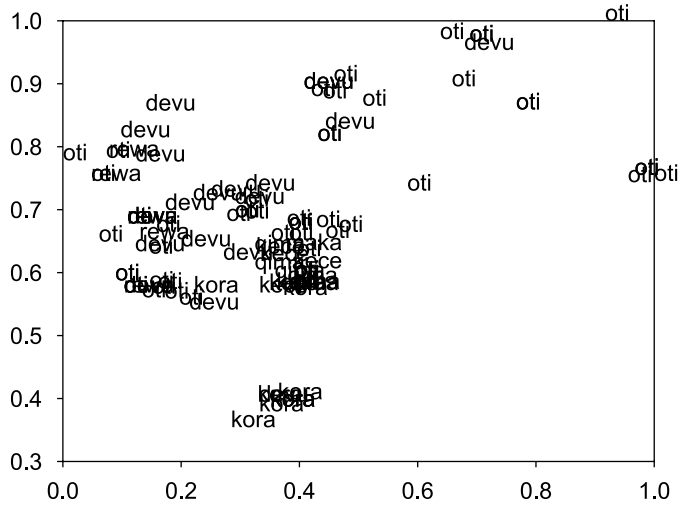
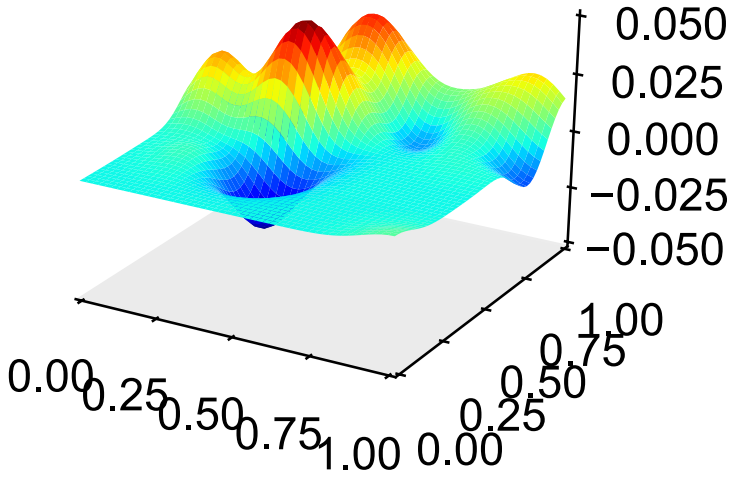
ノンパラメトリックベイズ統計と自然言語処理





持橋大地
統計数理研究所
daichi@ism.ac.jp

Summer School 数理物理
2020-8-29 (日)

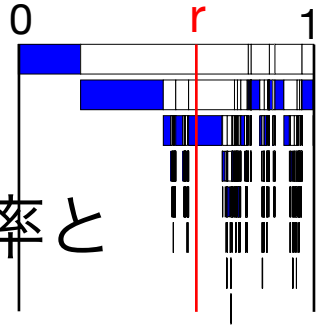
持橋担当分のスケジュール

- 1日目：概要と目次、ノンパラメトリックベイズ法
(離散的な場合; 無限モデル)
- 2日目：ガウス過程とその適用
(連続的な場合; ベイズ的関数回帰)
- 3日目：ノンパラメトリックベイズ法と自然言語処理への応用
(研究紹介)

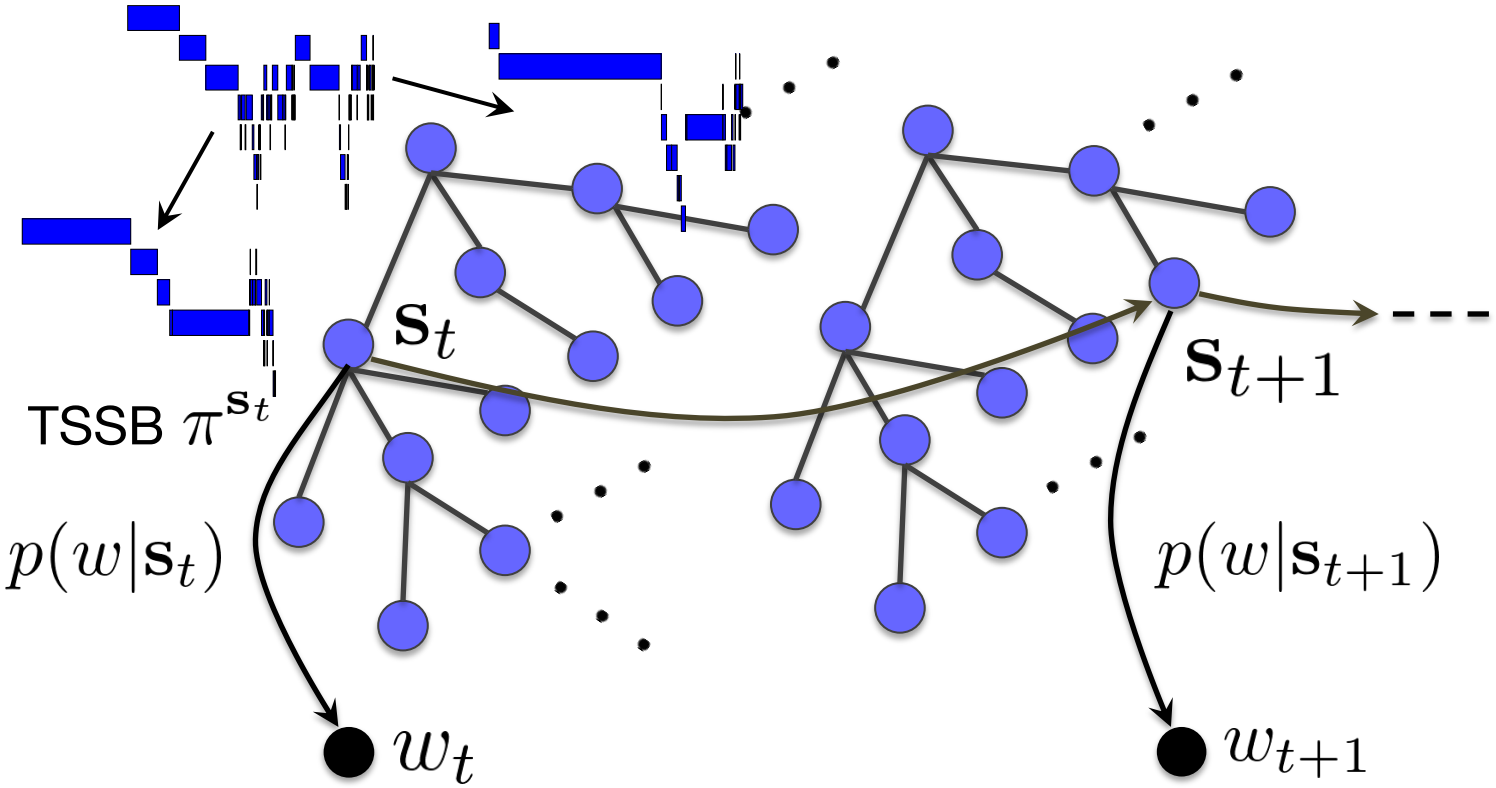
はじめに: 雑談

- 数理と実装について
  - パッケージ化と理解、データ構造

無限木構造HMM (iTHMM)

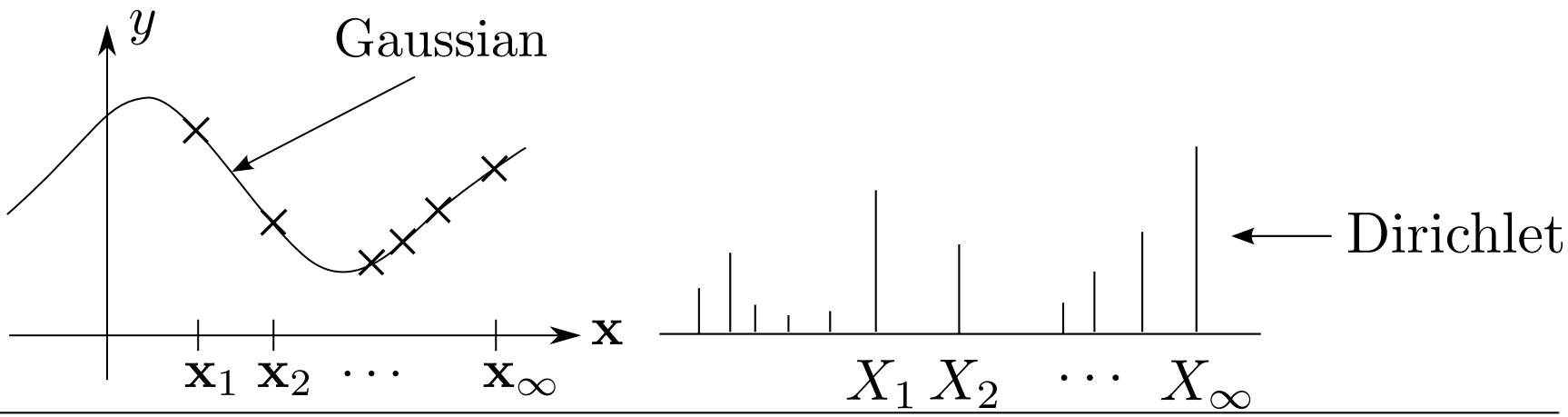


- HTSSBにより、無限木構造上の状態遷移確率とその事後分布が計算できる
  → HTSSB-HMM = Infinite Tree HMM (iTHMM)



ディリクレ過程(DP)とガウス過程(GP)

- 二者の定義は、ほとんど一緒
  - GP: どんな入力の集合 $(\mathbf{x}_1, \cdots, \mathbf{x}_N)$ をとってきても、対応する $(y_1, \cdots, y_N)$ がガウス分布に従う
  - DP: 空間のどんな離散化 $(X_1, \cdots, X_k)$ についても、対応する離散分布がディリクレ分布 $\mathrm{Dir}(\alpha(X_1), \cdots, \alpha(X_k))$ に従う
- どちらも、無限次元の smoother になっている



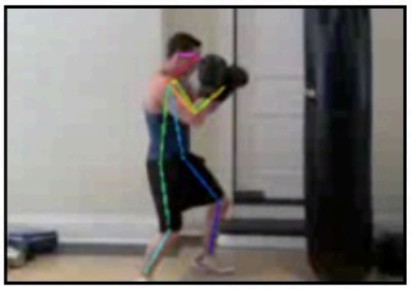
(1)「副詞」の学習

人間の動作からの副詞の理解

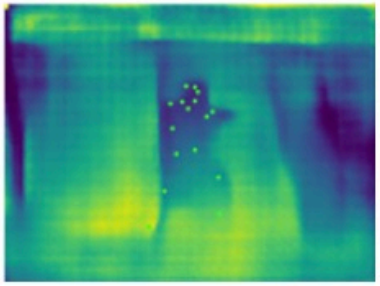
- “しっかりと”、“さっと”、“慎重に”といった副詞の理解は、特に介護などで今後とても重要
  - これらは静的な画像ではなく、動作の動的な時系列の性質に関連している
  - 適当なニューラル手法では、動的な性質を十分捉えることができない
- 『スペクトル混合カーネルとガウス過程に基づく動画からの副詞の意味理解』(言語処理学会2021)
  - 谷口巴(お茶大)、持橋大地(統数研) 他

動画からの副詞の理解

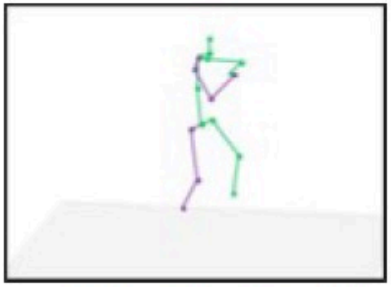
- 動作を表す動画に、「そわそわ」「速足で」「こっそりと」など副詞をクラウドソーシングで
  タグ付けしてもらう
- 動画からOpenPose+前処理で骨格を抽出



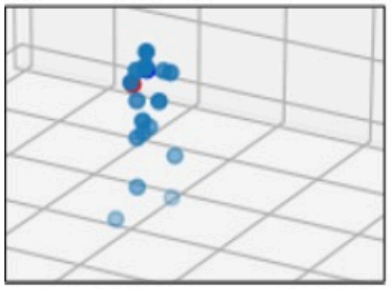
(a) Openpose による画面座標推定



(b) FCRN-depth による深度推定



(c) 3 次元の骨格座標の推定



(d) 回転行列による方向正規化

- 48次元の関節座標の時系列データが得られる

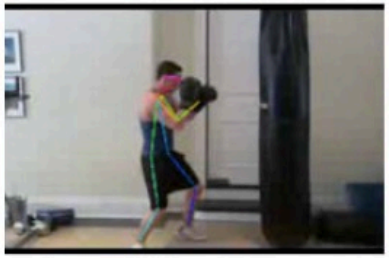
動画からの副詞の理解 (2)

- さまざまな動作を表す関節座標の時系列姿勢データと、
- それを記述する副詞の集合

がペアになったデータセットが得られた



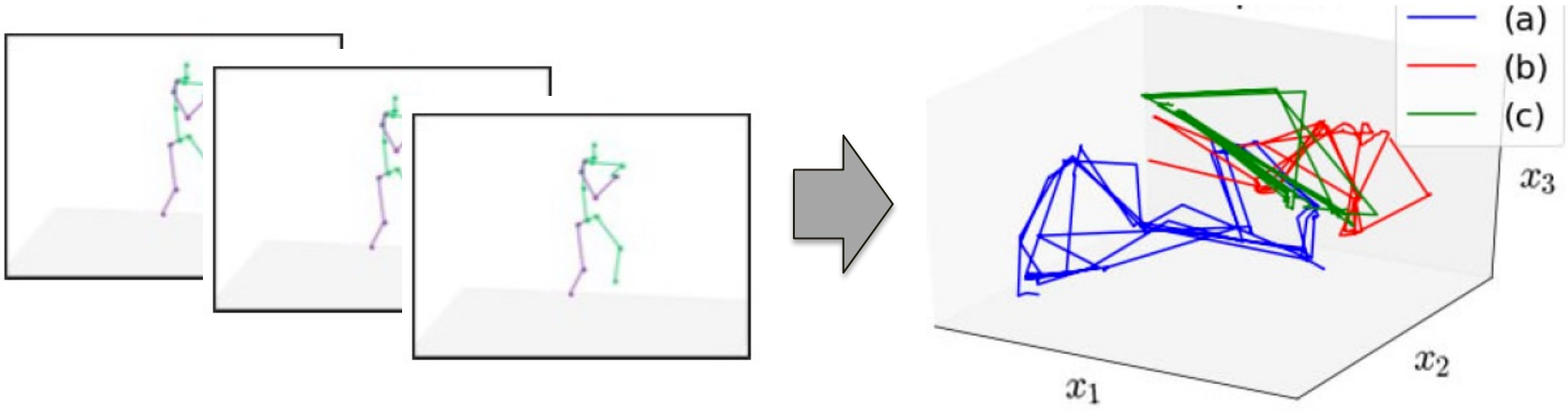
サッサッと, 力強く　重々しく, 確実に　素早く, 正確に




速く, 重々しく, 鋭く　軽く, ゆっくりと, 忍んで

動画からの副詞の理解 (3)

- 48次元の姿勢データを、GPLVMでP次元の潜在空間に非線形圧縮 (P=3)



- この潜在空間の時系列データの特徴をどう捉えるか?
  - 速い、遅い、微妙な震え、...
  - 「関数の特徴を捉える」問題

スペクトル混合カーネル

- “Gaussian Process Kernels for Pattern Discovery and Extrapolation”, Andrew Gordon Wilson, Ryan Prescott Adams, *ICML 2013*.
- ガウス過程で使うカーネルを、RBFのような既存のカーネルおよびその組み合わせに限定せず、フーリエ領域で混合ガウス分布を考えることでデータから自動的に学習できる (！)

スペクトル混合カーネル (2)

- ガウス過程のカーネルとして、値が $\tau = x - x'$ だけに依存する、定常カーネル $k(\tau)$ を考える
  - RBF (ガウス)カーネルもこの仲間
- ボホナーの定理により、任意の$k(\tau)$は

$$k(\tau) = \int_{\mathbb{R}^D} e^{2\pi i\, s^T \tau} \psi(ds)$$

の形に表せる (逆フーリエ変換)

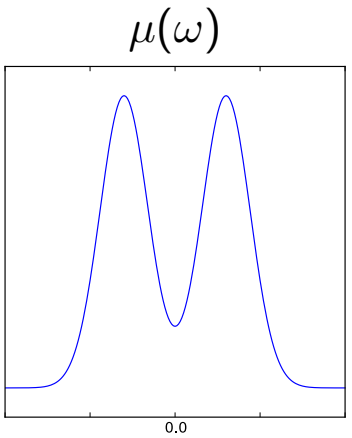
- $\psi(s)$が、周波数領域での$k(\tau)$の等価な表現

↓

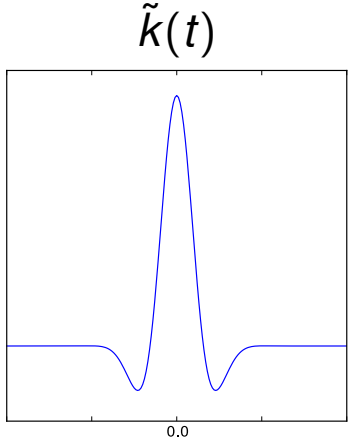
$k(\tau)$を確率分布で表せる！

スペクトル混合カーネルのイメージ

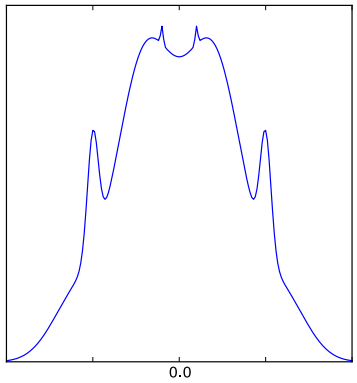
周波数空間
(確率分布)



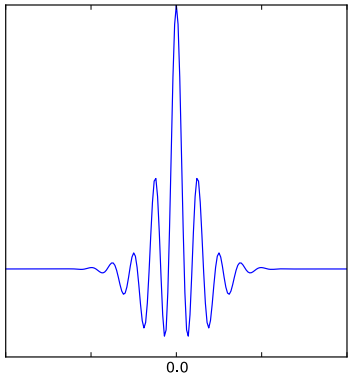
$\xrightarrow[\mathcal{F}^{-1}]{}$



カーネル
(関数)



$\xrightarrow[\mathcal{F}^{-1}]{}$



Gaussian process summer school 2013の資料
“Kernel Design” (N. Durrande) より引用

スペクトル混合カーネル (3)

- 通常のガウスカーネル

$$k(x, x') = \exp\left(-\frac{1}{2}(x - x')^2/\ell^2\right)$$

の周波数表現は、フーリエ変換すると

$$S(s) = (2\pi\ell^2)^{1/2} \exp(-2\pi^2\ell^2 s^2)$$

  - 中心0のガウス分布!

