Learning Linear Dynamical Systems from Multivariate Time Series: A Matrix Factorization Based Framework

Zitao Liu and Milos Hauskrecht
{ztliu, milos}@cs.pitt.edu
Department of Computer Science, University of Pittsburgh

Introduction

The linear dynamical system (LDS) model is arguably the most commonly used time series model for real-world engineering and financial applications due to its:

(1) relative simplicity
(2) mathematically predictable behavior
(3) exact inference and predictions for the model can be done efficiently

In this work, we propose a new generalized LDS framework, gLDS, for learning LDS models from a collection of multivariate time series (MTS) data based on matrix factorization, which is different from traditional EM learning and spectral learning algorithms.

Contribution

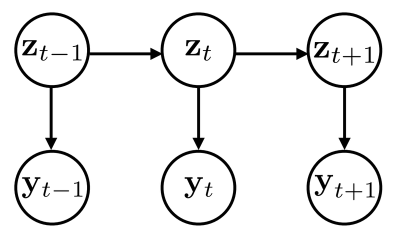
✓ Learn LDS from a collection of many different MTS sequences compared to spectral learning algorithm.
✓ Parameter optimization in gLDS can be done efficiently in each iteration compared to EM algorithm.
✓ Constraints can be easily incorporated into our generalized framework.
✓ A novel temporal smoothing regularization and a smooth LDS model are proposed.

Background

Linear Dynamical Systems(LDS)

The LDS is arguably the most commonly used MTS model for real-world applications.

$$\mathbf{z}_t = A\mathbf{z}_{t-1} + \epsilon_t, \quad \mathbf{y}_t = C\mathbf{z}_t + \zeta_t$$



- Real-valued observations $\{\mathbf{y}_t \in \mathbb{R}^n\}_{t=1}^T$.
- Hidden states $\{\mathbf{z}_t \in \mathbb{R}^d\}_{t=1}^T$.
- Transition matrix $A \in \mathbb{R}^{d\times d}$.
- Emission matrix $C \in \mathcal{R}^{n\times d}$.
- $\{\epsilon_t\}_{t=1}^T \sim \mathcal{N}(\mathbf{0}, Q)$.
- $\{\zeta_t\}_{t=1}^T \sim \mathcal{N}(\mathbf{0}, R)$.
- Initial state $\mathbf{z}_1 \sim \mathcal{N}(\boldsymbol{\xi}, \Psi)$.
- LDS parameters $\Omega = \{A, C, Q, R, \boldsymbol{\xi}, \Psi\}$.

Regularized LDS(rLDS)

Learn LDS models with a low-rank transition matrix from a limited number of MTS sequences. rLDS is able to:

(1) find the intrinsic dimensionality of the hidden state.
(2) prevent the overfitting problem whenever the amount of MTS data is small.

Stable LDS

LDS with dynamic matrix A is stable if all of A's eigenvalues have magnitude at most 1. Stable LDS is able to:

(1) simulate long sequences from LDS models in order to generate representative data.
(2) infer stretches of missing values.

The Generalized LDS Learning Framework

We propose a generalized framework, gLDS, for learning LDS models based on matrix factorization. In gLDS,

(1) the LDS models can be learned from multiple MTS sequences.
(2) various constraints can be easily incorporated into the learning process.

Notation

Given a collection of *N* multivariate time series sequences $\{\mathbf{Y}^1, \mathbf{Y}^2, \cdots, \mathbf{Y}^N\}$,

