NEC Laboratories America
Relentless passion for innovation

混合分布学習の基礎から最先端

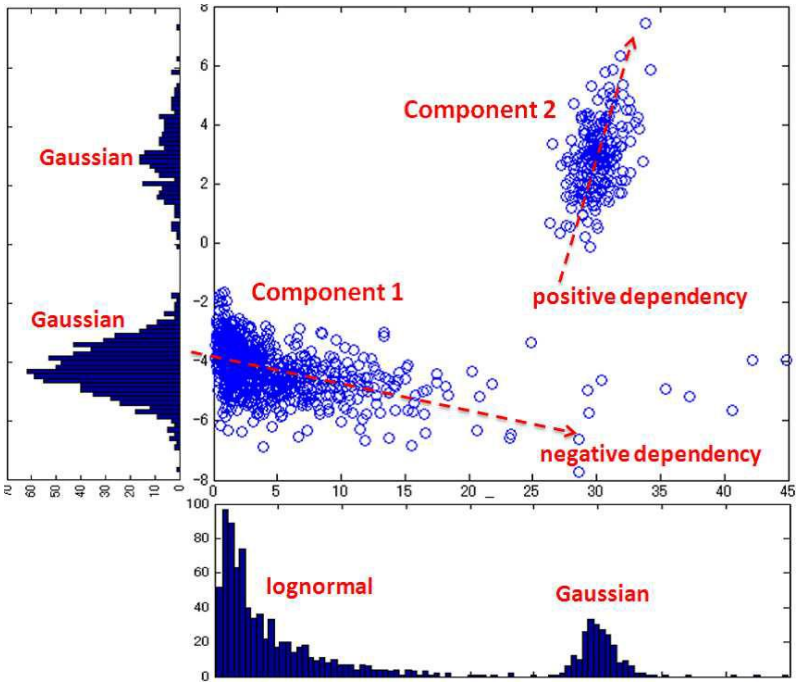
第1部：混合分布の基礎（藤巻）
第2部：ノンパラメトリックベイズ法（佐藤）
第3部：異種混合学習と因子化漸近ベイズ（藤巻）

Ryohei Fujimaki (NEC Laboratories America)
joint Tutorial with
Issei Sato (University of Tokyo)

NEC 基幹技術フォーラム
2013/4/4

NEC Laboratories America
Relentless passion for innovation

第3部　異種混合学習と因子化漸近ベイズ（藤巻）



NEC Laboratories America
Relentless passion for innovation

第3部の構成

- 異種混合学習とモデル選択
- 因子化情報量基準と因子化漸近ベイズ推論
- 因子化漸近ベイズ推論の性質
- 実験
- 発展モデルへ

分析案件を思い返してみると。。。

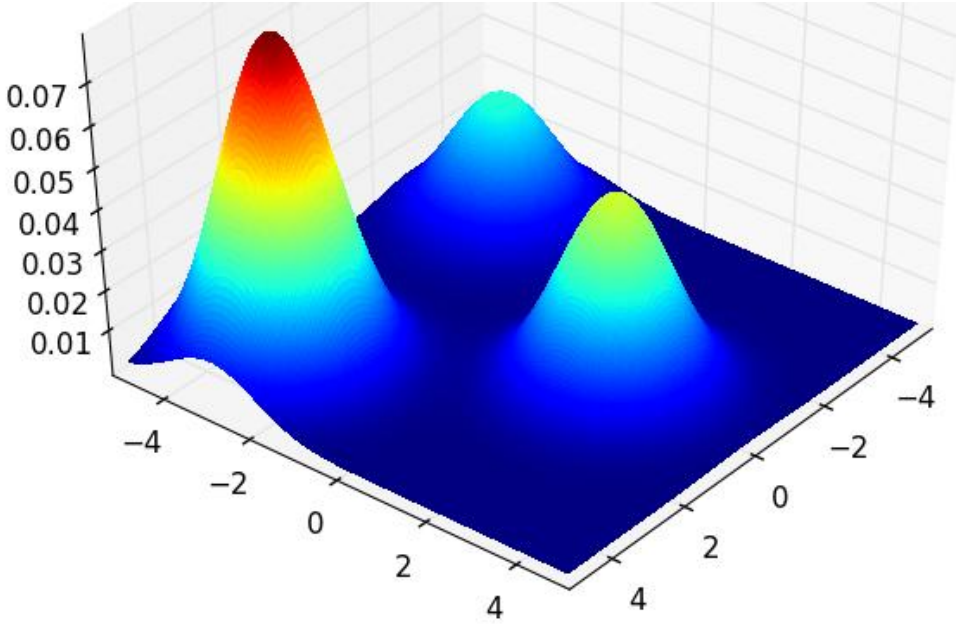
- 自動車センサデータ
  - 車種、走行環境（天候、道路状況。。）、走行状態（加減速、停止、カーブ。。）、、、
- 健康診断データ
  - 生活習慣、治療・通院状況、遺伝的リスク、、、
- サーバートラフィック分析
  - サーバー特性、運用（データ移動、サーバー追加。。）、周期（平日、週末、祝日。。）、、、
- 金融リスク分析
  - リスク（高頻度低リスク、低頻度高リスク）、銀行タイプ（メガバンク、地銀。。）、、、

- データは複雑怪奇に混ざっている。うまくモデルを選択しないと沈没

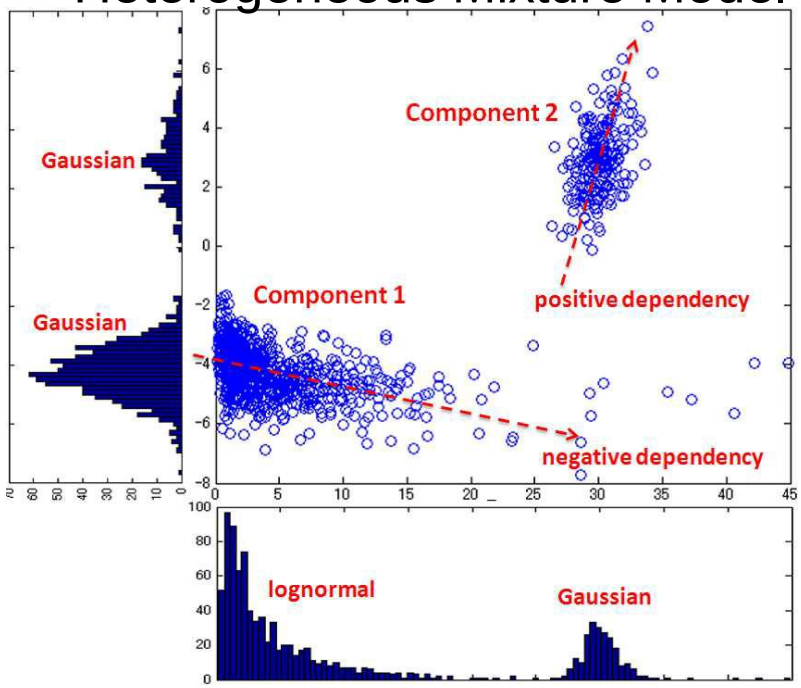
Mixture Distributions

- One of the most basic models in machine learning
  - Density estimation, clustering, anomaly detection, feature learning, …

Gaussian Mixture Model



Heterogeneous Mixture Model



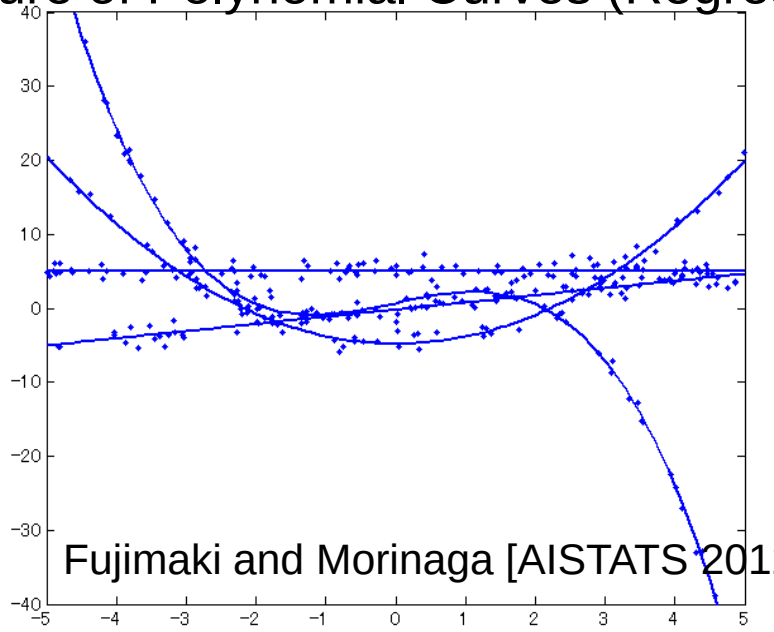
Fujimaki et al. [KDD2011]

- Learning Problem
  - How many distributions are mixed?
  - What are their parameters?
  - (What types of distributions are mixed?)

