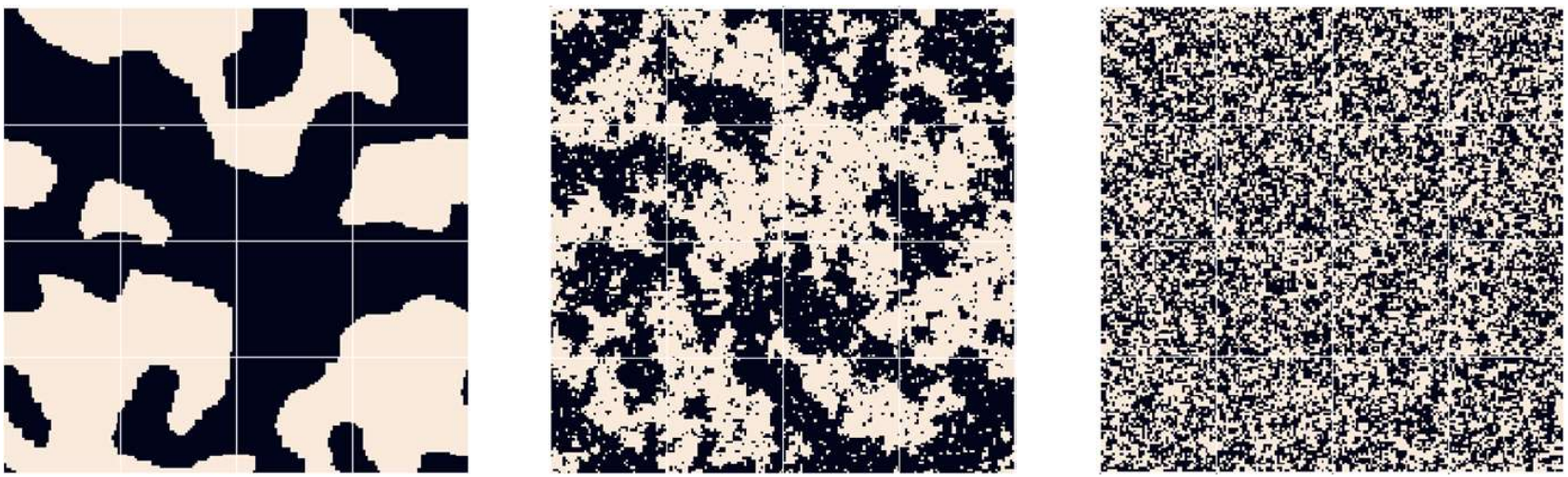


Figure: MCMC simulation of Ising model on 200 by 200 square lattice at temperature $T = 0.5$ (left), $T = 2.26$ (middle), and $T = 5$ (right).

Learning features from MCMC trajectories - Ising model

- The two dimensional Ising model (1920) is one of the most well-known spin systems in the physics literature, which models ferromagnetism.
- For each temperature parameter $T > 0$, Define a probability distribution π on the set $\{-1, 1\}^{\mathbb{Z}^2}$ of spin configurations by

$$\pi_T(\mathbf{x}) \propto \exp\left(-\frac{1}{T}\sum_{i\sim j}\mathbf{x}(i)\mathbf{x}(j)\right).$$

