2018 Intelligent Systems (知的システム構成論)

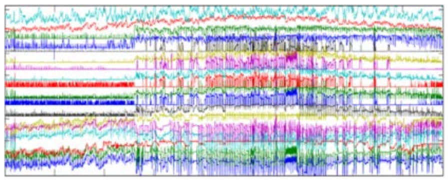
Canonical Correlation Analysis and Spectral Learning
- *What is the state ?* -

Jul.5, 2018

Takehisa YAIRI（矢入健久）
E-mail: yairi@ailab.t.u-tokyo.ac.jp

Review of Subspace Identification

Data



block Hankel matrix

$$Y_f = \begin{bmatrix} y_k & y_{k+1} & \cdots & y_{k+N-1} \\ y_{k+1} & y_{k+2} & \cdots & y_{k+N} \\ \vdots & \vdots & \ddots & \vdots \\ y_{k+f-1} & y_{k+f} & \cdots & y_{k+f-2} \end{bmatrix}$$

System

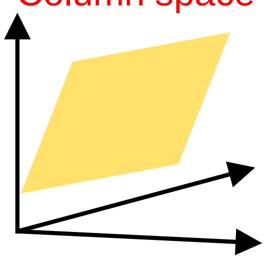
State Space Model

$$\boldsymbol{x}_{t+1} = \boldsymbol{A}\boldsymbol{x}_t + \boldsymbol{v}$$
$$\boldsymbol{y}_t = \boldsymbol{C}\boldsymbol{x}_t + \boldsymbol{w}$$

extended observability matrix

$$\boldsymbol{\Gamma}_f = \begin{bmatrix} \boldsymbol{C} \\ \boldsymbol{CA} \\ \vdots \\ \boldsymbol{CA}^{f-1} \end{bmatrix}$$

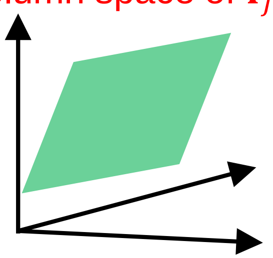
Column space of $\boldsymbol{Y}_f$



Equal under certain conditions

Identify

Column space of $\boldsymbol{\Gamma}_f$



- Advantage:
  - Global optimum is obtained without iteration by linear algebraic operations

Subspace Identification : CCA Approach*

$$\underbrace{\boldsymbol{Y}_f \boldsymbol{\Pi}_{\boldsymbol{U}_f}^{\perp}}_{\text{Known}} = \underbrace{\boldsymbol{\Gamma}_f \boldsymbol{L}_z}_{\text{Unknown}} \underbrace{\boldsymbol{Z}_p \boldsymbol{\Pi}_{\boldsymbol{U}_f}^{\perp}}_{\text{Known}} + \underbrace{\boldsymbol{G}_f \boldsymbol{E}_f}_{\text{Residual}}$$

Variable set 1

Variable set 2

Extended state space representation

$$\boldsymbol{Y}_f = \boldsymbol{\Gamma}_f \boldsymbol{X}_k + \boldsymbol{H}_f \boldsymbol{U}_f + \boldsymbol{G}_f \boldsymbol{E}_f$$

States estimated from past inputs and outputs

$$\boldsymbol{X}_k \equiv \boldsymbol{L}_z \boldsymbol{Z}_p$$

Obtained by Canonical Correlation Analysis (CCA):

Define $\boldsymbol{W}_r = \left(\boldsymbol{Y}_f \boldsymbol{\Pi}_{\boldsymbol{U}_f}^{\perp} \boldsymbol{Y}_f{}^{\boldsymbol{T}}\right)^{-1/2}$ and $\boldsymbol{W}_c = \left(\boldsymbol{Z}_p \boldsymbol{\Pi}_{\boldsymbol{U}_f}^{\perp} \boldsymbol{Z}_p{}^{\boldsymbol{T}}\right)^{-1/2}$