スペクトル混合カーネル (4)

- k(τ)は周波数領域での確率密度 $\psi(s)$ と等価なので、$\psi(s)$ に関して混合ガウス分布を考える
  - 0に関して対称なので、正だけ考えて鏡映

$$\phi(s \,|\, \mu, \sigma^2) = \mathcal{N}(s \,|\, \mu, \sigma^2)$$
$$S(s) = (\phi(s) + \phi(-s))/\,2$$

- ガウス分布の各要素は、もとの領域では以下のカーネル関数を考えていることと等価

$$k(\tau|\sigma, \mu) = \exp(-2\pi^2\sigma^2\tau^2)\cos(2\pi\mu\tau)$$

スペクトル混合カーネル (5)

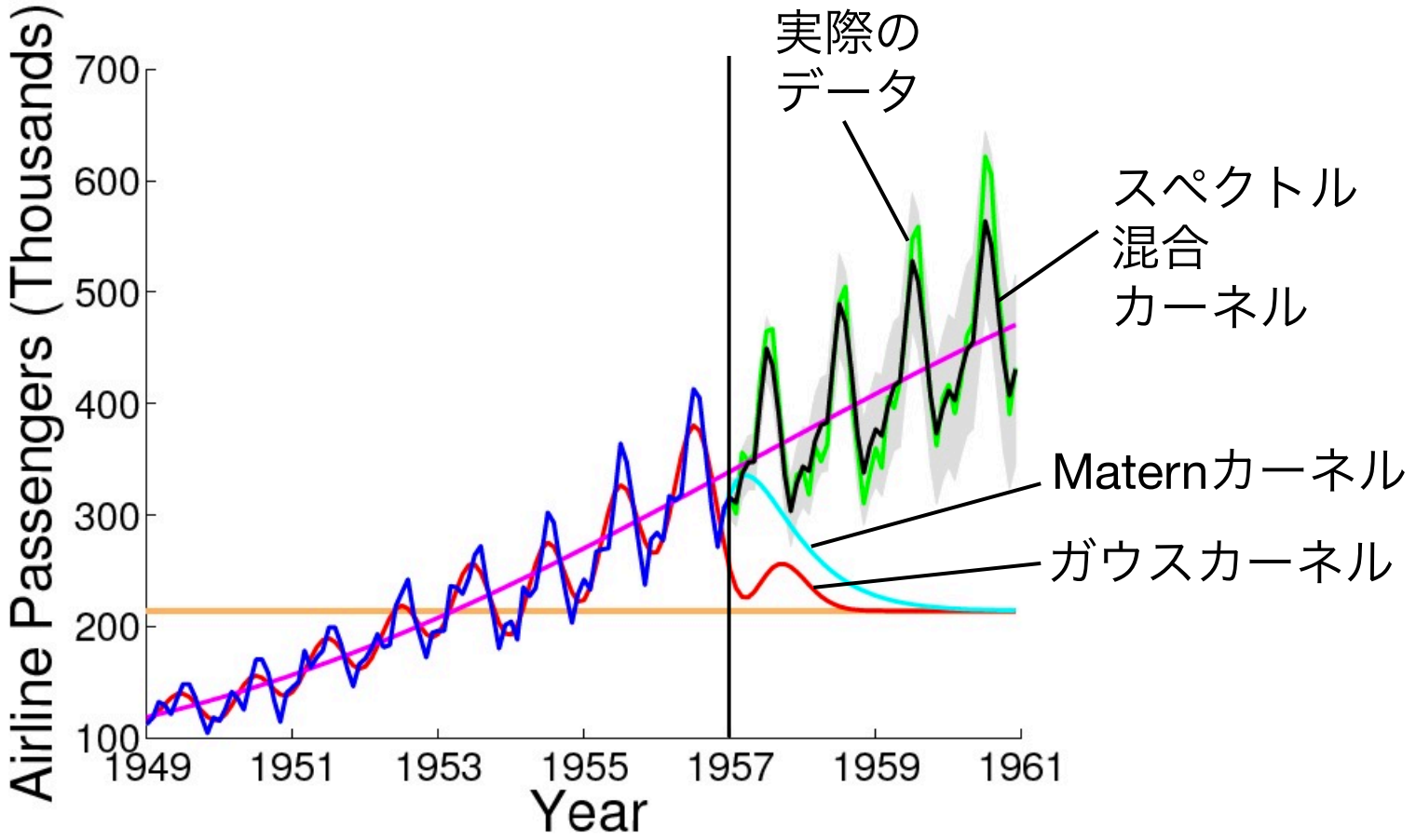
- すなわち、これはカーネルとして、次の混合を考えていることになる (Spectral Mixture kernel)

$$k(\boldsymbol{\tau}) = \sum_{p=1}^{P} w_p \cos(2\pi \boldsymbol{\tau}^T \boldsymbol{\mu}^{(p)}) \exp\left(-\sum_{d=1}^{D} 2\pi^2 \sigma_d^{(p)} \tau_d^2\right)$$

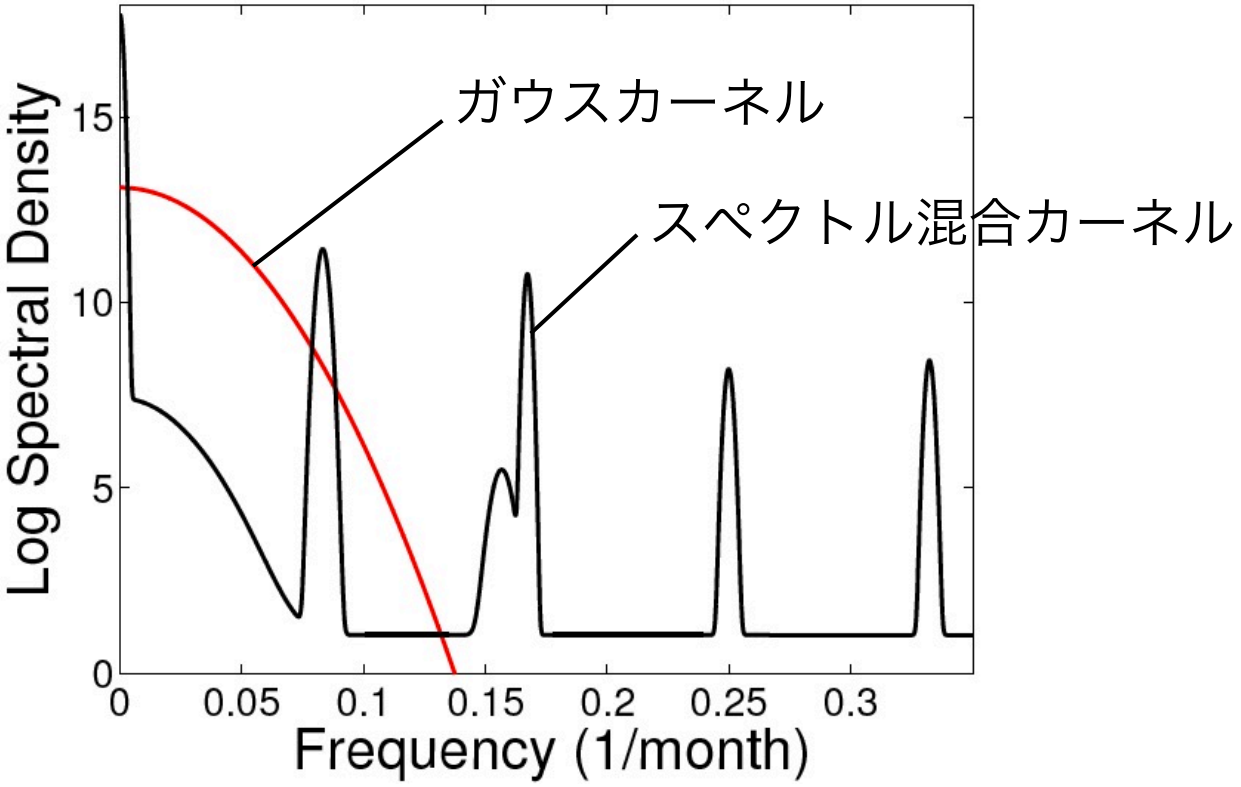
  - パラメータw、μ、σは通常のハイパーパラメータ最適化で学習できる
  - ARD事前分布を使うことで、不要なガウス分布を除去している

Airline Passengerデータ

- 1949-1961の毎月の航空乗客数のうち、最初の8年を学習に使って残りの4年分を予測



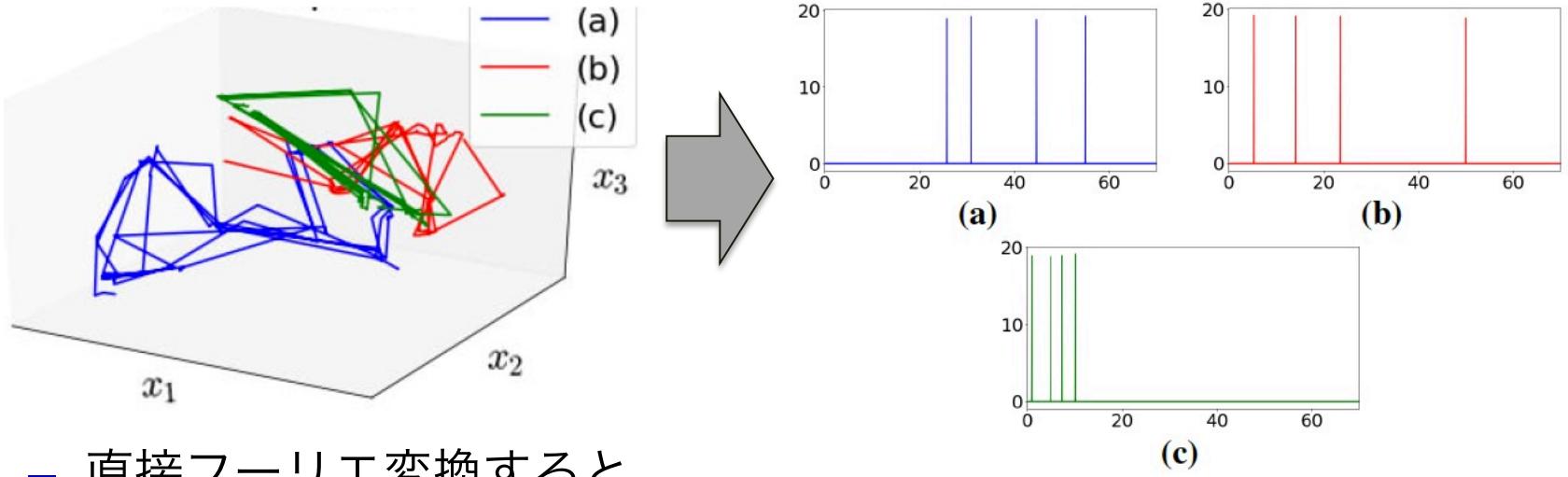
Airline Passengerデータ (2)



- 線形トレンド(一番左)に加え、12, 6, 4, 3ヶ月の周期性とその細かな違いを捉えている
  - 通常のガウスカーネル(赤)はまったくの大雑把

動画からの副詞の理解 (3)

- スペクトル混合カーネルの適用：カーネルの周波数空間での特徴を抽出



  - 直接フーリエ変換すると、
    関数の位相に依存＆FFTにはデータが不足

- 一方で、各動作には「しっかり」「堂々と」などの副詞が付与されている

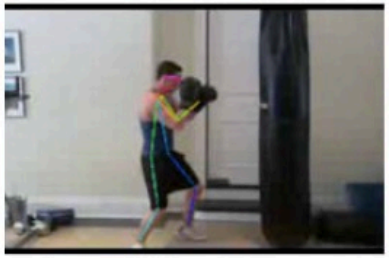
動画からの副詞の理解 (再掲)

- さまざまな動作を表す関節座標の時系列姿勢データと、
- それを記述する副詞の集合

がペアになったデータセットが得られた



サッサッと, 力強く　重々しく, 確実に　素早く, 正確に



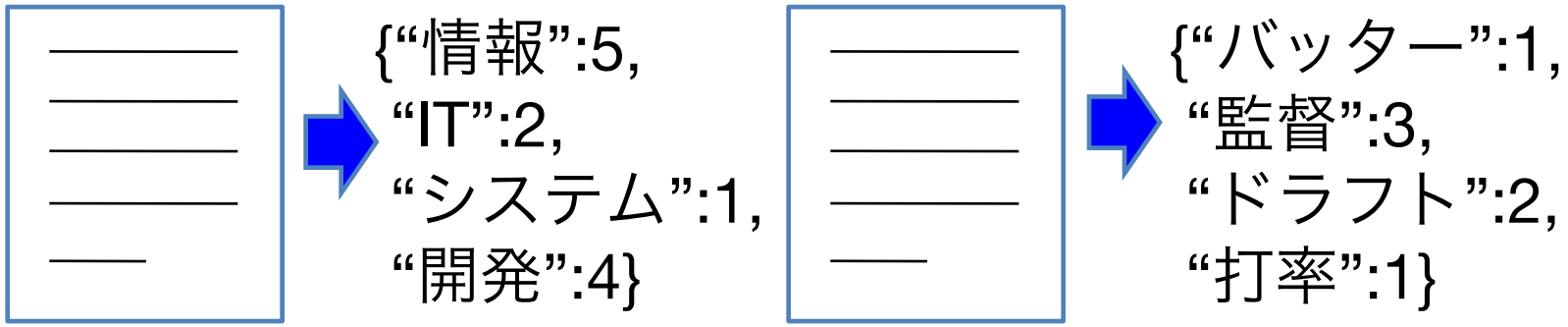

速く, 重々しく, 鋭く　軽く, ゆっくりと, 忍んで

テキストの潜在トピックモデル

- Hofmann (1999), Blei+ (2001) に始まる
- 文章の「意味」を自動的に解釈するためのモデル
- 解釈性の高さから、社会科学などでは特に
  現在でも使われている

“Bag of words” データ

- 文書の中に同時に現れる単語群には相関がある
  → 文書の中に現れた“単語”とその頻度



- 一般には、文書＝観測ユニット、単語＝離散データと読み替えてOK
  - 例：文書＝人、単語＝購入した商品

トピックモデルの学習例 (1) (Blei+ 2003)より

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

- それぞれの語がどんな「話題」(トピック)に属するのかが、大量のデータだけからわかる
  - 実際には、話題の確率がわかる
- 文書=人、単語=その人の買った商品、URL、….

トピックモデルの学習例 (2)

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

(Blei+ 2003) より

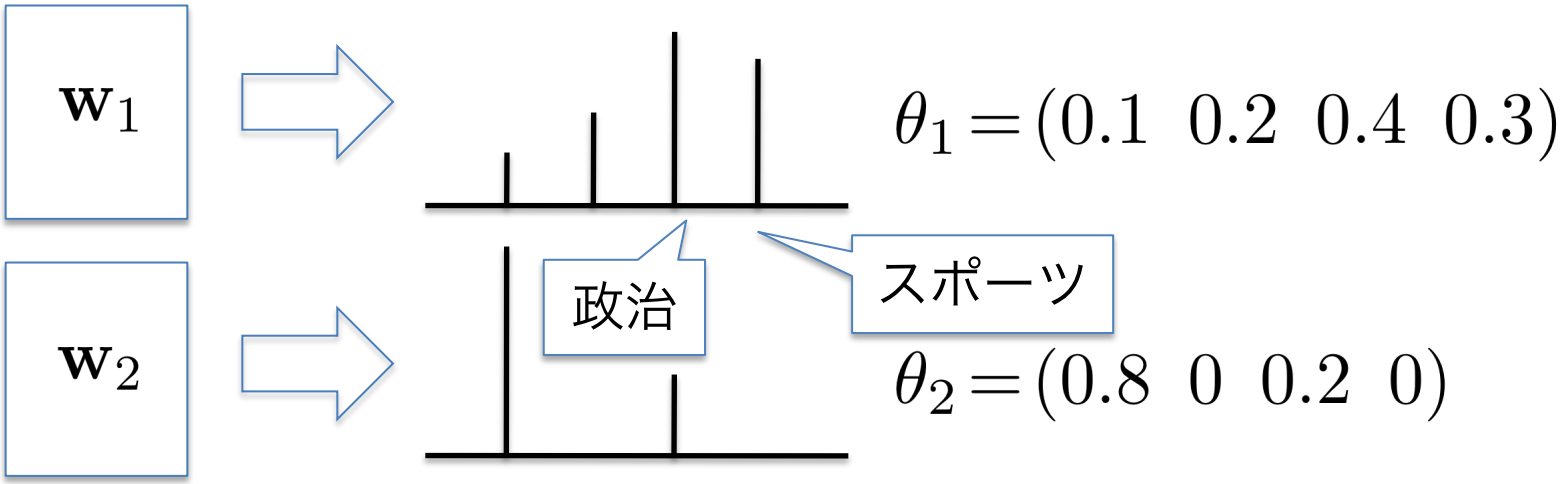
- コーパスから完全に自動的に関連語を抽出できる

LDAとは

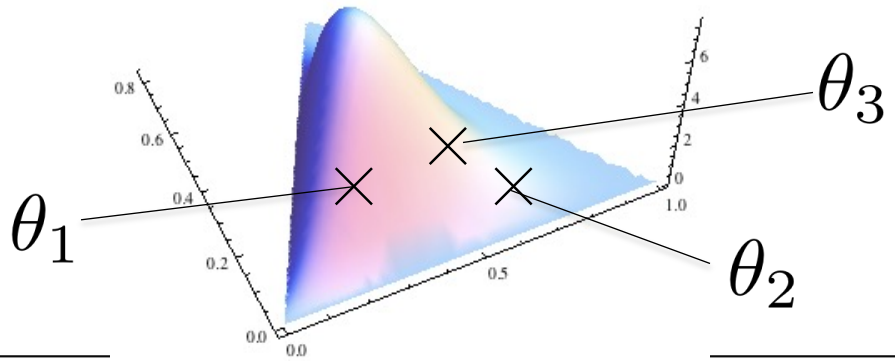
- トピックモデルの最も基本となるモデル
- Blei+ (NIPS 2001, JMLR 2003)で提案
- 基本的な考え方は、単語ごとに潜在トピックがあること (文書毎の混合モデル)

LDAの生成モデル

- 文書$\mathbf{w}$を話題(トピック)の混合で表現

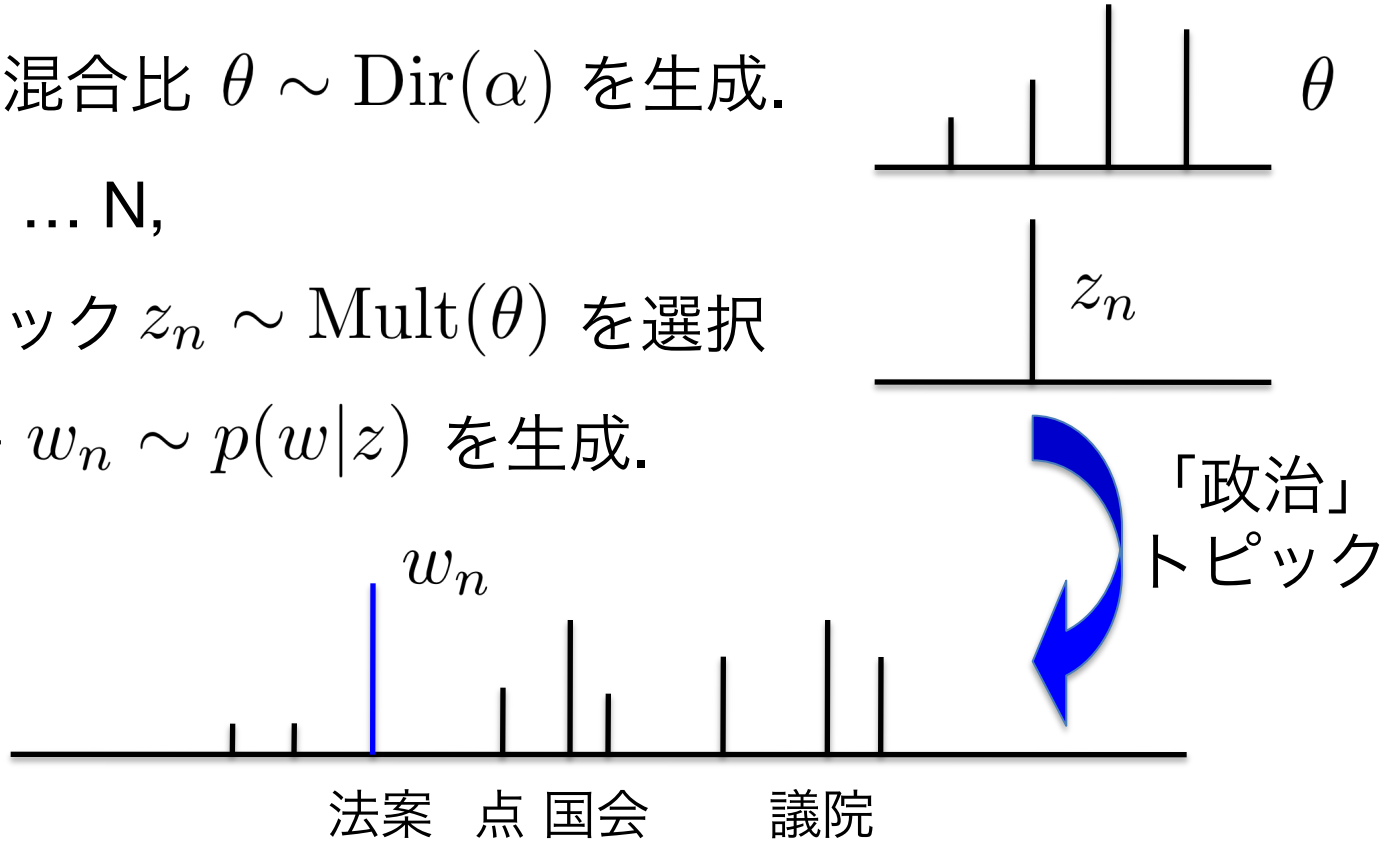


- 混合比 θ をディリクレ事前分布から生成



LDAの生成モデル (2)