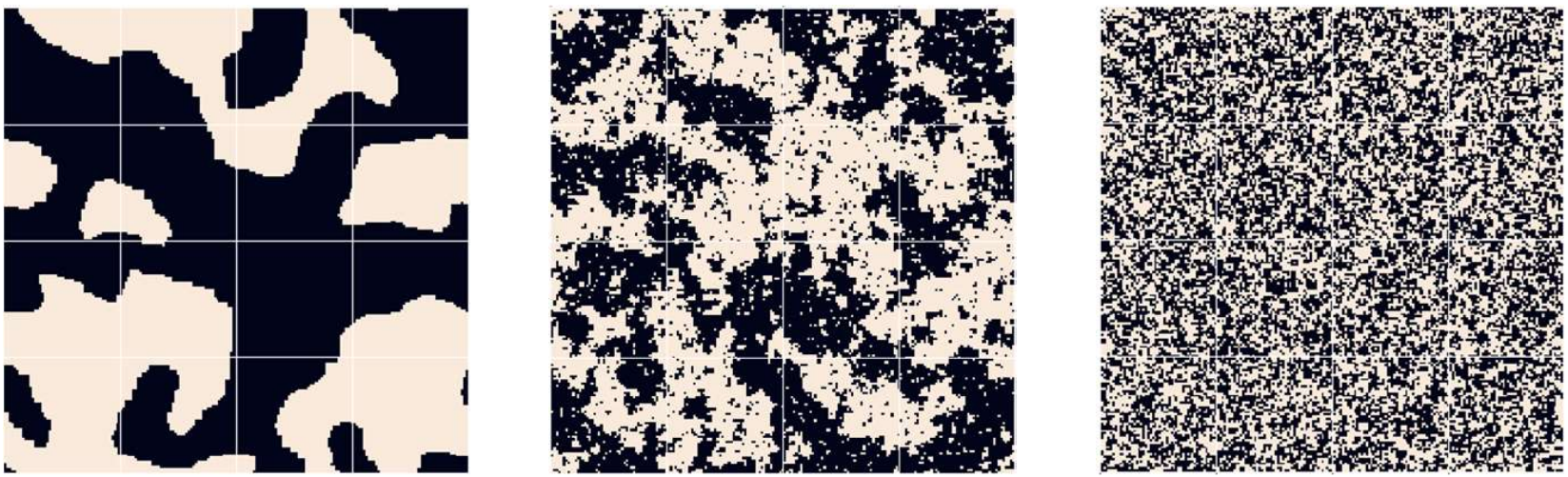
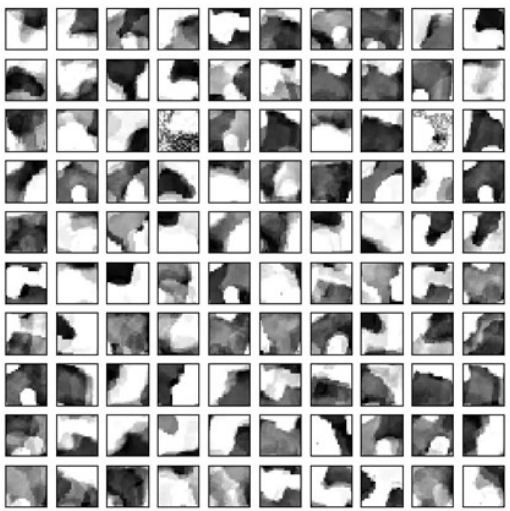


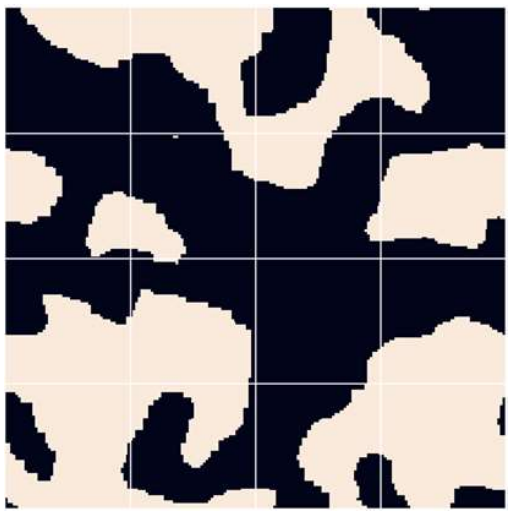
Figure: MCMC simulation of Ising model on 200 by 200 square lattice at temperature $T = 0.5$ (left), $T = 2.26$ (middle), and $T = 5$ (right).

Learning features from MCMC trajectories - Ising model

Dictionary patches of size 20
learned from MCMC trajectory



An Ising spin configuration
sampled at temperature $T = 0.5$

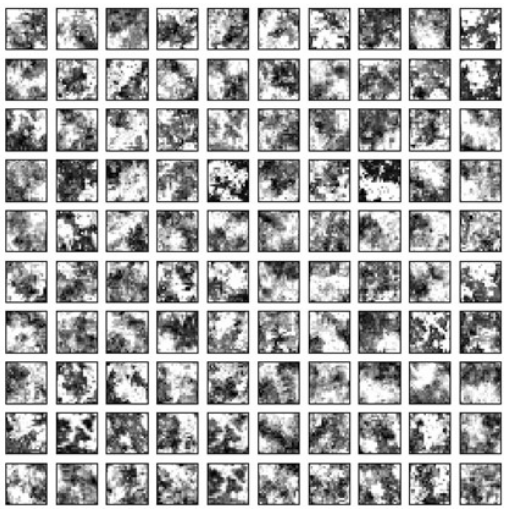


Reconstructed spin configuration
using learned dictionary patches

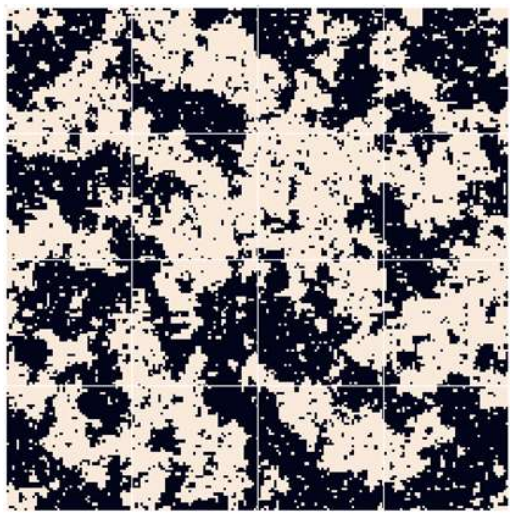
Figure: (Left) 100 learned dictionary patches from a MCMC Gibbs sampler for the Ising model on 200×200 square lattice at a subcritical temperature ($T = 0.5$). (Middle) A sampled spin configuration at $T = 0.5$. (Right) Reconstruction of the original spin configuration in the middle using the dictionary patches on the left.