Perform SVD on $\boldsymbol{W}_r \boldsymbol{Y}_f \boldsymbol{\Pi}_{\boldsymbol{U}_f}^{\perp} \boldsymbol{Z}_p{}^{\boldsymbol{T}} \boldsymbol{W}_c \approx \boldsymbol{U}_d \boldsymbol{S}_d \boldsymbol{V}_d^{T}$

Then, $\hat{\boldsymbol{\Gamma}}_f \leftarrow \boldsymbol{W}_r{}^{-1} \boldsymbol{U}_d \boldsymbol{S}_d{}^{1/2}$

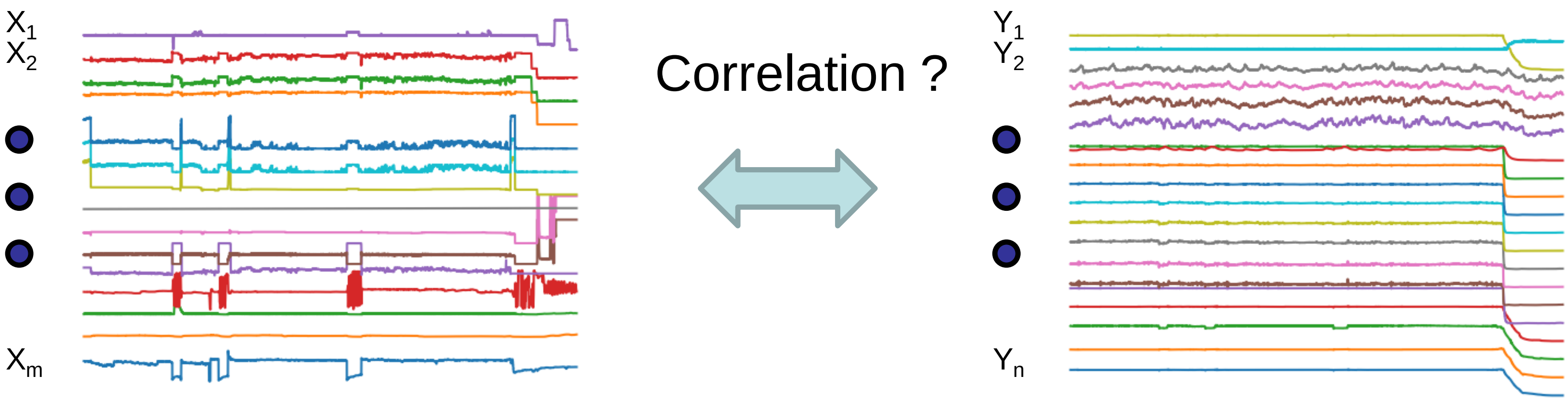
(*) Also known as canonical variate analysis (CVA)

Canonical Correlation Analysis (CCA)

What is CCA ?

Assume there are two sets of variables (signals)

Variable set 1 Variable set 2



Q : How are the two sets of variables correlated ?

- It's easy to examine the correlation of any pair of variables
- But variables in each set have correlations themselves
- Something like "normalization" and "orthogonalization" is necessary

Canonical Correlation Analysis (1)

There are two random vectors $\boldsymbol{x}$ and $\boldsymbol{y}$

$$\boldsymbol{x} \in R^m \ , \ \ \boldsymbol{y} \in R^n$$

For simplicity, assume both $\boldsymbol{x}$ and $\boldsymbol{y}$ are "centered", i.e., their means are zero.

$$E[\boldsymbol{x}] = \boldsymbol{0}_m \ , \ \ E[\boldsymbol{y}] = \boldsymbol{0}_n$$

