Matrix-valued Time Series in High Dimension

Nayel Bettache
Supervised by Cristina Butucea

July 5, 2024

Table of Contents

Table of Contents

Model

Observe repeatedly and independently n samples $(X_1, \ldots, X_n)$ of a $\mathbb{R}$-valued time series of length p.

Model

Observe repeatedly and independently n samples $(X_1, \ldots, X_n)$ of a $\mathbb{R}$-valued time series of length p.

- Given a long stationary time series extract blocs of length p sufficiently far apart to assume independence.

Model

Observe repeatedly and independently n samples $(X_1, \ldots, X_n)$ of a $\mathbb{R}$-valued time series of length p.

- Given a long stationary time series extract blocs of length p sufficiently far apart to assume independence.

Consider X a generic $p-$dimensional gaussian vector such that $X \sim \mathcal{N}_p(0, \Sigma)$.

Model

Observe repeatedly and independently n samples $(X_1, \dots, X_n)$ of a $\mathbb{R}$-valued time series of length p.

- Given a long stationary time series extract blocs of length p sufficiently far apart to assume independence.

Consider X a generic $p-$dimensional gaussian vector such that $X \sim \mathcal{N}_p(0, \Sigma)$.

- $\Sigma \in \mathcal{S}_p^{++}$ has a Toeplitz structure .

$$\Sigma := \begin{pmatrix} \sigma_0 & \sigma_1 & \sigma_2 & \sigma_3 & \sigma_4 & \cdots & \sigma_{p-1} \\ \sigma_1 & \sigma_0 & \sigma_1 & \sigma_2 & \sigma_3 & \cdots & \sigma_{p-2} \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \sigma_1 & \sigma_0 & \sigma_1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \sigma_{p-2} & \cdots & \sigma_3 & \sigma_2 & \sigma_1 & \sigma_0 & \sigma_1 \\ \sigma_{p-1} & \cdots & \sigma_4 & \sigma_3 & \sigma_2 & \sigma_1 & \sigma_0 \end{pmatrix}$$

Objective

- The objective is to test $H_0 : \Sigma = I_p$ against a set of one-sided $\mathcal{F}_+$ or two-sided $\mathcal{F}$ sparse alternatives and provide non asymptotic upper bounds of the maximal testing risk.

Objective

- The objective is to test $H_0 : \Sigma = I_p$ against a set of one-sided $\mathcal{F}_+$ or two-sided $\mathcal{F}$ sparse alternatives and provide non asymptotic upper bounds of the maximal testing risk.
- The test procedure needs to be very sensitive to:

Objective

- The objective is to test $H_0 : \Sigma = I_p$ against a set of one-sided $\mathcal{F}_+$ or two-sided $\mathcal{F}$ sparse alternatives and provide non asymptotic upper bounds of the maximal testing risk.
- The test procedure needs to be very sensitive to:
  1. The moderately sparse case: a relatively large number of very small but significant covariance values.

Objective

- The objective is to test $H_0 : \Sigma = I_p$ against a set of one-sided $\mathcal{F}_+$ or two-sided $\mathcal{F}$ sparse alternatives and provide non asymptotic upper bounds of the maximal testing risk.
- The test procedure needs to be very sensitive to:
    1. The moderately sparse case: a relatively large number of very small but significant covariance values.
    2. The highly sparse case: a very small number of significant covariance values.

Objective

- The objective is to test $H_0 : \Sigma = I_p$ against a set of one-sided $\mathcal{F}_+$ or two-sided $\mathcal{F}$ sparse alternatives and provide non asymptotic upper bounds of the maximal testing risk.
- The test procedure needs to be very sensitive to:
  1. The moderately sparse case: a relatively large number of very small but significant covariance values.
  2. The highly sparse case: a very small number of significant covariance values.
- This is analogous to but more general than the detection of sparse Gaussian means: Ingster 2001, 2002 (Math. Methods Statist.) and Donoho, Jin 2004 (Ann. Statist.)

Objective

- The objective is to test $H_0 : \Sigma = I_p$ against a set of one-sided $\mathcal{F}_+$ or two-sided $\mathcal{F}$ sparse alternatives and provide non asymptotic upper bounds of the maximal testing risk.
- The test procedure needs to be very sensitive to:
  1. The moderately sparse case: a relatively large number of very small but significant covariance values.
  2. The highly sparse case: a very small number of significant covariance values.
- This is analogous to but more general than the detection of sparse Gaussian means: Ingster 2001, 2002 (Math. Methods Statist.) and Donoho, Jin 2004 (Ann. Statist.)
- We also develop a procedure that selects non-null correlation coefficients.

Objective

- The objective is to test $H_0 : \Sigma = I_p$ against a set of one-sided $\mathcal{F}_+$ or two-sided $\mathcal{F}$ sparse alternatives and provide non asymptotic upper bounds of the maximal testing risk.
- The test procedure needs to be very sensitive to:
  1. The moderately sparse case: a relatively large number of very small but significant covariance values.
  2. The highly sparse case: a very small number of significant covariance values.
- This is analogous to but more general than the detection of sparse Gaussian means: Ingster 2001, 2002 (Math. Methods Statist.) and Donoho, Jin 2004 (Ann. Statist.)
- We also develop a procedure that selects non-null correlation coefficients.
- Numerical results illustrate the excellent behaviour of the test procedures and the support selector.

Testing problems

- The one-sided test problem is

$$H_0 : \Sigma = I_p, \quad \text{vs. } H_1 : \Sigma \in \mathcal{F}_+(s, S, \sigma),$$

where

$$\mathcal{F}_+(s, S, \sigma) = \Big\{ \Sigma \in \mathcal{S}_p^{++} \cap \mathcal{T}_p \text{ and } \exists \mathcal{C} \subseteq \{1, \dots, S\}, \\ |\mathcal{C}| = s, \ \forall j \in \{1, p-1\}, \begin{matrix} \sigma_j \geq \sigma > 0, & j \in \mathcal{C}, \\ \sigma_j = 0, & j \notin \mathcal{C} \end{matrix} \Big\}$$

Testing problems

- The one-sided test problem is

$$H_0 : \Sigma = I_p, \quad \text{vs. } H_1 : \Sigma \in \mathcal{F}_+(s, S, \sigma),$$

where

$$\mathcal{F}_+(s, S, \sigma) = \Big\{ \Sigma \in \mathcal{S}_p^{++} \cap \mathcal{T}_p \text{ and } \exists \mathcal{C} \subseteq \{1, \ldots, S\}, \\ |\mathcal{C}| = s, \ \forall j \in \{1, p-1\}, \ \begin{matrix} \sigma_j \geq \sigma > 0, & j \in \mathcal{C}, \\ \sigma_j = 0, & j \notin \mathcal{C} \end{matrix} \Big\}$$

- The two-sided test problem is

$$H_0 : \Sigma = I_p, \quad \text{vs. } H_1 : \Sigma \in \mathcal{F}(s, S, \sigma),$$

where $\mathcal{F}(s, S, \sigma)$ is defined similarly as $\mathcal{F}_+(s, S, \sigma)$ by considering the absolute values of the covariance elements.

Moderately sparse case in the one-sided alternative

- When the alternative is $\mathcal{F}_+(s, S, \sigma)$, we consider for some threshold $t_{n,p}^{MS+}$ the test procedure

$$\Delta_n^{MS+} = \mathbb{1}\left(Sum_{\{1:S\}}^+(\Sigma_n - I_p) \geq t_{n,p}^{MS+}\right),$$

where for an arbitrary set $\mathcal{C} \subseteq \{1, \ldots, S\}$,

$$Sum_{\mathcal{C}}^+(\Sigma_n) := \sum_{j \in \mathcal{C}} \mathrm{Tr}(A_j \Sigma_n) = \sum_{j \in \mathcal{C}} \hat{\sigma}_j.$$

Moderately sparse case in the one-sided alternative

- When the alternative is $\mathcal{F}_+(s, S, \sigma)$, we consider for some threshold $t_{n,p}^{MS+}$ the test procedure

$$\Delta_n^{MS+} = \mathbb{1}\left(Sum^+_{\{1:S\}}(\Sigma_n - I_p) \geq t_{n,p}^{MS+}\right),$$

where for an arbitrary set $\mathcal{C} \subseteq \{1, \ldots, S\}$,

$$Sum^+_{\mathcal{C}}(\Sigma_n) := \sum_{j \in \mathcal{C}} \mathrm{Tr}(A_j \Sigma_n) = \sum_{j \in \mathcal{C}} \hat{\sigma}_j.$$

- When the alternative is $\mathcal{F}(s, S, \sigma)$, we consider for some threshold $t_{n,p}^{MS}$ a test Δ_n^{MS} that sums the absolute values of the first S covariance elements of $\Sigma_n - I_p$ and compare it to $t_{n,p}^{MS}$.

Moderately sparse case in the one-sided alternative

- When the alternative is $\mathcal{F}_+(s, S, \sigma)$, we consider for some threshold $t_{n,p}^{MS+}$ the test procedure

$$\Delta_n^{MS+} = \mathbb{1}\left(Sum_{\{1:S\}}^{+}(\Sigma_n - I_p) \geq t_{n,p}^{MS+} \right),$$

where for an arbitrary set $\mathcal{C} \subseteq \{1, \dots, S\}$,

$$Sum_{\mathcal{C}}^{+}(\Sigma_n) := \sum_{j \in \mathcal{C}} \mathrm{Tr}(A_j \Sigma_n) = \sum_{j \in \mathcal{C}} \hat{\sigma}_j.$$

- When the alternative is $\mathcal{F}(s, S, \sigma)$, we consider for some threshold $t_{n,p}^{MS}$ a test Δ_n^{MS} that sums the absolute values of the first S covariance elements of $\Sigma_n - I_p$ and compare it to $t_{n,p}^{MS}$.

Theorem (B., Butucea, Sorba 2022)

For $u > 0$, consider $t_{n,p}^{MS+} = \max\left\{ \sqrt{\frac{u \cdot S}{n(p-S)}}, \frac{2u \cdot S}{n(p-S)} \right\}$. Then $R(\Delta_n^{MS+}, \mathcal{F}_+) \leq 2\exp\left(-\frac{u}{4}\right)$ provided that $\sigma \geq \frac{2(s+1)}{s} t_{n,p}^{MS+}$.