- $\mathbf{Y}^m = [\mathbf{y}_1^m, \cdots, \mathbf{y}_t^m, \cdots, \mathbf{y}_{T_m}^m]$, $\mathbf{Y}^m \in \mathcal{R}^{n\times T_m}$, $\mathbf{y}_t^m \in \mathcal{R}^{n\times 1}$.
- $\mathbf{Z}^m = [\mathbf{z}_1^m, \cdots, \mathbf{z}_t^m, \cdots, \mathbf{y}_{T_m}^m]$, $\mathbf{Z}^m \in \mathcal{R}^{d\times T_m}$, $\mathbf{z}_t^m \in \mathcal{R}^{d\times 1}$.
- *n* is the number of variables; T_m is the length of *m*thsequence; d is the dimension of hidden state.
- $\mathbf{Z}_+^m = [\mathbf{z}_2^m, \mathbf{z}_3^m, \cdots, \mathbf{z}_{T_m}^m]$ and $\mathbf{Z}_-^m = [\mathbf{z}_1^m, \mathbf{z}_2^m, \cdots, \mathbf{z}_{T_m-1}^m]$.
- We use $\mathbf{Y}$, $\mathbf{Z}$, $\mathbf{Z}_+$, and $\mathbf{Z}_-$ to denote the horizontal concatenations of $\{\mathbf{Y}^m\}$, $\{\mathbf{Z}^m\}$, $\{\mathbf{Z}_+^m\}$, and $\{\mathbf{Z}_-^m\}$.
- $\mathbf{Y} \in \mathcal{R}^{n\times T}$, $\mathbf{Z} \in \mathcal{R}^{d\times T}$, $\mathbf{Z}_+ \in \mathcal{R}^{d\times(T-N)}$, $\mathbf{Z}_- \in \mathcal{R}^{d\times(T-N)}$. $T = \sum_{m=1}^N T_m$.

gLDS Framework

Linear assumption in LDS: the sequential observation vector is generated by the linear emission transformation C from hidden states at each time stamp.

$$\min_{C,\mathbf{Z}} \|\mathbf{Y} - C\mathbf{Z}\|_F^2 \tag{1}$$

Hidden factors are time dependent and they evolve with time. We estimate the transition matrix A by solving another least square problem as follows:

$$\min_{A,\mathbf{Z}} \|\mathbf{Z}_+ - A\mathbf{Z}_-\|_F^2 \tag{2}$$

Therefore, we jointly optimize both eq.(1) and eq.(2). Furthermore, in order to incorporate constraints into the learned LDS models, we add regularizations for C, A and $\mathbf{Z}$ into the objective function, shown as follows:

$$\min_{A,C,\mathbf{Z}} \|\mathbf{Y} - C\mathbf{Z}\|_F^2 + \lambda\|\mathbf{Z}_+ - A\mathbf{Z}_-\|_F^2 + \alpha\mathcal{R}_C(C) + \beta\mathcal{R}_\mathbf{Z}(\mathbf{Z}) + \gamma\mathcal{R}_A(A) \tag{3}$$

Intuitively, this formulation of the problem aims to find an LDS model that is able to fit as accurately as possible the time series in the training data by using a simple (less complex) model.

Learning - Optimization of A, C, and $\mathbf{Z}$

We apply the alternating minimization techniques to eq.(3), which leads to the following three optimization problems:

$$\min_{\mathbf{Z}} \|\mathbf{Y} - C\mathbf{Z}\|_F^2 + \lambda\|\mathbf{Z}_+ - A\mathbf{Z}_-\|_F^2 + \beta\mathcal{R}_\mathbf{Z}(\mathbf{Z}); \quad \min_{C} \|\mathbf{Y} - C\mathbf{Z}\|_F^2 + \alpha\mathcal{R}_C(C)$$

$$\min_{A} \|\mathbf{Z}_+ - A\mathbf{Z}_-\|_F^2 + \gamma/\lambda\mathcal{R}_A(A) \tag{4}$$