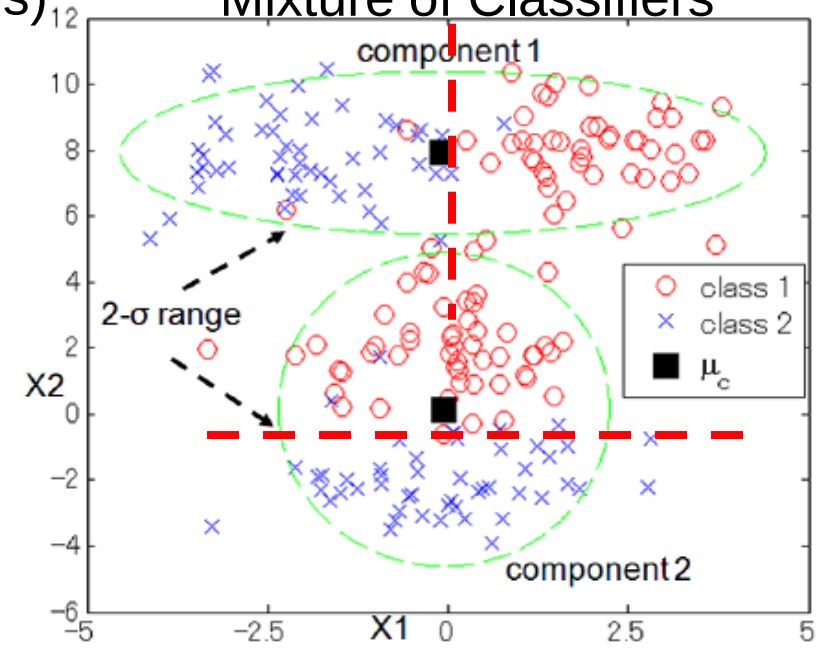
Mixture Experts

- Basic ensemble learners
  - Multimodality in classification, regression, etc.

Mixture of Polynomial Curves (Regressors)



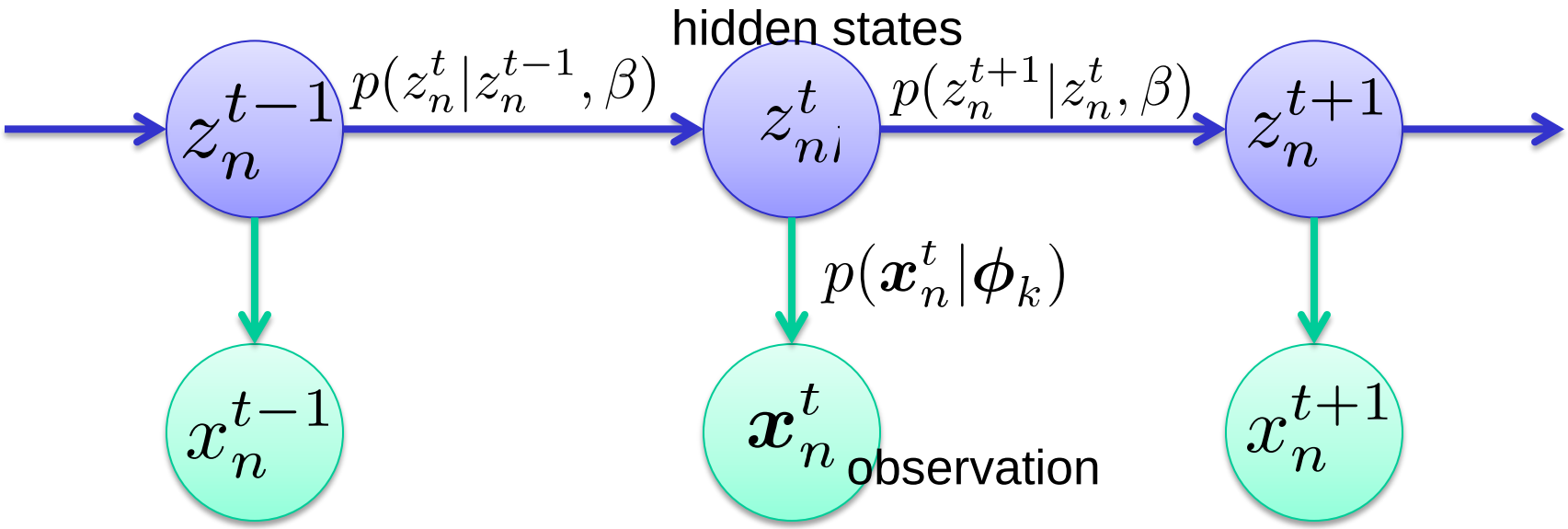
Mixture of Classifiers



- Learning Problem
  - How many experts are mixed?
  - What are their parameters?
  - What are appropriate levels of their regularizations? (or how many features does each expert use?)

Hidden Markov Models

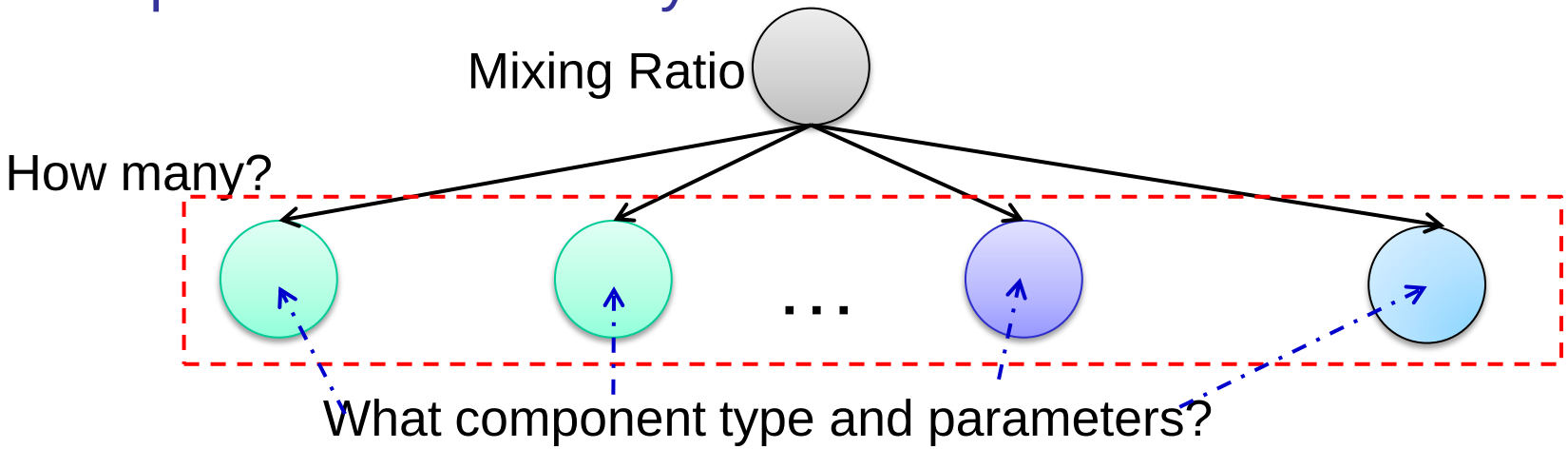
- One of the most important models for sequential data.
  - time series analysis, speech recognition, text indexing, etc.



- Learning Problem
  - How many hidden states are necessary?
  - What are the transition and observation parameters?

Challenges

- Computational Efficiency



- Identifiability issue of the maximum likelihood (ML) estimators
  - Ex. Mixture of two identical standard normal distributions

equivalent class (same likelihood)

$$0.1 * N(0,1) + 0.9 * N(0,1)$$
$$0.5 * N(0,1) + 0.5 * N(0,1)$$
$$\vdots$$
$$a * N(0,1) + b * N(0,1)$$
$$= N(0,1)$$

Mapping is NOT 1-to-1 in likelihood space (non-identifiable)

  - Generalization performance of a non-identifiable estimator is much worse than those of identifiable estimators (S.Watanabe 2000)

Bayesian Model Selection

- Bayesian Approximation Inference

$$M^* = \arg\max_M \log p(\boldsymbol{x}^N | M)$$

$$= \arg\max_M \log \int_{\boldsymbol{\theta}} p(\boldsymbol{x}^N | \boldsymbol{\theta}) p(\boldsymbol{\theta} | M) d\boldsymbol{\theta}$$

computationally/analytically intractable

  - Bayesian marginalization automatically takes model complexity into account.
- Non-parametric Bayesian Modeling
  - Plug-in ``infinite prior'' (Dirichlet process, hierarchical Diriclet process, Pitman-Yaw process, etc.)
  - Prior controls model complexity.