Highly sparse case in the one-sided alternative

- When the alternative is $\mathcal{F}_+(s, S, \sigma)$, we consider for some threshold $t_{n,p}^{HS+}$ the test procedure

$$\Delta_n^{HS+} = \max_{\mathcal{C} \subseteq \{1,\dots,S\}, \#\mathcal{C}=s} \mathbb{1}\left(Sum_{\mathcal{C}}^{+}(\Sigma_n - I_p) \geq t_{n,p}^{HS+} \right).$$

Highly sparse case in the one-sided alternative

- When the alternative is $\mathcal{F}_+(s, S, \sigma)$, we consider for some threshold $t_{n,p}^{HS+}$ the test procedure

$$\Delta_n^{HS+} = \max_{\mathcal{C} \subseteq \{1,\ldots,S\}, \#\mathcal{C}=s} \mathbb{1}\left(Sum_{\mathcal{C}}^{+}(\Sigma_n - I_p) \geq t_{n,p}^{HS+}\right).$$

- When the alternative is $\mathcal{F}(s, S, \sigma)$, we examine the same procedure by considering the absolute values of the empirical covariance elements.

Highly sparse case in the one-sided alternative

- When the alternative is $\mathcal{F}_+(s, S, \sigma)$, we consider for some threshold $t_{n,p}^{HS+}$ the test procedure

$$\Delta_n^{HS+} = \max_{\mathcal{C} \subseteq \{1,\dots,S\}, \#\mathcal{C}=s} \mathbb{1}\left(Sum_{\mathcal{C}}^{+}(\Sigma_n - I_p) \geq t_{n,p}^{HS+}\right).$$

- When the alternative is $\mathcal{F}(s, S, \sigma)$, we examine the same procedure by considering the absolute values of the empirical covariance elements.

Theorem (B., Butucea, Sorba 2022)

For $u > 1$, consider $t_{n,p}^{HS+} = \max\left\{\sqrt{\frac{4u\cdot s\log\binom{S}{s}}{n(p-S)}}, \frac{8u\cdot s\log\binom{S}{s}}{n(p-S)}\right\}$. *Then*
$R(\Delta_n^{HS+}, \mathcal{F}^+) \leq \exp\left(-(u-1)\log\binom{S}{s}\right) + \exp\left(-\frac{u}{4}\right)$ *provided that*
$\sigma \geq \frac{1}{s}\left(t_{n,p}^{HS+} + (2s+1)\max\left\{\sqrt{\frac{u\cdot s}{n(p-S)}}, \frac{2u\cdot s}{n(p-S)}\right\}\right)$.

Table of Contents

Two-sided matrix regression

Two-sided matrix regression (2MR):

Two-sided matrix regression

Two-sided matrix regression (2MR):

- **2MR**: Consider an observed target matrix $Y \in \mathbb{R}^{n \times p}$ and an observed design matrix $X \in \mathbb{R}^{m \times q}$ following:

$$Y = A^* X B^* + E,$$

where $(A^*, B^*) \in \mathbb{R}^{n \times m} \times \mathbb{R}^{q \times p}$ are low-rank matrix parameters.

Two-sided matrix regression

Two-sided matrix regression (2MR):

- **2MR**: Consider an observed target matrix $Y \in \mathbb{R}^{n \times p}$ and an observed design matrix $X \in \mathbb{R}^{m \times q}$ following:

$$Y = A^* X B^* + E,$$

where $(A^*, B^*) \in \mathbb{R}^{n \times m} \times \mathbb{R}^{q \times p}$ are low-rank matrix parameters.
The noise matrix E is assumed to have independent centered $\sigma-$sub-Gaussian entries.

Two-sided matrix regression

Two-sided matrix regression (2MR):

- **2MR**: Consider an observed target matrix $Y \in \mathbb{R}^{n \times p}$ and an observed design matrix $X \in \mathbb{R}^{m \times q}$ following:

$$Y = A^* X B^* + E,$$

  where $(A^*, B^*) \in \mathbb{R}^{n \times m} \times \mathbb{R}^{q \times p}$ are low-rank matrix parameters. The noise matrix E is assumed to have independent centered $\sigma-$sub-Gaussian entries.
- **Objective**: Learning the signal $A^* X B^*$.

Two-sided matrix regression

Two-sided matrix regression (2MR):

- **2MR**: Consider an observed target matrix $Y \in \mathbb{R}^{n \times p}$ and an observed design matrix $X \in \mathbb{R}^{m \times q}$ following:

$$Y = A^* X B^* + E,$$

  where $(A^*, B^*) \in \mathbb{R}^{n \times m} \times \mathbb{R}^{q \times p}$ are low-rank matrix parameters. The noise matrix E is assumed to have independent centered $\sigma-$sub-Gaussian entries.

- **Objective**: Learning the signal $A^* X B^*$.
  The problem is not convex !

Two-sided matrix regression

Two-sided matrix regression (2MR):

- **2MR**: Consider an observed target matrix $Y \in \mathbb{R}^{n\times p}$ and an observed design matrix $X \in \mathbb{R}^{m\times q}$ following:

$$Y = A^* X B^* + E,$$

  where $(A^*, B^*) \in \mathbb{R}^{n\times m} \times \mathbb{R}^{q\times p}$ are low-rank matrix parameters.
  The noise matrix E is assumed to have independent centered $\sigma-$sub-Gaussian entries.
- **Objective**: Learning the signal $A^* X B^*$.
  The problem is not convex !
  Without additional assumptions, the problem is not identifiable.

Two-sided matrix regression

Two-sided matrix regression (2MR):

- **2MR**: Consider an observed target matrix $Y \in \mathbb{R}^{n \times p}$ and an observed design matrix $X \in \mathbb{R}^{m \times q}$ following:

$$Y = A^* X B^* + E,$$

  where $(A^*, B^*) \in \mathbb{R}^{n \times m} \times \mathbb{R}^{q \times p}$ are low-rank matrix parameters.
  The noise matrix E is assumed to have independent centered $\sigma-$sub-Gaussian entries.

- **Objective**: Learning the signal $A^* X B^*$.
  The problem is not convex !
  Without additional assumptions, the problem is not identifiable.
  Different structured matrix estimation is studied in Klopp, Lu, Tsybakov, Zhou 2019 (Bernoulli)

Beyond the limits of the MR

The two-sided matrix regression (2MR) extends the one-sided matrix regression (MR)

Beyond the limits of the MR

The two-sided matrix regression (2MR) extends the one-sided matrix regression (MR)

- **MR**: Consider an observed target matrix $Y \in \mathbb{R}^{n \times p}$ and an observed design matrix $X \in \mathbb{R}^{n \times q}$ following:

$$Y = XB^* + E,$$

  where $B^* \in \mathbb{R}^{q \times p}$.

Beyond the limits of the MR

The two-sided matrix regression (2MR) extends the one-sided matrix regression (MR)

- **MR**: Consider an observed target matrix $Y \in \mathbb{R}^{n \times p}$ and an observed design matrix $X \in \mathbb{R}^{n \times q}$ following:

$$Y = XB^* + E,$$

where $B^* \in \mathbb{R}^{q \times p}$.
Without any constraint on the structure of B^* (full rank), the MR is equivalent to performing p independent linear regressions.

Beyond the limits of the MR

The two-sided matrix regression (2MR) extends the one-sided matrix regression (MR)

- **MR**: Consider an observed target matrix $Y \in \mathbb{R}^{n\times p}$ and an observed design matrix $X \in \mathbb{R}^{n\times q}$ following:

$$Y = XB^* + E,$$

where $B^* \in \mathbb{R}^{q\times p}$.
Without any constraint on the structure of B^* (full rank), the MR is equivalent to performing p independent linear regressions.
It ignores the multivariate nature of the response, *i.e.* the possible correlations among columns of Y.

Beyond the limits of the MR

The two-sided matrix regression (2MR) extends the one-sided matrix regression (MR)

- **MR**: Consider an observed target matrix $Y \in \mathbb{R}^{n \times p}$ and an observed design matrix $X \in \mathbb{R}^{n \times q}$ following:

$$Y = XB^* + E,$$

where $B^* \in \mathbb{R}^{q \times p}$.
Without any constraint on the structure of B^* (full rank), the MR is equivalent to performing p independent linear regressions.
It ignores the multivariate nature of the response, *i.e.* the possible correlations among columns of Y.
Solution: impose a low-rank structure on B^*.

Beyond the limits of the MR

The two-sided matrix regression (2MR) extends the one-sided matrix regression (MR)

- **MR**: Consider an observed target matrix $Y \in \mathbb{R}^{n \times p}$ and an observed design matrix $X \in \mathbb{R}^{n \times q}$ following:

  $$Y = XB^* + E,$$

  where $B^* \in \mathbb{R}^{q \times p}$.
  Without any constraint on the structure of B^* (full rank), the MR is equivalent to performing p independent linear regressions.
  It ignores the multivariate nature of the response, *i.e.* the possible correlations among columns of Y.
  Solution: impose a low-rank structure on B^*.
  Bunea, She, Wegkamp 2011 (Ann. Statist.), Giraud 2011 (Electron. J. Statist.)

Beyond the limits of the MR

The two-sided matrix regression (2MR) extends the one-sided matrix regression (MR)

- **MR**: Consider an observed target matrix $Y \in \mathbb{R}^{n\times p}$ and an observed design matrix $X \in \mathbb{R}^{n\times q}$ following:

$$Y = XB^* + E,$$

  where $B^* \in \mathbb{R}^{q\times p}$.
  Without any constraint on the structure of B^* (full rank), the MR is equivalent to performing p independent linear regressions.
  It ignores the multivariate nature of the response, *i.e.* the possible correlations among columns of Y.
  Solution: impose a low-rank structure on B^*.
  Bunea, She, Wegkamp 2011 (Ann. Statist.), Giraud 2011 (Electron. J. Statist.)
- **Limits**: MR cannot handle possible correlations among the rows of Y.

Beyond the limits of the MR

The two-sided matrix regression (2MR) extends the one-sided matrix regression (MR)

- **MR**: Consider an observed target matrix $Y \in \mathbb{R}^{n \times p}$ and an observed design matrix $X \in \mathbb{R}^{n \times q}$ following:

  $$Y = XB^* + E,$$

  where $B^* \in \mathbb{R}^{q \times p}$.
  Without any constraint on the structure of B^* (full rank), the MR is equivalent to performing p independent linear regressions.
  It ignores the multivariate nature of the response, *i.e.* the possible correlations among columns of Y.
  Solution: impose a low-rank structure on B^*.
  Bunea, She, Wegkamp 2011 (Ann. Statist.), Giraud 2011 (Electron. J. Statist.)
- **Limits**: MR cannot handle possible correlations among the rows of Y.
  Need for another matrix parameter A^* that left multiplies the signal XB^*.

Minimization of the squared Frobenius norm

If $r := \operatorname{rank} A^* X B^*$ is given, it can be exploited.

