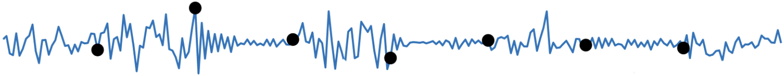
Learning Mixtures of Linear Dynamical Systems

Yanxi Chen, H. Vincent Poor

PRINCETON UNIVERSITY

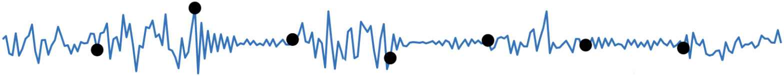
Mixtures of time-series models

Mixture models are powerful in the face of *heterogeneous and complex* time-series data



Mixtures of time-series models

Mixture models are powerful in the face of *heterogeneous and complex* time-series data



- Higher accuracy of fitting the data
- Better interpretability: reveal cluster structures

Numerous applications

(Bulteel et al., 2016) Time-series measurements of certain psychological symptoms for multiple patients $\rightarrow$ identify *subgroups of patients*, provide tailored treatments.

Numerous applications

(Bulteel et al., 2016) Time-series measurements of certain psychological symptoms for multiple patients $\rightarrow$ identify *subgroups of patients*, provide tailored treatments.

(Hallac et al., 2017) Sensory data of a car under *a few driving modes* (e.g. “driving straight”, “slowing down”, “turning”, etc.)

Numerous applications

(Bulteel et al., 2016) Time-series measurements of certain psychological symptoms for multiple patients $\rightarrow$ identify *subgroups of patients*, provide tailored treatments.

(Hallac et al., 2017) Sensory data of a car under *a few driving modes* (e.g. "driving straight", "slowing down", "turning", etc.)

(Brunskill et al., 2009) Sensory data of a robot navigating *a complex environment* (e.g. with areas of grass, sand, carpets, rocks, etc.)

......

Problem formulation

Recap: linear dynamical system (LDS). A d-dimensional time-series trajectory $\{\boldsymbol{x}_t\}$ generated by an LDS model, i.e. the $d \times d$ state transition matrix $\boldsymbol{A}$ and noise covariance $\boldsymbol{W}$:

Problem formulation

Recap: linear dynamical system (LDS). A d-dimensional time-series trajectory $\{\boldsymbol{x}_t\}$ generated by an LDS model, i.e. the $d \times d$ state transition matrix $\boldsymbol{A}$ and noise covariance $\boldsymbol{W}$:

$$\boldsymbol{x}_{t+1} = \boldsymbol{A}\boldsymbol{x}_t + \boldsymbol{w}_t, \quad \text{where} \quad \mathbb{E}[\boldsymbol{w}_t] = \boldsymbol{0}, \ \mathrm{cov}(\boldsymbol{w}_t) = \boldsymbol{W} \succ \boldsymbol{0}.$$

Problem formulation

Recap: linear dynamical system (LDS). A d-dimensional time-series trajectory $\{\boldsymbol{x}_t\}$ generated by an LDS model, i.e. the $d \times d$ state transition matrix $\boldsymbol{A}$ and noise covariance $\boldsymbol{W}$:

$$\boldsymbol{x}_{t+1} = \boldsymbol{A}\boldsymbol{x}_t + \boldsymbol{w}_t, \quad \text{where} \quad \mathbb{E}[\boldsymbol{w}_t] = \boldsymbol{0},\ \mathsf{cov}(\boldsymbol{w}_t) = \boldsymbol{W} \succ \boldsymbol{0}.$$

Mixed LDSs. K models $\{\boldsymbol{A}^{(k)}, \boldsymbol{W}^{(k)}\}_{1 \le k \le K}$, M trajectories $\{\boldsymbol{X}_m\}_{1 \le m \le M}$, where $\boldsymbol{X}_m = \{\boldsymbol{x}_{m,t}\}$ is generated by the k_m-th model:

$$\boldsymbol{x}_{m,t+1} = \boldsymbol{A}^{(k_m)}\boldsymbol{x}_{m,t} + \boldsymbol{w}_{m,t}, \quad \mathsf{cov}(\boldsymbol{w}_{m,t}) = \boldsymbol{W}^{(k_m)}$$