Due to the asymmetric positions of different $\mathbf{z}_t^m$s in $\mathbf{Z}^m$, we decompose the optimization into three parts:

$$\min_{\mathbf{z}_1^m} \|\mathbf{y}_1^m - C\mathbf{z}_1^m\|_2^2 + \lambda\|\mathbf{z}_2^m - A\mathbf{z}_1^m\|_2^2 + \beta\mathcal{R}_\mathbf{Z}(\mathbf{z}_1^m); \quad \min_{\mathbf{z}_{T_m}^m} \|\mathbf{y}_{T_m}^m - C\mathbf{z}_{T_m}^m\|_2^2 + \lambda\|\mathbf{z}_{T_m}^m - A\mathbf{z}_{T_m-1}^m\|_2^2 + \beta\mathcal{R}_\mathbf{Z}(\mathbf{z}_{T_m}^m)$$

$$\min_{\mathbf{z}_t^m} \|\mathbf{y}_t^m - C\mathbf{z}_t^m\|_2^2 + \lambda\|\mathbf{z}_t^m - A\mathbf{z}_{t-1}^m\|_2^2 + \lambda\|\mathbf{z}_{t+1}^m - A\mathbf{z}_t^m\|_2^2 + \beta\mathcal{R}_\mathbf{Z}(\mathbf{z}_t^m)$$

Learning - Optimization of R, Q, $\boldsymbol{\xi}$ and Ψ

Once we obtain A, C and $\mathbf{Z}$, the rest of LDS's parameters, R, Q, $\boldsymbol{\xi}$ and Ψ, can be analytically estimated as follows:

$$\hat{Q} = \frac{1}{T-N}(\hat{\mathbf{Z}}_+ - \hat{A}\hat{\mathbf{Z}}_-)(\hat{\mathbf{Z}}_+ - \hat{A}\hat{\mathbf{Z}}_-)^\top$$

$$\hat{\boldsymbol{\xi}} = \frac{1}{N}\sum_{m=1}^N \hat{\mathbf{z}}_1^m$$

$$\hat{R} = \frac{1}{T}(\mathbf{Y} - \hat{C}\hat{\mathbf{Z}})(\mathbf{Y} - \hat{C}\hat{\mathbf{Z}})^\top$$

$$\hat{\Psi} = \frac{1}{N}\sum_{m=1}^N \hat{\mathbf{z}}_1^m(\hat{\mathbf{z}}_1^m)^\top$$

The Ridge Model (gLDS-ridge)

In our gLDS framework, we achieve the ridge model (gLDS-ridge) by setting $\mathcal{R}_C(C)$, $\mathcal{R}_A(A)$, and $\mathcal{R}_\mathbf{Z}(\mathbf{Z})$ to the square of Frobenius norm, i.e., $\|\cdot\|_F^2$. Due to the differentiability of ridge regularization, we can take the partial derivatives, set them to zero and solve. The results are shown as follows:

$$\hat{A} = (\mathbf{Z}_+\mathbf{Z}_-^\top)(\mathbf{Z}_-\mathbf{Z}_-^\top + \gamma/\lambda I_d)^{-1}; \quad \hat{C} = (\mathbf{Y}\mathbf{Z}^\top)(\mathbf{Z}\mathbf{Z}^\top + \alpha I_d)^{-1}; \quad \hat{\mathbf{z}}_{T_m}^m = (G + \lambda I_d)^{-1}F_{T_m}^m$$

$$\hat{\mathbf{z}}_1^m = (G + \lambda A^\top A)^{-1}(C^\top\mathbf{y}_1^m + \lambda A^\top\mathbf{z}_2^m); \quad \hat{\mathbf{z}}_t^m = (G + \lambda A^\top A + \lambda I_d)^{-1}(F_t^m + \lambda A^\top\mathbf{z}_{t+1}^m)$$

where $G = C^\top C + \beta I_d$ and $F_t^m = C^\top\mathbf{y}_t^m + \lambda A\mathbf{z}_{t-1}^m$.