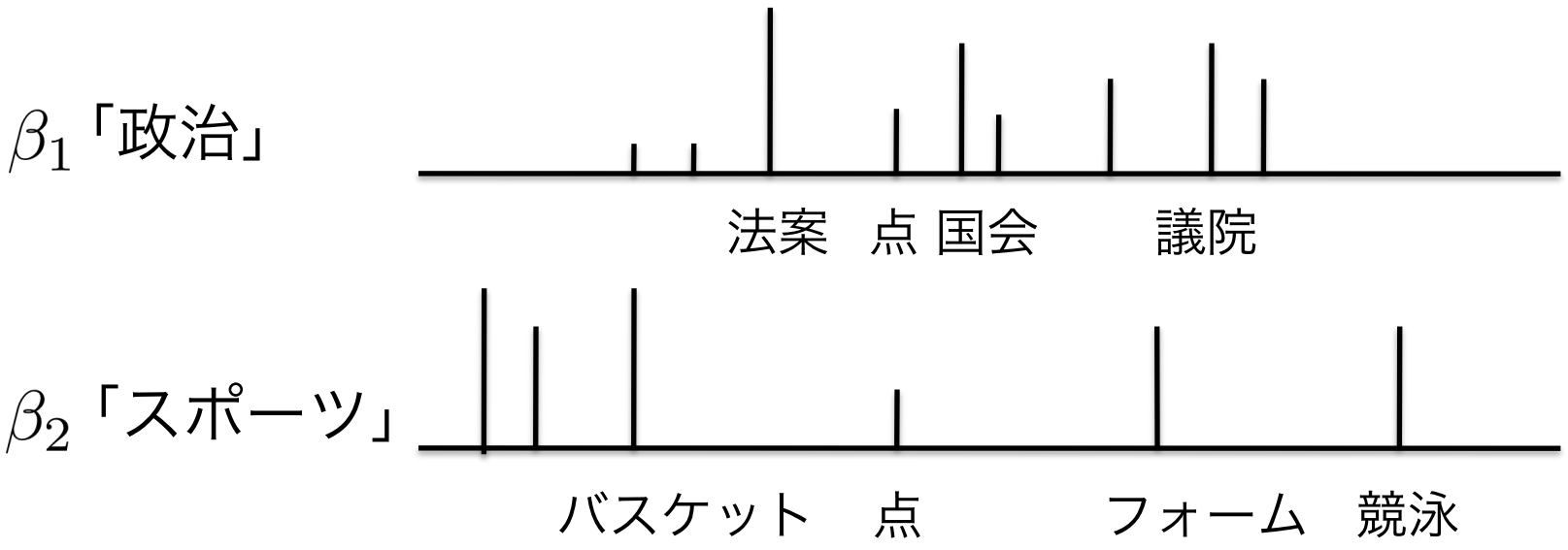
1. トピック混合比 $\theta \sim \mathrm{Dir}(\alpha)$ を生成.
2. For n = 1 … N,
   a. トピック $z_n \sim \mathrm{Mult}(\theta)$ を選択
   b. 単語 $w_n \sim p(w|z)$ を生成.



LDAの生成モデル (3)

- 「話題」とは?→単語の生起確率分布 $\beta_k = \{ p(w|k) \}$

$(w = 1 \cdots V)$



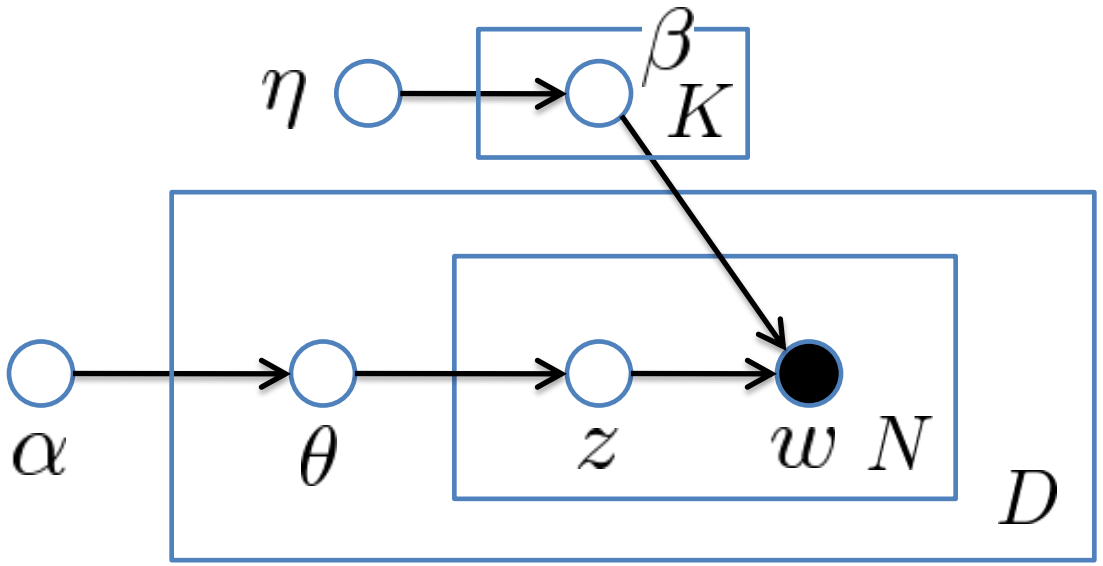
LDAの学習例

- 川端康成「雪国」の冒頭

  国境の長いトンネルを抜けると雪国であった。
  夜の底が白くなった。信号所に汽車が止まった。
  向側の座席から娘が立って来て、島村の前のガラス
  窓を落した。雪の冷気が流れこんだ。…

  – 2000年度毎日新聞記事全文 (2,887万語) で学習したモデルで分析
    - 青色のトピックは冬に関係する
    - 緑色のトピックは電車に関係する
    - 黒色は地の文

LDAのグラフィカルモデル



- $\theta \to z \to w$ の順で単語 w を生成
- θ は D 回(文書数)、z は N 回(単語数) 生成
  - $\beta = (\beta_1, \cdots, \beta_K)$ はトピック別単語分布 $p(w|k)$

29

LDAの確率モデル

$$p(w, z, \theta) = p(w|z)p(z|\theta)p(\theta|\alpha)$$

観測単語　トピック　トピック分布

よって、文書 $\mathbf{w} = w_1 w_2 \cdots w_N$ について

$$p(\mathbf{w}, z, \theta) = p(\theta|\alpha) \prod_n p(w_n|z_n)p(z_n|\theta)$$

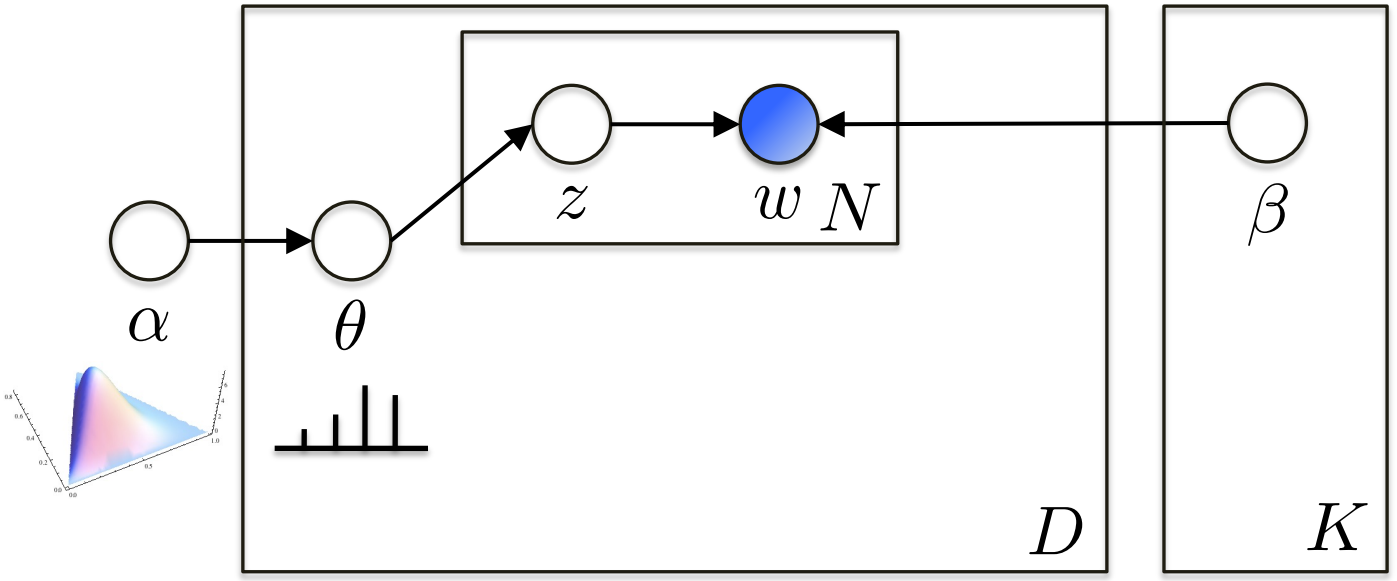
$$p(\mathbf{w}) = \int \sum_z p(\mathbf{w}, z, \theta) d\theta$$

$$= \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \int \left(\prod_k \theta_k^{\alpha_k - 1} \right) \prod_n \sum_k p(w_n|k)\theta_k d\theta$$

- パラメータは α と $\beta = \{ p(w|k) \}$

Spectral Mixture LDA

- 動画に付与された副詞 w の集合と、

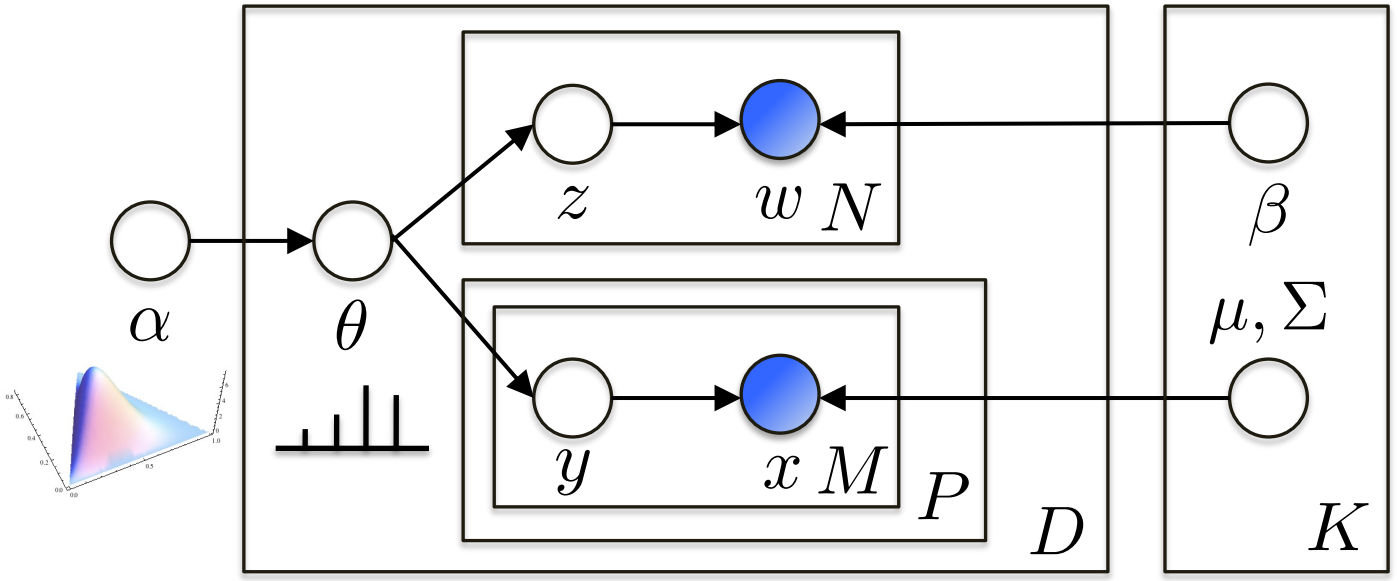


w：副詞
z：トピック
θ：潜在トピック分布

- 各トピックは、副詞の確率分布を持つ

Spectral Mixture LDA

- 動画に付与された副詞の集合と、
  動作の周波数空間での成分の同時分布をモデル化
  → スペクトル混合潜在ディリクレ配分法 (SMLDA)

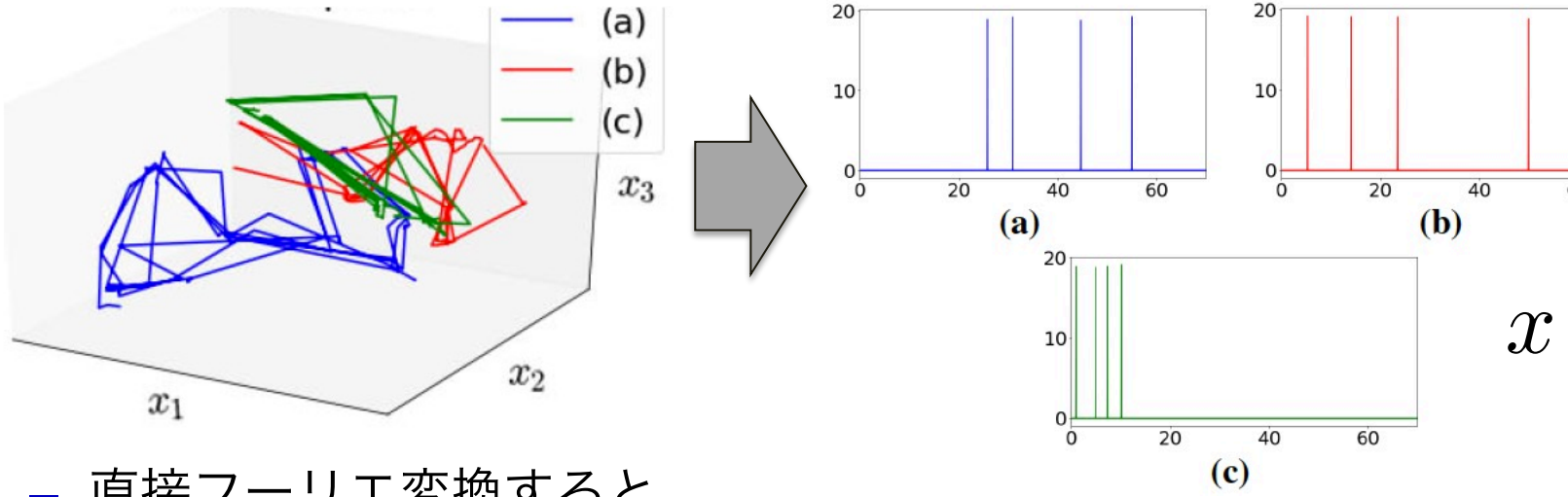


w：
副詞
x：
周波数
z, y：
トピック
θ：
潜在トピック分布

- 各トピックは、副詞の確率分布と周波数空間でのガウス分布の両方を持つ

動画からの副詞の理解 (再掲)

- スペクトル混合カーネルの適用：カーネルの周波数空間での特徴を抽出



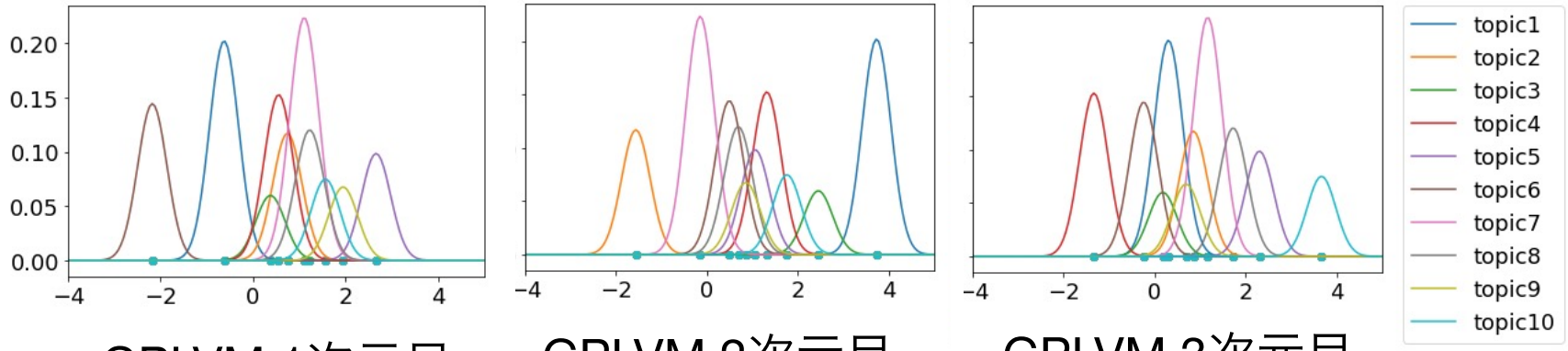
  - 直接フーリエ変換すると、
    関数の位相に依存＆FFTにはデータが不足

- 一方で、各動作には「しっかり」「堂々と」などの副詞が付与されている

Spectral Mixture LDA (2)

- 推定された副詞のトピックと、対応する対数周波数

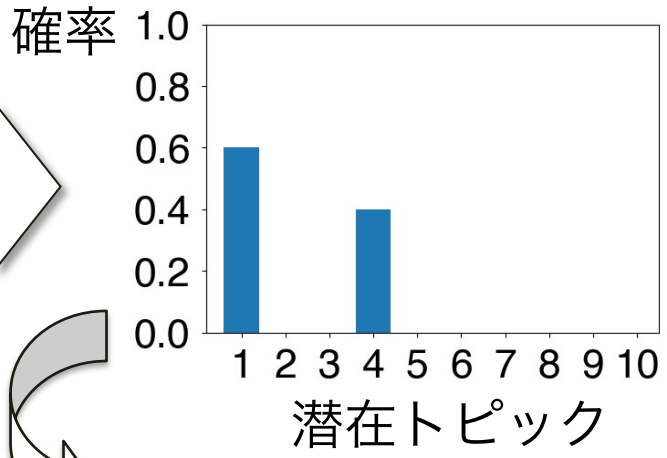
Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
硬く	だらだら	大股で	痛そうに	楽しそうに	堂々と	ゆっくりと
奇妙に	元気なく	威嚇して	クタクタな	リズミカル	サッサッ	じっくりと
ふざけて	気だるく	どっしりと	足が痛い	調子よく	普通に	静々と
ぎこちなく	脱力して	忍んで	辛そうに	きびきび	力強く	確実に
不自然に	辛い	偉そうな	よろよろ	軽快に	颯爽と	一歩ずつ



GPLVM 1次元目　GPLVM 2次元目　GPLVM 3次元目

Spectral Mixture LDA (3)