Problem formulation

Recap: linear dynamical system (LDS). A d-dimensional time-series trajectory $\{\boldsymbol{x}_t\}$ generated by an LDS model, i.e. the $d \times d$ state transition matrix $\boldsymbol{A}$ and noise covariance $\boldsymbol{W}$:

$$\boldsymbol{x}_{t+1} = \boldsymbol{A}\boldsymbol{x}_t + \boldsymbol{w}_t, \quad \text{where} \quad \mathbb{E}[\boldsymbol{w}_t] = \boldsymbol{0},\ \mathsf{cov}(\boldsymbol{w}_t) = \boldsymbol{W} \succ \boldsymbol{0}.$$

Mixed LDSs. K models $\{\boldsymbol{A}^{(k)}, \boldsymbol{W}^{(k)}\}_{1 \le k \le K}$, M trajectories $\{\boldsymbol{X}_m\}_{1 \le m \le M}$, where $\boldsymbol{X}_m = \{\boldsymbol{x}_{m,t}\}$ is generated by the k_m-th model:

$$\boldsymbol{x}_{m,t+1} = \boldsymbol{A}^{(k_m)}\boldsymbol{x}_{m,t} + \boldsymbol{w}_{m,t}, \quad \mathsf{cov}(\boldsymbol{w}_{m,t}) = \boldsymbol{W}^{(k_m)}$$

Note that the labels $\{k_m\}$ are **unknown**!

Literature: lack of **provable guarantees** for model estimation

Literature: lack of **provable guarantees** for model estimation

Major challenges:

- Latent variables are not observed;

Literature: lack of **provable guarantees** for model estimation

Major challenges:

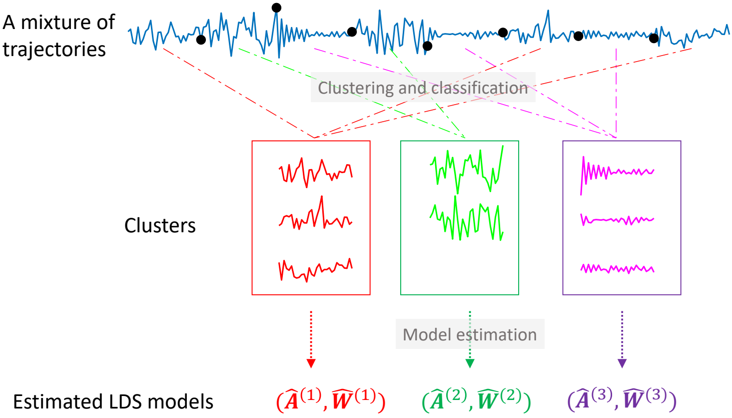
- Latent variables are not observed;
- Short trajectories might have lengths much smaller than the model dimension d;

Literature: lack of **provable guarantees** for model estimation

Major challenges:

- Latent variables are not observed;
- Short trajectories might have lengths much smaller than the model dimension d;
- Temporal dependence inherent to time series (in contrast to mixed regression problems).

Outline of our solution



A two-stage approach

Stage 1: coarse estimation

- **Subspace estimation**
- **Clustering** of trajectories (assisted by variance reduction)
- Initial **model estimation** within each cluster

A two-stage approach

Stage 1: coarse estimation

- **Subspace estimation**
- **Clustering** of trajectories (assisted by variance reduction)
- Initial **model estimation** within each cluster

Stage 2: refined estimation

- **Classification** of additional trajectories
- Refined **model estimation** within each cluster

A two-stage approach

Stage 1: coarse estimation

- **Subspace estimation**
- **Clustering** of trajectories (assisted by variance reduction)
- Initial **model estimation** within each cluster

Stage 2: refined estimation

- **Classification** of additional trajectories
- Refined **model estimation** within each cluster

The algorithm outline is largely inspired by the works on meta-learning for mixed linear regression (Kong et al., 2020a;b), but the detailed implementations are substantially different due to temporal dependence in mixed LDSs; see Section 2 of paper for detailed algorithms.

Assumptions for simplification

Initial state: each trajectory starts at $\boldsymbol{x}_{m,0} = \boldsymbol{0}$. (Another canonical case is when the short trajectories are segments of a single continuous trajectory; the main results are slightly different, and included in the paper.)

Assumptions for simplification

Initial state: each trajectory starts at $\boldsymbol{x}_{m,0} = \mathbf{0}$. (Another canonical case is when the short trajectories are segments of a single continuous trajectory; the main results are slightly different, and included in the paper.)

Balanced clusters: each LDS model accounts for (order-wise) $1/K$ proportion of data.

Assumptions for simplification

Initial state: each trajectory starts at $\boldsymbol{x}_{m,0} = \mathbf{0}$. (Another canonical case is when the short trajectories are segments of a single continuous trajectory; the main results are slightly different, and included in the paper.)