Bayes Information Criterion (BIC)
[Shwarz 1978]

NEC Laboratories America
Relentless passion for innovation

- Bayes Information Criterion

parameter dimensionality

$$BIC(M|\boldsymbol{x}^N) = \log p(\boldsymbol{x}^N|\boldsymbol{\theta}_{ML}) + \frac{\mathcal{D}_M}{2}\log N$$

model complexity

- Non-regularity Issue and Non-Identifiability issues

Fisher information matrix is singular around ML estimator
-> BIC's complexity looses theoretical justification
(In fact, BIC over-estimates the complexity*)

ML estimator is not unique and non-identifiable
-> Generalization performance becomes significantly worse**

*S. Watanabe (Neural Computation 2001)
**K. Yamazaki and S. Watanabe (Neural Networks 2003)

Variational Bayesian Inference
[MacKay 1997, Beal 2003]

- Maximize lower bound instead of marginal log-likelihood.

$$\log p(\boldsymbol{x}^N|M) \geq \sum_{\boldsymbol{z}^N} \int q(\boldsymbol{z}^N)q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \log \frac{p(\boldsymbol{x}^N, \boldsymbol{z}^N|\boldsymbol{\theta})p(\boldsymbol{\theta}|M)}{q(\boldsymbol{z}^N)q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} d\boldsymbol{\theta}$$

equality does NOT hold in general

$$q(\boldsymbol{z}^N, \boldsymbol{\theta}) = q(\boldsymbol{z}^N)q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$$

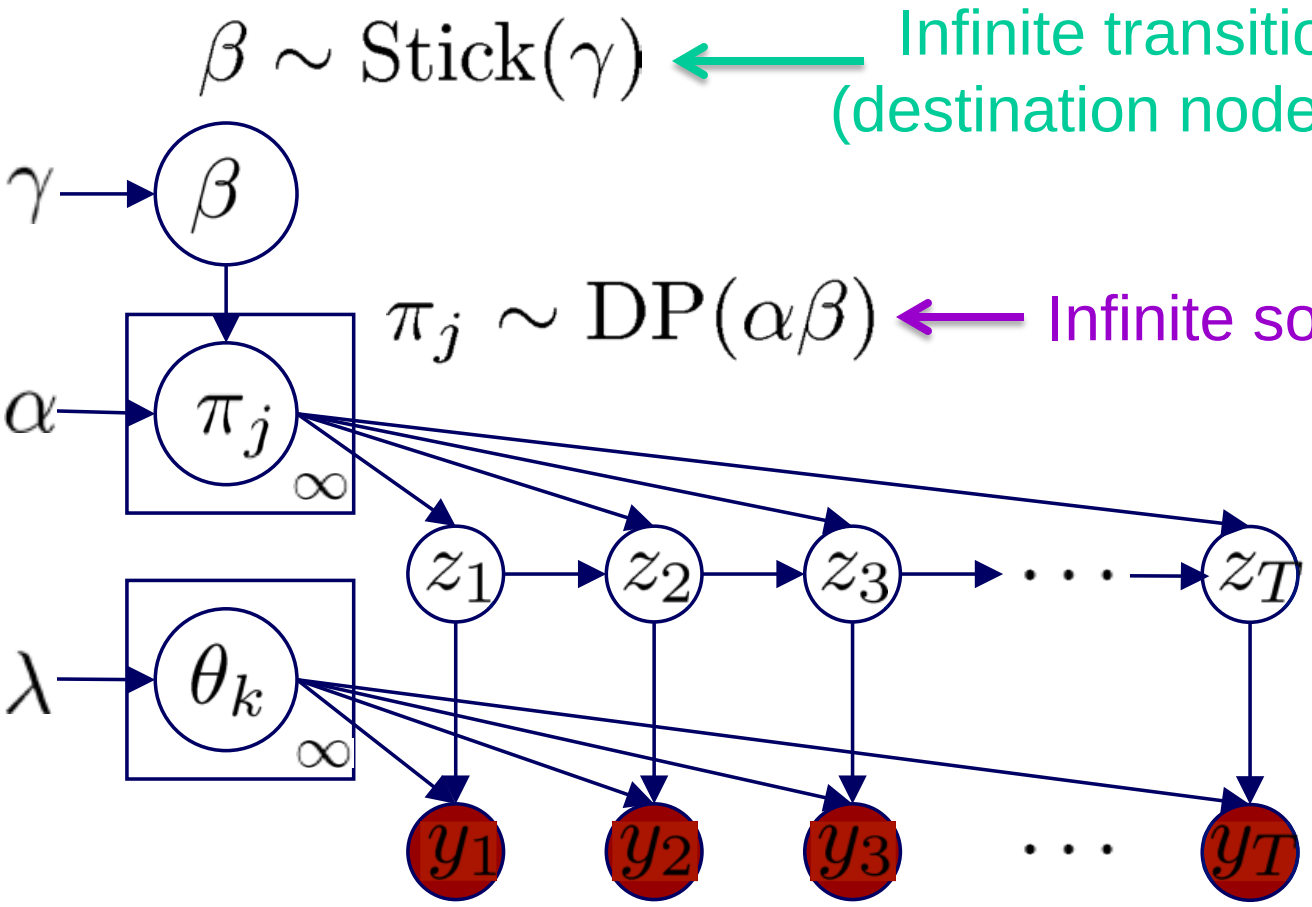
Assume independency between hidden states and parameters

- VB EM algorithm
  - VB Estep (VB forward-backward algorithm): maximization of the lower bound in terms of $q(\boldsymbol{z}^N)$
  - VB Mstep: maximization of the lower bound in terms of $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$

Non-parametric Bayesian Method (infinite HMMs) [Beal 2002, van Gael 2008]

NEC Laboratories America
Relentless passion for innovation

- Hierarchical Dirichlet Process Prior



- Efficient MCMC method : Beam Sampling

第3部の構成

- 異種混合学習とモデル選択
- 因子化情報量基準と因子化漸近ベイズ推論
- 因子化漸近ベイズ推論の性質
- 実験
- 発展モデルへ

Key Idea

- An alternative representation

$$\log p(\boldsymbol{x}^N|M) = \max_q \Big\{ \sum_{\boldsymbol{z}^N} q(\boldsymbol{z}^N) \log \Big(\frac{p(\boldsymbol{x}^N, \boldsymbol{z}^N|M)}{q(\boldsymbol{z}^N)} \Big) \Big\}$$

parameter is marginalized out

- ``Factorized'' Laplace Method

complete likelihood is regular

$$p(\boldsymbol{x}^N, \boldsymbol{z}^N|M) = \int p(\boldsymbol{z}^N|\boldsymbol{\alpha}) \prod_{c=1}^{C} p_c(\boldsymbol{x}^N|\boldsymbol{z}_c^N, \boldsymbol{\phi}_c) p(\boldsymbol{\theta}|M) d\boldsymbol{\theta}$$

Laplace approximation Laplace approximation

Each Laplace approximation is done around the ML estimator of complete likelihood

Key Idea

- An alternative representation

$$\log p(\boldsymbol{x}^N|M) = \max_q \Big\{ \sum_{\boldsymbol{z}^N} q(\boldsymbol{z}^N) \log \Big(\frac{p(\boldsymbol{x}^N, \boldsymbol{z}^N|M)}{q(\boldsymbol{z}^N)} \Big) \Big\}$$

parameter is marginalized out

- ``Factorized'' Laplace Method

complete ML estimator

$$p(\boldsymbol{x}^N, \boldsymbol{z}^N|M) \approx p(\boldsymbol{x}^N, \boldsymbol{z}^N|\bar{\boldsymbol{\theta}}) \frac{(2\pi)^{\mathcal{D}_{\boldsymbol{\alpha}}/2}}{N^{\mathcal{D}_{\boldsymbol{\alpha}}/2} |\bar{\mathcal{F}}_Z|^{1/2}} \times \prod_{c=1}^{C} \frac{(2\pi)^{\mathcal{D}_c/2}}{(\sum_{n=1}^N z_{nc})^{\mathcal{D}_c/2} |\bar{\mathcal{F}}_c|^{1/2}}$$

Fisher information matrices

parameter dimensionalities

Factorized Information Criterion (FIC)