Covariance matrices *of* and *between* $\boldsymbol{x}$ and $\boldsymbol{y}$

$$\mathrm{var}(\boldsymbol{x}) = E[\boldsymbol{x}\boldsymbol{x}^T] = \boldsymbol{\Sigma}_{xx} \ , \ \mathrm{var}(\boldsymbol{y}) = E[\boldsymbol{y}\boldsymbol{y}^T] = \boldsymbol{\Sigma}_{yy}$$

$$\mathrm{cov}(\boldsymbol{x}, \boldsymbol{y}) = E[\boldsymbol{x}\boldsymbol{y}^T] = \boldsymbol{\Sigma}_{xy} \quad \Rightarrow \quad \boldsymbol{\Sigma}_{yx} = {\boldsymbol{\Sigma}_{xy}}^T$$

Canonical Correlation Analysis (2)

Think of constructing synthetic variables u and v by linear combinations of $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively.

$$u = \boldsymbol{a}^T \boldsymbol{x} = a_1 x_1 + a_2 x_2 + \cdots + a_m x_m$$

$$v = \boldsymbol{b}^T \boldsymbol{y} = b_1 \boldsymbol{y} + b_2 y_2 + \cdots + b_n y_n$$

Problem: Find $\boldsymbol{a}$ and $\boldsymbol{b}$ so that the correlation between u and v is maximized

$$\rho = \mathrm{cor}(u, v) = \frac{\mathrm{cov}(u, v)}{\sqrt{\mathrm{var}(u) \cdot \mathrm{var}(v)}}$$

$$= \frac{\boldsymbol{a}^T \boldsymbol{\Sigma}_{xy} \boldsymbol{b}}{\sqrt{\boldsymbol{a}^T \boldsymbol{\Sigma}_{xx} \boldsymbol{a}} \sqrt{\boldsymbol{b}^T \boldsymbol{\Sigma}_{yy} \boldsymbol{b}}}$$

Canonical Correlation Analysis (3)

Impose constraints on $\boldsymbol{a}$ and $\boldsymbol{b}$, so that the variances of u and v become 1

$$\mathrm{var}(u) = \mathrm{var}(\boldsymbol{a}^T \boldsymbol{x}) = \boldsymbol{a}^T \boldsymbol{\Sigma}_{xx} \boldsymbol{a} = 1$$

$$\mathrm{var}(v) = \mathrm{var}(\boldsymbol{b}^T \boldsymbol{y}) = \boldsymbol{b}^T \boldsymbol{\Sigma}_{yy} \boldsymbol{b} = 1$$

The problem is formulated as,

$$(\boldsymbol{a}_1, \boldsymbol{b}_1) = \underset{\boldsymbol{a}^T \boldsymbol{\Sigma}_{xx} \boldsymbol{a} = 1\ , \boldsymbol{b}^T \boldsymbol{\Sigma}_{yy} \boldsymbol{b} = 1}{\mathrm{argmax}} \boldsymbol{a}^T \boldsymbol{\Sigma}_{xy} \boldsymbol{b}$$

It can be solved by Lagrange multiplier, but there is a more elegant method.

Canonical Correlation Analysis (4)

Let square root matrices of Σ_{xx} and Σ_{yy} be ${\Sigma_{xx}}^{1/2}$ and ${\Sigma_{yy}}^{1/2}$, respectively. Then, define $\boldsymbol{c}$ and $\boldsymbol{d}$ as,

Note that Σ_{xx} and Σ_{yy} are positive definite matrices

$$\boldsymbol{c} = {\boldsymbol{\Sigma}_{xx}}^{1/2}\boldsymbol{a} \quad \text{, and } \boldsymbol{d} = {\boldsymbol{\Sigma}_{yy}}^{1/2}\boldsymbol{b}$$

The problem turns to be

$$(\boldsymbol{c}_1, \boldsymbol{d}_1) = \underset{\|\boldsymbol{c}\|=1\,,\|\boldsymbol{d}\|=1}{\operatorname{argmax}} \boldsymbol{c}^T \left(\boldsymbol{\Sigma}_{xx}^{-T/2}\boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{xx}^{-1/2}\right)\boldsymbol{d}$$