- テストデータの動画からトピック分布と副詞を予測

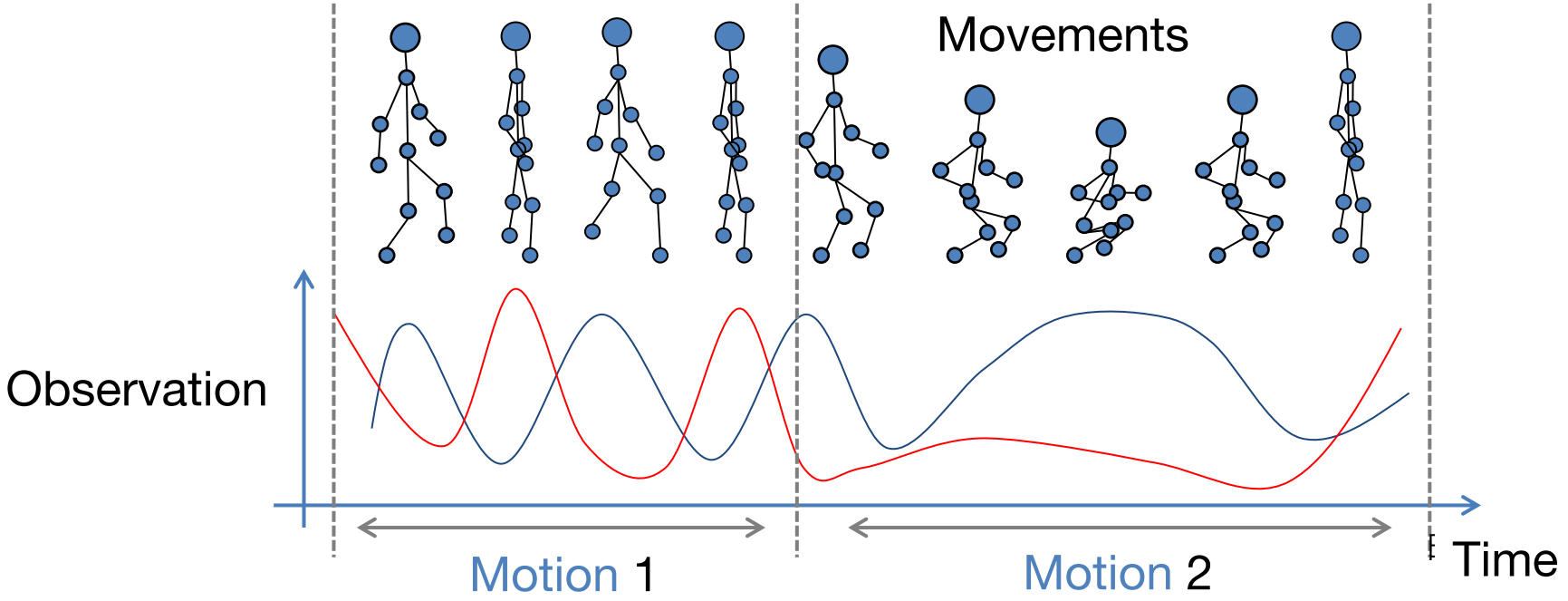


副詞	予測確率
ぎこちなく	0.056
痛そうに	0.054
不自然に	0.048
ゆっくり	0.036
ふざけて	0.035

予測パープレキシティ：
149 (ユニグラム)
→ 32 (スペクトル混合LDA)

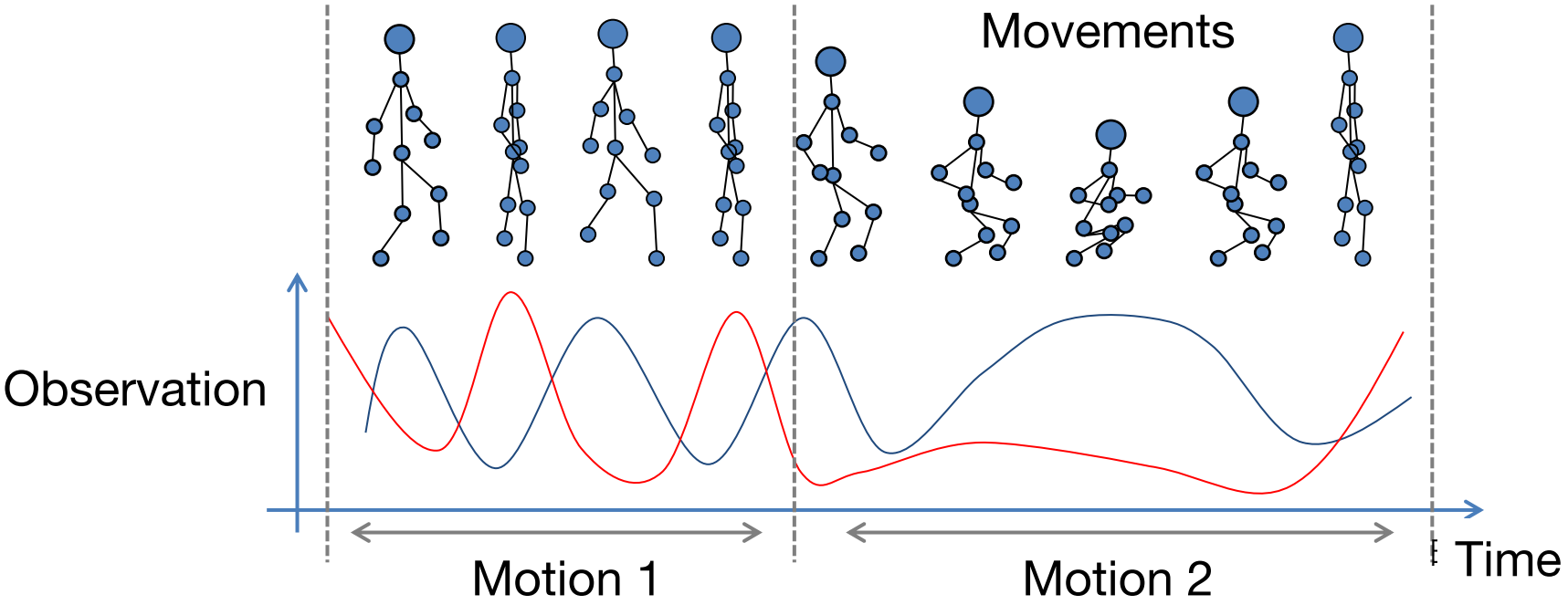
(2) ガウス過程と階層ディリクレ過程による「動作」の学習

From Movements to Actions



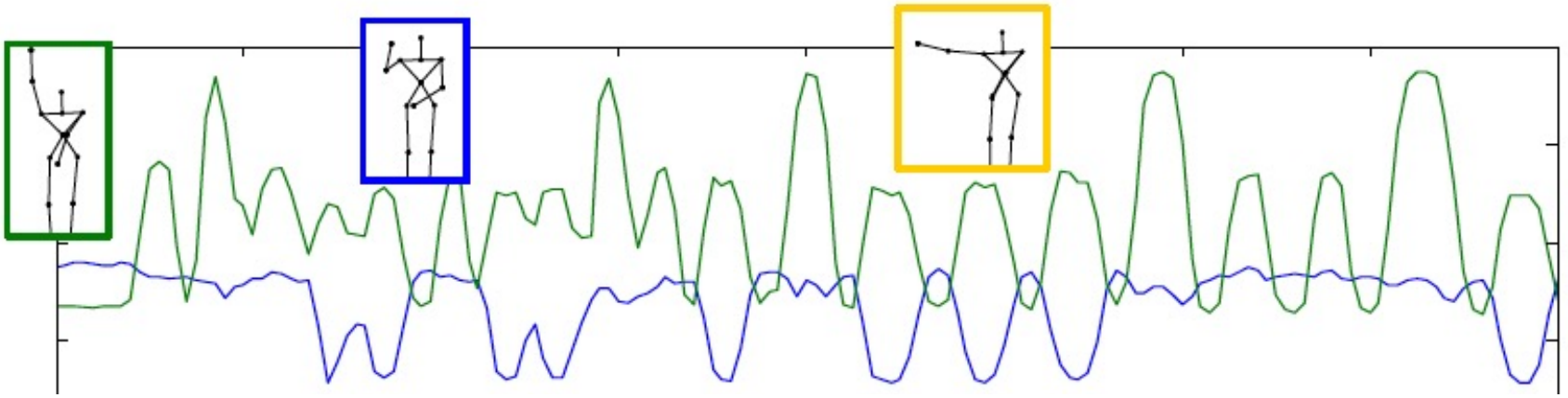
- Crucial for *high-level recognition and planning of actions of robots*
  - Planning for preparing dishes: wash->cut->boil->...

Model of movements



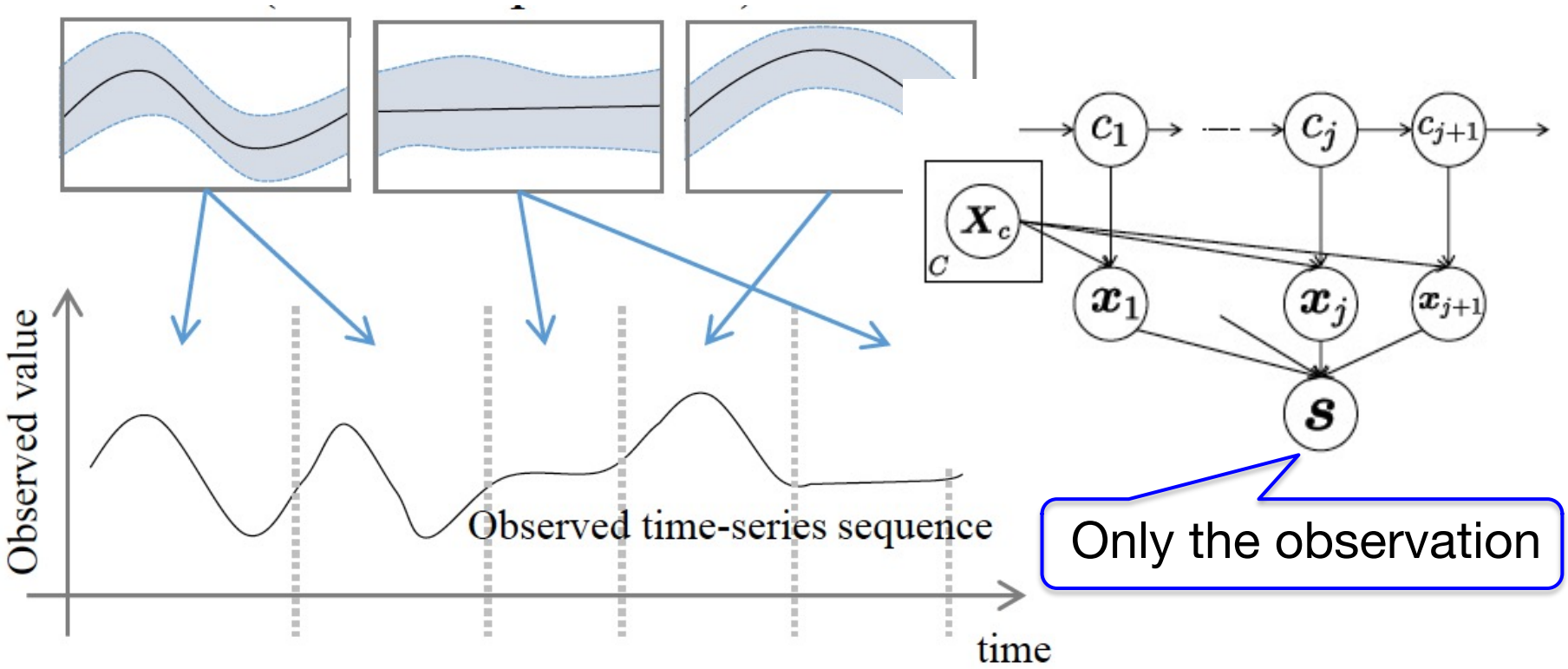
- Observations are continuous→ Stochastic models for trajectories are necessary
- We leverage the Gaussian processes

Actual data of movements



- Time series from two angles (knee and shoulder)
- How to induce “motion” from this data?

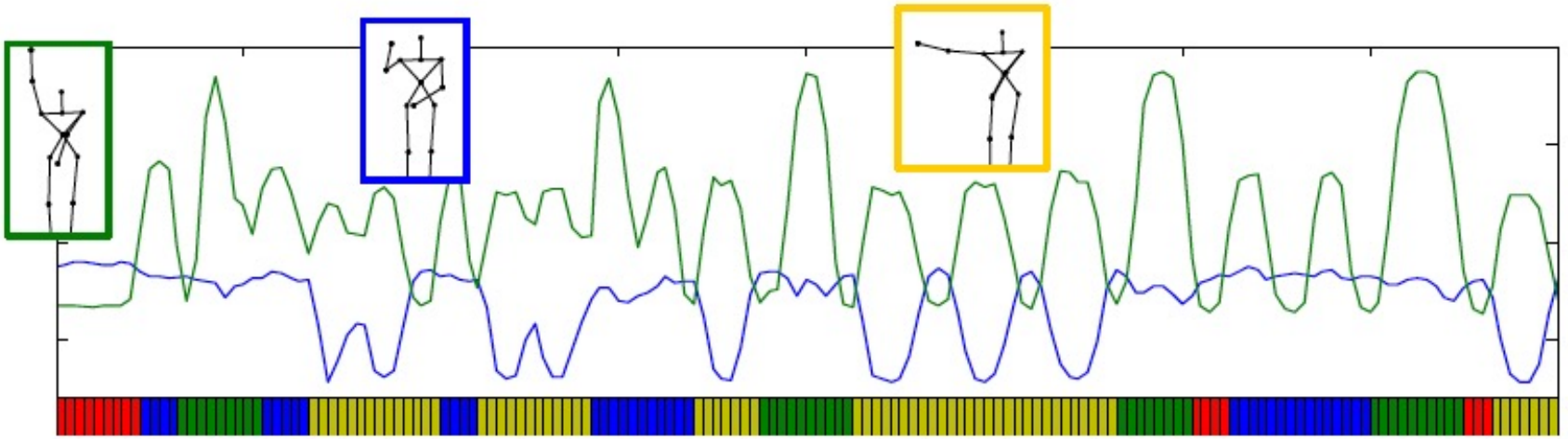
Model using Gaussian processes



- Hidden state (motion class): $c_j \sim \mathrm{Mult}(c|c_{j-1})$
- Generate trajectory: $\mathbf{x}_j \sim \mathrm{GP}(\mathbf{x} \,|\, \mathbf{X}_{c_j})$
- HMM, but segmentations are unknown

Application to Robotics

- Hidden semi-Markov model with Gaussian process observations to model the time series of joint angles, Forward-Backward Bayesian learning (infer “words”)
- Segmentation results using simple arm motions :



Segmentation result

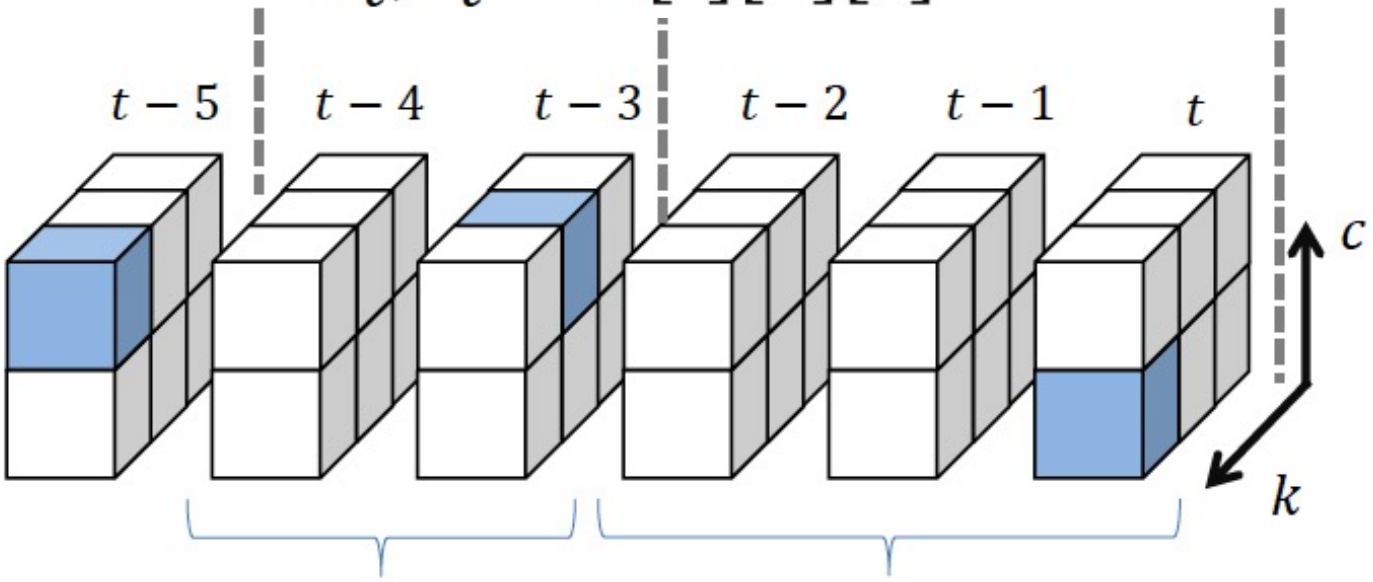
Dynamic programming MCMC

- Forward filtering:

$$\alpha[t][k][c] = GP(\boldsymbol{s}_{t-k:k}|\boldsymbol{X}_c)\sum_k\sum_c p(c|c')\alpha[t-k][k'][c']$$

- Backward sampling:

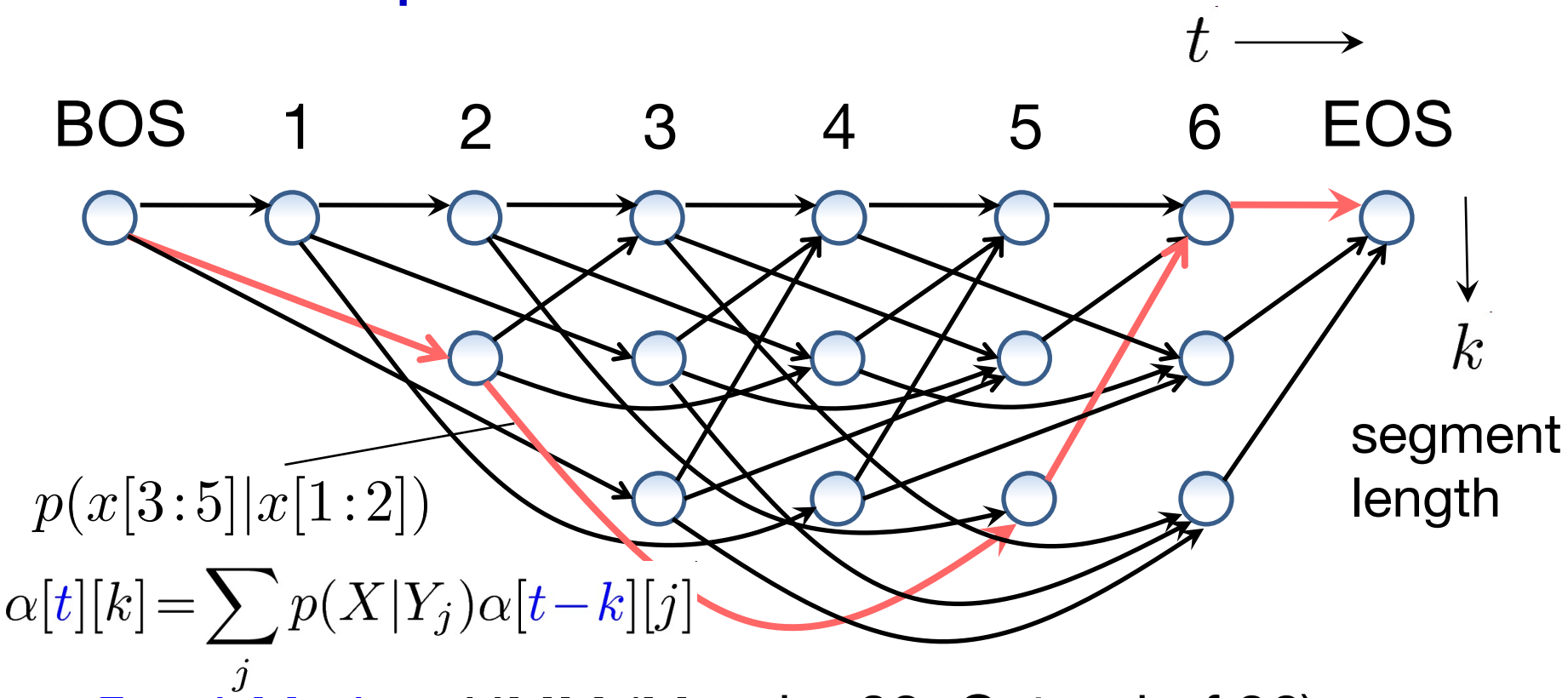
$$k_t, c_t \sim \alpha[t][k][c]$$



$k_{t-3}, c_{t-3} \sim \alpha[t-3][k][c]$
$(k_{t-3} = 2, c_{t-3} = 2)$

$k_t, c_t \sim \alpha[t][k][c]$
$(k_t = 3, c_t = 1)$

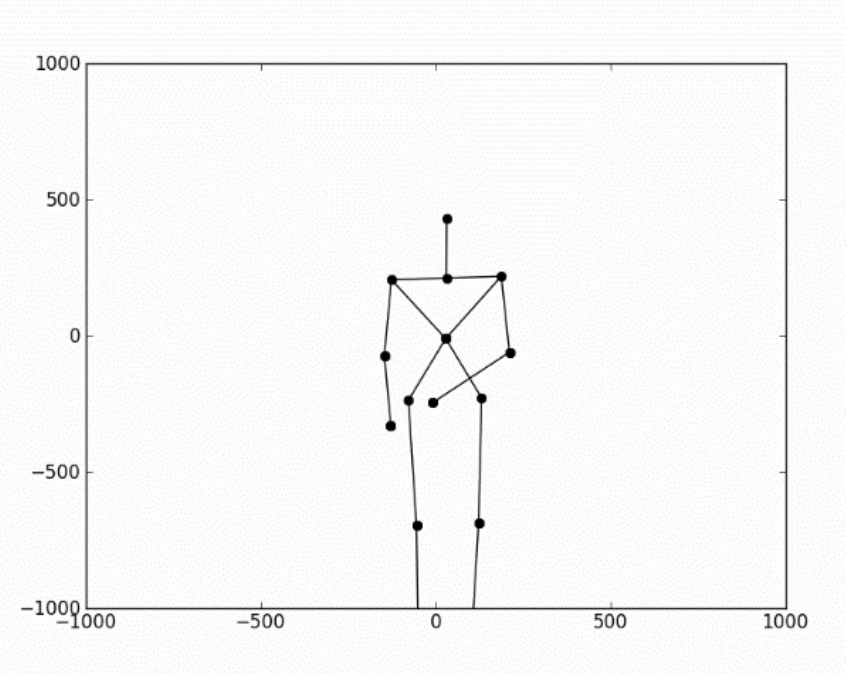
Lattice representation



- Semi-Markov HMM (Murphy 02, Ostendorf 96)
  - We do not know “correct path” beforehand (as HMM)
- Draw high-probability path via dynamic programming