- FIC as an approximation of marginal log-likelihood

$$\log p(\boldsymbol{x}^N|M) \approx FIC(\boldsymbol{x}^N, M) \equiv \max_q \left\{ \mathcal{J}(q, \bar{\boldsymbol{\theta}}, \boldsymbol{x}^N) \right\}$$

$$\mathcal{J}(q, \bar{\boldsymbol{\theta}}, \boldsymbol{x}^N) = \sum_{\boldsymbol{z}^N} q(\boldsymbol{z}^N) \Big(\log p(\boldsymbol{x}^N, \boldsymbol{z}^N | \bar{\boldsymbol{\theta}}) - \frac{\mathcal{D}_{\boldsymbol{\alpha}}}{2} \log N$$

complete ML estimator is not available

$$- \sum_{c=1}^{C} \frac{\mathcal{D}_c}{2} \log \Big(\sum_{n=1}^{N} z_{nc} \Big) - \log q(\boldsymbol{z}^N) \Big) \Big\}$$

complexity of the c-th component is computationally intractable

Theorem 1 *$FIC(\boldsymbol{x}^N, M)$ is asymptotically consistent with $\log p(\boldsymbol{x}^N|M)$.*

Factorized Asymptotic Bayesian Inference (FAB)

- Lower bound of FIC

$$FIC(\boldsymbol{x}^N, M) \geq \mathcal{G}(q, \tilde{q}, \theta, \boldsymbol{x}^N)$$

$$\equiv \sum_{\boldsymbol{z}^N} q(\boldsymbol{z}^N) \Big(\log p(\boldsymbol{x}^N, \boldsymbol{z}^N | \boldsymbol{\theta}) - \frac{\mathcal{D}_{\boldsymbol{\alpha}}}{2} \log N$$

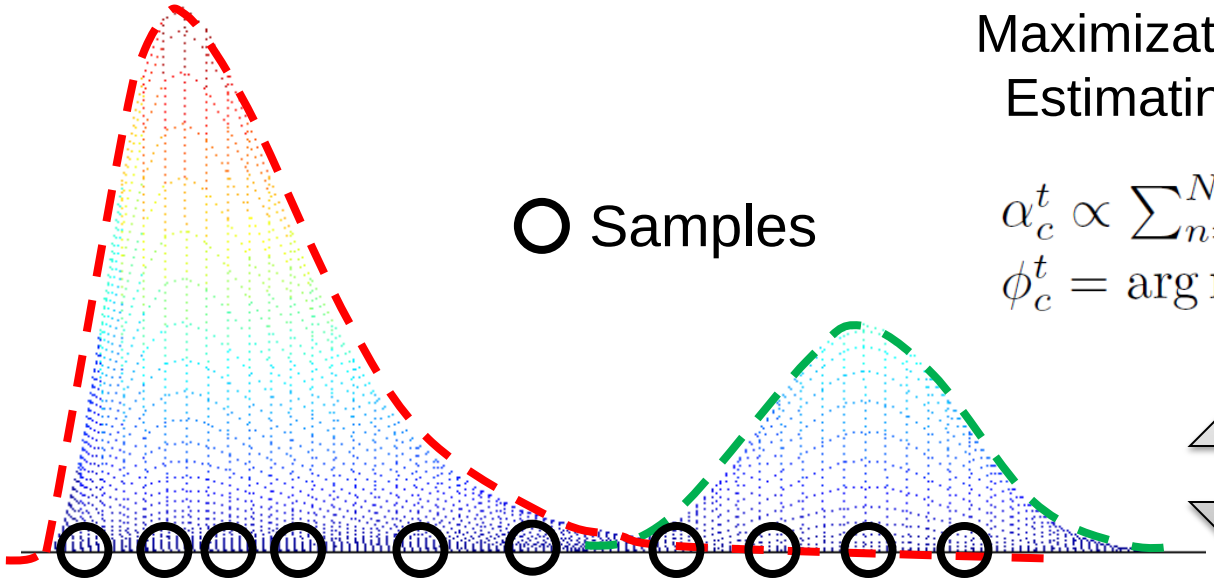
$$- \sum_{c=1}^{C} \frac{\mathcal{D}_c}{2} \mathcal{L}(\sum_{n=1}^{N} z_{nc}, \sum_{n=1}^{N} \tilde{q}(z_{nc})) - \log q(\boldsymbol{z}^N) \Big)$$

- FAB solves model selection problems by maximizing the lower bound of FIC

$$M^*, q^*, \boldsymbol{\theta}^*, \tilde{q}^* = \arg \max_{M, q, \boldsymbol{\theta}, \tilde{q}} \mathcal{G}(q, \tilde{q}, \boldsymbol{\theta}, \boldsymbol{x}^N)$$

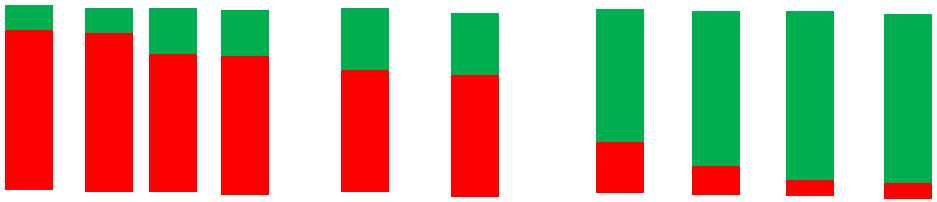
Brief Reminder of the EM algorithm for MMs

- Parameter estimation of latent variable models



Maximization step (M-step):
Estimating the component parameters

$$\alpha_c^t \propto \sum_{n=1}^{N} q_{nc}^t$$
$$\phi_c^t = \arg\max_\phi \sum_{n=1}^{N} q_{nc}^t \log p(\boldsymbol{x}_n|\phi_c^{t-1})$$



Expectation step (E-step):
Estimating expected assignments

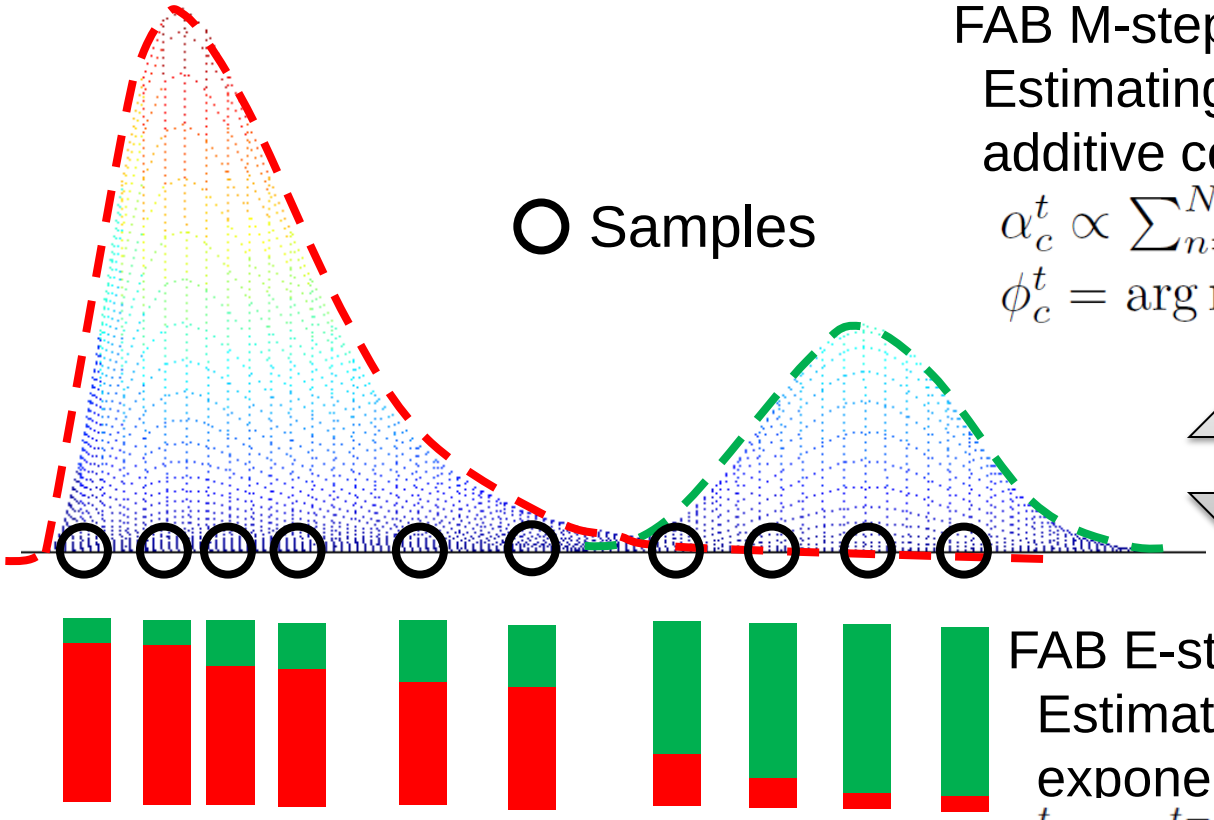
$$q_{nc}^t \propto \alpha_c^{t-1} p(\boldsymbol{x}_n|\phi_c^{t-1})$$

- Convergence to the ML estimator (non-identifiable)
- Additional model selection (e.g., CV, BIC, …)

Factorized Asymptotic Bayesian Inference for MMs

NEC Laboratories America
Relentless passion for innovation

- Simultaneous model selection and parameter estimation



FAB M-step:

Estimating component models with additive component-wise regularization.

$$\alpha_c^t \propto \sum_{n=1}^N q_{nc}^t$$

$$\phi_c^t = \arg\max_\phi \sum_{n=1}^N q_{nc}^t \log p(\boldsymbol{x}_n|\phi_c^{t-1}) - \frac{\mathcal{D}_c}{2}\log(\sum_{n=1}^N q_{nc}^t)$$

FAB E-step:

Estimating variational posterior with exponentiated regularization.

$$q_{nc}^t \propto \alpha_c^{t-1} p(\boldsymbol{x}_n|\phi_c^{t-1}) \exp(-D_c/2\alpha_c^{(t-1)}N)$$

- Identifiable estimation based on model-induced regularization
- No hand-determined model selection parameters.