Minimization of the squared Frobenius norm

If $r := \operatorname{rank} A^* X B^*$ is given, it can be exploited. Fix $r \in [n \wedge p \wedge r_X]$ where $r_X = \operatorname{rank} X$.

Minimization of the squared Frobenius norm

If $r := \operatorname{rank} A^* X B^*$ is given, it can be exploited. Fix $r \in [n \wedge p \wedge r_X]$ where $r_X = \operatorname{rank} X$.

- **Procedure**: Build r-dependent explicit predictors satisfying the non-convex constrained minimization problem:

$$(\hat{A}_r, \hat{B}_r) \in \underset{\substack{A,B: \\ \operatorname{rank} A \wedge \operatorname{rank} B \leq r}}{\arg\min} \|Y - AXB\|_F^2.$$

Minimization of the squared Frobenius norm

If $r := \text{rank}\, A^* X B^*$ is given, it can be exploited. Fix $r \in [n \wedge p \wedge r_X]$ where $r_X = \text{rank}\, X$.

- **Procedure**: Build r-dependent explicit predictors satisfying the non-convex constrained minimization problem:

$$(\hat{A}_r, \hat{B}_r) \in \underset{\substack{A,B: \\ \text{rank}\, A \wedge \text{rank}\, B \leq r}}{\arg\min} \|Y - AXB\|_F^2.$$

- Note: $\text{rank}\, A^* X B^* \leq \min(\text{rank}\, A^*, \text{rank}\, X, \text{rank}\, B^*)$.

Minimization of the squared Frobenius norm

If $r := \operatorname{rank} A^* X B^*$ is given, it can be exploited. Fix $r \in [n \wedge p \wedge r_X]$ where $r_X = \operatorname{rank} X$.

- **Procedure**: Build r-dependent explicit predictors satisfying the non-convex constrained minimization problem:

$$(\hat{A}_r, \hat{B}_r) \in \underset{\substack{A,B: \\ \operatorname{rank} A \wedge \operatorname{rank} B \leq r}}{\arg\min} \|Y - AXB\|_F^2.$$

- Note: $\operatorname{rank} A^* X B^* \leq \min(\operatorname{rank} A^*, \operatorname{rank} X, \operatorname{rank} B^*)$.
- Global idea: $Y \longrightarrow Y_r \longrightarrow \hat{A}_r X \hat{B}_r$.

Minimization of the squared Frobenius norm

If $r := \operatorname{rank} A^* X B^*$ is given, it can be exploited. Fix $r \in [n \wedge p \wedge r_X]$ where $r_X = \operatorname{rank} X$.

- **Procedure**: Build r-dependent explicit predictors satisfying the non-convex constrained minimization problem:

$$(\hat{A}_r, \hat{B}_r) \in \underset{\substack{A,B: \\ \operatorname{rank} A \wedge \operatorname{rank} B \leq r}}{\arg\min} \|Y - AXB\|_F^2.$$

- Note: $\operatorname{rank} A^* X B^* \leq \min(\operatorname{rank} A^*, \operatorname{rank} X, \operatorname{rank} B^*)$.
- Global idea: $Y \longrightarrow Y_r \longrightarrow \hat{A}_r X \hat{B}_r$.
- Identifiability: The predictors are not uniquely defined in this setting. Without further strong assumptions, we cannot hope to learn parameters from a non identifiable model.

Diagonal 2MR

The model can be re-written using the SVD of Y and X. This leads to the Diagonal 2MR (D2MR).

Diagonal 2MR

The model can be re-written using the SVD of Y and X. This leads to the Diagonal 2MR (D2MR).

$$Y = A^* X B^* + E$$

Diagonal 2MR

The model can be re-written using the SVD of Y and X. This leads to the Diagonal 2MR (D2MR).

$$Y = A^* X B^* + E$$

$$U_Y \Sigma_Y V_Y^\top = A^* U_X \Sigma_X V_X^\top B^* + E$$

Diagonal 2MR

The model can be re-written using the SVD of Y and X. This leads to the Diagonal 2MR (D2MR).

$$Y = A^* X B^* + E$$

$$U_Y \Sigma_Y V_Y^\top = A^* U_X \Sigma_X V_X^\top B^* + E$$

$$\Sigma_Y = U_Y^\top A^* U_X \Sigma_X V_X^\top B^* V_Y + U_Y^\top E V_Y$$

Diagonal 2MR

The model can be re-written using the SVD of Y and X. This leads to the Diagonal 2MR (D2MR).

$$Y = A^* X B^* + E$$

$$U_Y \Sigma_Y V_Y^\top = A^* U_X \Sigma_X V_X^\top B^* + E$$

$$\Sigma_Y = U_Y^\top A^* U_X \Sigma_X V_X^\top B^* V_Y + U_Y^\top E V_Y$$

$$\Sigma_Y = \left(U_Y^\top A^* U_X\right) \Sigma_X \left(V_X^\top B^* V_Y\right) + U_Y^\top E V_Y$$

Diagonal 2MR

The model can be re-written using the SVD of Y and X. This leads to the Diagonal 2MR (D2MR).

$$Y = A^* X B^* + E$$

$$U_Y \Sigma_Y V_Y^\top = A^* U_X \Sigma_X V_X^\top B^* + E$$

$$\Sigma_Y = U_Y^\top A^* U_X \Sigma_X V_X^\top B^* V_Y + U_Y^\top E V_Y$$

$$\Sigma_Y = \underbrace{\left(U_Y^\top A^* U_X\right)}_{A_0^*} \Sigma_X \underbrace{\left(V_X^\top B^* V_Y\right)}_{B_0^*} + \underbrace{U_Y^\top E V_Y}_{E_0}$$

Diagonal 2MR

The model can be re-written using the SVD of Y and X. This leads to the Diagonal 2MR (D2MR).

$$Y = A^* X B^* + E$$

$$U_Y \Sigma_Y V_Y^\top = A^* U_X \Sigma_X V_X^\top B^* + E$$

$$\Sigma_Y = U_Y^\top A^* U_X \Sigma_X V_X^\top B^* V_Y + U_Y^\top E V_Y$$

$$\Sigma_Y = \underbrace{\left(U_Y^\top A^* U_X\right)}_{A_0^*} \Sigma_X \underbrace{\left(V_X^\top B^* V_Y\right)}_{B_0^*} + \underbrace{U_Y^\top E V_Y}_{E_0}$$

$$\Sigma_Y = A_0^* \Sigma_X B_0^* + E_0$$

Diagonal 2MR

The model can be re-written using the SVD of Y and X. This leads to the Diagonal 2MR (D2MR).

$$Y = A^* X B^* + E$$

$$U_Y \Sigma_Y V_Y^\top = A^* U_X \Sigma_X V_X^\top B^* + E$$

$$\Sigma_Y = U_Y^\top A^* U_X \Sigma_X V_X^\top B^* V_Y + U_Y^\top E V_Y$$

$$\Sigma_Y = \underbrace{\left(U_Y^\top A^* U_X\right)}_{A_0^*} \Sigma_X \underbrace{\left(V_X^\top B^* V_Y\right)}_{B_0^*} + \underbrace{U_Y^\top E V_Y}_{E_0}$$

$$\Sigma_Y = A_0^* \Sigma_X B_0^* + E_0$$

One to one mapping between A^*/A_0^* and B^*/B_0^*.

Diagonal 2MR

The model can be re-written using the SVD of Y and X. This leads to the Diagonal 2MR (D2MR).

$$Y = A^* X B^* + E$$

$$U_Y \Sigma_Y V_Y^\top = A^* U_X \Sigma_X V_X^\top B^* + E$$

$$\Sigma_Y = U_Y^\top A^* U_X \Sigma_X V_X^\top B^* V_Y + U_Y^\top E V_Y$$

$$\Sigma_Y = \underbrace{\left(U_Y^\top A^* U_X\right)}_{A_0^*} \Sigma_X \underbrace{\left(V_X^\top B^* V_Y\right)}_{B_0^*} + \underbrace{U_Y^\top E V_Y}_{E_0}$$

$$\Sigma_Y = A_0^* \Sigma_X B_0^* + E_0$$

One to one mapping between A^*/A_0^* and B^*/B_0^*. E_0 and E share the same singular values.

D2MR is equivalent to 2MR

For any matrices A, B, there is:

D2MR is equivalent to 2MR

For any matrices A, B, there is:

$$\|Y - AXB\|_F^2 = \|\Sigma_Y - U_Y^\top A U_X \Sigma_X V_X^T B V_Y\|_F^2,$$

because the Frobenius norm being invariant by multiplication of orthogonal matrices.

D2MR is equivalent to 2MR

For any matrices A, B, there is:

$$\|Y - AXB\|_F^2 = \|\Sigma_Y - U_Y^\top A U_X \Sigma_X V_X^T B V_Y\|_F^2,$$

because the Frobenius norm being invariant by multiplication of orthogonal matrices.

This leads to:

$$\|Y - AXB\|_F^2 = \|\Sigma_Y - A_0 \Sigma_X B_0\|_F^2,$$

where $A_0 = U_Y^\top A U_X$ and $B_0 = V_X^\top B V_Y$.

D2MR is equivalent to 2MR

For any matrices A, B, there is:

$$\|Y - AXB\|_F^2 = \|\Sigma_Y - U_Y^\top A U_X \Sigma_X V_X^T B V_Y\|_F^2,$$

because the Frobenius norm being invariant by multiplication of orthogonal matrices.
This leads to:

$$\|Y - AXB\|_F^2 = \|\Sigma_Y - A_0 \Sigma_X B_0\|_F^2,$$

where $A_0 = U_Y^\top A U_X$ and $B_0 = V_X^\top B V_Y$.
A and A_0 have the same rank, idem for B and B_0.