Existing Models in gLDS Framework

Learning Regularized LDS (gLDS-low-rank)

In gLDS framework, a low-rank transition matrix A can be easily obtained by setting $\mathcal{R}_A(A) = \|A\|_F^2 + \frac{\lambda}{\gamma}\gamma_A\|A\|_*$, which leads to the following objective function:

$$\min_A g(A) + \gamma_A\|A\|_* \quad \text{where} \quad g(A) = \|\mathbf{Z}_+ - A\mathbf{Z}_-\|_F^2 + \gamma/\lambda\|A\|_F^2$$

Since $g(A)$ is convex and differentiable with respect to *A*, we can adopt the generalized gradient descent algorithm to minimize the above equation.

Learning Stable LDS (gLDS-stable)

[1] proposes a novel method for learning stable LDS models by formulating the problem as a quadratic program. In gLDS framework, by setting $\mathcal{R}_A(A) = \emptyset$, we can easily transform our optimization to the same objective function in [1]. Furthermore, we can apply the following theorem to change eq.(4) into the standard quadratic program formulation.

Theorem 1. Minimizing A from eq.(4) with $\mathcal{R}_A(A) = \emptyset$ is equivalent to minimizing the following problem:

$$\min_a a^\top Ba - 2q^\top a \tag{5}$$

where $a = \mathrm{vec}(A^\top)$, $B = I_d \otimes (\mathbf{Z}_-\mathbf{Z}_-^\top)$, $q = (I_d \otimes \mathbf{Z}_-\mathbf{Z}_+^\top)\,\mathrm{vec}(I_d)$.

The Smooth Model (gLDS-smooth)

We propose a novel temporal smoothing regularization, which penalizes the difference of predictive results from the learned model during the learning phase, to achieve smooth forecasts from the learned LDS models.

Temporal Smoothing Regularization

We propose a temporal smoothing regularization term $\mathcal{R}_\mathcal{T}^m$ for each MTS sequence *m*:

$$\mathcal{R}_\mathcal{T}^m = \frac{1}{2}\sum_{i=1}^{T_m}\sum_{j=1}^{T_m} w_{i,j}^m\|\hat{\mathbf{y}}_i^m - \hat{\mathbf{y}}_j^m\|_2^2 = \sum_{l=1}^n \hat{\mathbf{Y}}^m(l,:)(D^m - W^m)\hat{\mathbf{Y}}^m(l,:)^\top = Tr[C\mathbf{Z}^m L^m(\mathbf{Z}^m)^\top C^\top]$$

In order to learn a smooth LDS model, we apply the temporal smooth regularization to each MTS sequence in the training data, which leads to the following compact form of regularization:

$$\mathcal{R}_\mathcal{T} = \sum_{m=1}^N \mathcal{R}_\mathcal{T}^m = Tr[C\mathbf{Z}P\mathbf{Z}^\top C^\top] \tag{6}$$

where *P* is the $T\times T$ block diagonal matrix with *N* blocks and the *m*th block component is the Laplacian matrix L^m for mth MTS sequence.

Learning Smooth LDS

We incorporate the temporal smooth regularization (eq.(6)) into the objective function (eq.(3)). Here similar to gLDS-ridge, we set $\mathcal{R}_C(C)$, $\mathcal{R}_A(A)$, and $\mathcal{R}_\mathbf{Z}(\mathbf{Z})$ to the ridge regularizations (square of Frobenius norm), which leads to the following new learning objective function:

$$\min_{A,C,\mathbf{Z}} \|\mathbf{Y} - C\mathbf{Z}\|_F^2 + \lambda\|\mathbf{Z}_+ - A\mathbf{Z}_-\|_F^2 + \alpha\|C\|_F^2 + \beta\|\mathbf{Z}\|_F^2 + \gamma\|A\|_F^2 + \delta Tr[C\mathbf{Z}P\mathbf{Z}^\top C^\top] \tag{7}$$