Learning features from MCMC trajectories - Ising model

Dictionary patches of size 20
learned from MCMC trajectory



An Ising spin configuration
sampled at temperature $T = 2.26$



Reconstructed spin configuration
using learned dictionary patches

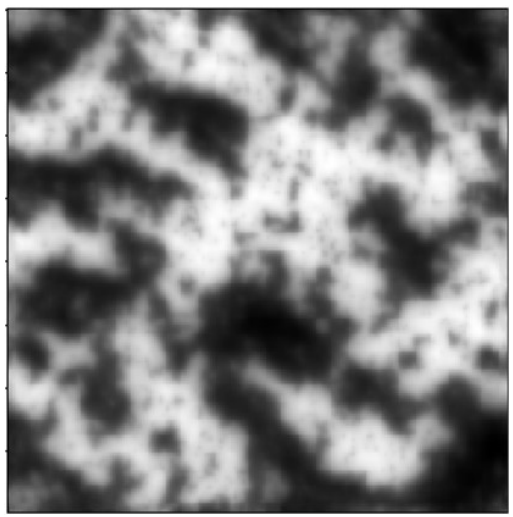
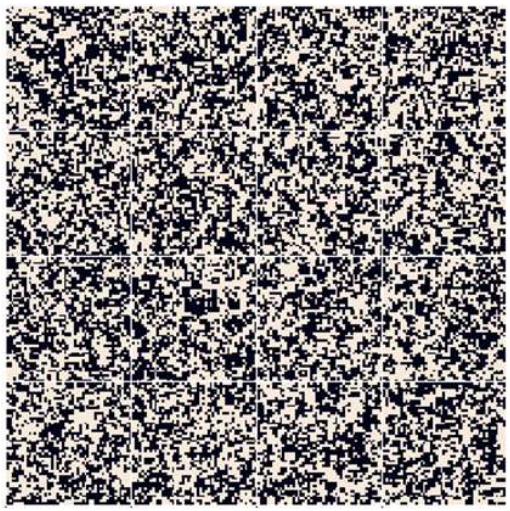


Figure: (Left) 100 learned dictionary patches from a MCMC Gibbs sampler for the Ising model on 200×200 square lattice near the critical temperature ($T = 2.26$). (Middle) A sampled spin configuration at $T = 2.26$. (Right) Reconstruction of the original spin configuration in the middle using the dictionary patches on the left.

Learning features from MCMC trajectories - Ising model

Dictionary patches of size 20 learned from MCMC trajectory

An Ising spin configuration sampled at temperature $T = 5$



Reconstructed spin configuration using learned dictionary patches

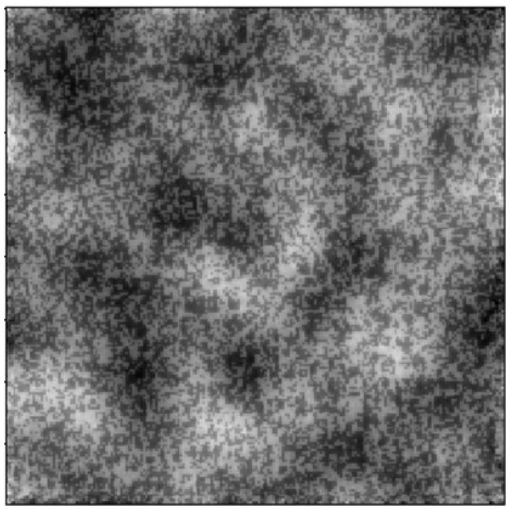


Figure: (Left) 100 learned dictionary patches from a MCMC Gibbs sampler for the Ising model on 200×200 square lattice at a supercritical temperature ($T = 5$). (Middle) A sampled spin configuration at $T = 5$. (Right) Reconstruction of the original spin configuration in the middle using the dictionary patches on the left.

Further research directions

- ONMF for variable number of dictionaries (added optimization dimension)

Further research directions

- ONMF for variable number of dictionaries (added optimization dimension)
- ONMF for non-stationary data matrices (what do we want to learn in this case?)

Further research directions

- ONMF for variable number of dictionaries (added optimization dimension)
- ONMF for non-stationary data matrices (what do we want to learn in this case?)
- Dynamic topic modeling using ONMF

Further research directions

- ONMF for variable number of dictionaries (added optimization dimension)
- ONMF for non-stationary data matrices (what do we want to learn in this case?)
- Dynamic topic modeling using ONMF
- Extension to online Tensor Factorization (no convergence result known even for the i.i.d. case)

Thank you for listening!

- deanna@math.ucla.edu
- math.ucla.edu/~deanna

Thanks!