D2MR is equivalent to 2MR

For any matrices A, B, there is:

$$\|Y - AXB\|_F^2 = \|\Sigma_Y - U_Y^\top A U_X \Sigma_X V_X^T B V_Y\|_F^2,$$

because the Frobenius norm being invariant by multiplication of orthogonal matrices.
This leads to:

$$\|Y - AXB\|_F^2 = \|\Sigma_Y - A_0 \Sigma_X B_0\|_F^2,$$

where $A_0 = U_Y^\top A U_X$ and $B_0 = V_X^\top B V_Y$.
A and A_0 have the same rank, idem for B and B_0.
The initial problem is equivalent to finding predictors satisfying

$$(\hat{A}_{0r}, \hat{B}_{0r}) \in \underset{\substack{A_0, B_0: \\ \text{rank } A_0 \wedge \text{rank } B_0 \leq r}}{\arg\min} \|\Sigma_Y - A_0 \Sigma_X B_0\|_F^2.$$

Solution of D2MR

- **Objective**: Under the constraint $\text{rank}(A_0) \leq r$ and $\text{rank}(B_0) \leq r$, minimize:

$$\left\| \underbrace{\begin{pmatrix} \sigma_1(Y) & & & \\ & \ddots & & \\ & & \sigma_{r_Y}(Y) & \\ & & & 0 \end{pmatrix}}_{n \times p} - A_0 \underbrace{\begin{pmatrix} \sigma_1(X) & & & \\ & \ddots & & \\ & & \sigma_{r_X}(X) & \\ & & & 0 \end{pmatrix}}_{m \times q} B_0 \right\|_F^2 .$$

Solution of D2MR

- **Objective**: Under the constraint $\text{rank}(A_0) \le r$ and $\text{rank}(B_0) \le r$, minimize:

$$\left\| \underbrace{\begin{pmatrix} \sigma_1(Y) & & & \\ & \ddots & & \\ & & \sigma_{r_Y}(Y) & \\ & & & 0 \end{pmatrix}}_{n \times p} - A_0 \underbrace{\begin{pmatrix} \sigma_1(X) & & & \\ & \ddots & & \\ & & \sigma_{r_X}(X) & \\ & & & 0 \end{pmatrix}}_{m \times q} B_0 \right\|_F^2 .$$

- **Solution**:

$$\hat{A}_{0r} = \underbrace{\begin{pmatrix} \sigma_1(Y) & & & \\ & \ddots & & \\ & & \sigma_{r \wedge r_Y}(Y) & \\ & & & 0 \end{pmatrix}}_{n \times m}, \quad \hat{B}_{0r} = \underbrace{\begin{pmatrix} \sigma_1(X)^{-1} & & & \\ & \ddots & & \\ & & \sigma_r(X)^{-1} & \\ & & & 0 \end{pmatrix}}_{q \times p}.$$

Solution of D2MR

- **Objective**: Under the constraint $\text{rank}(A_0) \leq r$ and $\text{rank}(B_0) \leq r$, minimize:

$$\left\| \underbrace{\begin{pmatrix} \sigma_1(Y) & & & \\ & \ddots & & \\ & & \sigma_{r_Y}(Y) & \\ & & & 0 \end{pmatrix}}_{n \times p} - A_0 \underbrace{\begin{pmatrix} \sigma_1(X) & & & \\ & \ddots & & \\ & & \sigma_{r_X}(X) & \\ & & & 0 \end{pmatrix}}_{m \times q} B_0 \right\|_F^2 .$$

- **Solution**:

$$\hat{A}_{0r} = \underbrace{\begin{pmatrix} \sigma_1(Y) & & & \\ & \ddots & & \\ & & \sigma_{r \wedge r_Y}(Y) & \\ & & & 0 \end{pmatrix}}_{n \times m}, \quad \hat{B}_{0r} = \underbrace{\begin{pmatrix} \sigma_1(X)^{-1} & & & \\ & \ddots & & \\ & & \sigma_r(X)^{-1} & \\ & & & 0 \end{pmatrix}}_{q \times p} .$$

- How far is the predictor $\hat{A}_{0r} \Sigma_X \hat{B}_{0r}$ from the signal $A^* X B^*$?

Oracle inequality in the fixed rank case

Theorem (B., Butucea 2023)

The predictor $\hat{A}_{0r}\Sigma_X\hat{B}_{0r}$ satisfies for $C > 0$ and for any $t > 0$:

$$\|A_0^*\Sigma_X B_0^* - \hat{A}_{0r}\Sigma_X\hat{B}_{0r}\|_F^2 \leq 9 \inf_{\substack{A_0, B_0: \\ \text{rank } A_0 \wedge \text{rank } B_0 \leq r}} \|A_0^*\Sigma_X B_0^* - A_0\Sigma_X B_0\|_F^2 + C\sigma^2(1+t)^2 \cdot r(n+p),$$

with probability larger than $1 - 2\exp(-t^2(\sqrt{n}+\sqrt{p})^2)$.

Oracle inequality in the fixed rank case

Theorem (B., Butucea 2023)

The predictor $\hat{A}_{0r}\Sigma_X\hat{B}_{0r}$ *satisfies for* $C>0$ *and for any* $t>0$*:*

$$\|A_0^*\Sigma_X B_0^* - \hat{A}_{0r}\Sigma_X\hat{B}_{0r}\|_F^2 \leq 9 \inf_{\substack{A_0,B_0:\\ \text{rank } A_0 \wedge \text{rank } B_0 \leq r}} \|A_0^*\Sigma_X B_0^* - A_0\Sigma_X B_0\|_F^2 + C\sigma^2(1+t)^2 \cdot r(n+p),$$

with probability larger than $1-2\exp(-t^2(\sqrt{n}+\sqrt{p})^2)$.

- $\inf_{\substack{A,B:\\ \text{rank } A \wedge \text{rank } B \leq r}} \|A^*XB^* - AXB\|_F^2 = \sum_{k=r+1}^{r^*} \sigma_k(A^*XB^*)^2 \cdot \mathbf{1}_{r<r^*}$.

Oracle inequality in the fixed rank case

Theorem (B., Butucea 2023)

The predictor $\hat{A}_{0r}\Sigma_X\hat{B}_{0r}$ satisfies for $C > 0$ and for any $t > 0$:

$$\|A_0^*\Sigma_X B_0^* - \hat{A}_{0r}\Sigma_X\hat{B}_{0r}\|_F^2 \leq 9 \inf_{\substack{A_0, B_0: \\ \text{rank } A_0 \wedge \text{rank } B_0 \leq r}} \|A_0^*\Sigma_X B_0^* - A_0\Sigma_X B_0\|_F^2 + C\sigma^2(1+t)^2 \cdot r(n+p),$$

with probability larger than $1 - 2\exp(-t^2(\sqrt{n}+\sqrt{p})^2)$.

- $\inf_{\substack{A,B: \\ \text{rank } A \wedge \text{rank } B \leq r}} \|A^*XB^* - AXB\|_F^2 = \sum_{k=r+1}^{r^*} \sigma_k(A^*XB^*)^2 \cdot \mathbf{1}_{r<r^*}.$
- $\mathcal{O}(r(n+p))$ is the minimax optimal rate in the (one-sided) *matrix regression* (MR) model.

Oracle inequality in the fixed rank case

Theorem (B., Butucea 2023)

The predictor $\hat{A}_{0r}\Sigma_X\hat{B}_{0r}$ satisfies for $C > 0$ and for any $t > 0$:

$$\|A_0^*\Sigma_X B_0^* - \hat{A}_{0r}\Sigma_X\hat{B}_{0r}\|_F^2 \leq 9 \inf_{\substack{A_0, B_0: \\ \text{rank } A_0 \wedge \text{rank } B_0 \leq r}} \|A_0^*\Sigma_X B_0^* - A_0\Sigma_X B_0\|_F^2 + C\sigma^2(1+t)^2 \cdot r(n+p),$$

with probability larger than $1 - 2\exp(-t^2(\sqrt{n}+\sqrt{p})^2)$.

- $$\inf_{\substack{A, B: \\ \text{rank } A \wedge \text{rank } B \leq r}} \|A^*XB^* - AXB\|_F^2 = \sum_{k=r+1}^{r^*} \sigma_k(A^*XB^*)^2 \cdot \mathbf{1}_{r<r^*}.$$
- $\mathcal{O}\left(r(n+p)\right)$ is the minimax optimal rate in the (one-sided) *matrix regression* (MR) model.
- From the explicit solutions $(\hat{A}_{0r}, \hat{B}_{0r})$ we deduce $(\hat{A}_r, \hat{B}_r)$ solution to the initial problem:

$$\hat{A}_r = U_Y\hat{A}_{0r}U_X^\top \quad \text{and} \quad \hat{B}_r = V_X\hat{B}_{0r}V_Y^\top.$$

Further results

We derive a rank-adaptive procedure.

Further results

We derive a rank-adaptive procedure.

- It retrieves the true rank of the signal with high probability.

Further results

We derive a rank-adaptive procedure.

- It retrieves the true rank of the signal with high probability.
- Then we derive predictors that exhibit almost oracle deviation.

Further results

We derive a rank-adaptive procedure.

- It retrieves the true rank of the signal with high probability.
- Then we derive predictors that exhibit almost oracle deviation.
- Theoretical guarantees require a σ^2 dependent lower bound on a hyper parameter λ.

Further results

We derive a rank-adaptive procedure.

- It retrieves the true rank of the signal with high probability.
- Then we derive predictors that exhibit almost oracle deviation.
- Theoretical guarantees require a σ^2 dependent lower bound on a hyper parameter λ.
- What can we do if σ is unknown ?

Further results

We derive a rank-adaptive procedure.

- It retrieves the true rank of the signal with high probability.
- Then we derive predictors that exhibit almost oracle deviation.
- Theoretical guarantees require a σ^2 dependent lower bound on a hyper parameter λ.
- What can we do if σ is unknown ?

We derive a data-driven rank-adaptive procedure free of σ with the same rate as in the oracle case.

Further results

We derive a rank-adaptive procedure.

- It retrieves the true rank of the signal with high probability.
- Then we derive predictors that exhibit almost oracle deviation.
- Theoretical guarantees require a σ^2 dependent lower bound on a hyper parameter λ.
- What can we do if σ is unknown ?

We derive a data-driven rank-adaptive procedure free of σ with the same rate as in the oracle case.
Simulation results confirm the good prediction and the rank consistency results under data-driven explicit choices of the tuning parameters and the scaling parameter of the noise.

Table of Contents

Topic models

Given a dictionary of p words we observe n documents.

Topic models

Given a dictionary of p words we observe n documents.

- A document is modeled by Y_j, a probability vector in the simplex

$$\mathcal{S}_{p-1} = \{v \in \mathbb{R}_+^p : \sum_j v_j = 1\}.$$

Each Y_j contains the frequencies of N_j words. For simplicity, $N_j = N$, for all $j = 1, \ldots, n$.

Topic models

Given a dictionary of p words we observe n documents.

- A document is modeled by Y_j, a probability vector in the simplex

$$\mathcal{S}_{p-1} = \{v \in \mathbb{R}_+^p : \sum_j v_j = 1\}.$$

Each Y_j contains the frequencies of N_j words. For simplicity, $N_j = N$, for all $j = 1, \ldots, n$. We assume that for $\pi_j \in \mathcal{S}_{p-1}$:

$$N \cdot Y_j \sim Multinomial_p(N, \pi_j).$$

.

Topic models

Given a dictionary of p words we observe n documents.