This problem can be solved by SVD !

$$\boldsymbol{\Sigma}_{xx}^{-T/2}\boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{xx}^{-1/2} = \boldsymbol{CSD}^T \Rightarrow \begin{cases} \boldsymbol{c}_1 \leftarrow \text{1st column of } \boldsymbol{C} \\ \boldsymbol{d}_1 \leftarrow \text{1st column of } \boldsymbol{D} \end{cases}$$

$$\Rightarrow \quad \boldsymbol{a}_1 = {\boldsymbol{\Sigma}_{xx}}^{1/2}\boldsymbol{c}_1, \text{ and } \boldsymbol{b}_1 = {\boldsymbol{\Sigma}_{yy}}^{1/2}\boldsymbol{d}_1$$

Canonical Correlation Analysis (4)

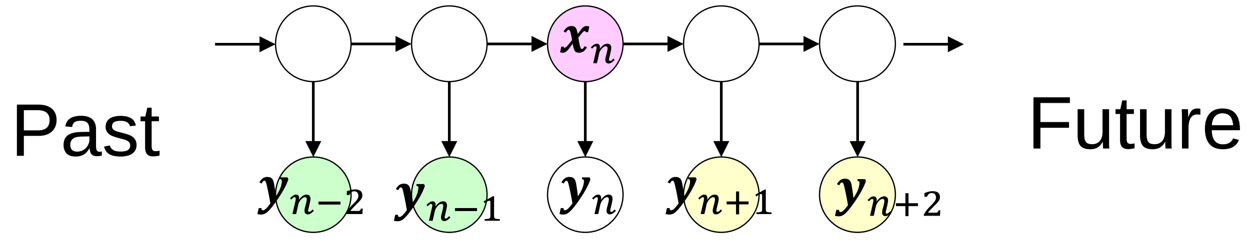
- The (1st) canonical correlation is the largest singular value σ_1 , where $\boldsymbol{S} = diag(\sigma_1, \sigma_2, \ldots)$.
- Similarly, the 2nd and higher canonical correlations are obtained from the other singular values.
- CCA has many interesting properties and close connections to PCA, PLS (partial least squares), CA (correspondence analysis), etc.

Canonical Correlation Analysis of Time-series [Akaike 74][Akaike 75]

- Time-series of observation: $\{\cdots, \boldsymbol{y}_{n-1}, \boldsymbol{y}_n, \boldsymbol{y}_{n+1}, \cdots\}$
- "State" space representation of the system:

$$\boldsymbol{x}_{n+1} = F\boldsymbol{x}_n + \boldsymbol{w}_n$$
$$\boldsymbol{y}_n = H\boldsymbol{x}_n$$

- The "state" $\boldsymbol{x}_n$ can be interpreted in two ways:
  – Random variables that contain full information of the future to be expressed by the present and past
  – Random variables that contain full information of the past to be expressed by the present and future



Canonical Correlation Analysis of Time-Series (cont.)

- "State" vector $\boldsymbol{x}_n$ is the information interface between future $\boldsymbol{y}_n, \boldsymbol{y}_{n+1}, \cdots$ and past $\boldsymbol{y}_{n-1}, \boldsymbol{y}_{n-2}, \cdots$
- Predict future by past :

$$\underbrace{\boldsymbol{y}_{n-1}, \boldsymbol{y}_{n-2}, \cdots}_{\text{Past}} \longrightarrow \boldsymbol{u}_n \longrightarrow \underbrace{\boldsymbol{y}_n, \boldsymbol{y}_{n+1}, \cdots}_{\text{Future}}$$