5. Proof of convergence

Notations

- Fix $\lambda > 0$ and define the following the **quadratic loss function**

$$\ell(X, W) = \inf_{H \in \mathbb{R}^{r \times n}} \|X - WH\|_F^2 + \lambda \|H\|_1,$$

Define the **expected loss** and **empirical loss** functions

$$f(W) = \mathbb{E}_{X \sim \pi}[\ell(X, W)], \qquad f_t(W) = \frac{1}{t} \sum_{s=1}^{t} \ell(X_s, W)$$

Notations

- Fix $\lambda > 0$ and define the following the **quadratic loss function**

$$\ell(X, W) = \inf_{H \in \mathbb{R}^{r \times n}} \|X - WH\|_F^2 + \lambda \|H\|_1,$$

Define the **expected loss** and **empirical loss** functions

$$f(W) = \mathbb{E}_{X \sim \pi}[\ell(X, W)], \qquad f_t(W) = \frac{1}{t} \sum_{s=1}^{t} \ell(X_s, W)$$

- **ONMF algorithm**:

Upon arrival of X_t:

$$\begin{cases} H_t = \text{argmin}_{H \in \mathbb{R}^{r \times n}_{\geq 0}} \|X_t - W_{t-1} H\|_F^2 + \lambda \|H\|_1 \\ W_t = \text{argmin}_{W \in \mathcal{C}} \hat{f}_t(W), \end{cases}$$

where

$$\hat{f}_t(W) = \frac{1}{t} \sum_{s=1}^{t} (\|X_s - WH_s\|_F^2 + \lambda \|H_s\|_1).$$

Notations

- Fix $\lambda > 0$ and define the following the **quadratic loss function**

$$\ell(X, W) = \inf_{H \in \mathbb{R}^{r \times n}} \|X - WH\|_F^2 + \lambda \|H\|_1,$$

Define the **expected loss** and **empirical loss** functions

$$f(W) = \mathbb{E}_{X \sim \pi}[\ell(X, W)], \qquad f_t(W) = \frac{1}{t} \sum_{s=1}^{t} \ell(X_s, W)$$

- **ONMF algorithm**:

$$\textit{Upon arrival of } X_t\textit{:} \qquad \begin{cases} H_t = \operatorname{argmin}_{H \in \mathbb{R}_{\geq 0}^{r \times n}} \|X_t - W_{t-1} H\|_F^2 + \lambda \|H\|_1 \\ W_t = \operatorname{argmin}_{W \in \mathcal{C}} \hat{f}_t(W), \end{cases}$$

where

$$\hat{f}_t(W) = \frac{1}{t} \sum_{s=1}^{t} (\|X_s - WH_s\|_F^2 + \lambda \|H_s\|_1).$$

- WTS: W_t converges to the set of critical points of the expected loss f

Proposition

Let $(W_{t-1}, H_t)_{t\geq 1}$ be a solution to the ONMF algorithm. Then for each $t \geq 0$, the followings hold almost surely:

(i) $\hat{f}_{t+1}(W_{t+1}) - \hat{f}_t(W_t) \leq \frac{1}{t+1}\left(\ell(X_{t+1}, W_t) - f_t(W_t)\right)$.

(ii) $0 \leq \frac{1}{t+1}\left(\hat{f}_t(W_t) - f_t(W_t)\right) \leq \frac{1}{t+1}\left(\ell(X_{t+1}, W_t) - f_t(W_t)\right) + \hat{f}_t(W_t) - \hat{f}_{t+1}(W_{t+1})$.

- Recall that $\hat{f}_t \geq f_t$ for all $t \geq 0$.

Proposition

Let $(W_{t-1}, H_t)_{t\geq 1}$ be a solution to the ONMF algorithm. Then for each $t \geq 0$, the followings hold almost surely:

(i) $\hat{f}_{t+1}(W_{t+1}) - \hat{f}_t(W_t) \leq \frac{1}{t+1}\left(\ell(X_{t+1}, W_t) - f_t(W_t)\right)$.

(ii) $0 \leq \frac{1}{t+1}\left(\hat{f}_t(W_t) - f_t(W_t)\right) \leq \frac{1}{t+1}\left(\ell(X_{t+1}, W_t) - f_t(W_t)\right) + \hat{f}_t(W_t) - \hat{f}_{t+1}(W_{t+1})$.

- Recall that $\hat{f}_t \geq f_t$ for all $t \geq 0$.
- Also note that $\hat{f}_{t+1}(W_t) = \frac{1}{t+1}\left(t\hat{f}_t(W_t) + \ell(X_{t+1}, W_t)\right)$.

Proposition

Let $(W_{t-1}, H_t)_{t\geq 1}$ be a solution to the ONMF algorithm. Then for each $t \geq 0$, the followings hold almost surely:

(i) $\hat{f}_{t+1}(W_{t+1}) - \hat{f}_t(W_t) \leq \frac{1}{t+1}\left(\ell(X_{t+1}, W_t) - f_t(W_t)\right)$.