- A document is modeled by Y_j, a probability vector in the simplex

$$\mathcal{S}_{p-1} = \{v \in \mathbb{R}_+^p : \sum_j v_j = 1\}.$$

Each Y_j contains the frequencies of N_j words. For simplicity, $N_j = N$, for all $j = 1, \ldots, n$. We assume that for $\pi_j \in \mathcal{S}_{p-1}$:

$$N \cdot Y_j \sim Multinomial_p(N, \pi_j).$$

.

- **Topic model**: There is $K \ll \min(n, p)$ such that a Non-negative Matrix Factorization (NMF) takes place on $\Pi^* := (\pi_1^*, \ldots, \pi_n^*)$:

$$\Pi^* = A^* W^*,$$

where $A^* \in \mathbb{R}^{p \times K}$ has columns in $\mathcal{S}_{p-1}$, $W^* \in \mathbb{R}^{K \times n}$ has columns in $\mathcal{S}_{K-1}$.

Topic models

Given a dictionary of p words we observe n documents.

- A document is modeled by Y_j, a probability vector in the simplex

  $$\mathcal{S}_{p-1} = \{v \in \mathbb{R}_+^p : \sum_j v_j = 1\}.$$

  Each Y_j contains the frequencies of N_j words. For simplicity, $N_j = N$, for all $j = 1, \dots, n$. We assume that for $\pi_j \in \mathcal{S}_{p-1}$:

  $$N \cdot Y_j \sim Multinomial_p(N, \pi_j).$$

  .
- **Topic model**: There is $K \ll \min(n, p)$ such that a Non-negative Matrix Factorization (NMF) takes place on $\Pi^* := (\pi_1^*, \dots, \pi_n^*)$:

  $$\Pi^* = A^* W^*,$$

  where $A^* \in \mathbb{R}^{p \times K}$ has columns in $\mathcal{S}_{p-1}$, $W^* \in \mathbb{R}^{K \times n}$ has columns in $\mathcal{S}_{K-1}$.
- **Interpretation**:

  $$\mathbb{P}(\text{word } i | \text{document } j) = \sum_{k=1}^{K} \mathbb{P}(\text{word } i | \text{topic } k) \mathbb{P}(\text{topic } k | \text{document } j)$$

Identifiability

Objective: Estimate A^* and/or W^*.

Identifiability

Objective: Estimate A^* and/or W^*.

Assumptions are required to ensure the uniqueness of the representation and the existence of fast-performing algorithms.

Identifiability

Objective: Estimate A^* and/or W^*.

Assumptions are required to ensure the uniqueness of the representation and the existence of fast-performing algorithms.

- **Anchor word assumption**: For each topic $k \in [K]$, there exists at least one word j such that $[A^*]_{jk} > 0$ and $[A^*]_{jl} = 0$ for $l \in [K] \setminus \{k\}$.

Identifiability

Objective: Estimate A^* and/or W^*.

Assumptions are required to ensure the uniqueness of the representation and the existence of fast-performing algorithms.

- **Anchor word assumption**: For each topic $k \in [K]$, there exists at least one word j such that $[A^*]_{jk} > 0$ and $[A^*]_{jl} = 0$ for $l \in [K] \backslash \{k\}$.
- W^* **is full rank**: $\text{rank}(W^*) = K$.

Bibliography

NMF bibliography:

Bibliography

NMF bibliography:

- NMF is NP-hard: Vavasis 2010 (SIAM J Optim.).

Bibliography

NMF bibliography:

- NMF is NP-hard: Vavasis 2010 (SIAM J Optim.).
- Separability assumption = Anchor word ass. and full-rank ass. on W^* provides unique NMF: Donoho, Stodden 2003 (NeurIPS).

Bibliography

NMF bibliography:

- NMF is NP-hard: Vavasis 2010 (SIAM J Optim.).
- Separability assumption = Anchor word ass. and full-rank ass. on W^* provides unique NMF: Donoho, Stodden 2003 (NeurIPS).
- Given the index set $\mathbf{J}$ of anchor words and the number of topics K, A^* can be uniquely retrieved via Π^*: Arora, Ge, Moitra 2012 (IEEE Annual Symposium FoCS)

Bibliography

NMF bibliography:

- NMF is NP-hard: Vavasis 2010 (SIAM J Optim.).
- Separability assumption = Anchor word ass. and full-rank ass. on W^* provides unique NMF: Donoho, Stodden 2003 (NeurIPS).
- Given the index set $\mathbf{J}$ of anchor words and the number of topics K, A^* can be uniquely retrieved via Π^*: Arora, Ge, Moitra 2012 (IEEE Annual Symposium FoCS)
- When $\mathbf{J}$ is unknown but K is known, A^* and W^* can be uniquely retrieved via Π^*: Recht, Re, Tropp, Bittorf 2012 (NeurIPS)

Bibliography

NMF bibliography:

- NMF is NP-hard: Vavasis 2010 (SIAM J Optim.).
- Separability assumption = Anchor word ass. and full-rank ass. on W^* provides unique NMF: Donoho, Stodden 2003 (NeurIPS).
- Given the index set $\mathbf{J}$ of anchor words and the number of topics K, A^* can be uniquely retrieved via Π^*: Arora, Ge, Moitra 2012 (IEEE Annual Symposium FoCS)
- When $\mathbf{J}$ is unknown but K is known, A^* and W^* can be uniquely retrieved via Π^*: Recht, Re, Tropp, Bittorf 2012 (NeurIPS)
- Standard procedures leverage the simplex structure in the matrix Π^*, Ding, Rohban, Ishwar, Saligrama 2013 (ICML) or in the singular vectors of Π^*, Ke, Wang 2024 (JASA).

Bibliography

NMF bibliography:

- NMF is NP-hard: Vavasis 2010 (SIAM J Optim.).
- Separability assumption = Anchor word ass. and full-rank ass. on W^* provides unique NMF: Donoho, Stodden 2003 (NeurIPS).
- Given the index set $\mathbf{J}$ of anchor words and the number of topics K, A^* can be uniquely retrieved via Π^*: Arora, Ge, Moitra 2012 (IEEE Annual Symposium FoCS)
- When $\mathbf{J}$ is unknown but K is known, A^* and W^* can be uniquely retrieved via Π^*: Recht, Re, Tropp, Bittorf 2012 (NeurIPS)
- Standard procedures leverage the simplex structure in the matrix Π^*, Ding, Rohban, Ishwar, Saligrama 2013 (ICML) or in the singular vectors of Π^*, Ke, Wang 2024 (JASA).
- When K is unknown: Bing, Bunea, Wegkamp 2020 (Bernoulli) provides another procedure that is not a variation of simplex algorithms.

Bibliography

NMF bibliography:

- NMF is NP-hard: Vavasis 2010 (SIAM J Optim.).
- Separability assumption = Anchor word ass. and full-rank ass. on W^* provides unique NMF: Donoho, Stodden 2003 (NeurIPS).
- Given the index set $\mathbf{J}$ of anchor words and the number of topics K, A^* can be uniquely retrieved via Π^*: Arora, Ge, Moitra 2012 (IEEE Annual Symposium FoCS)
- When $\mathbf{J}$ is unknown but K is known, A^* and W^* can be uniquely retrieved via Π^*: Recht, Re, Tropp, Bittorf 2012 (NeurIPS)
- Standard procedures leverage the simplex structure in the matrix Π^*, Ding, Rohban, Ishwar, Saligrama 2013 (ICML) or in the singular vectors of Π^*, Ke, Wang 2024 (JASA).
- When K is unknown: Bing, Bunea, Wegkamp 2020 (Bernoulli) provides another procedure that is not a variation of simplex algorithms.
- Direct estimation of W^* is studied in Klopp, Panov, Sigalla, Tsybakov 2023 (Ann. Statist.) under the anchor document assumption.

Dynamic Topic Model

Batches of n documents are collected in T steps over time.

Dynamic Topic Model

Batches of n documents are collected in T steps over time.

- **Model**: the topic-document probability matrix W^* follows a simplex-valued stationary autoregressive model of order one and A^* stays constant.

Dynamic Topic Model

Batches of n documents are collected in T steps over time.

- **Model**: the topic-document probability matrix W^* follows a simplex-valued stationary autoregressive model of order one and A^* stays constant.

$$\boldsymbol{W}^{t+1} = (1 - c^*) \cdot \boldsymbol{W}^t + c^* \cdot \boldsymbol{\Delta}^t, \quad t = 1, \ldots, T-1,$$

where $c^* \in (0, 1)$, and each $\boldsymbol{\Delta}^t \in \mathbb{R}^{K \times n}$ has i.i.d. columns sampled from the Dirichlet distribution $\mathcal{D}(\theta^*)$ with $\theta^* \in \mathbb{R}_+^K$.

Dynamic Topic Model

Batches of n documents are collected in T steps over time.

- **Model**: the topic-document probability matrix W^* follows a simplex-valued stationary autoregressive model of order one and A^* stays constant.

$$\boldsymbol{W}^{t+1} = (1 - c^*) \cdot \boldsymbol{W}^t + c^* \cdot \boldsymbol{\Delta}^t, \quad t = 1, \ldots, T-1,$$

  where $c^* \in (0,1)$, and each $\boldsymbol{\Delta}^t \in \mathbb{R}^{K \times n}$ has i.i.d. columns sampled from the Dirichlet distribution $\mathcal{D}(\theta^*)$ with $\theta^* \in \mathbb{R}_+^K$.

- **Objective**: Estimation of c^*, $\quad \tilde{\theta}^* := \frac{\theta^*}{\alpha} \in \mathcal{S}_{K-1}$ and $\alpha := \|\theta^*\|_1$.

Dynamic Topic Model

Batches of n documents are collected in T steps over time.

- **Model**: the topic-document probability matrix W^* follows a simplex-valued stationary autoregressive model of order one and A^* stays constant.

$$\boldsymbol{W}^{t+1} = (1 - c^*) \cdot \boldsymbol{W}^t + c^* \cdot \boldsymbol{\Delta}^t, \quad t = 1, \ldots, T-1,$$

  where $c^* \in (0,1)$, and each $\boldsymbol{\Delta}^t \in \mathbb{R}^{K \times n}$ has i.i.d. columns sampled from the Dirichlet distribution $\mathcal{D}(\theta^*)$ with $\theta^* \in \mathbb{R}_+^K$.

- **Objective**: Estimation of c^*, $\quad \tilde{\theta}^* := \frac{\theta^*}{\alpha} \in \mathcal{S}_{K-1}$ and $\quad \alpha := \|\theta^*\|_1$.
- **Dynamic Latent Factors**: $\boldsymbol{\Pi}^{1:T} := \left(\boldsymbol{\Pi}^1, \ldots, \boldsymbol{\Pi}^T\right)$ is available where

$$\boldsymbol{\Pi}^t = A^* \boldsymbol{W}^t, \quad t = 1, \ldots, T-1.$$

Dynamic Topic Model

Batches of n documents are collected in T steps over time.