Since the temporal smoothing regularization only involves C and $\mathbf{Z}$, the update rules for A, R, Q, $\boldsymbol{\xi}$ and $\boldsymbol{\Psi}$ remain the same. The update rules for C and $\mathbf{Z}$ are shown as follows:

$$\hat{C} = (\mathbf{Y}\mathbf{Z}^\top)(\mathbf{Z}\mathbf{Z}^\top + \delta\mathbf{Z}P\mathbf{Z}^\top + \alpha I_d)^{-1}; \quad \hat{\mathbf{z}}_1^m = \left(\Gamma_1^m + \lambda A^\top A\right)^{-1}\left(\Phi_1^m + \lambda A^\top\mathbf{z}_2^m\right)$$

$$\hat{\mathbf{z}}_t^m = \left(\Gamma_t^m + \lambda A^\top A + \lambda I_d\right)^{-1}\left(\Phi_t^m + \lambda A^\top\mathbf{z}_{t+1}^m + \lambda A\mathbf{z}_{t-1}^m\right); \quad \hat{\mathbf{z}}_{T_m}^m = \left(\Gamma_{T_m}^m + \lambda I_d\right)^{-1}\left(\Phi_{T_m}^m + \lambda A\mathbf{z}_{T_m-1}^m\right)$$

where

$$\Gamma_t^m = (1 + \delta L_{t,t} - \delta W_{t,t})C^\top C + \beta I_d; \quad \Phi_t^m = C^\top\mathbf{y}_t^m + \delta C^\top C\sum_{j\neq t} W_{t,j}\mathbf{z}_j^m$$

Experiments

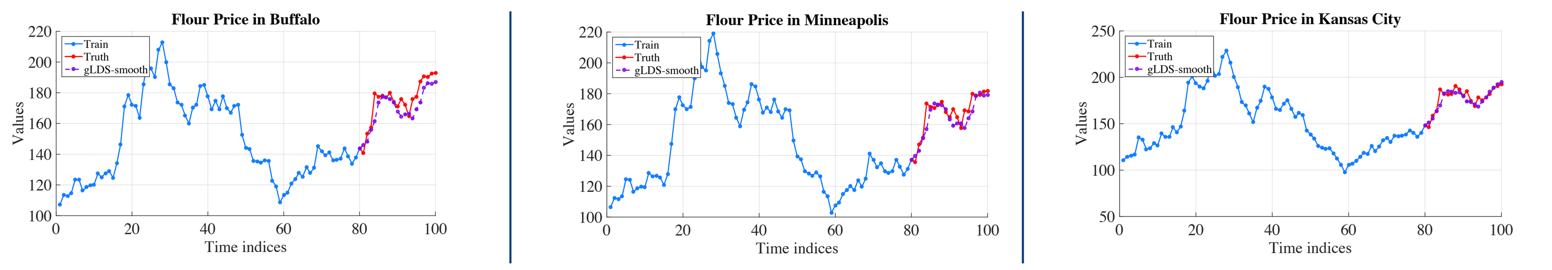
In the following experiments, we smooth only the pairs of two consecutive forecasts, i.e., $w_{ij}^m = 1$ if $|i-j| = 1$; otherwise, $w_{ij}^m = 0$. Experiments are conducted on the following four data sets:

- Flour price data (*flourprice*). It is a monthly flour price indices data, which contains the flour price series in Buffalo, Minneapolis and Kansas City, from August 1972 to November 1980.
- Evap data (*evap*). The evaporation data contains the daily amounts of water evaporated, temperature, and barometric pressure from 10/11/1692 to 09/11/1693.
- H2O evap data (*h2o evap*). It contains six MTS variables: the amount of evaporation, total global radiation, estimated net radiation, saturation deficit at max temperature, mean daily wind speed and saturation deficit at mean temperature.
- Clinical data (*clinical*). A MTS clinical data obtained from electronic health records of post-surgical cardiac patients in PCP database [2].

Evaluation Metric

$$\text{Average Mean Absolute Percentage Error (Avg-MAPE)} = \frac{1}{nTN}\sum_{m=1}^N\sum_{l=1}^n\sum_{t=1}^{T_m}|1 - \hat{y}_{l,t}^m/y_{l,t}^m| \times 100\%$$

Qualitative Prediction Analysis



Quantitative Prediction Analysis

We quantitatively compute and compare the prediction accuracy of the proposed methods (gLDS-ridge and gLDS-smooth) with the standard LDS learning approaches: EM and spectral algorithms.