(ii) $0 \leq \frac{1}{t+1}\left(\hat{f}_t(W_t) - f_t(W_t)\right) \leq \frac{1}{t+1}\left(\ell(X_{t+1}, W_t) - f_t(W_t)\right) + \hat{f}_t(W_t) - \hat{f}_{t+1}(W_{t+1})$.

- Recall that $\hat{f}_t \geq f_t$ for all $t \geq 0$.
- Also note that $\hat{f}_{t+1}(W_t) = \frac{1}{t+1}\left(t\hat{f}_t(W_t) + \ell(X_{t+1}, W_t)\right)$.
- $$\begin{aligned}\hat{f}_{t+1}(W_{t+1}) - \hat{f}_t(W_t) &= \hat{f}_{t+1}(W_{t+1}) - \hat{f}_{t+1}(W_t) + \hat{f}_{t+1}(W_t) - \hat{f}_t(W_t) \\ &= \left[\hat{f}_{t+1}(W_{t+1}) - \hat{f}_{t+1}(W_t)\right] + \frac{1}{t+1}\left(\ell(X_{t+1}, W_t) - f_t(W_t)\right) + \left[\frac{f_t(W_t) - \hat{f}_t(W_t)}{t+1}\right] \\ &\leq \frac{1}{t+1}\left(\ell(X_{t+1}, W_t) - f_t(W_t)\right).\end{aligned}$$

  This shows (i). Using the second equality above and the fact that $\hat{f}_{t+1}(W_{t+1}) \leq \hat{f}_{t+1}(W_t)$, this also shows (ii).

Proposition

Let $(W_{t-1}, H_t)_{t\geq 1}$ be a solution to the ONMF algorithm. Then for each $t \geq 0$, the followings hold almost surely:

(i) $\hat{f}_{t+1}(W_{t+1}) - \hat{f}_t(W_t) \leq \frac{1}{t+1}\left(\ell(X_{t+1}, W_t) - f_t(W_t)\right)$.

(ii) $0 \leq \frac{1}{t+1}\left(\hat{f}_t(W_t) - f_t(W_t)\right) \leq \frac{1}{t+1}\left(\ell(X_{t+1}, W_t) - f_t(W_t)\right) + \hat{f}_t(W_t) - \hat{f}_{t+1}(W_{t+1})$.

Sketch of main argument:

Proposition

Let $(W_{t-1}, H_t)_{t\geq 1}$ be a solution to the ONMF algorithm. Then for each $t \geq 0$, the followings hold almost surely:

(i) $\hat{f}_{t+1}(W_{t+1}) - \hat{f}_t(W_t) \leq \frac{1}{t+1}\left(\ell(X_{t+1}, W_t) - f_t(W_t)\right)$.

(ii) $0 \leq \frac{1}{t+1}\left(\hat{f}_t(W_t) - f_t(W_t)\right) \leq \frac{1}{t+1}\left(\ell(X_{t+1}, W_t) - f_t(W_t)\right) + \hat{f}_t(W_t) - \hat{f}_{t+1}(W_{t+1})$.

Sketch of main argument:

- By bounding the sum of $\frac{1}{t+1}\left(\ell(X_{t+1}, W_t) - f_t(W_t)\right)$ in expectation, (i) will show that $\mathbb{E}[\hat{f}_t(W_t)]$ converges.

Proposition

Let $(W_{t-1}, H_t)_{t\geq 1}$ be a solution to the ONMF algorithm. Then for each $t \geq 0$, the followings hold almost surely:

(i) $\hat{f}_{t+1}(W_{t+1}) - \hat{f}_t(W_t) \leq \frac{1}{t+1}\left(\ell(X_{t+1}, W_t) - f_t(W_t)\right)$.

(ii) $0 \leq \frac{1}{t+1}\left(\hat{f}_t(W_t) - f_t(W_t)\right) \leq \frac{1}{t+1}\left(\ell(X_{t+1}, W_t) - f_t(W_t)\right) + \hat{f}_t(W_t) - \hat{f}_{t+1}(W_{t+1})$.

Sketch of main argument:

- By bounding the sum of $\frac{1}{t+1}\left(\ell(X_{t+1}, W_t) - f_t(W_t)\right)$ in expectation, (i) will show that $\mathbb{E}[\hat{f}_t(W_t)]$ converges.
- Then (ii) will show that $\hat{f}_t(W_t) - f_t(W_t) \to 0$ as $t \to \infty$.

Proposition

Let $(W_{t-1}, H_t)_{t\geq 1}$ be a solution to the ONMF algorithm. Then for each $t \geq 0$, the followings hold almost surely:

(i) $\hat{f}_{t+1}(W_{t+1}) - \hat{f}_t(W_t) \leq \frac{1}{t+1}\left(\ell(X_{t+1}, W_t) - f_t(W_t)\right)$.