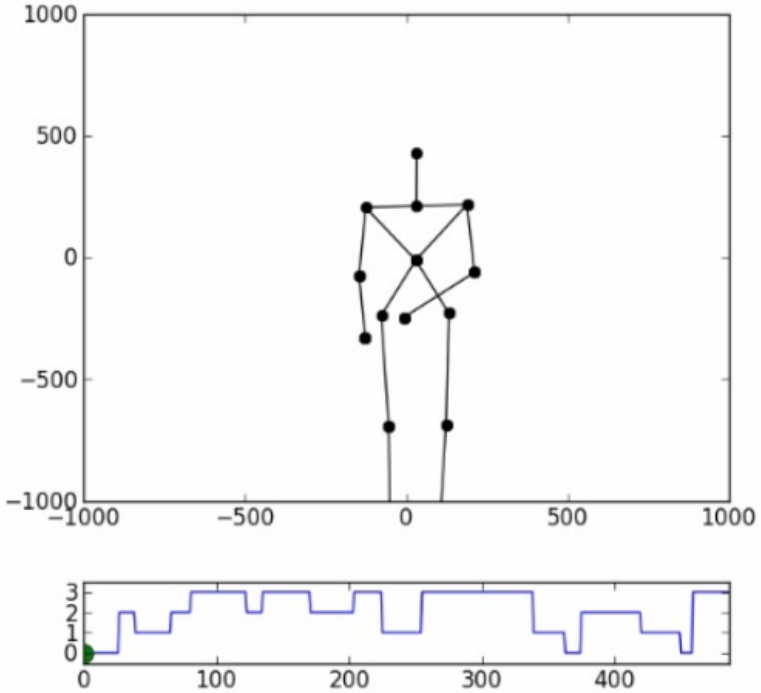
The first experiment

- Unsupervised segmentation of arm movements measured by Kinect
- 2D coordinates (x,y) of the right hand
  - Assume x and y are independently generated: x ~ GP(c), y ~ GP(c)
- Motions involved:
  - Move the hand to right
  - Raise the hand high
  - Raise the hand slightly

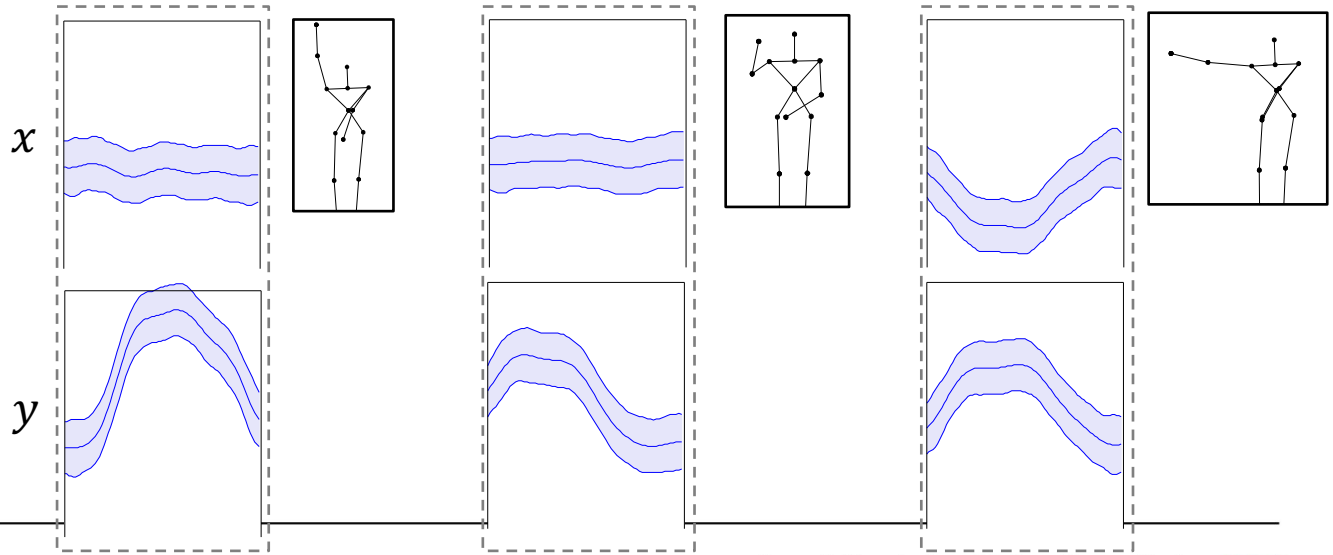


Results

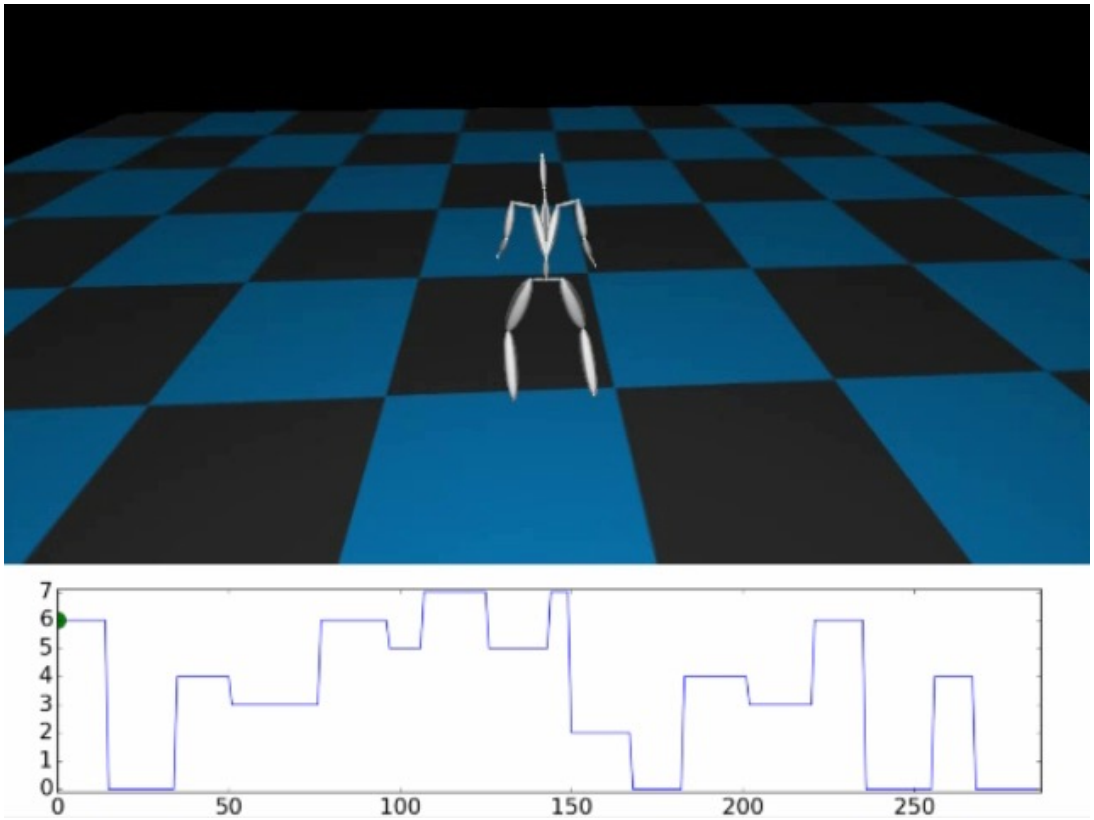
- Correctly recovered the three motions



- Learned Gaussian processes:



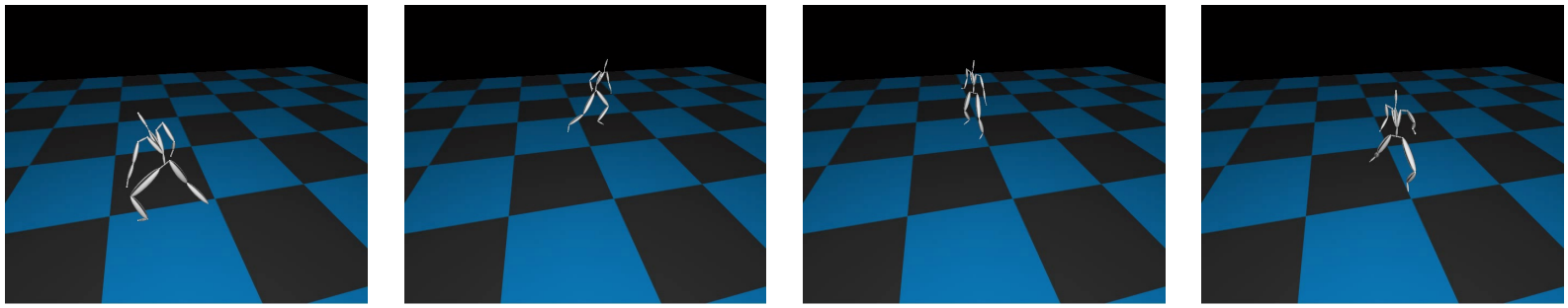
Motion segmentation



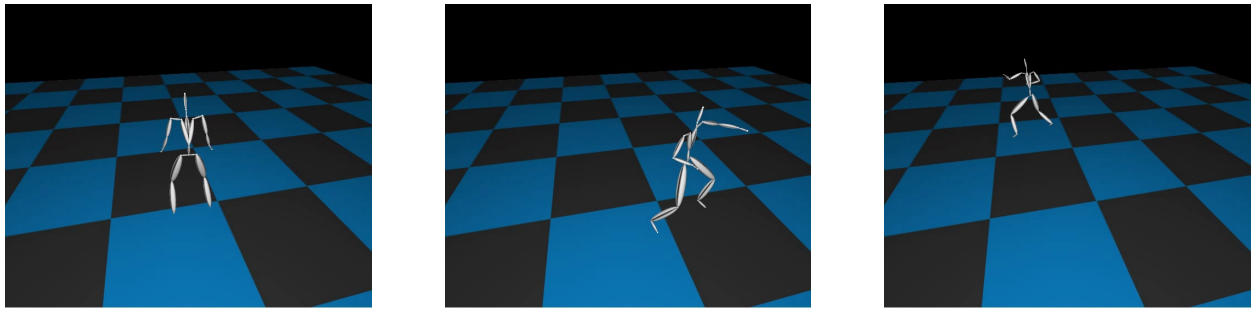
- Correctly recognized Karate motions from observations

Inferred motions (excerpt)

- Class 0: Right punch with an additional step

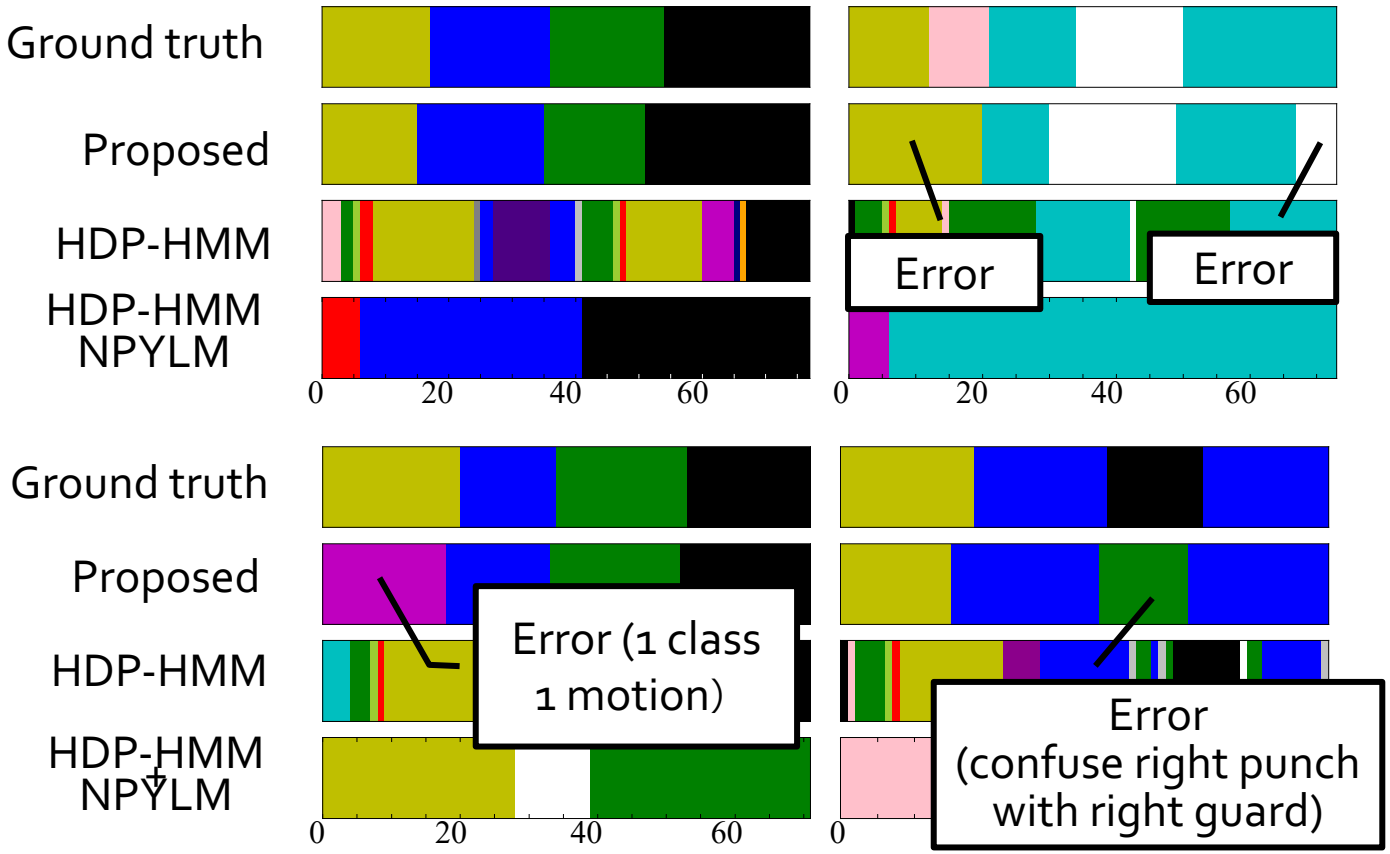


- Class 6: Left lower guarding



Comparison with other methods

- GP-HSMM vs. HDP-HMM, HDP-HMM+NPYLM



Number of Motions

- Number of hidden states (motions) can also be estimated using hierarchical Dirichlet processes
  - HDP-GP-HSMM
  - “Sequence Pattern Extraction by Segmenting Time Series Data Using GP-HSMM with Hierarchical Dirichlet Process”, Nagano+, IROS 2018

TABLE V

SEGMENTATION RESULTS FOR THE EXERCISE MOTION.

	Hamming distance	Precision	Recall	F-measure	# of estimated classes
HDP-GP-HSMM	0.31	0.38	0.95	0.55	10
HDP-HMM	0.82	0.070	1.0	0.13	14
HDP-HMM+NPYLM	0.63	0.61	1.0	0.76	26
BP-HMM	0.23	0.25	1.0	0.40	18
Autoplait	0.61	0.67	0.18	0.28	5

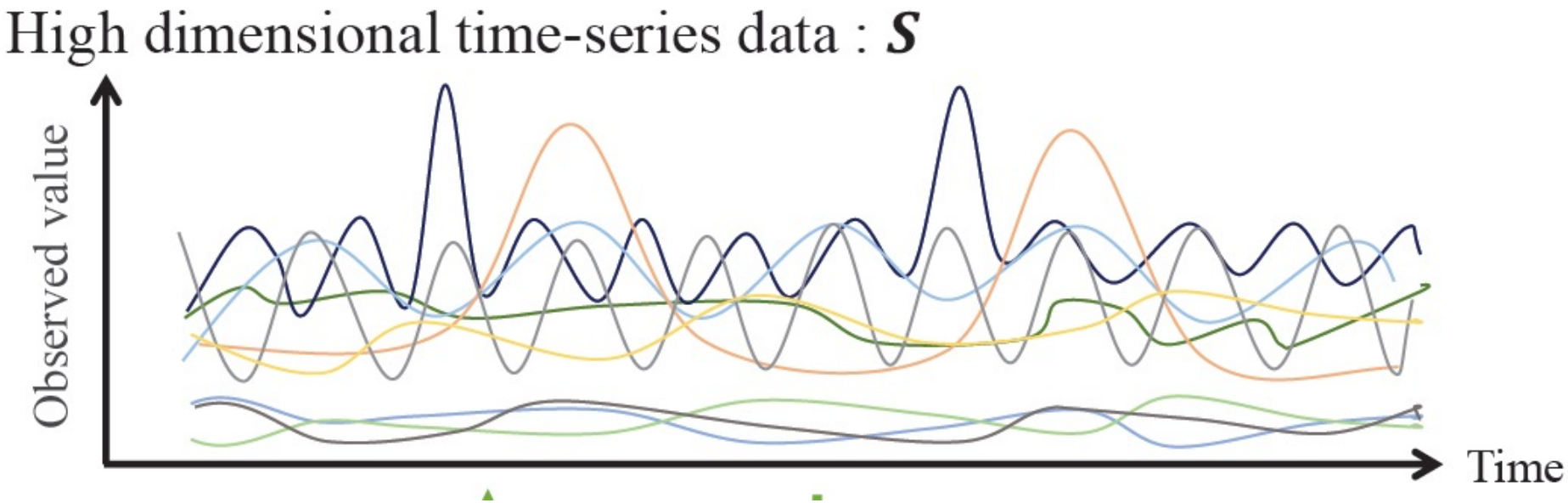
Ground truth: 11

Number of Motions (2)

- We leveraged infinite HMM (Beal+ 2001, Teh+ 2006) with our semi-Markov structure
- Since the state space is infinite-dimensional, also used a Beam sampling (van Gael+2007) for slice sampling+dynamic programming
  - Internally,using a stick-breaking representation for possibly infinite state spaces

High-dimensional regime

- Actually, robot movements are quite high-dimensional
  - In our case, # of joint angles = 93
  - Cannot apply the method to whole data



Strategy (to appear at IROS 2019)

- Solution: dimensionality reduction
- Linear PCA → NG
- GPLVM (Gaussian process LVM) → OK, but inference is not stable