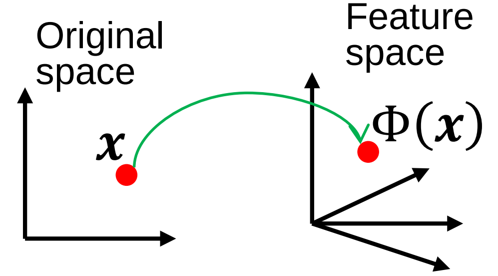
- Postdict past by future :

$$\underbrace{\boldsymbol{y}_n, \boldsymbol{y}_{n+1}, \cdots}_{\text{Future}} \longrightarrow \boldsymbol{v}_n \longrightarrow \underbrace{\boldsymbol{y}_{n-1}, \boldsymbol{y}_{n-2}, \cdots}_{\text{Past}}$$

- Canonical vectors $\boldsymbol{u}_n$ and $\boldsymbol{v}_n$ should be maximally correlated

Extension of CCA (1) : Kernel CCA

- Most of classical linear multivariate analysis methods can be non-linearized by (RKHS) kernel
  - Linear regression -> Kernel regression, SVR, GPR
  - Linear classifier -> SVM
  - Linear PCA -> Kernel PCA



- Kernel CCA [Akaho 01][Bach 02]

Linear CCA

$$u = \boldsymbol{a}^T \boldsymbol{x}$$
$$v = \boldsymbol{b}^T \boldsymbol{y}$$
$$\max_{\boldsymbol{a},\boldsymbol{b}} \operatorname{cor}(u, v)$$

Non-linear mapping

$$\boldsymbol{x}^\Phi = \Phi(\boldsymbol{x})$$
$$\boldsymbol{y}^\Psi = \Psi(\boldsymbol{y})$$
$$\boldsymbol{a}^\Phi = \sum_i \alpha_i \cdot \Phi(\boldsymbol{x}_i)$$
$$\boldsymbol{b}^\Psi = \sum_i \beta_i \cdot \Psi(\boldsymbol{y}_i)$$

Kernel CCA

$$u = {\boldsymbol{a}^\Phi}^T \boldsymbol{x}^\Phi$$
$$v = {\boldsymbol{b}^\Psi}^T \boldsymbol{y}^\Psi$$
$$\max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \operatorname{cor}(u, v)$$

Learning Non-linear Dynamical Systems by Kernel CCA

- Y. Kawahara, T. Yairi, and K. Machida, "A kernel subspace method by stochastic realization for learning nonlinear dynamical systems", NIPS-2006
- Kernel CCA between past and future data
- A pioneering work of spectral learning of dynamical systems in machine learning community
- But, it was too early ..

Non-linear dynamical system

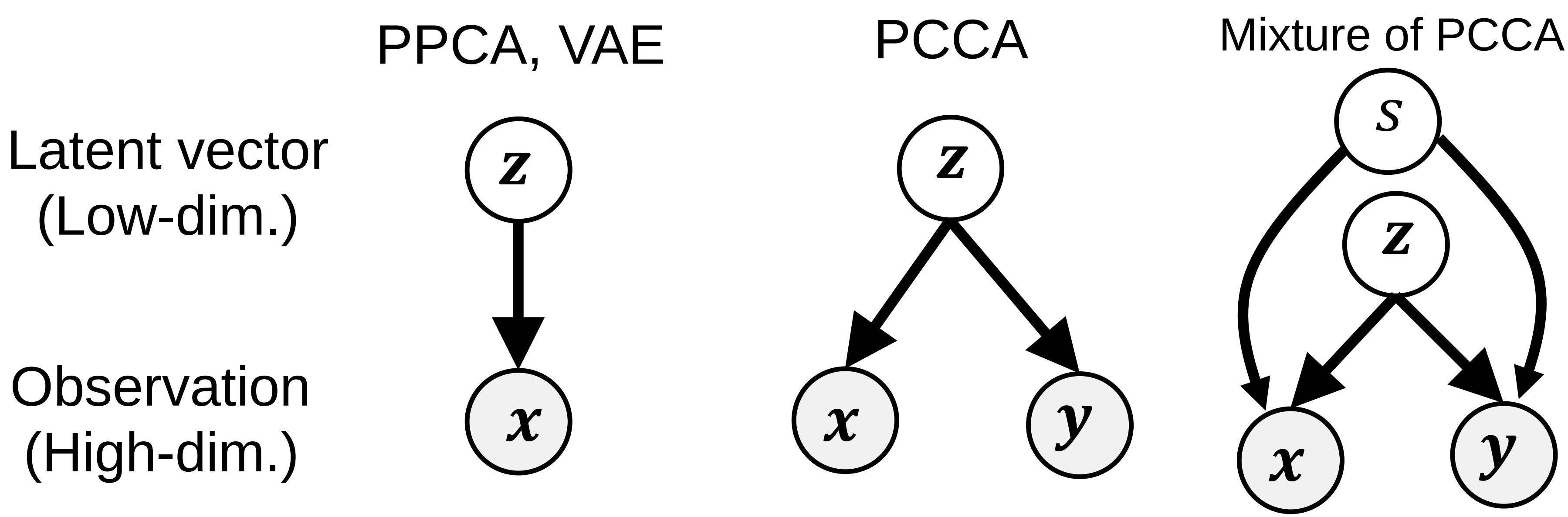
$$
\begin{aligned}
\boldsymbol{x}(t+1) &= \boldsymbol{g}(\boldsymbol{x}(t), \boldsymbol{u}(t)) + \boldsymbol{v} \\
\boldsymbol{y}(t) &= \boldsymbol{h}(\boldsymbol{x}(t), \boldsymbol{u}(t)) + \boldsymbol{w},
\end{aligned}
$$