Factorized Asymptotic Bayesian Inference (FAB)

- Gaussian Mixture Models

$$\min_{C,\alpha,\mu_c,\Sigma_c} \left\{ \sum_{n=1}^{N} -\log \sum_{c=1}^{C} \alpha_c \mathcal{N}(x_n|\mu_c,\Sigma_c) \right.$$

- E step : update posterior of latent variables

$$q^{(t)}(z_{nc}) \propto \alpha_c^{(t-1)} p(\boldsymbol{x}_n|\boldsymbol{\phi}_c^{(t-1)}) \exp\left(\frac{-\mathcal{D}_c}{2\alpha_c^{(t-1)} N}\right)$$

$$\sum_{c=1}^{C} q^{(t)}(z_{nc}) = 1 \quad \tilde{q} = q^{(t-1)}$$

Smaller/more complex component is more strongly regularized

- M step : update parameters

$$\alpha_c^{(t)} = \sum_{n=1}^{N} q^{(t)}(z_{nc})/N \qquad \boldsymbol{\phi}_c^{(t)} = \arg\max_{\boldsymbol{\phi}_c} \sum_{n=1}^{N} q^{(t)}(z_{nc}) \log p(\boldsymbol{x}_n|\boldsymbol{\phi}_c)$$

Factorized Asymptotic Bayesian Inference (FAB)

- Mixture of regressors

$$\min_{C,\alpha,w_c,\sigma_c} \left\{ \sum_{n=1}^{N} -\log \sum_{c=1}^{C} \alpha_c P(y_n|w_c^T x_n, \sigma_c^2) + \sum_{c=1}^{C} \lambda_c r(w_c) \right\}$$

- E step : update posterior of latent variables

$$q^{(t)}(z_{nc}) \propto \alpha_c^{(t-1)} p(y_n|\phi_c^{(t-1)}, x_n) \exp\left(\frac{-\mathcal{D}_c}{2\alpha_c^{(t-1)} N} \right)$$

- M step : update parameters

$$\alpha_c^{(t)} = \sum_{n=1}^{N} q^{(t)}(z_{nc})/N$$

$$\phi_c^{(t)} = \arg\max_{\phi_c} \sum_{n=1}^{N} q^{(t)}(z_{nc}) \log p(y_n|\phi_c, x_n) - \frac{\mathcal{D}_c}{2} \log\left(\sum_{n=1}^{N} q^{(t)}(z_{nc}) \right)$$

Component regularization is automatically determined

第3部の構成

- 異種混合学習とモデル選択
- 因子化情報量基準と因子化漸近ベイズ推論
- 因子化漸近ベイズ推論の性質
- 実験
- 発展モデルへ

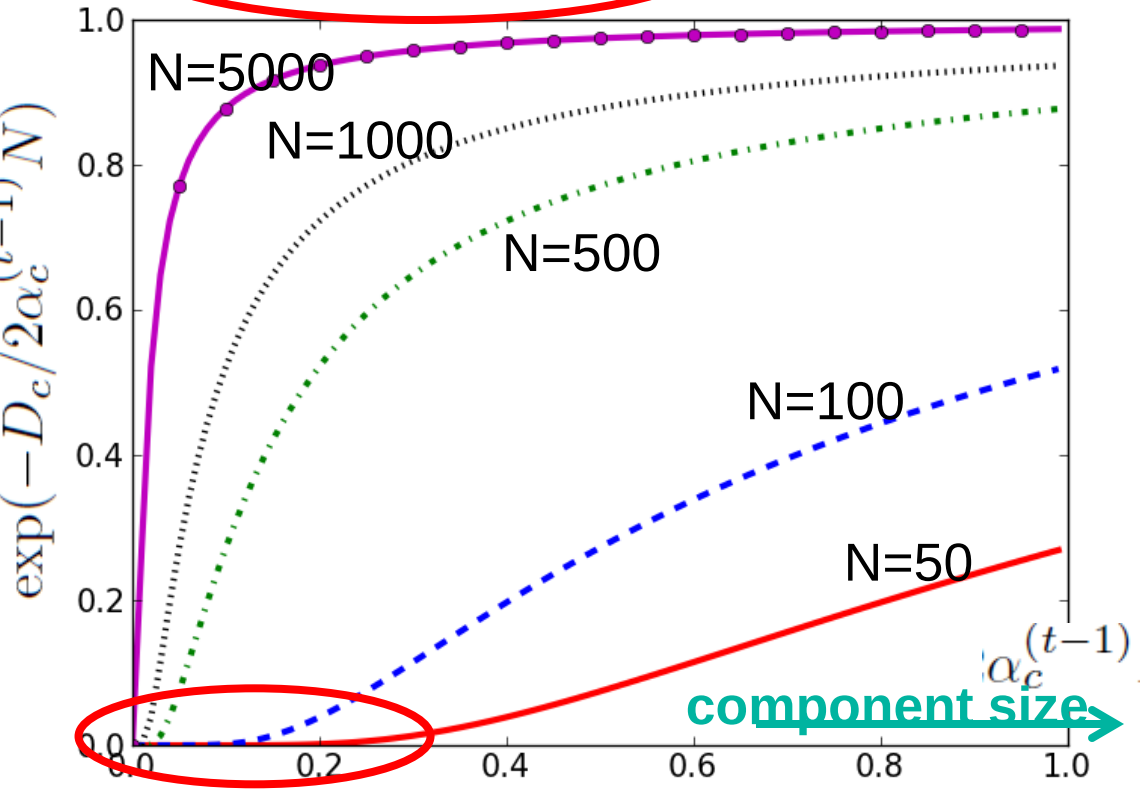
Automatic Component Shrinkage Mechanism

- FAB-unique exponentiated regularization in E-steps

$$q_{nc}^{t} \propto \alpha_c^{t-1} p(\boldsymbol{x}_n|\phi_c^{t-1}) \exp(-D_c/2\alpha_c^{(t-1)}N)$$

Gaussian Mixture Models
D=10 (Dc = 65)