- Using VAE as a surrogate of GPLVM

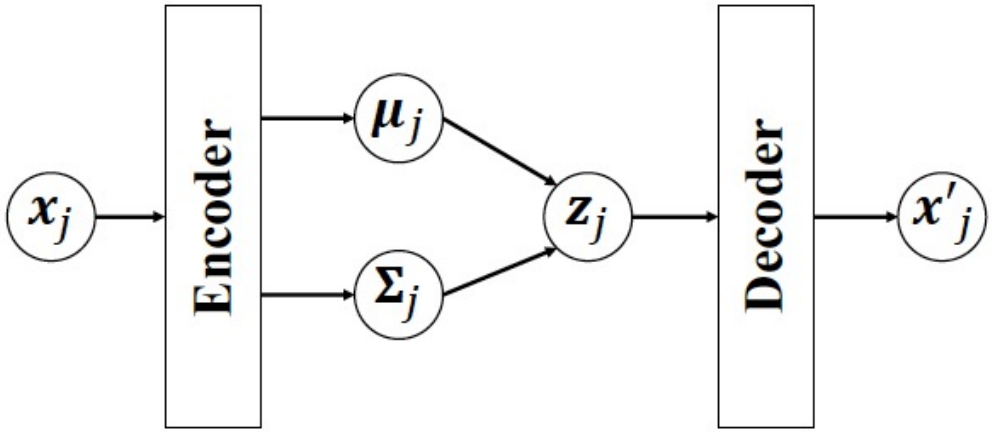
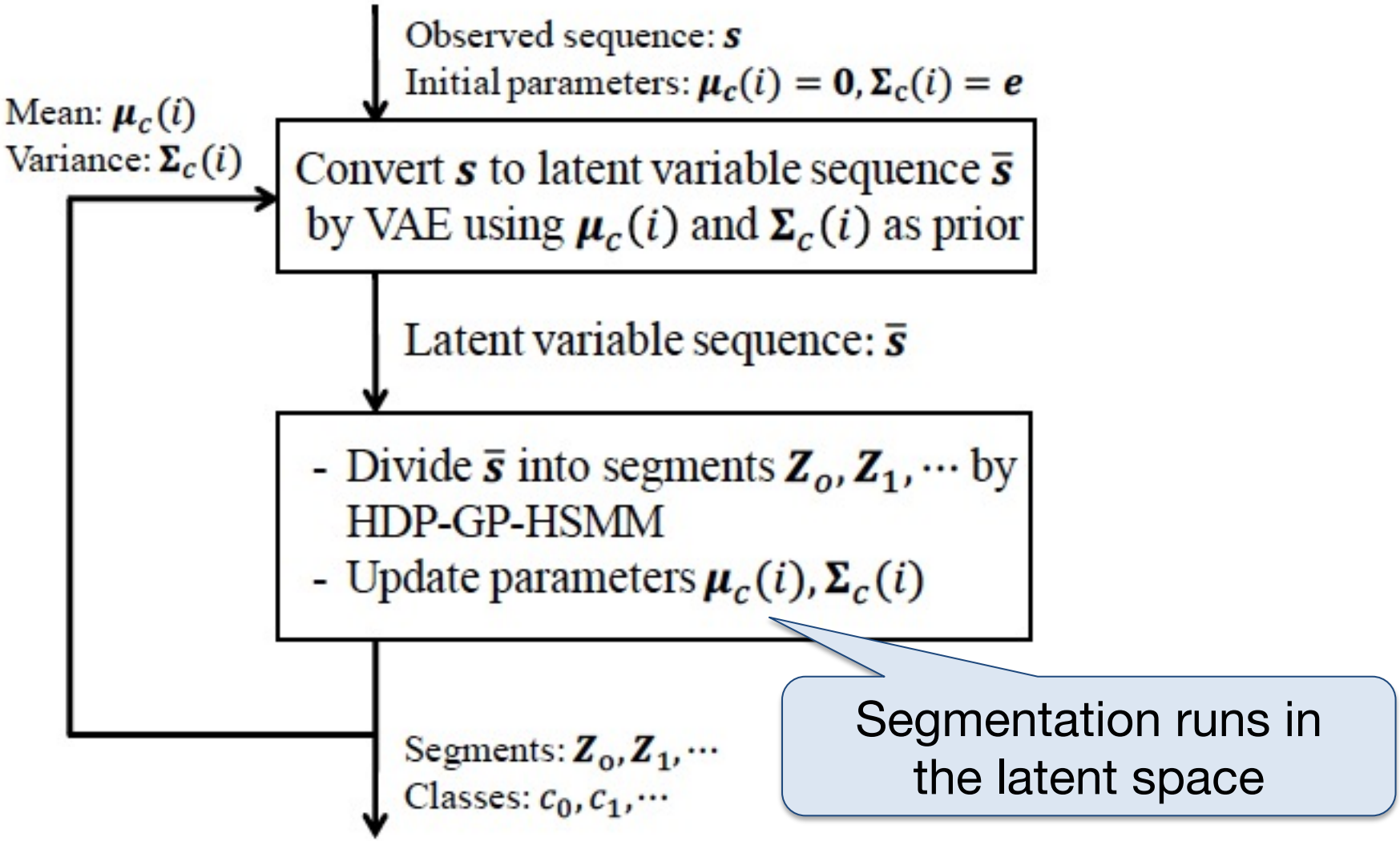


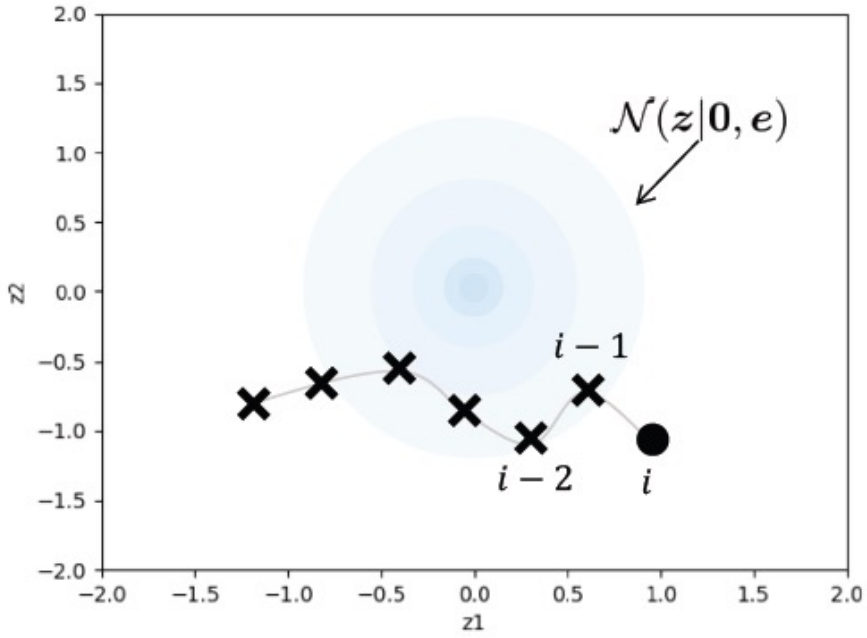
Fig. 3. Variational autoencoder (VAE) to obtain the latent low-dimensional representation z_j of observed time series x_j.

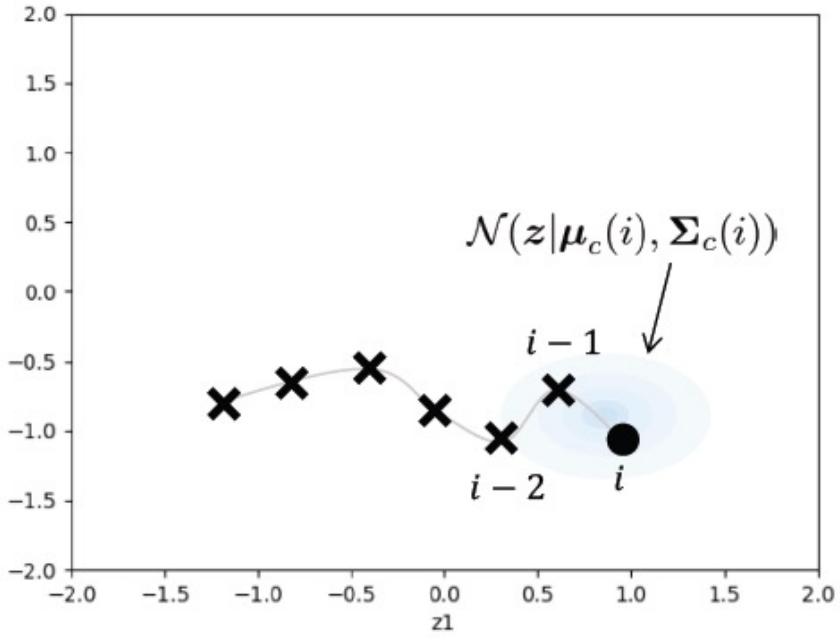
Simultaneous optimization



Note

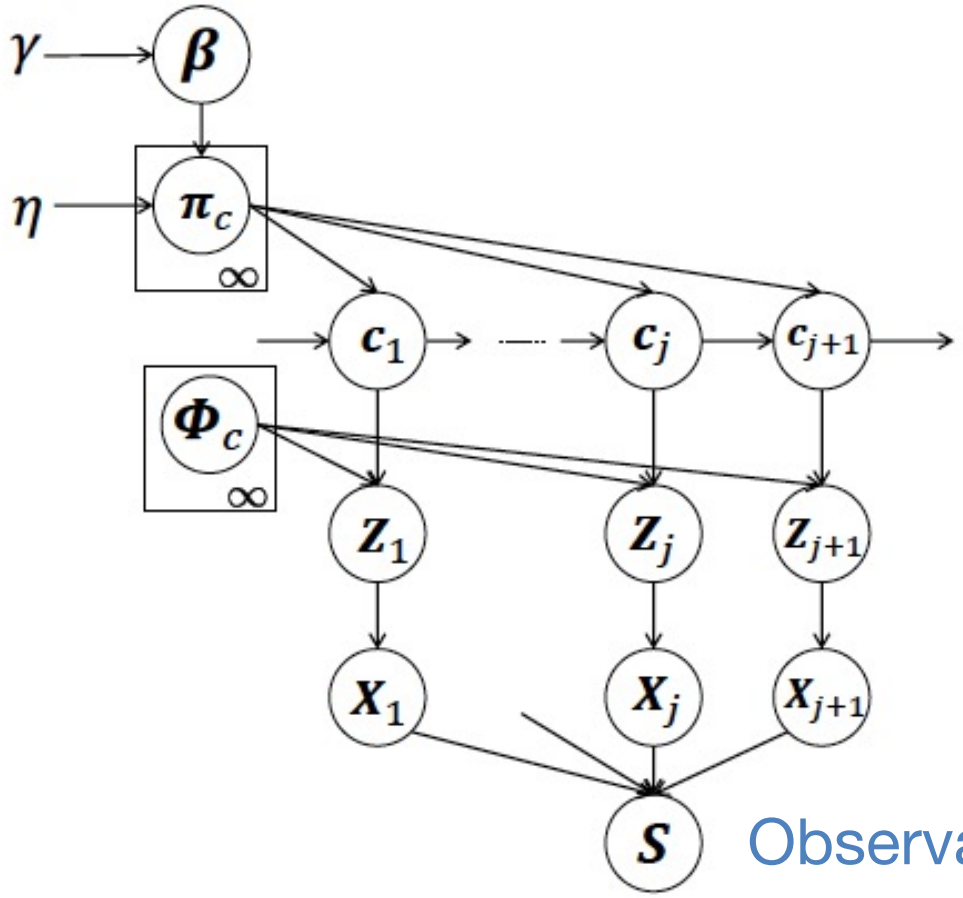
- In this case, each cluster (=motion) has its own VAE for compression
  - VAE priors are different for each cluster





Graphical model

- We observe only $\boldsymbol{s}$ (high-dimensional time series)



Hierarchical Dirichlet process-variational autoencoder-Gaussian process-hidden semi-Markov model (Long!)
=
HVGH

Experiments

- Dance exercises which include four and seven unit motions (labels are not used in learning)

図 4: Four unit motions included in the chicken dance: (a) beaks, (b) wings, (c) tail feathers, and (d) claps.

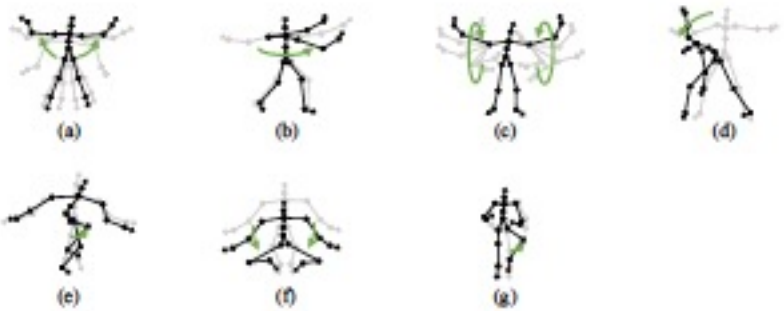


図 5: Seven unit motions included in the exercise motion1: (a) jumping jack, (b) twist, (c) arm circle, (d) bend over, (e) knee raise, (f) squatting, and (g) jogging

Results (1)

- Exercise containing four unit motions:

表 1: Segmentation results for the chicken dance.

	Hamming distance	Precision	Recall	F-measure	# of estimated classes
HVGH	0.23	0.86	0.86	0.86	4
VAE+HDP-GP-HSMM	0.31	1.0	0.71	0.83	4
VAE+HDP-HMM	0.74	0.15	1.0	0.26	11
VAE+ HDP-HMM+NPYLM	0.48	1.0	0.86	0.92	7
VAE+BP-HMM	0.34	1.0	0.86	0.92	3
VAE+Autoplait	0.66	0.0	0.0	0.0	1

VAE as a preprocessing

Results (2)

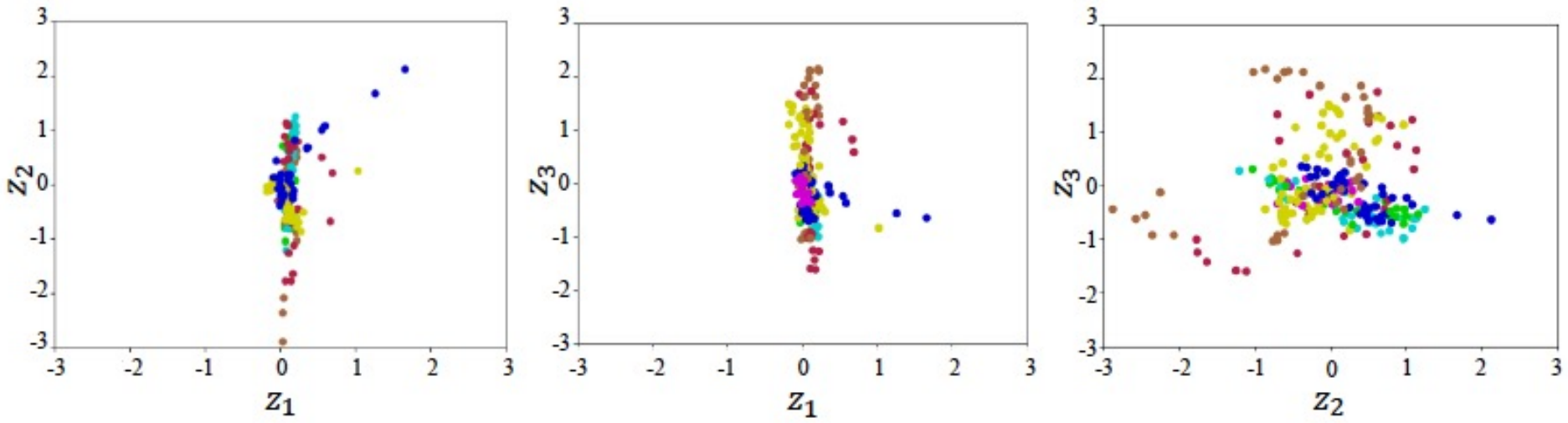
- Exercise containing seven unit motions:

表 2: Segmentation results for the exercise motion.

	Hamming distance	Precision	Recall	F-measure	# of estimated classes
HVGH	0.16	0.66	0.93	0.75	11
VAE+HDP-GP-HSMM	0.24	0.53	0.93	0.67	12
VAE+HDP-HMM	0.75	0.05	1.0	0.09	10
VAE+ HDP-HMM+NPYLM	0.61	0.30	1.0	0.45	28
VAE+BP-HMM	0.58	0.29	0.97	0.44	7
VAE+Autoplait	0.76	0.0	0.0	0.0	2

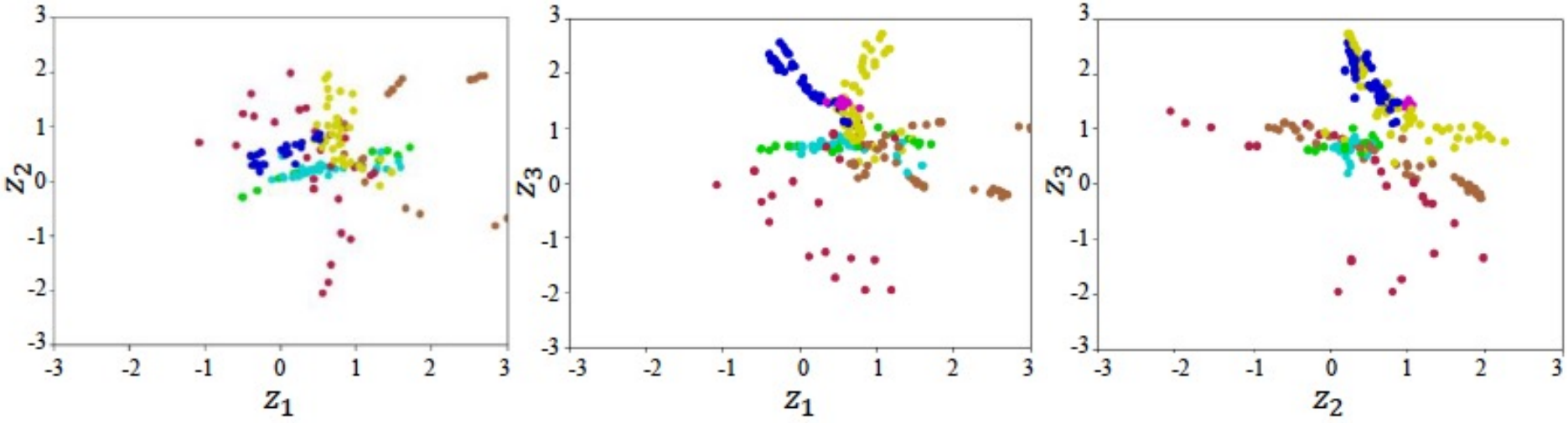
VAE as a preprocessing

Estimated latent space in VAE



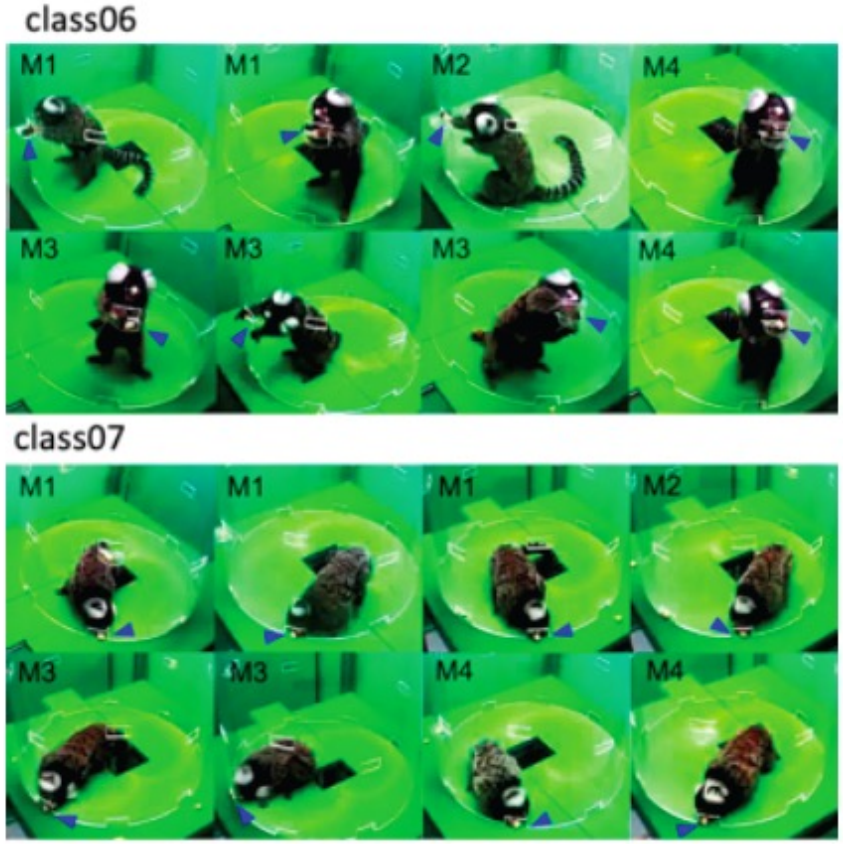
- When VAE is separately learned for compression
- Motions (in color) are mixed and not separated