	Training: 80%		Training: 90%	
# of states	5	10	5	10
Spectral	24.62	24.85	25.08	26.28
EM	17.68	14.45	16.32	17.35
gLDS-ridge	10.58	10.35	13.60	14.05
gLDS-smooth	**10.35**	**10.27**	**13.39**	**13.68**

Table 1: Average-MAPE on *evap* dataset.

	Training: 80%		Training: 90%	
# of states	5	10	5	10
Spectral	6.25	5.86	6.61	5.93
EM	3.62	4.15	3.63	3.94
gLDS-ridge	3.37	3.14	3.29	2.82
gLDS-smooth	**3.24**	**2.71**	**2.86**	**2.50**

Table 2: Average-MAPE on *flourprice* dataset.

	Training: 80%		Training: 90%	
# of states	5	10	5	10
Spectral	36.26	32.20	13.73	15.88
EM	39.53	68.68	17.33	17.46
gLDS-ridge	27.97	28.53	16.12	14.42
gLDS-smooth	**26.38**	**26.46**	**14.01**	**14.08**

Table 3: Average-MAPE on *h2o_evap* dataset.

Stability Effects of gLDS-stable

We show the stability effects of the gLDS-stable model learned using our framework by generating the simulated sequences in the future.

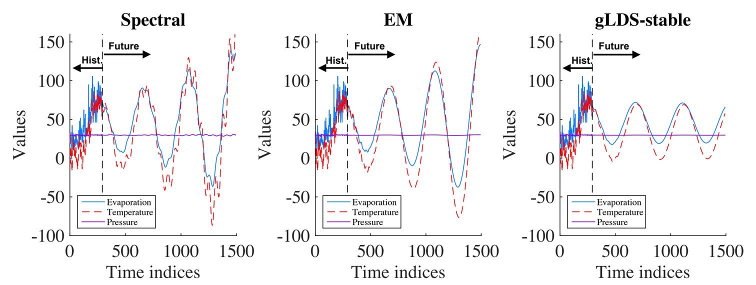


Figure 1: Training data and simulated sequences from gLDS-stable in *evap*.

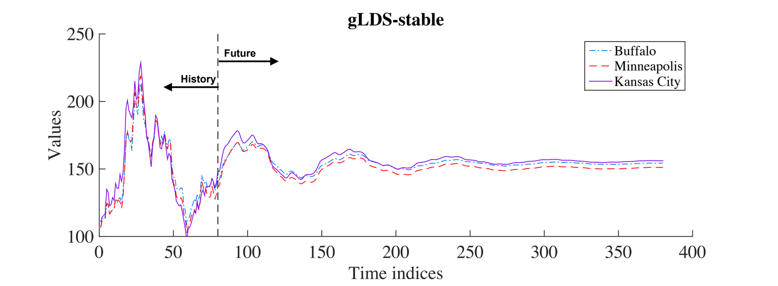


Figure 2: Training data and simulated sequences from gLDS-stable in *fourprice*.

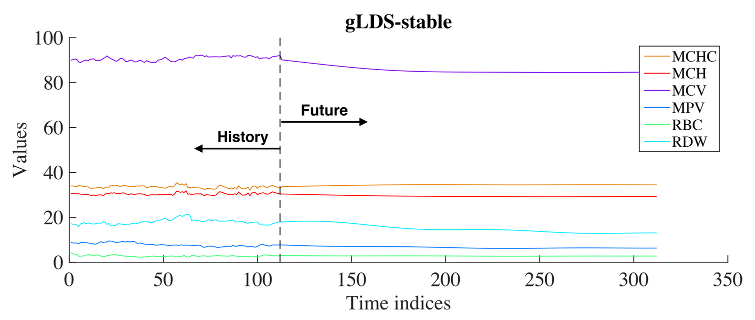


Figure 3: Training data and simulated sequences from gLDS-stable in *clinical*.