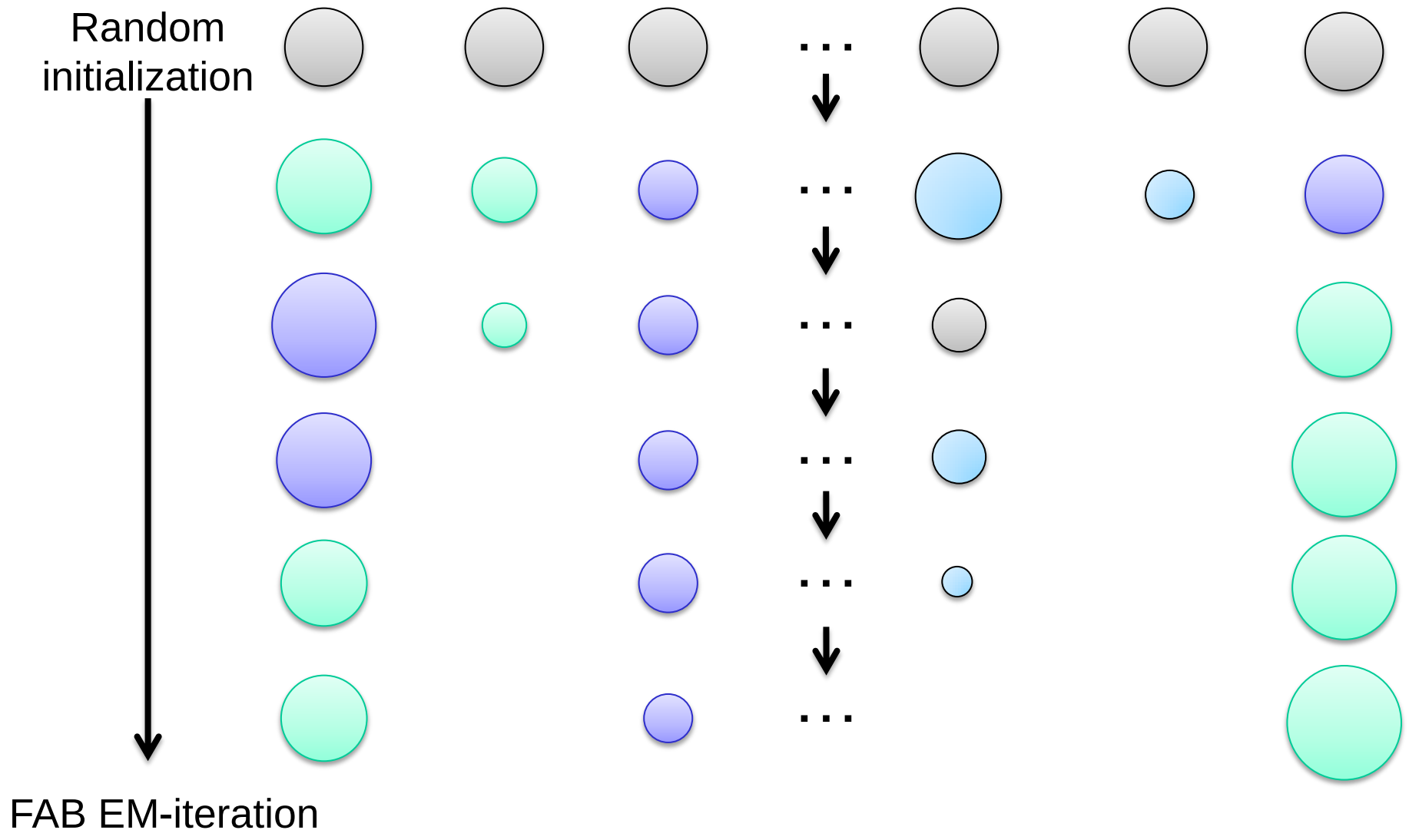
D: dimensionality of $\boldsymbol{x}_n$
Dc : dimensionality of $\phi_c^{(t-1)}$



Small Components are shrunk through the EM iterations.

Automatically controls the number of latent variables (complexity)

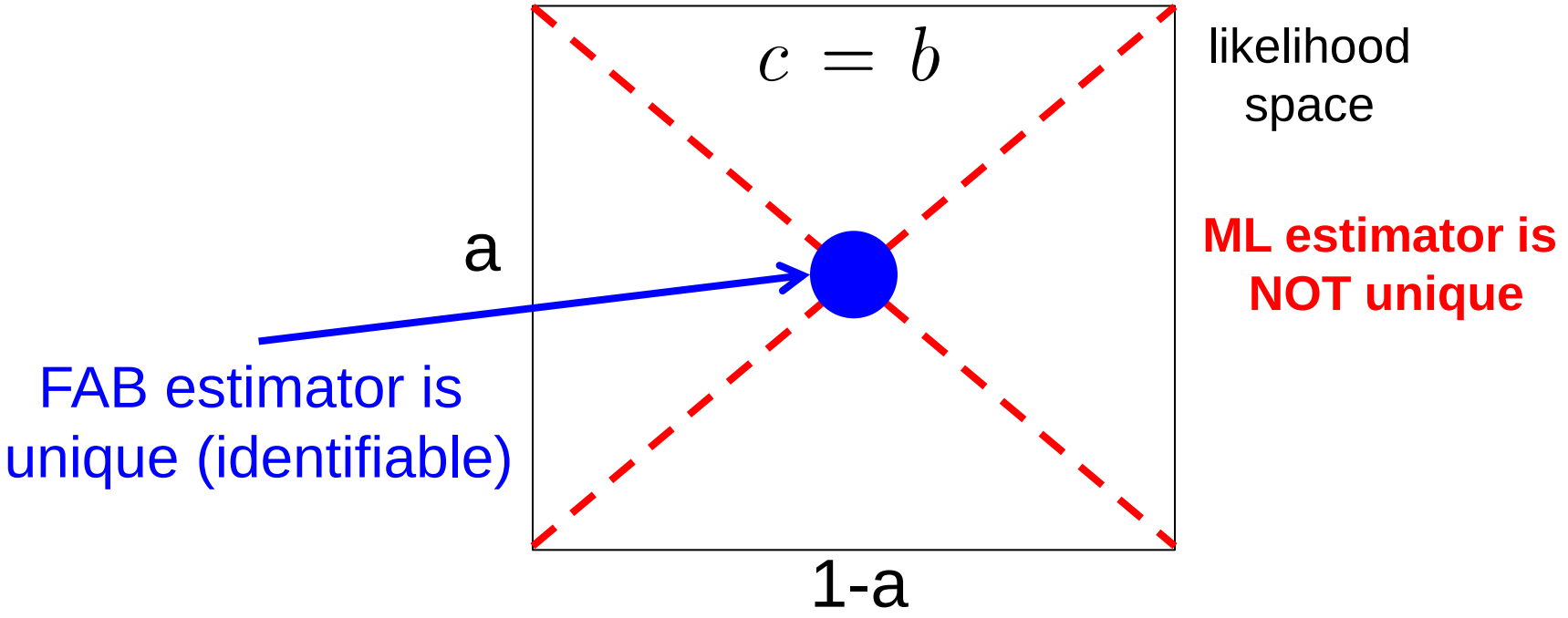
How FAB works?



BIC and Difficulties of Model Selection for MMs

- Non-regularity Issue and Non-Identifiability issues

$$p(X|a,b,c) = ap(X|b) + (1-a)p(X|c)$$



Identifiability Analysis on FAB

- Convergence point

$$\sum_{n=1}^{N} q^*(z_{nc}) = \alpha_c^* N = \sum_{n=1}^{N} \alpha_c^* p(\boldsymbol{x}_n|\boldsymbol{\phi}_c^*) \exp(\frac{-\mathcal{D}_c}{2\alpha_c^* N})/Z_n$$

$$\Rightarrow \quad N = \sum_{n=1}^{N} p(\boldsymbol{x}_n|\boldsymbol{\phi}_c^*) \exp(\frac{-\mathcal{D}_c}{2\alpha_c^* N})/Z_n$$

- What happen if b = c $\quad p(X|a,b,c) = ap(X|b) + (1-a)p(X|c)$

$$\sum_{n=1}^{N} p(\boldsymbol{x}_n|\boldsymbol{\phi}_1^*) \exp(\frac{-\mathcal{D}_1}{2\alpha_1^* N})/Z_n = \sum_{n=1}^{N} p(\boldsymbol{x}_n|\boldsymbol{\phi}_2^*) \exp(\frac{-\mathcal{D}_2}{2\alpha_2^* N})/Z_n$$

$$\Rightarrow \quad \exp(\frac{-\mathcal{D}_1}{2\alpha_1^* N}) = \exp(\frac{-\mathcal{D}_2}{2\alpha_2^* N}) \quad \Rightarrow \quad \alpha_1^* = \alpha_2^*$$

  - Components having the same parameter have the same size
  - The maximum entropy distribution (in terms of qZ) is selected in the equivalent class

Non-Identifiability (EM case)

- Convergence point (EM case)

$$\sum_{n=1}^{N} q^*(z_{nc}) = \alpha_c^* N = \sum_{n=1}^{N} \alpha_c^* p(\boldsymbol{x}_n|\boldsymbol{\phi}_c^*) \boxed{} / Z_n$$

$$\Rightarrow \quad N = \sum_{n=1}^{N} p(\boldsymbol{x}_n|\boldsymbol{\phi}_c^*) \boxed{} / Z_n$$

- What happen if b = c $\quad p(X|a,b,c) = ap(X|b) + (1-a)p(X|c)$

$$\sum_{n=1}^{N} p(\boldsymbol{x}_n|\boldsymbol{\phi}_1^*) \boxed{} / Z_n = \sum_{n=1}^{N} p(\boldsymbol{x}_n|\boldsymbol{\phi}_2^*) \boxed{} / Z_n$$

- No ``identifiable'' conclusion

Comparison with BIC (HMM)

- FAB lower bound

$$FIC_{LB}^{(i)}(\boldsymbol{x}^N, M) = \sum_{n=1}^{N}\sum_{t=1}^{T_n} \log \zeta_n^{t(i)} + \sum_{n,t=1}^{N,T_n} \log \Delta^t - \frac{\mathcal{D}_{\boldsymbol{\alpha}}}{2}\log N$$

$$- \sum_{k=1}^{K}\left(\frac{\mathcal{D}_{\boldsymbol{\beta}_k}}{2}(\log(\sum_{n,t}^{N,T_n-1} \tilde{q}(z_{nk}^t)) - 1) + \frac{\mathcal{D}_{\boldsymbol{\phi}_k}}{2}(\log(\sum_{n,t}^{N,T_n} \tilde{q}(z_{nk}^t)) - 1)\right)$$