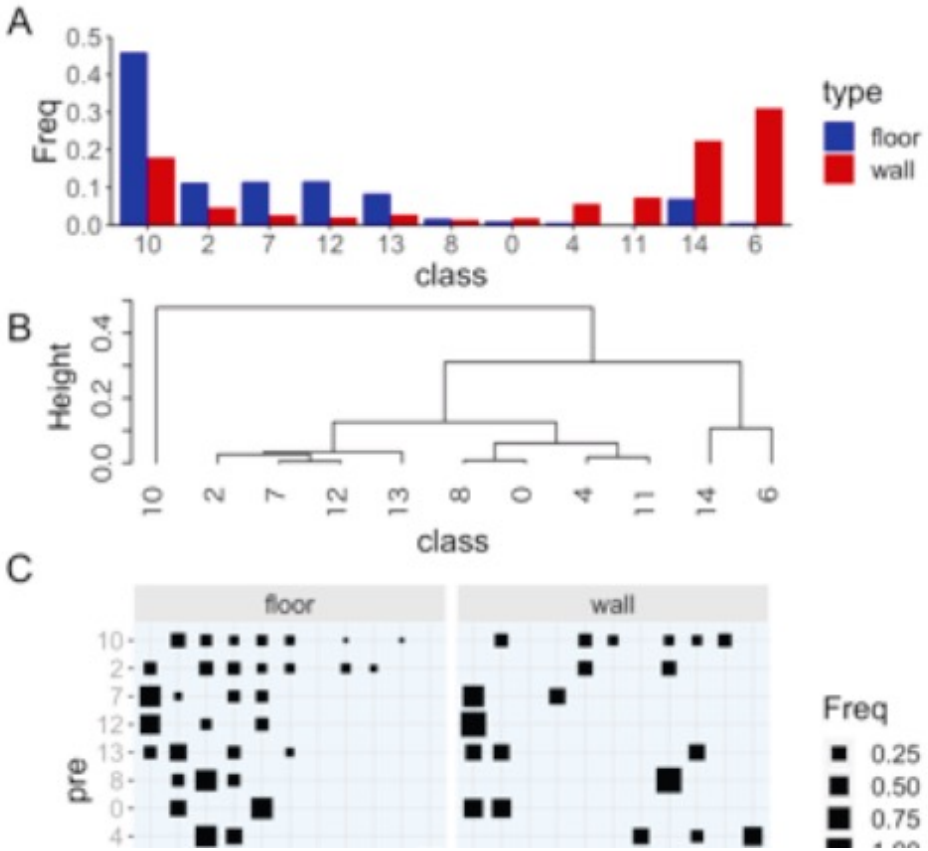
Estimated latent space in VAE (2)



- When simultaneously learned in HVGH
- Motions (in color) are clearly separated

Zoology and Brain science (Marmoset)





- Can recognize ape's motions *completely automatically*
- joint work with Koki Mimura (NCAP), JSAI 2019

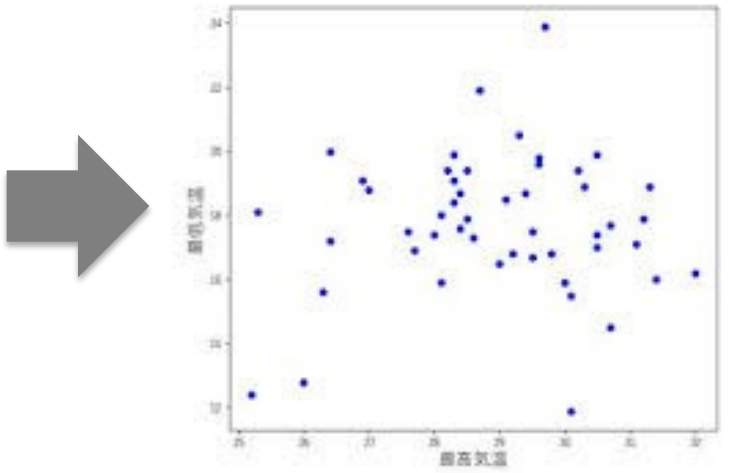
(3) Nonparametric Bayesian Deep Visualization

(ブリヂストン(株)との共同研究, IBISML 43 (2021)で発表済)

データ可視化

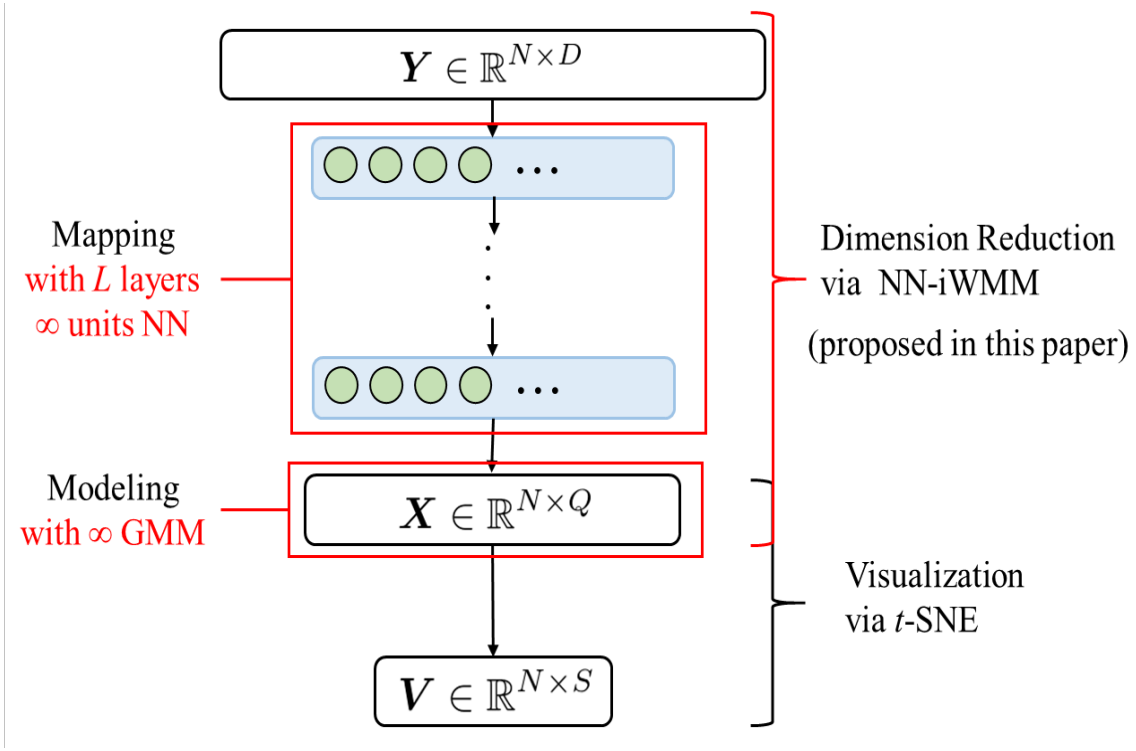
- 高次元データを2次元または3次元に変換して可視化
  - データ科学の最初のステップとして、依然として不可欠

個体ID	Y1	Y2	…	Y_D
1				
2				
・・・				
N				



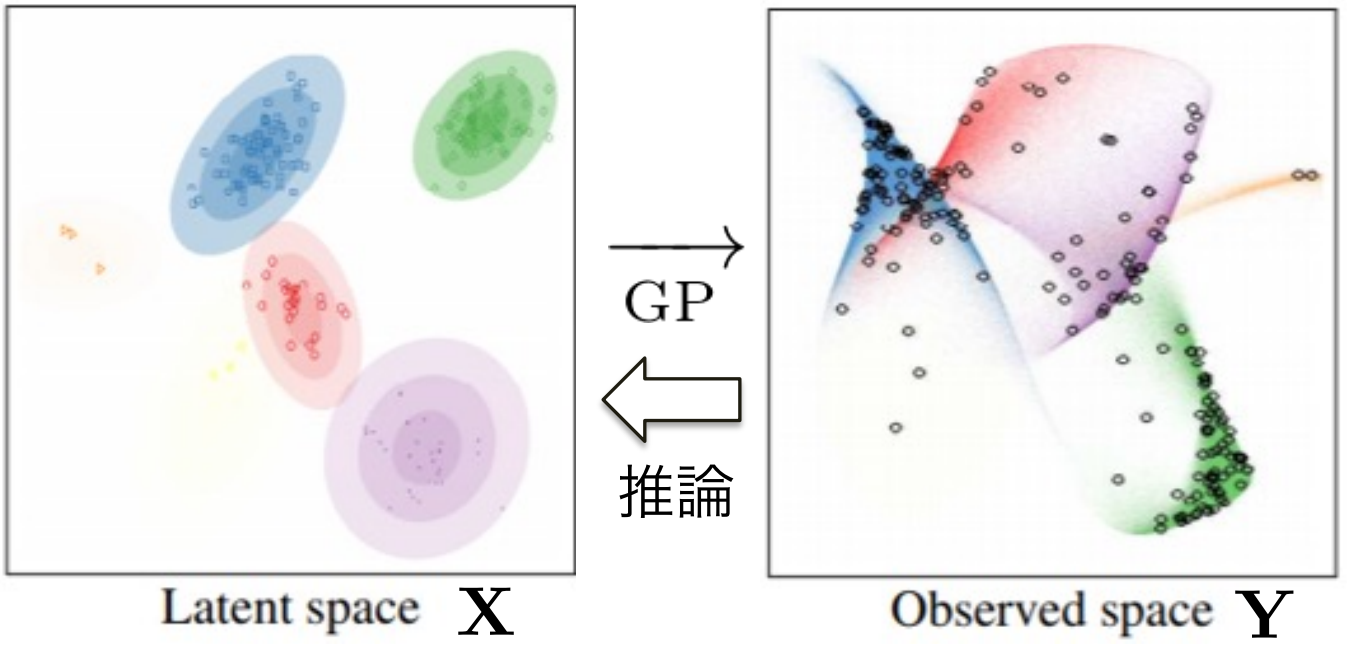
- 主成分分析(PCA)では限界がある…線形変換では捉えられない非線形な変換が必要な場合が多い

NPDV: NP Bayesian Deep Visualization



- $\boldsymbol{Y} \rightarrow \boldsymbol{X}$, $\boldsymbol{X} \rightarrow \boldsymbol{V}$への圧縮を同時に最適化
- $\boldsymbol{Y} \rightarrow \boldsymbol{X}$：NN-iWMMガウス過程によるニューラルネット
- $\boldsymbol{X} \rightarrow \boldsymbol{V}$：t-SNEによる可視化

無限ワープ混合モデル (iWMM)



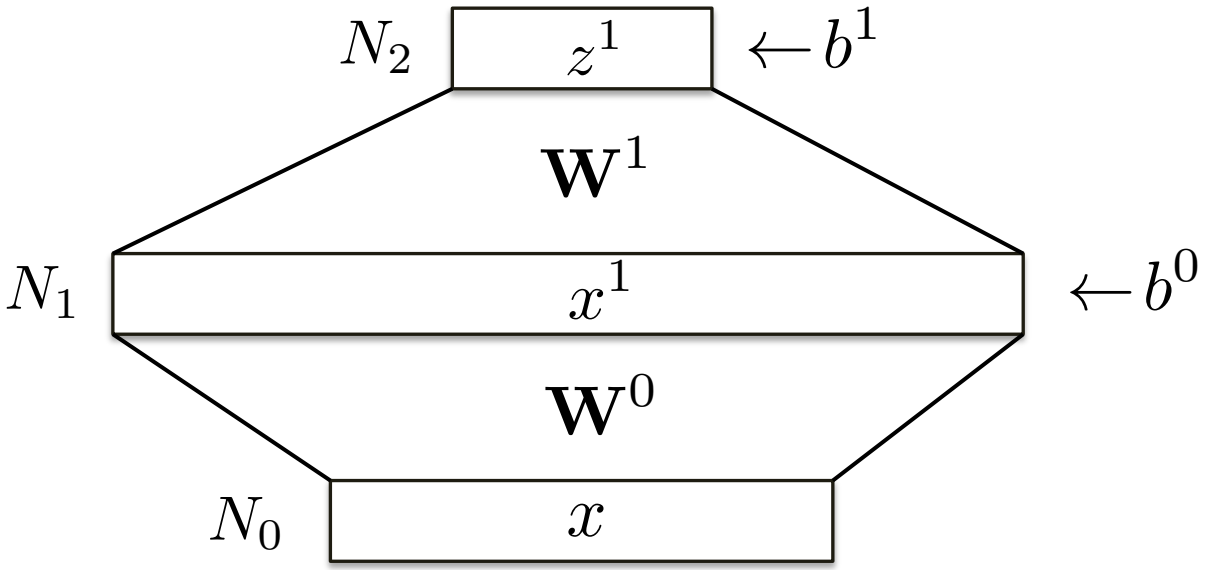
(Iwata+ 2013)
より引用

- 潜在空間での無限混合ガウス分布→ガウス過程で観測値へ変換
- 観測データYのd次元目 Y_d が、潜在空間Xでのカーネルを使ったGPに従う　$Y_d \sim \mathrm{GP}(\boldsymbol{\mu}, K_{\mathbf{X}})$

NNGPカーネルとニューラルネット

- 1層のニューラルネットで、入力を x とする
- i番目の出力 $z_i^1(x)$ は

$$z_i^1(x) = b_i^1 + \sum_{j=1}^{N_1} W_{ij}^1 x_j^1(x), \quad x_j^1(x) = \phi\left(b_j^0 + \sum_{k=1}^{N_0} W_{jk}^0 x_k\right)$$



結合重み

$W_{ij}^\ell \sim \mathcal{N}(0, \sigma_w/N_\ell)$

バイアス項

$b_i^\ell \sim \mathcal{N}(0, \sigma_b)$

と仮定する

NNGPカーネルとニューラルネット (2)

$$z_i^1(x) = b_i^1 + \sum_{j=1}^{N_1} W_{ij}^1 x_j^1(x), \quad x_j^1(x) = \phi\left(b_j^0 + \sum_{k=1}^{N_0} W_{jk}^0 x_k\right)$$

- 重み W_{ij}^ℓ とバイアス b_i^ℓ は独立なので、出力は独立な確率変数の和で、中心極限定理から正規分布に従う
  → 結合分布 $p(z_1^1(x), z_2^1(x), \cdots, z_{N_2}^1(x))$
  も多変量正規分布に従う…ガウス過程
  - 平均は上式から明らかに0
  - 共分散は

$$K^1(x, x') \equiv \mathbb{E}\left[z_i^1(x) z_i^1(x')\right]$$
$$= \sigma_b^2 + \sigma_w^2 \mathbb{E}\left[x_i^1(x) x_i^1(x')\right] = \sigma_b^2 + \sigma_w^2 C(x, x')$$

NNGPカーネルとニューラルネット (3)

- $\ell-1$ 層目の出力 $z_j^{\ell-1}$ がガウス過程だと仮定すると

$$z_i^\ell(x) = b_i^\ell + \sum_{j=1}^{N_\ell} W_{ij}^\ell \, x_j^\ell(x), \quad x_j^\ell(x) = \phi(z_j^{\ell-1}(x))$$

- 平均は同様に0, 共分散は

$$\begin{aligned} K^\ell(x, x') &\equiv \mathbb{E}\left[z_i^\ell(x) z_i^\ell(x')\right] \\ &= \sigma_b^2 + \sigma_w^2 \, \mathbb{E}_{z_i^{\ell-1} \sim \mathrm{GP}(0, K^{\ell-1})} \left[\phi(z_i^{\ell-1}(x)) \phi(z_i^{\ell-1}(x'))\right] \end{aligned}$$

  - この期待値は、(1) ガウス過程回帰 (2) 数値的近似で順に計算するか、(3) ReLUなど特定のφについては、解析的に求められる.

NNGPカーネルとニューラルネット

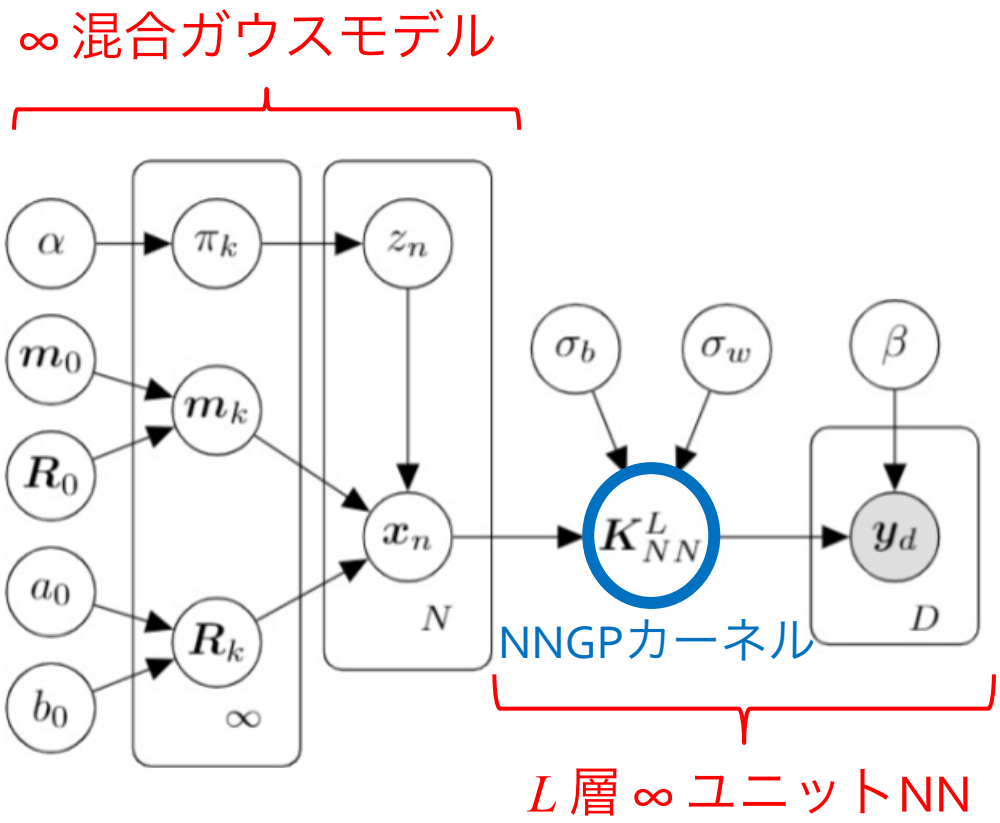
- φがReLUのとき (Cho&Saul 2009, Lee+ 2017)

$$K^{\ell}(x,x') = \sigma_b^2 + \frac{\sigma_w^2}{2\pi}\sqrt{K^{\ell-1}(x,x)K^{\ell-1}(x',x')} \times \left(\sin\theta_{x,x'}^{\ell-1} + (\pi - \theta_{x,x'}^{\ell-1})\cos\theta_{x,x'}^{\ell-1}\right)$$

$$\theta_{x,x'}^{\ell} = \cos^{-1}\left(\frac{K^{\ell}(x,x')}{\sqrt{K^{\ell}(x,x)K^{\ell}(x',x')}}\right)$$