Linear dynamical system in kernel feature space

$$
\begin{aligned}
\boldsymbol{x}(t+1) &= A^{\phi}\boldsymbol{x}(t) + B^{\phi}\boldsymbol{\phi}_u(\boldsymbol{u}(t)) + K^{\phi}\boldsymbol{e}(t), \\
\boldsymbol{\phi}_y(\boldsymbol{y}(t)) &= C^{\phi}\boldsymbol{x}(t) + D^{\phi}\boldsymbol{\phi}_u(\boldsymbol{u}(t)) + \boldsymbol{e}(t),
\end{aligned}
$$

Extension of CCA (2): Probabilistic CCA

- Probabilistic canonical correlation analysis (PCCA) [Bach & Jordan 06]
  - Probabilistic (latent variable) model
  - Similar to Probabilistic PCA, variational autoencoder, etc.

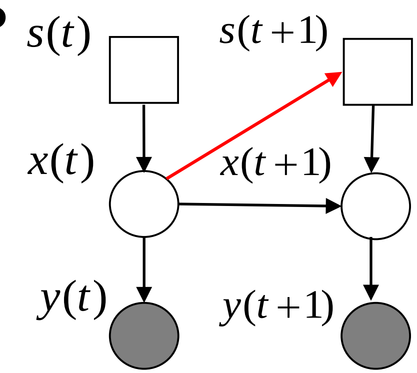


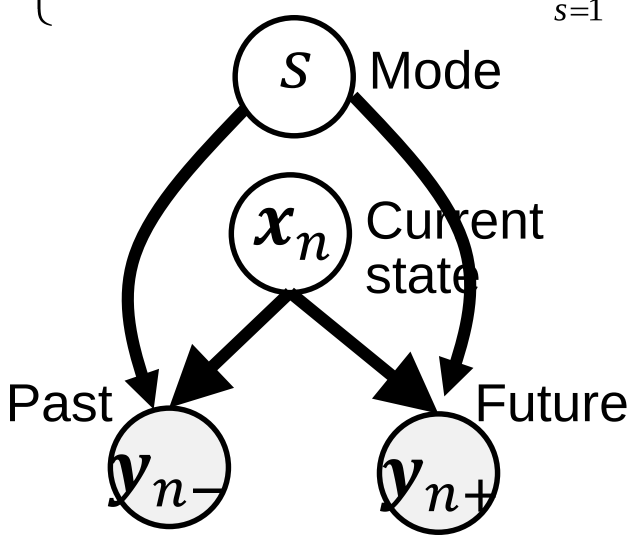
Learning Mixture of Linear Dynamical Systems