Balanced clusters: each LDS model accounts for (order-wise) $1/K$ proportion of data.

Sample splitting: M sample trajectories,

$$\{1, 2, \ldots, M\} = \mathcal{M}_{\text{subspace}} \cup \mathcal{M}_{\text{clustering}} \cup \mathcal{M}_{\text{classification}}.$$

Assume that each trajectory in $\mathcal{M}_{\text{o}}$ has the same length T_{o}, and denote the total sample size of $\mathcal{M}_{\text{o}}$ as $T_{\text{total,o}} = T_{\text{o}} \cdot |\mathcal{M}_{\text{o}}|$.

Essential assumptions

Mixing: for each $\boldsymbol{A} \in \{\boldsymbol{A}^{(k)}\}$ and all $t \geq 1$, $\|\boldsymbol{A}^t\| \leq \kappa_A \cdot \rho^t$ for some $0 \leq \rho < 1$; denote mixing time $t_{\mathsf{mix}} := 1/(1-\rho)$.

Essential assumptions

Mixing: for each $\boldsymbol{A} \in \{\boldsymbol{A}^{(k)}\}$ and all $t \geq 1$, $\|\boldsymbol{A}^t\| \leq \kappa_A \cdot \rho^t$ for some $0 \leq \rho < 1$; denote mixing time $t_{\mathsf{mix}} := 1/(1-\rho)$.

Stationary autocovariance matrices $\{\boldsymbol{\Gamma}^{(k)}, \boldsymbol{Y}^{(k)}\}$, where

$$\boldsymbol{\Gamma}(\boldsymbol{A}, \boldsymbol{W}) := \mathbb{E}\big[\boldsymbol{x}_t \boldsymbol{x}_t{}^\top | \boldsymbol{A}, \boldsymbol{W}\big],$$
$$\boldsymbol{Y}(\boldsymbol{A}, \boldsymbol{W}) := \mathbb{E}\big[\boldsymbol{x}_{t+1} \boldsymbol{x}_t{}^\top | \boldsymbol{A}, \boldsymbol{W}\big].$$

Essential assumptions

Mixing: for each $\boldsymbol{A} \in \{\boldsymbol{A}^{(k)}\}$ and all $t \geq 1$, $\|\boldsymbol{A}^t\| \leq \kappa_A \cdot \rho^t$ for some $0 \leq \rho < 1$; denote mixing time $t_{\mathsf{mix}} := 1/(1-\rho)$.

Stationary autocovariance matrices $\{\boldsymbol{\Gamma}^{(k)}, \boldsymbol{Y}^{(k)}\}$, where

$$\boldsymbol{\Gamma}(\boldsymbol{A}, \boldsymbol{W}) := \mathbb{E}\big[\boldsymbol{x}_t \boldsymbol{x}_t{}^\top | \boldsymbol{A}, \boldsymbol{W}\big],$$
$$\boldsymbol{Y}(\boldsymbol{A}, \boldsymbol{W}) := \mathbb{E}\big[\boldsymbol{x}_{t+1} \boldsymbol{x}_t{}^\top | \boldsymbol{A}, \boldsymbol{W}\big].$$

Model separation: there exist $\delta_{\Gamma,Y}, \delta_{A,W} > 0$ such that

$$\|\boldsymbol{\Gamma}^{(k)} - \boldsymbol{\Gamma}^{(\ell)}\|_{\mathrm{F}}^2 + \|\boldsymbol{Y}^{(k)} - \boldsymbol{Y}^{(\ell)}\|_{\mathrm{F}}^2 \geq d \cdot \delta_{\Gamma,Y}^2,$$
$$\|\boldsymbol{A}^{(k)} - \boldsymbol{A}^{(\ell)}\|_{\mathrm{F}}^2 + \|\boldsymbol{W}^{(k)} - \boldsymbol{W}^{(\ell)}\|_{\mathrm{F}}^2 \geq d \cdot \delta_{A,W}^2,$$

for all $1 \leq k < \ell \leq K$

Theorem. With high probability, the proposed two-stage method achieves exact clustering and classification of the sample trajectories, as well as final model estimation errors

$$\|\widehat{\boldsymbol{A}}^{(k)} - \boldsymbol{A}^{(k)}\| \le \epsilon, \quad \frac{\|\widehat{\boldsymbol{W}}^{(k)} - \boldsymbol{W}^{(k)}\|}{\|\boldsymbol{W}^{(k)}\|} \le \epsilon, \quad 1 \le k \le K,$$