- **Model**: the topic-document probability matrix W^* follows a simplex-valued stationary autoregressive model of order one and A^* stays constant.

$$\boldsymbol{W}^{t+1} = (1 - c^*) \cdot \boldsymbol{W}^t + c^* \cdot \boldsymbol{\Delta}^t, \quad t = 1, \ldots, T-1,$$

  where $c^* \in (0,1)$, and each $\boldsymbol{\Delta}^t \in \mathbb{R}^{K \times n}$ has i.i.d. columns sampled from the Dirichlet distribution $\mathcal{D}(\theta^*)$ with $\theta^* \in \mathbb{R}_+^K$.

- **Objective**: Estimation of c^*, $\quad \tilde{\theta}^* := \frac{\theta^*}{\alpha} \in \mathcal{S}_{K-1}$ and $\alpha := \|\theta^*\|_1$.
- **Dynamic Latent Factors**: $\boldsymbol{\Pi}^{1:T} := \left(\boldsymbol{\Pi}^1, \ldots, \boldsymbol{\Pi}^T\right)$ is available where

$$\boldsymbol{\Pi}^t = A^* \boldsymbol{W}^t, \quad t = 1, \ldots, T-1.$$

- **Dynamic Topic Model**: $\boldsymbol{Y}^{1:T} := \left(\boldsymbol{Y}^1, \ldots, \boldsymbol{Y}^T\right)$ is available where

$$N \boldsymbol{Y}_j^t | \boldsymbol{W}_j^t \sim \text{Multinomial}_p \left(N, A^* \boldsymbol{W}_j^t\right), \quad t = 1, \ldots, T-1.$$

Dynamic Topic Model

Batches of n documents are collected in T steps over time.

- **Model**: the topic-document probability matrix W^* follows a simplex-valued stationary autoregressive model of order one and A^* stays constant.

$$\boldsymbol{W}^{t+1} = (1 - c^*) \cdot \boldsymbol{W}^t + c^* \cdot \boldsymbol{\Delta}^t, \quad t = 1, \dots, T-1,$$

  where $c^* \in (0,1)$, and each $\boldsymbol{\Delta}^t \in \mathbb{R}^{K \times n}$ has i.i.d. columns sampled from the Dirichlet distribution $\mathcal{D}(\theta^*)$ with $\theta^* \in \mathbb{R}_+^K$.

- **Objective**: Estimation of c^*, $\quad \tilde{\theta}^* := \frac{\theta^*}{\alpha} \in \mathcal{S}_{K-1}$ and $\quad \alpha := \|\theta^*\|_1$.
- **Dynamic Latent Factors**: $\boldsymbol{\Pi}^{1:T} := \left(\boldsymbol{\Pi}^1, \dots, \boldsymbol{\Pi}^T\right)$ is available where

$$\boldsymbol{\Pi}^t = A^* \boldsymbol{W}^t, \quad t = 1, \dots, T-1.$$

- **Dynamic Topic Model**: $\boldsymbol{Y}^{1:T} := \left(\boldsymbol{Y}^1, \dots, \boldsymbol{Y}^T\right)$ is available where

$$N \boldsymbol{Y}_j^t | \boldsymbol{W}_j^t \sim \text{Multinomial}_p \left(N, A^* \boldsymbol{W}_j^t\right), \quad t = 1, \dots, T-1.$$

Double randomness: Dirichlet + Multinomial

Assumptions

- Let the **topic-topic overlapping matrix** measure the affinity of topics using the same words:

$$\Sigma_A := (A^*)^\top H^{-1} A^*,$$

where $H := \text{diag}(h_1, \ldots, h_p)$ and $h_i := \|A^*_{i.}\|_1$.

Assumptions

- Let the **topic-topic overlapping matrix** measure the affinity of topics using the same words:

$$\Sigma_A := \left(A^*\right)^\top H^{-1} A^*,$$

where $H := \mathrm{diag}(h_1, \ldots, h_p)$ and $h_i := \|A^*_{i\cdot}\|_1$.

Assume: $\lambda_K\left(\Sigma_A\right) \geq c, \quad \min_{k,l}\left[\Sigma_A\right]_{kl} \geq c$ and $\min_i h_i := h_{\min} \geq c\dfrac{K}{p}$.

Assumptions

- Let the **topic-topic overlapping matrix** measure the affinity of topics using the same words:
$$\Sigma_A := \left(A^*\right)^\top H^{-1} A^*,$$
where $H := \mathrm{diag}(h_1, \dots, h_p)$ and $h_i := \|A^*_{i.}\|_1$.
Assume: $\lambda_K\left(\Sigma_A\right) \geq c, \quad \min_{k,l}\left[\Sigma_A\right]_{kl} \geq c$ and $\min_i h_i := h_{\min} \geq c\dfrac{K}{p}$.
- Let the **topic-topic concurrence matrix**
$$\Sigma_W^{1:T} := \frac{1}{nT}\left(\boldsymbol{W}^{1:T}\right)\left(\boldsymbol{W}^{1:T}\right)^\top,$$
capture the affinity of topics to be covered together in the same document.

Assumptions

- Let the **topic-topic overlapping matrix** measure the affinity of topics using the same words:

$$\Sigma_A := \left(A^*\right)^\top H^{-1} A^*,$$

where $H := \mathrm{diag}(h_1, \ldots, h_p)$ and $h_i := \|A^*_{i\cdot}\|_1$.

Assume: $\lambda_K\left(\Sigma_A\right) \geq c, \quad \min_{k,l}\left[\Sigma_A\right]_{kl} \geq c$ and $\min_i h_i := h_{\min} \geq c\dfrac{K}{p}$.

- Let the **topic-topic concurrence matrix**

$$\Sigma_{\boldsymbol{W}}^{1:T} := \frac{1}{nT}\left(\boldsymbol{W}^{1:T}\right)\left(\boldsymbol{W}^{1:T}\right)^\top,$$

capture the affinity of topics to be covered together in the same document.
Assume: $\lambda_K(\Sigma_{\boldsymbol{W}}^{1:T}) \geq c > 0, \quad$ a.s..
Remark: if $\min_k \tilde{\theta}_k^* \geq c > 0$, this holds for large enough n, T with high probability.

Assumptions

Assume the following hold

- Anchor word assumption

Assumptions

Assume the following hold

- Anchor word assumption
- Assumptions on the topic-topic overlapping matrix and the topic-topic concurrence matrix.

Assumptions

Assume the following hold

- Anchor word assumption
- Assumptions on the topic-topic overlapping matrix and the topic-topic concurrence matrix.
- For $\underline{c}$ and $\overline{c}$ in $(0,1)$, c^* satisfies: $\underline{c} \leq c^* \leq \overline{c}$.

Assumptions

Assume the following hold

- Anchor word assumption
- Assumptions on the topic-topic overlapping matrix and the topic-topic concurrence matrix.
- For $\underline{c}$ and $\overline{c}$ in $(0,1)$, c^* satisfies: $\underline{c} \leq c^* \leq \overline{c}$.
- For $\underline{\theta}$ and m in $(0,1)$ and $\Sigma(\theta^*) = \frac{1}{\alpha+1}\left(\mathrm{diag}(\tilde{\theta}^*) - \tilde{\theta}^* \cdot (\tilde{\theta}^*)^\top\right)$, θ^* satisfies:

$$\min_{k\in[K]} \tilde{\theta}^*(k) \geq \underline{\theta} \text{ and } m \leq \mathrm{Tr}(\Sigma(\theta^*)) \leq 1.$$

Global Procedure

The estimation procedure unfolds as follows in the Dynamic Latent Factors model (resp. Dynamic Topic Model):

Global Procedure

The estimation procedure unfolds as follows in the Dynamic Latent Factors model (resp. Dynamic Topic Model):

- Recover A^* (resp. estimate A^* with $\hat{A}$).

Global Procedure

The estimation procedure unfolds as follows in the Dynamic Latent Factors model (resp. Dynamic Topic Model):

- Recover A^* (resp. estimate A^* with $\hat{A}$).
- Project $\mathbf{\Pi}^{1:T}$ (resp. $\boldsymbol{Y}^{1:T}$) on the linear space spanned by the columns of A^* (resp. $\hat{A}$) and get a proxy random matrix of the unobserved $\boldsymbol{W}^{1:T}$.

Global Procedure

The estimation procedure unfolds as follows in the Dynamic Latent Factors model (resp. Dynamic Topic Model):

- Recover A^* (resp. estimate A^* with $\hat{A}$).
- Project $\mathbf{\Pi}^{1:T}$ (resp. $\boldsymbol{Y}^{1:T}$) on the linear space spanned by the columns of A^* (resp. $\hat{A}$) and get a proxy random matrix of the unobserved $\boldsymbol{W}^{1:T}$.
- Build estimators of

$$c^*, \quad \tilde{\theta}^* := \frac{\theta^*}{\alpha} \in \mathcal{S}_{K-1} \quad \text{and} \quad \alpha := \|\theta^*\|_1.$$

Dynamic Latent Factors: recovering of A^*

Given $\mathbf{\Pi}^{1:T}$, A^* is exactly recovered following these steps (Ke, Wang 2024):

Dynamic Latent Factors: recovering of A^*

Given $\mathbf{\Pi}^{1:T}$, A^* is exactly recovered following these steps (Ke, Wang 2024):

- **Pre-SVD normalization**: Tackles word frequency heterogeneity.

Dynamic Latent Factors: recovering of A^*

Given $\mathbf{\Pi}^{1:T}$, A^* is exactly recovered following these steps (Ke, Wang 2024):

- **Pre-SVD normalization**: Tackles word frequency heterogeneity.
- **SVD**: Creates an embedding of the p rows of $\mathbf{\Pi}^{1:T}$ into $\mathbb{R}^K$. These p points are contained in a cone. The anchor words are located on its supporting rays.

Dynamic Latent Factors: recovering of A^*

Given $\mathbf{\Pi}^{1:T}$, A^* is exactly recovered following these steps (Ke, Wang 2024):

- **Pre-SVD normalization**: Tackles word frequency heterogeneity.
- **SVD**: Creates an embedding of the p rows of $\mathbf{\Pi}^{1:T}$ into $\mathbb{R}^K$. These p points are contained in a cone. The anchor words are located on its supporting rays.
- **Post-SVD normalization**: Normalize the p points to ensure they are now contained in a simplex.