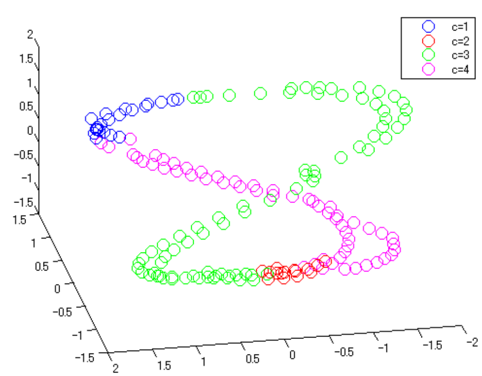
Masao Joko, Yoshinobu Kawahara, Takehisa Yairi, "Learning Non-linear Dynamical Systems by Alignment of Local Linear Models", 20th International Conference on Pattern Recognition (ICPR), pp. 1084-1087, 2010

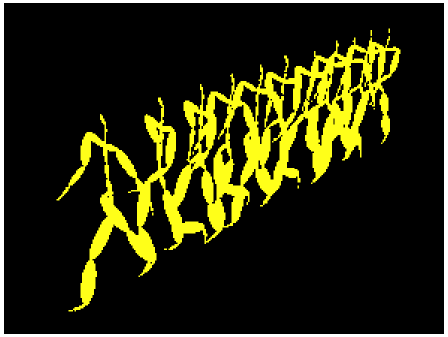
Mixture of locally linear dynamical systems

$$\begin{cases} p(x(t+1)\,|\,x(t)) = \sum_{s=1}^{C} p(s(t+1)\,|\,x(t))\,p(x(t+1)\,|\,x(t), s(t+1)) \\ p(y(t)\,|\,x(t)) = \sum_{s=1}^{C} p(s(t)\,|\,x(t))\,p(y(t)\,|\,x(t), s(t)) \end{cases}$$









Spectral Learning of HMM [Hsu 09]

Daniel J. Hsu, Sham M. Kakade, and Tong Zhang, "A Spectral Algorithm for Learning Hidden Markov Models", COLT 2009.

- For a long time, EM (Baum-Welch) algorithm was believed to be the only way to learn HMM
- Inspired by subspace identification
  - Consider the canonical correlation between past and future observation
  - (Latent) state sequence and transition/output probabilities are implicitly computed
- Limitations
  - Limited to discrete observations
  - Assumption of one-step observability

Spectral Learning of HMM (Cont.)