NN-iWMMの生成モデル

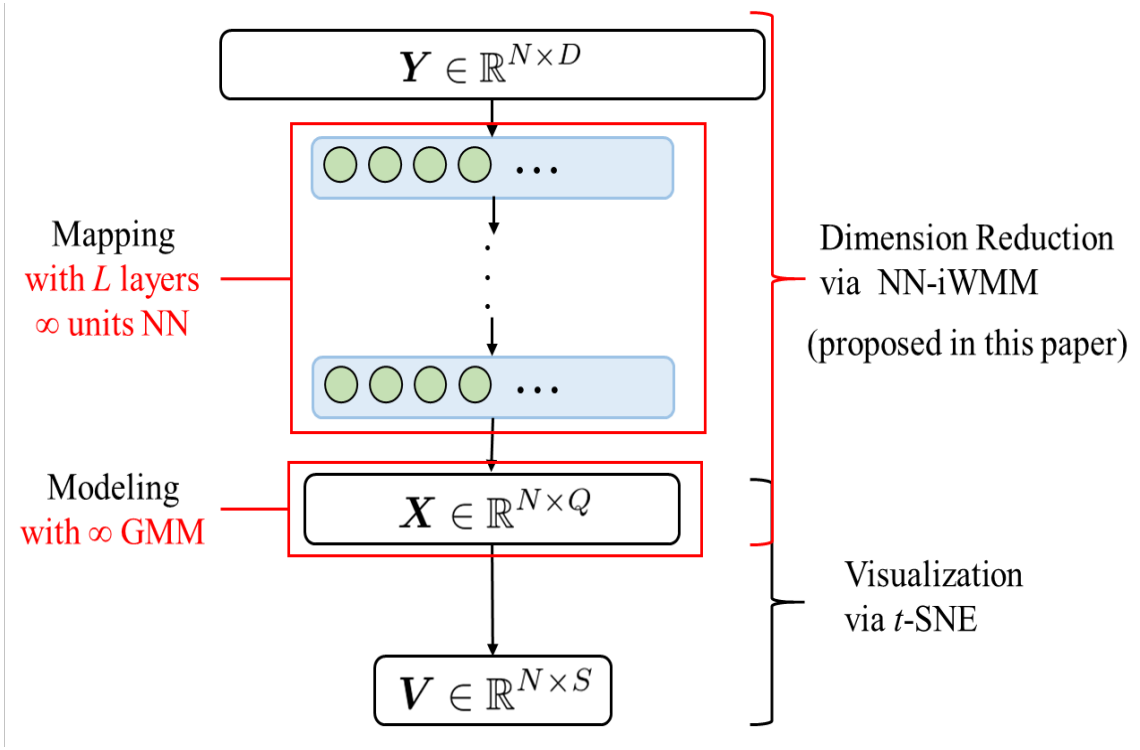
- ∞混合ガウス分布で潜在変数
  $$\boldsymbol{X} \equiv \{\boldsymbol{x}_i\}_{i=1}^{N}$$
  を生成
- $\boldsymbol{X}$ とカーネル関数のパラメータ σ_b, σ_w からNNGPカーネル行列 K_{NN}^{L} を生成
- 観測値 $\boldsymbol{Y}$ を次元dごとにNNGPから生成



クラスター数、ユニット数の指定が不要な教師なしNN

t-SNEとの統合

- 復習：



- 前提の違う2つのモデルをいかに組み合わせるか?

t-SNEとの統合: RegBayes

- 事後分布 $p(\boldsymbol{\theta}|\boldsymbol{Y})$: 次の最適化問題の解と一致 (A. Zellner, 1988)

$$\begin{cases} \min\limits_{q(\boldsymbol{\theta})} & \mathrm{KL}[q(\boldsymbol{\theta})\|p(\boldsymbol{\theta})] - \int q(\boldsymbol{\theta}) \log p(\boldsymbol{Y}|\boldsymbol{\theta}) d\boldsymbol{\theta} \\ \text{s.t.} & q(\boldsymbol{\theta}) \in \mathcal{P} \end{cases}$$

$\boldsymbol{\theta}$: parameters
$\mathcal{P}$: probability distribution

- RegBayes：事後分布への制約を考慮したベイズモデルを定式化 (J. Zhu+, 2014)

$$\begin{cases} \min\limits_{q(\boldsymbol{\theta})} & \mathrm{KL}[q(\boldsymbol{\theta})\|p(\boldsymbol{\theta})] - \int q(\boldsymbol{\theta}) \log p(\boldsymbol{Y}|\boldsymbol{\theta}) d\boldsymbol{\theta} \\ \text{s.t.} & E_{q(\boldsymbol{\theta})}[\mathcal{R}(\boldsymbol{\theta}, \boldsymbol{Y})] \leq 0, \ q(\boldsymbol{\theta}) \in \mathcal{P} \end{cases}$$

$\mathcal{R}(\boldsymbol{\theta}, \boldsymbol{Y})$: Regularization term

  - 制約を踏まえた最適分布 $q^*(\boldsymbol{\theta}) \propto p(\boldsymbol{Y}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) \exp(-\lambda \mathcal{R}(\boldsymbol{\theta}, \boldsymbol{Y}))$

$\boldsymbol{\theta}, \boldsymbol{Y}$ の同時分布に採用

NPDVの定式化

- NPDVの同時分布 ($q^*(\boldsymbol{\theta})$とNN-iWMM, t-SNEの対応)

$$q^*(\boldsymbol{\theta}) \propto \underline{p(\boldsymbol{Y}|\boldsymbol{\theta})p(\boldsymbol{\theta})} \exp(-\underline{\lambda \mathcal{R}(\boldsymbol{\theta}, \boldsymbol{Y})})$$

NN-iWMM　　t-SNE

- 分布の具体形

$$\begin{aligned}
&p(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{z}, \{\boldsymbol{m}_k, \boldsymbol{R}_k \pi_k\}_{k=1}^{\infty} | \boldsymbol{V}) \\
&\propto p(\boldsymbol{Y}|\boldsymbol{X}) p(\boldsymbol{X}|\boldsymbol{z}, \{\boldsymbol{m}_k, \boldsymbol{r}_k, \pi_k\}_{k=1}^{\infty}) p(\boldsymbol{z}|\{\pi_k\}_{k=1}^{\infty}) \\
&\times p(\{\boldsymbol{m}_k\}_{k=1}^{\infty}) p(\{\boldsymbol{r}_k\}_{k=1}^{\infty}) p(\{\pi_k\}_{k=1}^{\infty}) \\
&\times \exp(-\lambda \mathrm{KL}[\boldsymbol{p}^X \| \boldsymbol{p}^V])
\end{aligned}$$

NN-iWMM

$\boldsymbol{X}, \boldsymbol{V}$ 間の t-SNEのコスト関数

  - 可視化表現 $\boldsymbol{V}$、NNGPカーネルの σ_w, σ_b は非確率的なパラメータ
  - λ は超パラメータ　※ 本研究では $\lambda = ND$ で固定

NPDVの学習

- 目的関数：変分分布 Q を使った尤度の下界

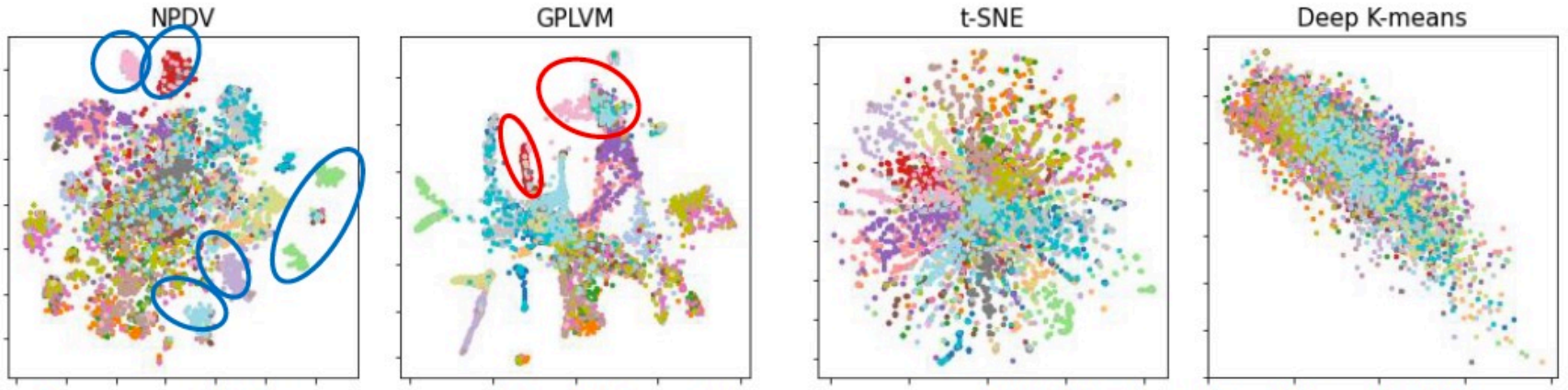
$$\mathcal{L} = \mathbb{E}_{q(\boldsymbol{X})}[\log p(\boldsymbol{Y}|\boldsymbol{X})] - \mathbb{E}_{q(\boldsymbol{X})}[\log q(\boldsymbol{X})]$$
$$+ \mathbb{E}_{q(\boldsymbol{X},\boldsymbol{z},\boldsymbol{m},\boldsymbol{\Sigma},\boldsymbol{\phi})}\left[\log \frac{p(\boldsymbol{X},\boldsymbol{z},\{\boldsymbol{m}_k,\boldsymbol{r}_k,\pi_k\}_{k=1}^K)}{q(\boldsymbol{z},\{\boldsymbol{m}_k,\boldsymbol{R}_k\}_k,\boldsymbol{\pi})}\right] - \lambda\mathbb{E}_{q(\boldsymbol{X})}[\mathrm{KL}[\boldsymbol{p}^X\|\boldsymbol{p}^V]]$$
$$= \underbrace{\mathcal{L}_1 + \mathcal{L}_2 - \lambda\mathcal{R}}_{(*)} + \underbrace{\mathcal{H}(q(\boldsymbol{X}))}_{\text{Gauss Entropy}}$$

  - Re-parametrization trickを使って $q(\boldsymbol{x}_n) = \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_n, \boldsymbol{S}_n)$ から正規乱数を生成し、(*)をモンテカルロ近似
  - 近似した後、勾配法ベースで学習

実験設定

- テキストデータ: 20newsgroups (英語)、Livedoor ニュースコーパス (日本語)
  - 20newsgroups: InternetのUSENETから取って来た20個のニュースグループのテキスト、18000記事
  - Livedoor: Livedoorニュースのテキスト、9カテゴリ、7300記事
- tf.idfで前処理、1,000次元に圧縮してから2次元に可視化
- t-SNE (Maaten+ 2009), Deep K-means (Phard+ 2020) と比較

実験結果 (20newsgroups:英語)



○：5つの明確なクラスタを確認

×：赤、桃クラスタと他のクラスタが重複

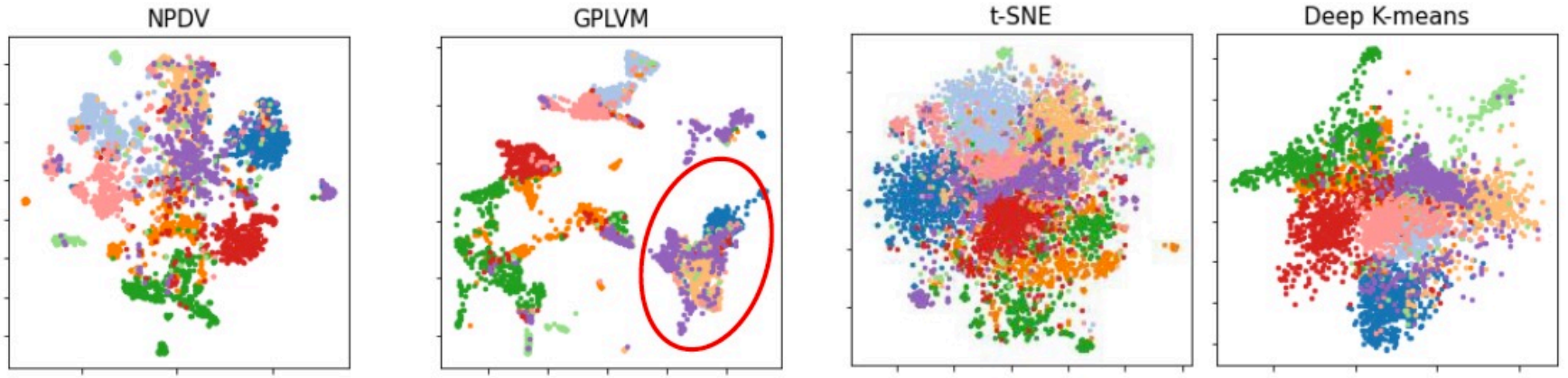
×：全体的にクラスターが互いに重複（特にDeep k-means）

- *k* 近傍法分類精度（**太字**：最高精度、*：次点の精度）

近傍数	Deep *k*-means	GPLVM	t-SNE	*t*-SNE-GPLVM	提案手法（次点との差分）
k = 10	0.383	0.527	0.559	0.593 *	**0.599** (+0.6%)
k = 20	0.347	0.496	0.512	0.529 *	**0.552** (+ 2.3%)
k = 30	0.330	0.487	0.492	0.500 *	**0.529** (+ 2.9%)

他手法よりも明確にクラスターを識別可能 & 最高精度を記録

実験結果 (Livedoor:日本語)

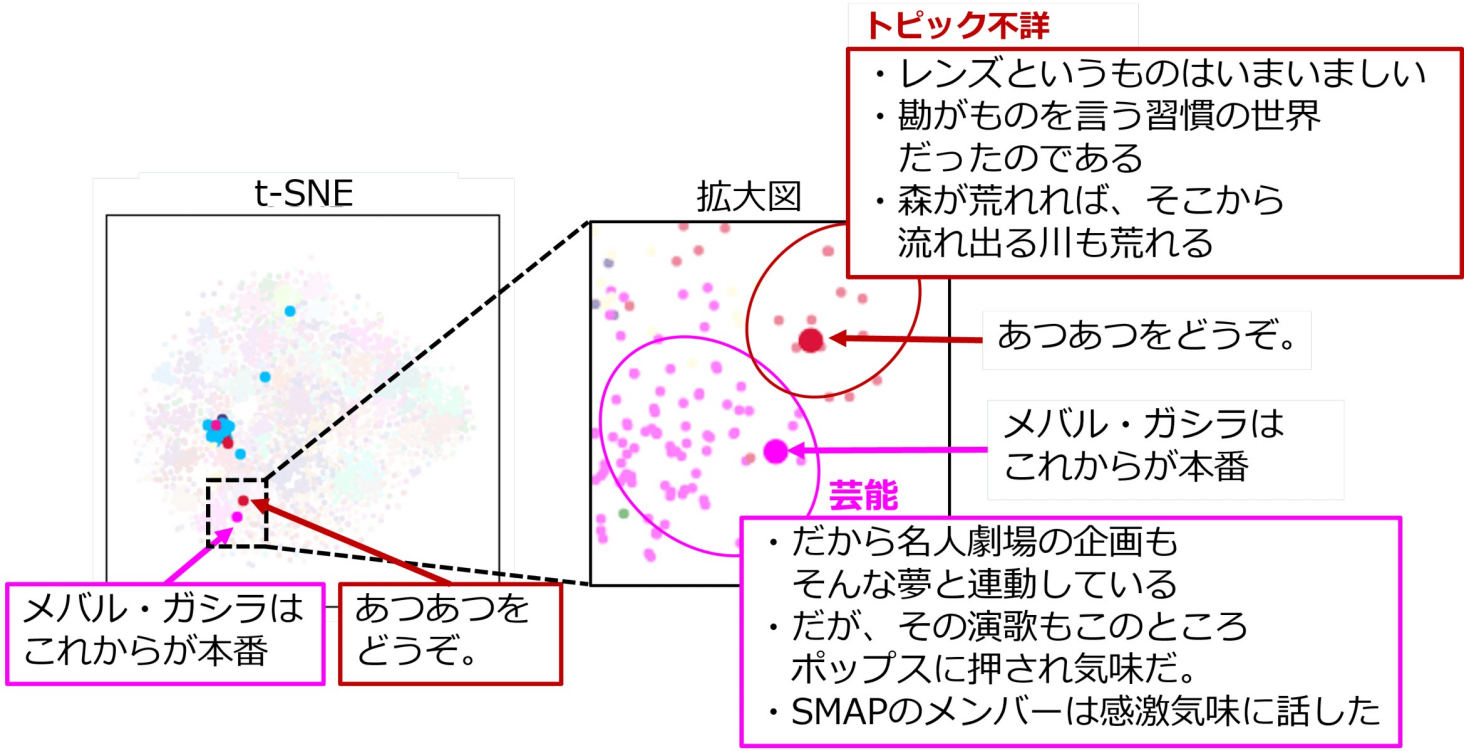


- *k* 近傍法分類精度（**太字**：最高精度、*：次点の精度）

近傍数	Deep *k*-means	*t*-SNE-GPLVM	t-SNE	GPLVM	提案手法（次点との差分）
k = 10	0.755	0.793	0.793	0.809 *	**0.844** (+ 3.5%)
k = 20	0.746	0.767	0.767	0.798 *	**0.822** (+ 2.4%)
k = 30	0.742	0.747	0.747	0.790 *	**0.809** (+ 1.9%)

文の教師なし可視化 (*t*-SNE)

- 観測データを*t*-SNEで直接クラスタリングする
  →ノイズに弱い、意味を考慮できない

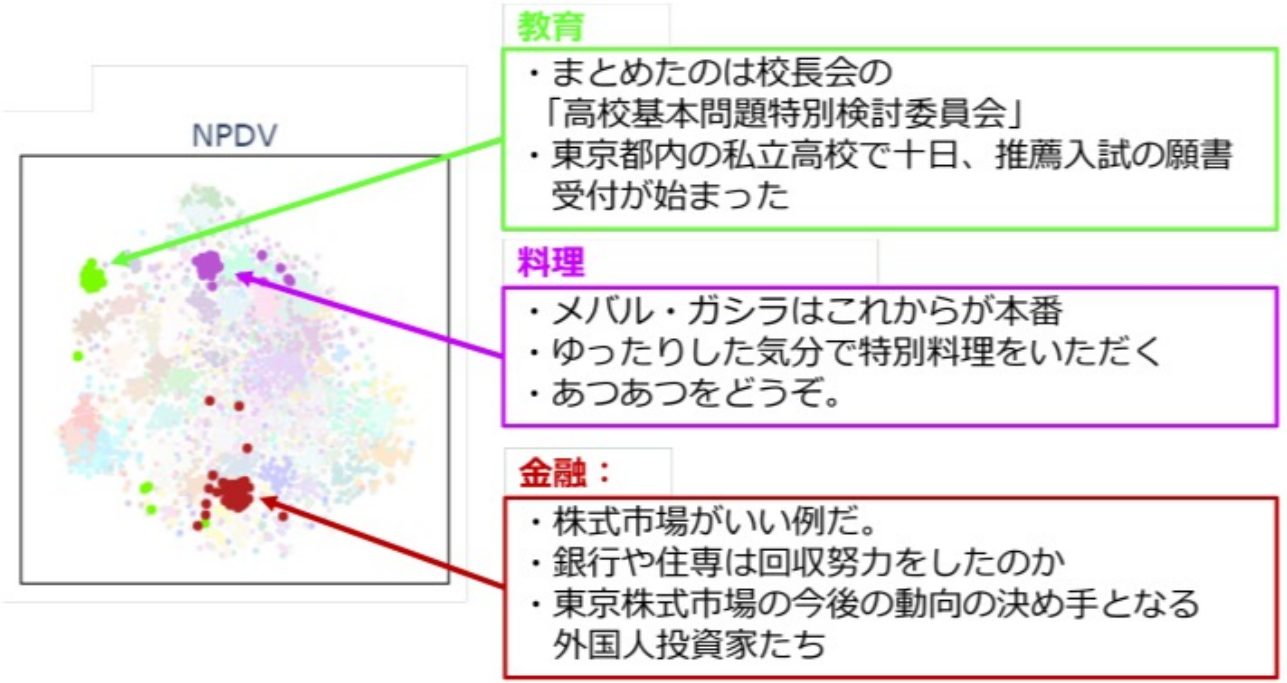


× NPDVでは料理クラスタになる文が無関係なクラスタに所属

文の教師なし可視化 (NPDV)

- 学習には教師データは不要だが、∞-GMMを学習しているため、内部的にクラスタリングを行っている

NPDVで得られたクラスター
潜在空間でのクラスター、可視化表現を同時に推定



データ：
京大コーパス
(毎日新聞1995年)から5000文

- 意味的類似性のあるクラスターを確認

まとめ

- 離散的なノンパラメトリックベイズ法である階層ディリクレ過程と、連続的なガウス過程の両方を自然言語に適用した講演者の研究を3つ紹介した
- ガウス過程を使うことで、解析的でありながらニューラルネットと等価な表現力を持つ
  - 多数の重みを学習する必要がない
  - 局所解や学習率等のハイパーパラメータ不要
- ディリクレ過程の使用により、内在的なカテゴリをその数とともに自動推定することができ、科学的な解釈に繋がる

持橋講演分のまとめ

- ノンパラメトリックな確率過程を用いて、自然言語を数学的に見通しよく解析するための基礎や講演者の研究について紹介しました
- 自然言語のモデル化の問題であり、数学的解析は追い付いていない部分がある
  - 階層的TSSBは、本来は自己相似過程の一種
  - 潜在空間で文や文書がどう埋め込まれているかは、研究最前線の課題
- 言語に限らず、離散データ一般に適用できる統計理論(実装も重要)