Dynamic Latent Factors: recovering of A^*

Given $\mathbf{\Pi}^{1:T}$, A^* is exactly recovered following these steps (Ke, Wang 2024):

- **Pre-SVD normalization**: Tackles word frequency heterogeneity.
- **SVD**: Creates an embedding of the p rows of $\mathbf{\Pi}^{1:T}$ into $\mathbb{R}^K$. These p points are contained in a cone. The anchor words are located on its supporting rays.
- **Post-SVD normalization**: Normalize the p points to ensure they are now contained in a simplex.
- **Vertex Hunting**: Recover the simplex by computing the convex hull of the p points.

Dynamic Latent Factors: recovering of A^*

Given $\mathbf{\Pi}^{1:T}$, A^* is exactly recovered following these steps (Ke, Wang 2024):

- **Pre-SVD normalization**: Tackles word frequency heterogeneity.
- **SVD**: Creates an embedding of the p rows of $\mathbf{\Pi}^{1:T}$ into $\mathbb{R}^K$. These p points are contained in a cone. The anchor words are located on its supporting rays.
- **Post-SVD normalization**: Normalize the p points to ensure they are now contained in a simplex.
- **Vertex Hunting**: Recover the simplex by computing the convex hull of the p points.
- **Word-topic matrix recovery**: Using that each column of A^* has unit L_1 norm allows the recovery.

Dynamic Latent Factors: recovering of A^*

Given $\boldsymbol{\Pi}^{1:T}$, A^* is exactly recovered following these steps (Ke, Wang 2024):

- **Pre-SVD normalization**: Tackles word frequency heterogeneity.
- **SVD**: Creates an embedding of the p rows of $\boldsymbol{\Pi}^{1:T}$ into $\mathbb{R}^K$. These p points are contained in a cone. The anchor words are located on its supporting rays.
- **Post-SVD normalization**: Normalize the p points to ensure they are now contained in a simplex.
- **Vertex Hunting**: Recover the simplex by computing the convex hull of the p points.
- **Word-topic matrix recovery**: Using that each column of A^* has unit L_1 norm allows the recovery.

Then $\boldsymbol{W}^{1:T}$ is recovered by projection of $\boldsymbol{\Pi}^{1:T}$ onto the span of A^*.

Dynamic Latent Factors: recovering of A^*

Given $\mathbf{\Pi}^{1:T}$, A^* is exactly recovered following these steps (Ke, Wang 2024):

Dynamic Latent Factors: recovering of A^*

Given $\mathbf{\Pi}^{1:T}$, A^* is exactly recovered following these steps (Ke, Wang 2024):

- *Pre-SVD normalization:* Compute $\mathbf{\Pi}_* := \boldsymbol{M}_*^{-1/2}\mathbf{\Pi}^{1:T}$ where

$$\boldsymbol{M}_* = (nT)^{-1}\mathrm{diag}\left(\mathbf{\Pi}^{1:T} \cdot 1_{nT}\right).$$

Dynamic Latent Factors: recovering of A^*

Given $\mathbf{\Pi}^{1:T}$, A^* is exactly recovered following these steps (Ke, Wang 2024):

- *Pre-SVD normalization:* Compute $\mathbf{\Pi}_* := \boldsymbol{M}_*^{-1/2}\mathbf{\Pi}^{1:T}$ where

$$\boldsymbol{M}_* = (nT)^{-1}\text{diag}\left(\mathbf{\Pi}^{1:T} \cdot 1_{nT}\right).$$

- *SVD:* of $\mathbf{\Pi}_* := \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ which satisfies $\text{rank}(\mathbf{\Pi}_*) = K$ a.s..

Dynamic Latent Factors: recovering of A^*

Given $\mathbf{\Pi}^{1:T}$, A^* is exactly recovered following these steps (Ke, Wang 2024):

- *Pre-SVD normalization:* Compute $\mathbf{\Pi}_* := \boldsymbol{M}_*^{-1/2}\mathbf{\Pi}^{1:T}$ where

$$\boldsymbol{M}_* = (nT)^{-1}\text{diag}\left(\mathbf{\Pi}^{1:T} \cdot 1_{nT}\right).$$

- *SVD:* of $\mathbf{\Pi}_* := \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ which satisfies $\text{rank}(\mathbf{\Pi}_*) = K$ a.s..
  Perron-Frobenius's theorem guarantees that $[\boldsymbol{U}]_{.1}$ does not possess any null entry a.s.. The SVD creates a low dimensional word embedding into $\mathbb{R}^K$ but these vectors do not directly lead to the recovery of A^*.

Dynamic Latent Factors: recovering of A^*

Given $\mathbf{\Pi}^{1:T}$, A^* is exactly recovered following these steps (Ke, Wang 2024):

- *Pre-SVD normalization:* Compute $\mathbf{\Pi}_* := \boldsymbol{M}_*^{-1/2}\mathbf{\Pi}^{1:T}$ where

$$\boldsymbol{M}_* = (nT)^{-1}\text{diag}\left(\mathbf{\Pi}^{1:T} \cdot 1_{nT}\right).$$

- *SVD:* of $\mathbf{\Pi}_* := \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ which satisfies $\text{rank}(\mathbf{\Pi}_*) = K$ a.s..
- *Post-SVD normalization:* Compute $\boldsymbol{R} \in \mathbb{R}^{p\times(K-1)}$: for $i \in [p]$ and $k \in [K-1]$,

$$[\boldsymbol{R}]_{ik} = \frac{[\boldsymbol{U}]_{i(k+1)}}{[\boldsymbol{U}]_{i1}}.$$

Dynamic Latent Factors: recovering of A^*

Given $\mathbf{\Pi}^{1:T}$, A^* is exactly recovered following these steps (Ke, Wang 2024):

- *Pre-SVD normalization:* Compute $\mathbf{\Pi}_* := \boldsymbol{M}_*^{-1/2}\mathbf{\Pi}^{1:T}$ where

$$\boldsymbol{M}_* = (nT)^{-1}\text{diag}\left(\mathbf{\Pi}^{1:T} \cdot 1_{nT}\right).$$

- *SVD:* of $\mathbf{\Pi}_* := \boldsymbol{U\Sigma V}^\top$ which satisfies $\text{rank}(\mathbf{\Pi}_*) = K$ a.s..
- *Post-SVD normalization:* Compute $\boldsymbol{R} \in \mathbb{R}^{p\times(K-1)}$: for $i \in [p]$ and $k \in [K-1]$,

$$[\boldsymbol{R}]_{ik} = \frac{[\boldsymbol{U}]_{i(k+1)}}{[\boldsymbol{U}]_{i1}}.$$

$[\boldsymbol{R}]_{1.}, \ldots, [\boldsymbol{R}]_{p.}$ are located in a simplex

$$G_{\boldsymbol{\eta}} := \left\{x : x = \sum_{k=1}^{K} \alpha_k \boldsymbol{\eta}_k,\ \forall k \in [K],\ \alpha_k \geq 0\ \sum_{k=1}^{K} \alpha_k = 1\right\}.$$

Dynamic Latent Factors: recovering of A^*

Given $\mathbf{\Pi}^{1:T}$, A^* is exactly recovered following these steps:

- *Pre-SVD normalization*
- *SVD*
- *Post-SVD normalization*
- *Vertex Hunting:* The vertices $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_K$ of $G_{\boldsymbol{\eta}}$ are recovered by computing the convex hull of the point cloud $[\boldsymbol{R}]_{1\cdot}, \ldots, [\boldsymbol{R}]_{p\cdot}$.

Dynamic Latent Factors: recovering of A^*

Given $\mathbf{\Pi}^{1:T}$, A^* is exactly recovered following these steps:

- *Pre-SVD normalization*
- *SVD*
- *Post-SVD normalization*
- *Vertex Hunting:* The vertices $\boldsymbol{\eta}_1, \dots, \boldsymbol{\eta}_K$ of $G_{\boldsymbol{\eta}}$ are recovered by computing the convex hull of the point cloud $[\boldsymbol{R}]_{1\cdot}, \dots, [\boldsymbol{R}]_{p\cdot}$.
  Define $\mathbf{\Lambda} \in \mathbb{R}^{p \times K}$ by solving for all $i \in [p]$,

$$[\boldsymbol{R}]_{i\cdot} = \sum_{k=1}^{K} [\mathbf{\Lambda}]_{ik} \boldsymbol{\eta}_k,$$

  s.t. $\sum_{k=1}^{K} [\mathbf{\Lambda}]_{ik} = 1$ and $[\mathbf{\Lambda}]_{ik} \geq 0$, for $k \in [K]$.

Dynamic Latent Factors: recovering of A^*

Given $\mathbf{\Pi}^{1:T}$, A^* is exactly recovered following these steps:

- *Pre-SVD normalization*
- *SVD*
- *Post-SVD normalization*
- *Vertex Hunting:* The vertices $\boldsymbol{\eta}_1, \dots, \boldsymbol{\eta}_K$ of $G_{\boldsymbol{\eta}}$ are recovered by computing the convex hull of the point cloud $[\boldsymbol{R}]_{1\cdot}, \dots, [\boldsymbol{R}]_{p\cdot}$.
  Define $\mathbf{\Lambda} \in \mathbb{R}^{p \times K}$ by solving for all $i \in [p]$,

$$[\boldsymbol{R}]_{i\cdot} = \sum_{k=1}^{K} [\mathbf{\Lambda}]_{ik} \boldsymbol{\eta}_k,$$

  s.t. $\sum_{k=1}^{K} [\mathbf{\Lambda}]_{ik} = 1$ and $[\mathbf{\Lambda}]_{ik} \geq 0$, for $k \in [K]$.
- *Word-topic matrix estimation:* Define $\mathbf{\Gamma} := \boldsymbol{M}_*^{1/2} \mathrm{diag}([\boldsymbol{U}]_{\cdot 1}) \mathbf{\Lambda}$. Normalize each column of $\mathbf{\Gamma}$ by its $\mathbb{L}_1$ norm. The resulting matrix is A^*.