Theorem. With high probability, the proposed two-stage method achieves exact clustering and classification of the sample trajectories, as well as final model estimation errors

$$\|\widehat{\boldsymbol{A}}^{(k)} - \boldsymbol{A}^{(k)}\| \le \epsilon, \quad \frac{\|\widehat{\boldsymbol{W}}^{(k)} - \boldsymbol{W}^{(k)}\|}{\|\boldsymbol{W}^{(k)}\|} \le \epsilon, \quad 1 \le k \le K,$$

provided the following sample complexities:

$$T_{\text{subspace}} \gtrsim t_{\text{mix}}, \quad T_{\text{total,subspace}} \gtrsim t_{\text{mix}} d \left(\frac{K^4}{\delta_{\Gamma,Y}^4} + 1 \right),$$

$$T_{\text{clustering}} \gtrsim t_{\text{mix}} \left(\frac{1}{\delta_{\Gamma,Y}^2} \sqrt{\frac{K}{d}} + 1 \right), \quad T_{\text{total,clustering}} \gtrsim K d \left(\frac{1}{\delta_{A,W}^2} + 1 \right),$$

$$T_{\text{classification}} \gtrsim \frac{1}{d \delta_{A,W}^2} + 1, \quad T_{\text{total,clustering}} + T_{\text{total,classification}} \gtrsim \frac{Kd}{\epsilon^2}.$$

Theorem. With high probability, the proposed two-stage method achieves exact clustering and classification of the sample trajectories, as well as final model estimation errors

$$\|\widehat{\boldsymbol{A}}^{(k)} - \boldsymbol{A}^{(k)}\| \leq \epsilon, \quad \frac{\|\widehat{\boldsymbol{W}}^{(k)} - \boldsymbol{W}^{(k)}\|}{\|\boldsymbol{W}^{(k)}\|} \leq \epsilon, \quad 1 \leq k \leq K,$$

provided the following sample complexities:

$$T_{\text{subspace}} \gtrsim t_{\text{mix}}, \quad T_{\text{total,subspace}} \gtrsim t_{\text{mix}} d \left(\frac{K^4}{\delta_{\Gamma,Y}^4} + 1 \right),$$

$$T_{\text{clustering}} \gtrsim t_{\text{mix}} \left(\frac{1}{\delta_{\Gamma,Y}^2} \sqrt{\frac{K}{d}} + 1 \right), \quad T_{\text{total,clustering}} \gtrsim K d \left(\frac{1}{\delta_{A,W}^2} + 1 \right),$$

$$T_{\text{classification}} \gtrsim \frac{1}{d \delta_{A,W}^2} + 1, \quad T_{\text{total,clustering}} + T_{\text{total,classification}} \gtrsim \frac{Kd}{\epsilon^2}.$$

See Section 3 of paper for formal theorems.

Summary

- Problem formulation of mixed LDSs;
- A two-stage approach for solving it;
- Theoretical guarantees with non-asymptotic sample complexities.

Future works

- Strengthening the theoretical analysis and algorithm design.

Future works

- Strengthening the theoretical analysis and algorithm design.
- Learning mixtures of more general time-series models. (Our algorithms essentially require (1) mixing; (2) the existence of stationary autocovariance matrices; (3) well-specified parametric models, and sufficient separation among them.)

Future works

- Strengthening the theoretical analysis and algorithm design.
- Learning mixtures of more general time-series models. (Our algorithms essentially require (1) mixing; (2) the existence of stationary autocovariance matrices; (3) well-specified parametric models, and sufficient separation among them.)
- Applications in real-world problems.

Future works

- Strengthening the theoretical analysis and algorithm design.
- Learning mixtures of more general time-series models. (Our algorithms essentially require (1) mixing; (2) the existence of stationary autocovariance matrices; (3) well-specified parametric models, and sufficient separation among them.)
- Applications in real-world problems.
- Extensions to the cases with controlled inputs, e.g. LQR in control and latent MDP in reinforcement learning.

Future works

- Strengthening the theoretical analysis and algorithm design.
- Learning mixtures of more general time-series models. (Our algorithms essentially require (1) mixing; (2) the existence of stationary autocovariance matrices; (3) well-specified parametric models, and sufficient separation among them.)
- Applications in real-world problems.
- Extensions to the cases with controlled inputs, e.g. LQR in control and latent MDP in reinforcement learning.

Thank you!