- BIC (no theoretical justification)

$$BIC = \sum_{n,t=1}^{N,T_n} \log \zeta_n^t(ML) - \frac{\mathcal{D}}{2}\log \sum_{n=1}^{N} T_n$$

- Stochastic complexities of HMMs are theoretically shown to be much smaller than the BIC's complexity term (Yamazaki and Watanabe, 2005)

Comparison with VB (HMM)

- FAB lower bound

$$FIC_{LB}^{(i)}(\boldsymbol{x}^N, M) = \sum_{n=1}^{N}\sum_{t=1}^{T_n} \log \zeta_n^{t(i)} + \sum_{n,t=1}^{N,T_n} \log \Delta^t - \frac{\mathcal{D}_{\boldsymbol{\alpha}}}{2}\log N$$

$$- \sum_{k=1}^{K}\left(\frac{\mathcal{D}_{\boldsymbol{\beta}_k}}{2}(\log(\sum_{n,t}^{N,T_n-1} \tilde{q}(z_{nk}^t)) - 1) + \frac{\mathcal{D}_{\boldsymbol{\phi}_k}}{2}(\log(\sum_{n,t}^{N,T_n} \tilde{q}(z_{nk}^t)) - 1)\right)$$

- VB lower bound

$$\mathcal{F}_{VB} = \sum_{n,t=1}^{N,T_n} \log \zeta_n^t(VB) + \int d\boldsymbol{\alpha} q(\boldsymbol{\alpha}) \log \frac{p(\boldsymbol{\alpha})}{q(\boldsymbol{\alpha})}$$

$$+ \int d\boldsymbol{\beta} q(\boldsymbol{\beta}) \log \frac{p(\boldsymbol{\beta})}{q(\boldsymbol{\beta})} + \sum_{k=1}^{K} \int d\boldsymbol{\phi_k} q(\boldsymbol{\phi_k}) \log \frac{p(\boldsymbol{\phi_k})}{q(\boldsymbol{\phi_k})}$$

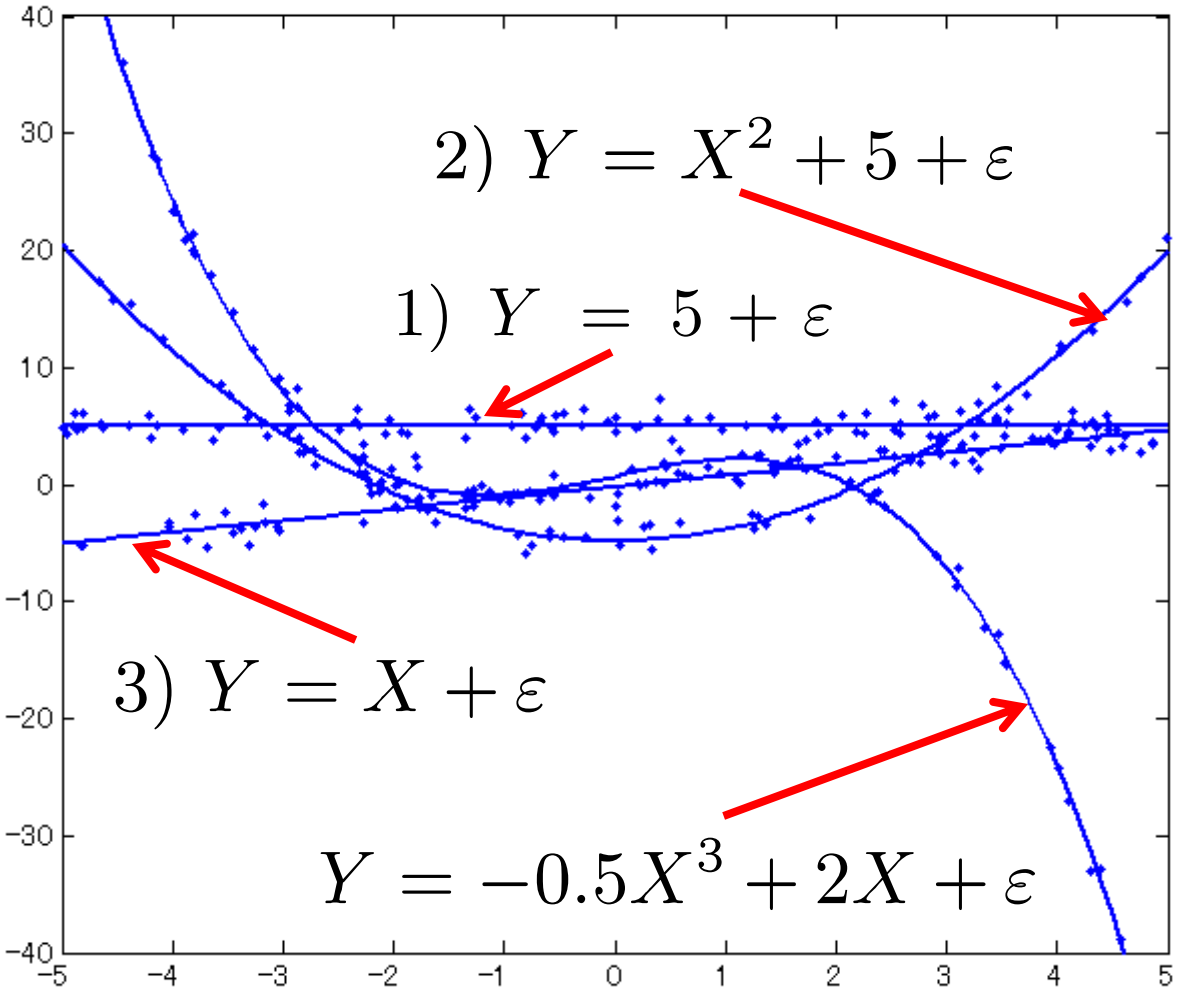
- Only FAB explicitly takes dependency between hidden states and model parameter into account.

第3部の構成

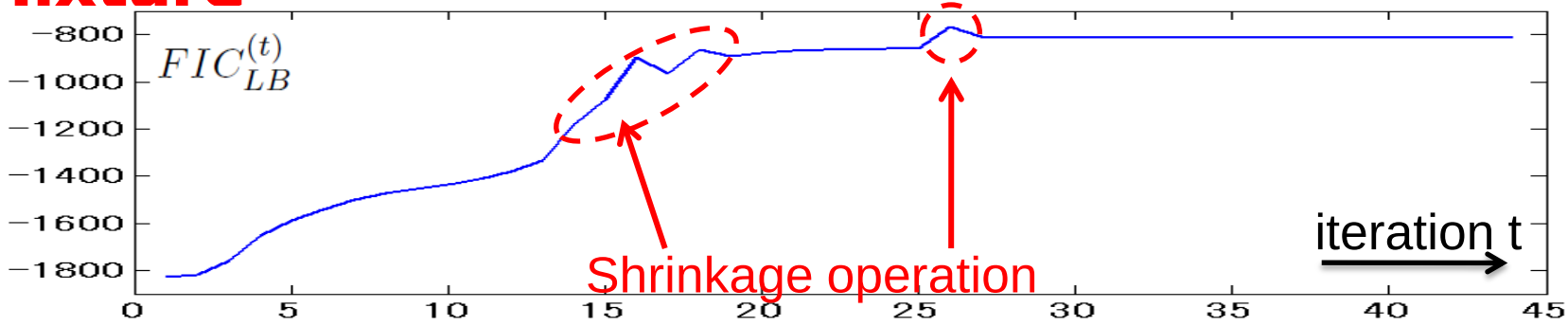
- 異種混合学習とモデル選択
- 因子化情報量基準と因子化漸近ベイズ推論
- 因子化漸近ベイズ推論の性質
- 実験
- 発展モデルへ