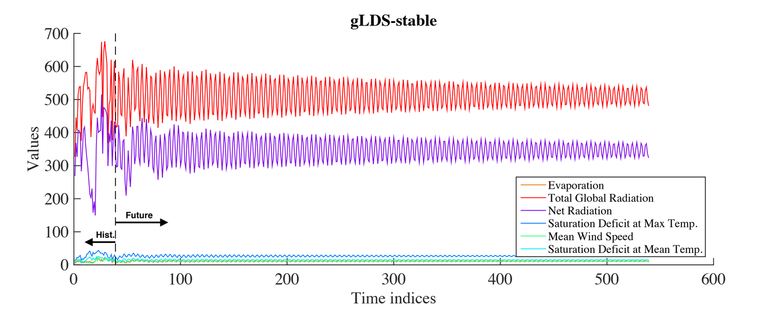


Figure 4: Training data and simulated sequences from gLDS-stable in *h2o evap*.

Sparsification Effects of gLDS-low-rank

We show the sparsification effects of the gLDS-low-rank model learned using our framework. The gLDS-low-rank model is able to identify the intrinsic dimensionality of the hidden state space.

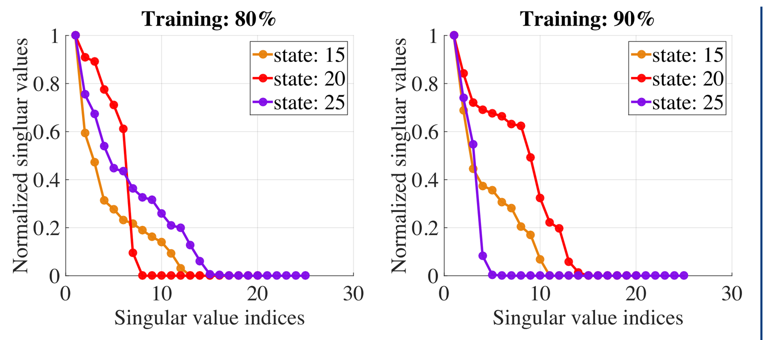


Figure 5: Intrinsic dimensionality recovery in *flourprice* dataset.

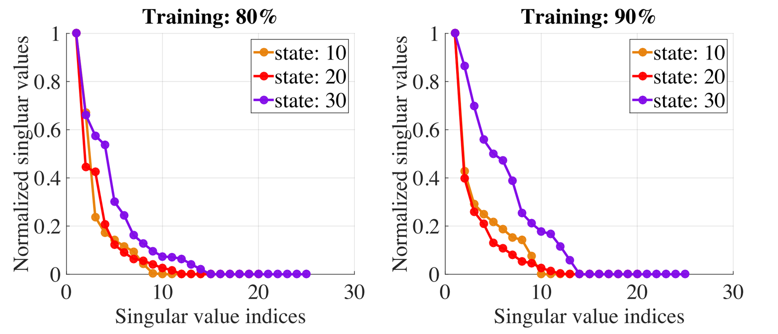


Figure 6: Intrinsic dimensionality recovery in *evap* dataset.

Acknowledgement

The work in this paper was supported by grant R01GM088224 from the NIH and by the Predoctoral Andrew Mellon Fellowship awarded to Zitao Liu for the school year 2015-2016. Its content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Reference

[1] Boots, Byron, Geoffrey J. Gordon, and Sajid M. Siddiqi. "A Constraint Generation Approach to Learning Stable Linear Dynamical Systems." Advances in Neural Information Processing Systems. 2007.
[2] Hauskrecht, Milos, et al. "Conditional outlier detection for clinical alerting." AMIA Annual Symposium Proceedings. Vol. 2010. American Medical Informatics Association, 2010.