(ii) $0 \leq \frac{1}{t+1}\left(\hat{f}_t(W_t) - f_t(W_t)\right) \leq \frac{1}{t+1}\left(\ell(X_{t+1}, W_t) - f_t(W_t)\right) + \hat{f}_t(W_t) - \hat{f}_{t+1}(W_{t+1})$.

Sketch of main argument:

- By bounding the sum of $\frac{1}{t+1}\left(\ell(X_{t+1}, W_t) - f_t(W_t)\right)$ in expectation, (i) will show that $\mathbb{E}[\hat{f}_t(W_t)]$ converges.
- Then (ii) will show that $\hat{f}_t(W_t) - f_t(W_t) \to 0$ as $t \to \infty$.
- Since $f_t \leq \hat{f}_t$ and every limit point of the sequence $(W_t)_{t\geq 0}$ is a critical point of $\hat{f}_\infty$, this will show that every limit point of $(W_t)_{t\geq 0}$ is also a critical point of f.

Key estimate in the i.i.d. case

- Suppose data matrices X_t are i.i.d. and let $\mathcal{F}_t$ denote the information up to time t. Then

$$\mathbb{E}\left[\ell(X_{t+1}, W_t) - f_t(W_t) \,\middle|\, \mathcal{F}_t\right] = \mathbb{E}_{X\sim\pi}[\ell(X, W_t)] - f_t(W_t) \tag{1}$$

$$= f(W_t) - f_t(W_t) \leq \|f - f_t\|_\infty \tag{2}$$

Key estimate in the i.i.d. case

- Suppose data matrices X_t are i.i.d. and let $\mathcal{F}_t$ denote the information up to time t. Then

$$\mathbb{E}\left[\ell(X_{t+1}, W_t) - f_t(W_t) \,\middle|\, \mathcal{F}_t\right] = \mathbb{E}_{X\sim\pi}[\ell(X, W_t)] - f_t(W_t) \tag{1}$$

$$= f(W_t) - f_t(W_t) \leq \|f - f_t\|_\infty \tag{2}$$

- $f(W) - f_t(W) \to 0$ by SLLN

Key estimate in the i.i.d. case

- Suppose data matrices X_t are i.i.d. and let $\mathcal{F}_t$ denote the information up to time t. Then

$$\mathbb{E}\left[\ell(X_{t+1}, W_t) - f_t(W_t) \,\middle|\, \mathcal{F}_t\right] = \mathbb{E}_{X\sim\pi}[\ell(X, W_t)] - f_t(W_t) \tag{1}$$

$$= f(W_t) - f_t(W_t) \le \|f - f_t\|_\infty \tag{2}$$

- $f(W) - f_t(W) \to 0$ by SLLN
- $\|f - f_t\|_\infty \to 0$ Glivenko-Cantelli Thm. ($W \in$ Compact set)

Key estimate in the i.i.d. case

- Suppose data matrices X_t are i.i.d. and let $\mathcal{F}_t$ denote the information up to time t. Then

$$\mathbb{E}\left[\ell(X_{t+1}, W_t) - f_t(W_t) \,\middle|\, \mathcal{F}_t\right] = \mathbb{E}_{X\sim\pi}[\ell(X, W_t)] - f_t(W_t) \tag{1}$$

$$= f(W_t) - f_t(W_t) \le \|f - f_t\|_\infty \tag{2}$$

- $f(W) - f_t(W) \to 0$ by SLLN
- $\|f - f_t\|_\infty \to 0$ Glivenko-Cantelli Thm. ($W \in$ Compact set)
- $\mathbb{E}[t^{1/2}\|f - f_t\|_\infty] = O(1)$ by uniform functional CLT

Key estimate in the i.i.d. case

- Suppose data matrices X_t are i.i.d. and let $\mathcal{F}_t$ denote the information up to time t. Then

$$\mathbb{E}\left[\ell(X_{t+1}, W_t) - f_t(W_t) \,\middle|\, \mathcal{F}_t\right] = \mathbb{E}_{X\sim\pi}[\ell(X, W_t)] - f_t(W_t) \tag{1}$$

$$= f(W_t) - f_t(W_t) \le \|f - f_t\|_\infty \tag{2}$$

- $f(W) - f_t(W) \to 0$ by SLLN
- $\|f - f_t\|_\infty \to 0$ Glivenko-Cantelli Thm. ($W \in$ Compact set)
- $\mathbb{E}[t^{1/2}\|f - f_t\|_\infty] = O(1)$ by uniform functional CLT
- Averaging over $\mathcal{F}_t$, this gives

$$\mathbb{E}\left[\left(\mathbb{E}\left[\frac{\ell(X_{t+1}, W_t) - f_t(W_t)}{t+1} \,\middle|\, \mathcal{F}_t\right]\right)^+\right] \le t^{-3/2}\mathbb{E}[t^{1/2}\|f - f_t\|_\infty] \tag{3}$$

$$= O(t^{-3/2}). \tag{4}$$

Key estimate in the Markovian case

- If $(X_t)_{t\geq 0}$ is Markovian, then $\mathbb{E}[\ell(X_{t+1}, W) \,|\, \mathcal{F}_t]$ could be very different from $f(W) = \mathbb{E}_{X\sim\pi}[\ell(X, W)]$.