The seminal work of [Hsu 09] was rapidly extended

- Frustration to EM algorithm

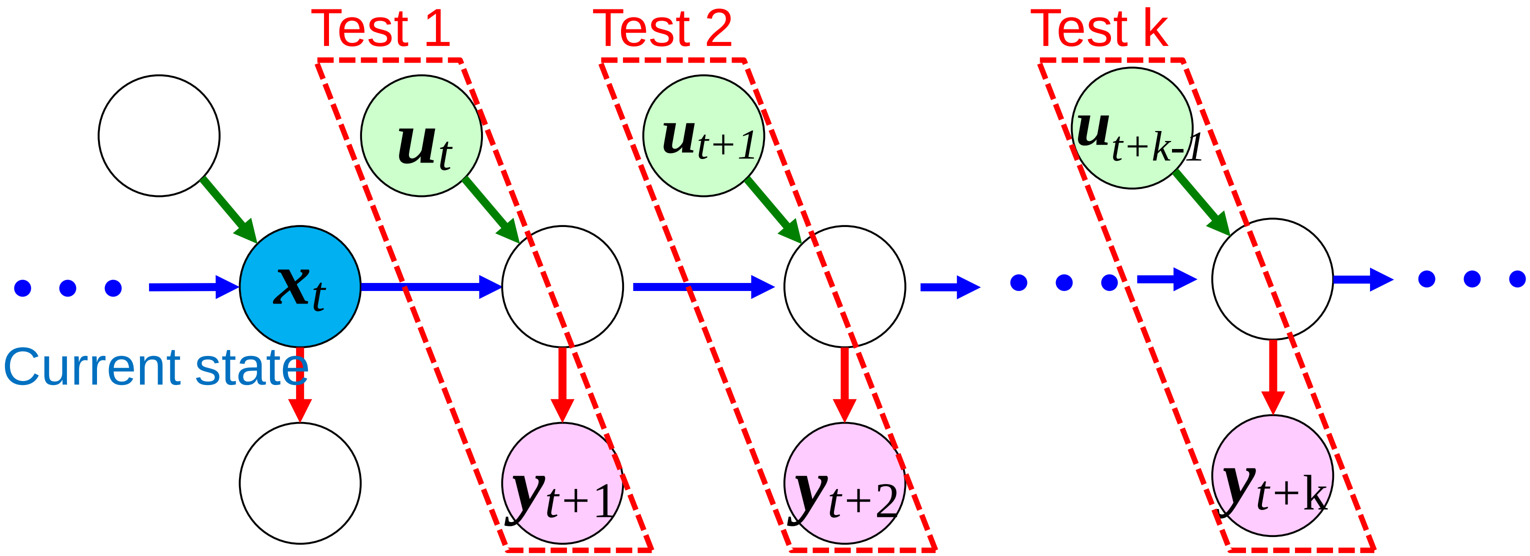
- [Siddiqi 10] Continuous observations
- [Song 10] Non-Gaussian continuous observation
- [Anandkumar 12] Generalization as a "method of moments"
- [Subakan 14] Mixture of HMM
- [Zhang 15] Latent states with tree-like structure
- [Kandasamy 16] Non-parametric observation model

Spectral Learning of Continuous State Space Models

- Spectral learning of HMM is "re-imported" to continuous state space models
  - [Buesing 12] Extention of Ho-Kalman realization algorithm to Poisson observation model
- Another idea : Use the past observation sequence, instead of estimating latent state explicitly
  - Reduced to supervised regression problems
  - [Langford 09], [Hefny 2015], [Sun 2016]
  - Predictive State Representation (PSR) [Littman 01]

Predictive State Representation（PSR）

- Originally developed as a state representation for partially observable environment [Littman 01] [Singh 04]
  - Extension of Observable operator models (OOM) [Jaeger 00]
- Instead of estimating the current state, predict a set of tests (pairs of input and output) in future
  - Predicting test results in future ≈ Guessing the current state



Predictive State Representation (Cont.)

- Sufficient statistic for future test results ≈ (current) state
- Predicting future test results based on past results ≈ State estimation (filtering)
- Transformed PSR [Rosencrantz 04] : Obtain a minimum set of bases necessary to predict any future test results
- Close relation to canonical variate [Akaike 75], subspace identification[Boots 09]

Epilogue

- We focused on the methods of "learning dynamical systems" or machine learning approaches to system identification problem
  - Learning a state space model from observation data
- They are roughly classified into two approaches
  1. Maximum-likelihood approach
  2. Spectral learning approach
- Topics not covered this time:
  - (Deep) neural networks for time-series (RNN, etc.)
  - Koopeman operator, dynamic mode decomposition

Final Task (of Yairi's Part)

- 適当な多変量時系列データに対して、システム同定・機械学習（動的システム学習）の手法を適用することによって、そのデータの発生源であるシステムの挙動モデルを推定し、その結果を考察せよ。用いた手法のアルゴリズム、データの出典や参考文献、等を明記すること。
- 締切: 2018年8月10日(金)
- 9月修了予定者は別途相談すること。
- 提出方法:ITC-LMS で提出
- 注:堀先生の課題も必ずやること