Dynamic Latent Factors: Estimators

- We define $\hat{\theta}$, estimator of $\tilde{\theta}^*$, as the empirical mean of the recovered $\left(W_j^{t+1}\right)_{j,t}$:

$$\hat{\theta} := \frac{1}{n(T-1)} \sum_{j=1}^{n} \sum_{t=1}^{T-1} W_j^t.$$

Dynamic Latent Factors: Estimators

- We define $\hat{\theta}$, estimator of $\tilde{\theta}^*$, as the empirical mean of the recovered $\left(W_j^{t+1}\right)_{j,t}$:

$$\hat{\theta} := \frac{1}{n(T-1)} \sum_{j=1}^{n} \sum_{t=1}^{T-1} W_j^t.$$

- We estimate $1 - c^*$ by the normalized sum of scalar products:

$$\widehat{(1-c)} := \frac{\sum_{t=1}^{T-1} \sum_{j=1}^{n} \left\langle W_j^{t+1} - \overline{W}^{+1};\ W_j^t - \overline{w} \right\rangle}{\sum_{t=1}^{T-1} \sum_{j=1}^{n} \left\| W_j^t - \overline{W} \right\|_2^2},$$

$$\overline{W}^{+1} := \frac{1}{n(T-1)} \sum_{t=1}^{T-1} \sum_{j=1}^{n} W_j^{t+1} \text{ and } \overline{W} := \frac{1}{n(T-1)} \sum_{t=1}^{T-1} \sum_{j=1}^{n} W_j^t.$$

Using the variance of the stationary sequence and the explicit expression of the matrix Σ, we see that:

$$\mathrm{Tr}(\mathbb{V}(w_j^t)) = \frac{c^*}{2 - c^*} \frac{1 - \|\tilde{\theta}^*\|_2^2}{\alpha + 1}.$$

Using the variance of the stationary sequence and the explicit expression of the matrix Σ, we see that:

$$\mathrm{Tr}(\mathbb{V}(w_j^t)) = \frac{c^*}{2-c^*}\frac{1-\|\tilde{\theta}^*\|_2^2}{\alpha+1}.$$

Thus, we plug-in estimators $\hat{\theta}$, $\hat{c}$ and the empirical variance to get an estimator $\hat{\alpha}$ of α:

Dynamic Latent Factors: Estimators

Using the variance of the stationary sequence and the explicit expression of the matrix Σ, we see that:

$$\mathrm{Tr}(\mathbb{V}(w_j^t)) = \frac{c^*}{2-c^*}\frac{1-\|\tilde{\theta}^*\|_2^2}{\alpha+1}.$$

Thus, we plug-in estimators $\hat{\theta}$, $\hat{c}$ and the empirical variance to get an estimator $\hat{\alpha}$ of α:

$$\hat{\alpha} = \frac{\hat{c}}{2-\hat{c}}\frac{1-\|\hat{\theta}\|_2^2}{\mathcal{V}} - 1, \quad \text{where} \quad \mathcal{V} := \frac{1}{n(T-1)}\sum_{t=1}^{T-1}\sum_{j=1}^{n}\left\|w_j^t - \overline{w}\right\|_2^2.$$

Dynamic Latent Factors: Theoretical guarantees

Theorem (B., Butucea, Ke 2024)

For any N, n and T large enough, with probability at least $1 - \frac{C_1}{nT}$:

$$\max\left\{\left\|\hat{\theta} - \tilde{\theta}^*\right\|_2, |\widehat{(1-c)} - (1-c^*)|, |\hat{\alpha} - \alpha^*|\right\} \leq C_2 \cdot \sqrt{\frac{\log(nT)}{n(T-1)}},$$

where C_1, $C_2 > 0$ are explicit constants, free of the dimensions appearing in the model.

Dynamic Topic Model: Plug-in

- Replace $\mathbf{\Pi}^{1:T}$ by the observed frequencies $\boldsymbol{Y}^{1:T}$.

Dynamic Topic Model: Plug-in

- Replace $\mathbf{\Pi}^{1:T}$ by the observed frequencies $\boldsymbol{Y}^{1:T}$.
- Estimate A^* by $\hat{A}$ and build proxy random variables

$$\hat{\boldsymbol{W}}^{1:T} = (\hat{A}^\top \hat{A})^{-1} \hat{A}^\top \cdot \boldsymbol{Y}^{1:T}.$$

Dynamic Topic Model: Plug-in

- Replace $\mathbf{\Pi}^{1:T}$ by the observed frequencies $\boldsymbol{Y}^{1:T}$.
- Estimate A^* by $\hat{A}$ and build proxy random variables

$$\hat{\boldsymbol{W}}^{1:T} = (\hat{A}^\top \hat{A})^{-1} \hat{A}^\top \cdot \boldsymbol{Y}^{1:T}.$$

- Build estimators of c^*, $\tilde{\theta}^*$, and $\alpha := \|\theta^*\|_1$, based on $\hat{\boldsymbol{W}}^{1:T}$.

Dynamic Topic Model: Plug-in

- Replace $\mathbf{\Pi}^{1:T}$ by the observed frequencies $\boldsymbol{Y}^{1:T}$.
- Estimate A^* by $\hat{A}$ and build proxy random variables

$$\hat{\boldsymbol{W}}^{1:T} = (\hat{A}^\top \hat{A})^{-1} \hat{A}^\top \cdot \boldsymbol{Y}^{1:T}.$$

- Build estimators of c^*, $\tilde{\theta}^*$, and $\alpha := \|\theta^*\|_1$, based on $\hat{\boldsymbol{W}}^{1:T}$.

Three noise inducing steps:

Dynamic Topic Model: Plug-in

- Replace $\mathbf{\Pi}^{1:T}$ by the observed frequencies $\boldsymbol{Y}^{1:T}$.
- Estimate A^* by $\hat{A}$ and build proxy random variables

$$\hat{\boldsymbol{W}}^{1:T} = (\hat{A}^\top \hat{A})^{-1} \hat{A}^\top \cdot \boldsymbol{Y}^{1:T}.$$

- Build estimators of c^*, $\tilde{\theta}^*$, and $\alpha := \|\theta^*\|_1$, based on $\hat{\boldsymbol{W}}^{1:T}$.

Three noise inducing steps:

1. *Deviation of* $\hat{M} := (nT)^{-1} diag\left(\boldsymbol{Y}^{1:T} 1_{nT}\right)$ *from* $\boldsymbol{M}_* := (nT)^{-1} diag\left(\mathbf{\Pi}^{1:T} 1_{nT}\right)$

Dynamic Topic Model: Plug-in

- Replace $\mathbf{\Pi}^{1:T}$ by the observed frequencies $\boldsymbol{Y}^{1:T}$.
- Estimate A^* by $\hat{A}$ and build proxy random variables

$$\hat{\boldsymbol{W}}^{1:T} = (\hat{A}^\top \hat{A})^{-1} \hat{A}^\top \cdot \boldsymbol{Y}^{1:T}.$$

- Build estimators of c^*, $\tilde{\theta}^*$, and $\alpha := \|\theta^*\|_1$, based on $\hat{\boldsymbol{W}}^{1:T}$.

Three noise inducing steps:

1. *Deviation of* $\hat{M} := (nT)^{-1} diag\left(\boldsymbol{Y}^{1:T} 1_{nT}\right)$ *from* $\boldsymbol{M}_* := (nT)^{-1} diag\left(\mathbf{\Pi}^{1:T} 1_{nT}\right)$
2. *Deviation of* $[\hat{U}]_{.1}, \ldots, [\hat{U}]_{.K}$ *from* $[\boldsymbol{U}]_{.1}, \ldots, [\boldsymbol{U}]_{.K}$

Dynamic Topic Model: Plug-in

- Replace $\boldsymbol{\Pi}^{1:T}$ by the observed frequencies $\boldsymbol{Y}^{1:T}$.
- Estimate A^* by $\hat{A}$ and build proxy random variables

$$\hat{\boldsymbol{W}}^{1:T} = (\hat{A}^\top \hat{A})^{-1} \hat{A}^\top \cdot \boldsymbol{Y}^{1:T}.$$

- Build estimators of c^*, $\tilde{\theta}^*$, and $\alpha := \|\theta^*\|_1$, based on $\hat{\boldsymbol{W}}^{1:T}$.

Three noise inducing steps:

1. *Deviation of* $\hat{M} := (nT)^{-1} diag\left(\boldsymbol{Y}^{1:T} 1_{nT}\right)$ *from* $\boldsymbol{M}_* := (nT)^{-1} diag\left(\boldsymbol{\Pi}^{1:T} 1_{nT}\right)$
2. *Deviation of* $[\hat{U}]_{.1}, \ldots, [\hat{U}]_{.K}$ *from* $[\boldsymbol{U}]_{.1}, \ldots, [\boldsymbol{U}]_{.K}$
3. *Behaviour of the vertex hunting algorithm with noisy entries.*

Dynamic Topic Model: Theoretical guarantees

Theorem (B., Butucea, Ke 2024)

For N, n and T large enough, there exists χ, a positive constant only depending on K, such that with probability at least $1 - \frac{8}{nT}$:

$$\sum_{i=1}^{p} \left\| [\hat{A}]_{i.} - [A^*]_{i.} \right\|_1 \leq \chi \sqrt{\frac{p \log(nT) + p^2}{nT(N-2)}} p(1+p)(1 + \max_{x \in \mathcal{G}_\eta} \|x\|_2).$$

Dynamic Topic Model: Theoretical guarantees

Theorem (B., Butucea, Ke 2024)

For N, n and T large enough, and fixed number of topics K and of the vocabulary size p, with probability at least $1 - \frac{C}{nT}$:

$$\max\left\{\left\|\hat{\theta} - \tilde{\theta}^*\right\|_2, |\widehat{(1-c)} - (1-c^*)|, |\hat{\alpha} - \alpha^*|\right\}$$
$$\leq \mathcal{O}\left(\sqrt{\frac{\log(nT)}{n(T-1)}} + \sqrt{\frac{\log(nT)}{N}}\right).$$

Dynamic Topic Model: Theoretical guarantees

Theorem (B., Butucea, Ke 2024)

For N, n and T large enough, and fixed number of topics K and of the vocabulary size p, with probability at least $1 - \frac{C}{nT}$:

$$\max\left\{\left\|\hat{\theta} - \tilde{\theta}^*\right\|_2, |\widehat{(1-c)} - (1-c^*)|, |\hat{\alpha} - \alpha^*|\right\} \leq \mathcal{O}\left(\sqrt{\frac{\log(nT)}{n(T-1)}} + \sqrt{\frac{\log(nT)}{N}}\right).$$

The convergence rates show an additive behavior of the noise contained at different levels in the model.

Dynamic Topic Model: Theoretical guarantees

Theorem (B., Butucea, Ke 2024)

For N, n and T large enough, and fixed number of topics K and of the vocabulary size p, with probability at least $1 - \frac{C}{nT}$:

$$\max\left\{\left\|\hat{\theta} - \tilde{\theta}^*\right\|_2, |\widehat{(1-c)} - (1-c^*)|, |\hat{\alpha} - \alpha^*|\right\}$$
$$\leq \mathcal{O}\left(\sqrt{\frac{\log(nT)}{n(T-1)}} + \sqrt{\frac{\log(nT)}{N}}\right).$$

The convergence rates show an additive behavior of the noise contained at different levels in the model.
The bounds are driven by the Dirichlet noise and by the multinomial noise.