Key estimate in the Markovian case

- If $(X_t)_{t\geq 0}$ is Markovian, then $\mathbb{E}[\ell(X_{t+1}, W) \,|\, \mathcal{F}_t]$ could be very different from $f(W) = \mathbb{E}_{X\sim\pi}[\ell(X, W)]$.
- Instead, **condition on a distant past** $\mathcal{F}_{t-N}$ and see how much the chain mixes to the stationary distribution during $[t-N, t]$.

Key estimate in the Markovian case

- If $(X_t)_{t\geq 0}$ is Markovian, then $\mathbb{E}[\ell(X_{t+1}, W) \,|\, \mathcal{F}_t]$ could be very different from $f(W) = \mathbb{E}_{X\sim\pi}[\ell(X, W)]$.
- Instead, **condition on a distant past** $\mathcal{F}_{t-N}$ and see how much the chain mixes to the stationary distribution during $[t-N, t]$.

$$\begin{aligned}
\mathbb{E}\left[\ell(X_{t+1}, W) \,\middle|\, \mathcal{F}_{t-N}\right] &= \sum_{\mathbf{x}'\in\Omega} \ell(\mathbf{x}', W)\, P^{N+1}(\mathbf{x}, \mathbf{x}') \\
&= \sum_{\mathbf{x}'\in\Omega} \ell(\mathbf{x}', W)\, \pi(\mathbf{x}') + \sum_{\mathbf{x}'\in\Omega} \ell(\mathbf{x}', W)(P^{N+1}(\mathbf{x}, \mathbf{x}') - \pi(\mathbf{x}')) \\
&\leq \sum_{\mathbf{x}'\in\Omega} \ell(\mathbf{x}', W)\, \pi(\mathbf{x}') + 2\|\ell(\cdot, W)\|_\infty \|P^{N+1}(\mathbf{x}, \cdot) - \pi\|_{TV} \\
&= f(W) + 2\|\ell(\cdot, W)\|_\infty \|P^{N+1}(\mathbf{x}, \cdot) - \pi\|_{TV}.
\end{aligned}$$

Proposition

Suppose (A1)'-(A2) and (M2). Fix $W \in \mathcal{C}$. Then for each $t \geq 0$ and $0 \leq N < t$, conditional on the information $\mathcal{F}_{t-N}$ up to time $t - N$,

$$\left| \mathbb{E}\left[\ell(X_{t+1}, W) - f_t(W) \,\middle|\, \mathcal{F}_{t-N} \right] \right| \leq |f(W) - f_{t-N}(W)| \tag{5}$$

$$+ \frac{N}{t} \left(f_{t-N}(W) + \|\ell(\cdot, W)\|_\infty \right) \tag{6}$$

$$+ 2\|\ell(\cdot, W)\|_\infty \sup_{\mathbf{x} \in \Omega} \|P^{N+1}(\mathbf{x}, \cdot) - \pi\|_{TV}. \tag{7}$$

Lemma

Suppose (A1)'-(A2) and (M1) hold.

(i) *Let $(a_t)_{t\geq 0}$ be a sequence of non-decreasing non-negative integers such that $a_t = o(t)$. Then there exists absolute constants $C_1, C_2, C_3 > 0$ such that for all sufficiently large $t \geq 0$,*

$$\mathbb{E}\left[\left|\mathbb{E}\left[\frac{\ell(X_{t+1}, W_t) - f_t(W_t)}{t+1} \,\middle|\, \mathcal{F}_{t-a_t}\right]\right|\right] \leq \frac{C_1}{t^{3/2}} + \frac{C_2}{t^2} a_t + \frac{C_3}{t} \sup_{\mathbf{x}\in\Omega} \|P^{a_t+1}(\mathbf{x}, \cdot) - \pi\|_{TV}.$$

Lemma

Suppose (A1)'-(A2) and (M1) hold.

(i) *Let $(a_t)_{t\geq 0}$ be a sequence of non-decreasing non-negative integers such that $a_t = o(t)$. Then there exists absolute constants $C_1, C_2, C_3 > 0$ such that for all sufficiently large $t \geq 0$,*

$$\mathbb{E}\left[\left\|\mathbb{E}\left[\frac{\ell(X_{t+1}, W_t) - f_t(W_t)}{t+1} \,\middle|\, \mathcal{F}_{t-a_t}\right]\right\|\right] \leq \frac{C_1}{t^{3/2}} + \frac{C_2}{t^2} a_t + \frac{C_3}{t} \sup_{\mathbf{x}\in\Omega} \|P^{a_t+1}(\mathbf{x}, \cdot) - \pi\|_{TV}.$$

(ii) *Further assume that (M2) holds. Then we have*

$$\sum_{t=0}^{\infty} \left(\mathbb{E}\left[\hat{f}_{t+1}(W_{t+1}) - \hat{f}_t(W_t)\right]\right)^+ \leq \sum_{t=0}^{\infty} \left|\mathbb{E}\left[\frac{\ell(X_{t+1}, W_t) - f_t(W_t)}{t+1}\right]\right| < \infty.$$

Proof of Thm 1

- Since $\sum_{t=0}^{\infty} \left(\mathbb{E}\left[\hat{f}_{t+1}(W_{t+1}) - \hat{f}_t(W_t) \right] \right)^+ < \infty$, $\mathbb{E}[\hat{f}_t(W_t)]$ converges.

Proof of Thm 1

- Since $\sum_{t=0}^{\infty} \left(\mathbb{E}\left[\hat{f}_{t+1}(W_{t+1}) - \hat{f}_t(W_t) \right] \right)^+ < \infty$, $\mathbb{E}[\hat{f}_t(W_t)]$ converges.
- By the earlier inequalities and estimates,

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=0}^{\infty} \frac{\hat{f}_t(W_t) - f_t(W_t)}{t+1} \right] &= \sum_{t=0}^{\infty} \frac{\mathbb{E}[\hat{f}_t(W_t)] - \mathbb{E}[f_t(W_t)]}{t+1} \\
&\leq \sum_{t=0}^{\infty} \left| \frac{\mathbb{E}[\ell(X_{t+1}, W_t)] - f_t(W_t)]}{t+1} \right| \\
&\quad - \sum_{t=0}^{\infty} \left(\mathbb{E}[\hat{f}_{t+1}(W_{t+1})] - \mathbb{E}[\hat{f}_t(W_t)] \right) \\
&< \infty.
\end{aligned}$$

Proof of Thm 1

- Since $\sum_{t=0}^{\infty} \left(\mathbb{E}\left[\hat{f}_{t+1}(W_{t+1}) - \hat{f}_t(W_t) \right] \right)^+ < \infty$, $\mathbb{E}[\hat{f}_t(W_t)]$ converges.
- By the earlier inequalities and estimates,

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=0}^{\infty} \frac{\hat{f}_t(W_t) - f_t(W_t)}{t+1} \right] &= \sum_{t=0}^{\infty} \frac{\mathbb{E}[\hat{f}_t(W_t)] - \mathbb{E}[f_t(W_t)]}{t+1} \\
&\leq \sum_{t=0}^{\infty} \left| \frac{\mathbb{E}[\ell(X_{t+1}, W_t)] - f_t(W_t)]}{t+1} \right| \\
&\quad - \sum_{t=0}^{\infty} \left(\mathbb{E}[\hat{f}_{t+1}(W_{t+1})] - \mathbb{E}[\hat{f}_t(W_t)] \right) \\
&< \infty.
\end{aligned}$$

- Hence $\sum_{t=0}^{\infty} \frac{\hat{f}_t(W_t) - f_t(W_t)}{t+1} < \infty$ a.s. This implies $\hat{f}_t(W_t) - f_t(W_t) \to 0$ a.s. as $t \to \infty$.

Proof of Thm 1

- Since $\sum_{t=0}^{\infty} \left(\mathbb{E}\left[\hat{f}_{t+1}(W_{t+1}) - \hat{f}_t(W_t) \right] \right)^+ < \infty$, $\mathbb{E}[\hat{f}_t(W_t)]$ converges.
- By the earlier inequalities and estimates,

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=0}^{\infty} \frac{\hat{f}_t(W_t) - f_t(W_t)}{t+1} \right] &= \sum_{t=0}^{\infty} \frac{\mathbb{E}[\hat{f}_t(W_t)] - \mathbb{E}[f_t(W_t)]}{t+1} \\
&\leq \sum_{t=0}^{\infty} \left| \frac{\mathbb{E}[\ell(X_{t+1}, W_t)] - f_t(W_t)]}{t+1} \right| \\
&\quad - \sum_{t=0}^{\infty} \left(\mathbb{E}[\hat{f}_{t+1}(W_{t+1})] - \mathbb{E}[\hat{f}_t(W_t)] \right) \\
&< \infty.
\end{aligned}$$

- Hence $\sum_{t=0}^{\infty} \frac{\hat{f}_t(W_t) - f_t(W_t)}{t+1} < \infty$ a.s. This implies $\hat{f}_t(W_t) - f_t(W_t) \to 0$ a.s. as $t \to \infty$.
- Since $\hat{f}_t \geq f_t$ and $\hat{f}_t(W_t) - f_t(W_t) \to 0$ a.s., argue that W_t has to converge to the set of critical points of $f = \lim_{t\to\infty} f_t$.