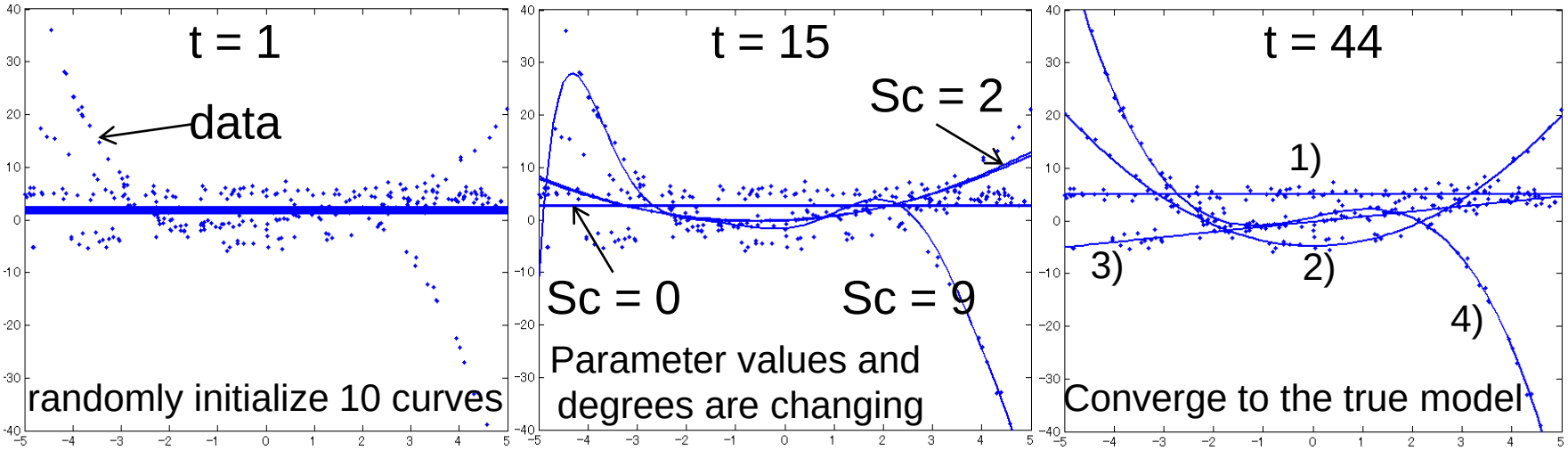
MM: Illustrative Polynomial Curve Mixture

- Artificially generate 4 curves



MM: Illustrative Polynomial Curve Mixture





polynomial curve with Gaussian noise

$$p(Y|X,\boldsymbol{\theta},\boldsymbol{S}) = \sum_{c=1}^{C} \alpha_c p(Y|X,\boldsymbol{\phi}_c,S_c)$$

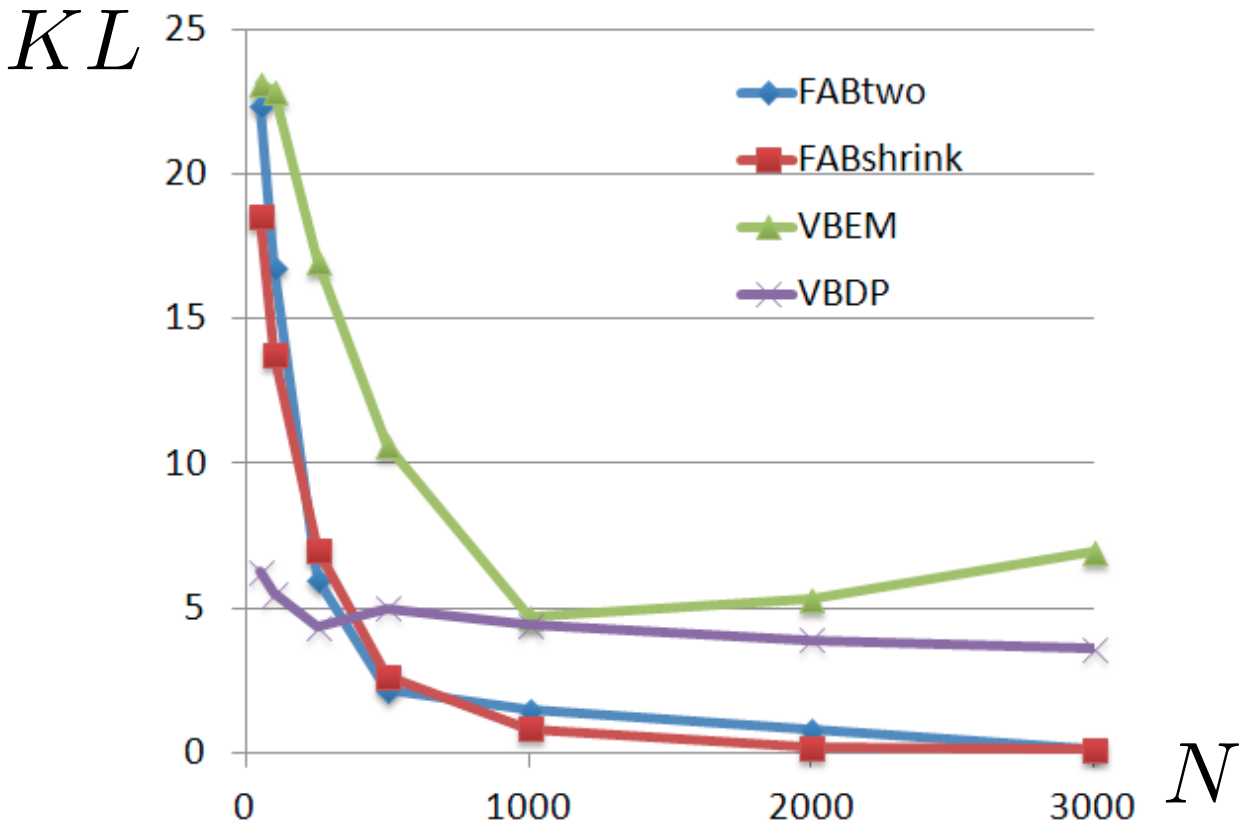
degree of curve

target variable

weight vector and noise level

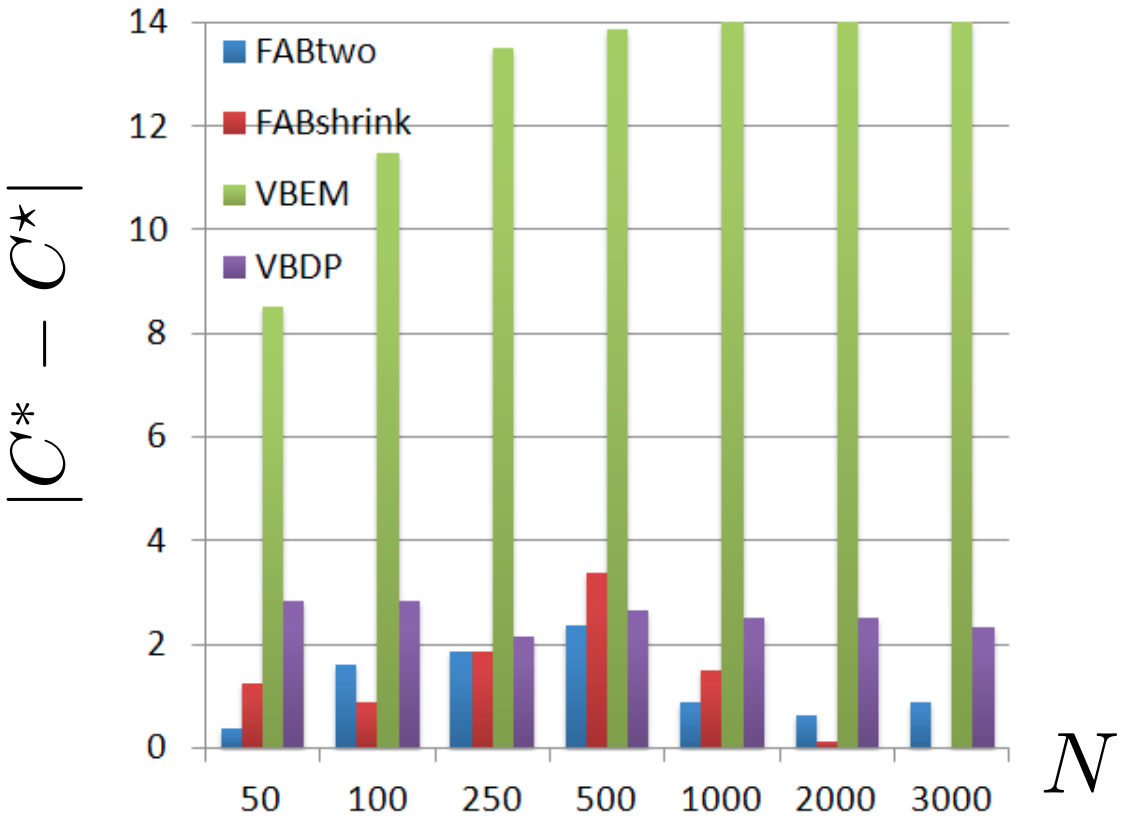
MM: Artificial GMM

- Comparison with VB and Non-parametric Bayes
  - VBEM (Bishop 2005): VBEM algorithm with Dirichlet prior (model selection via variational free energy)
  - VBDP (Blei and Jordan 2006): VB Dirichlet process mixture model (model selection via Dirichlet process prior)



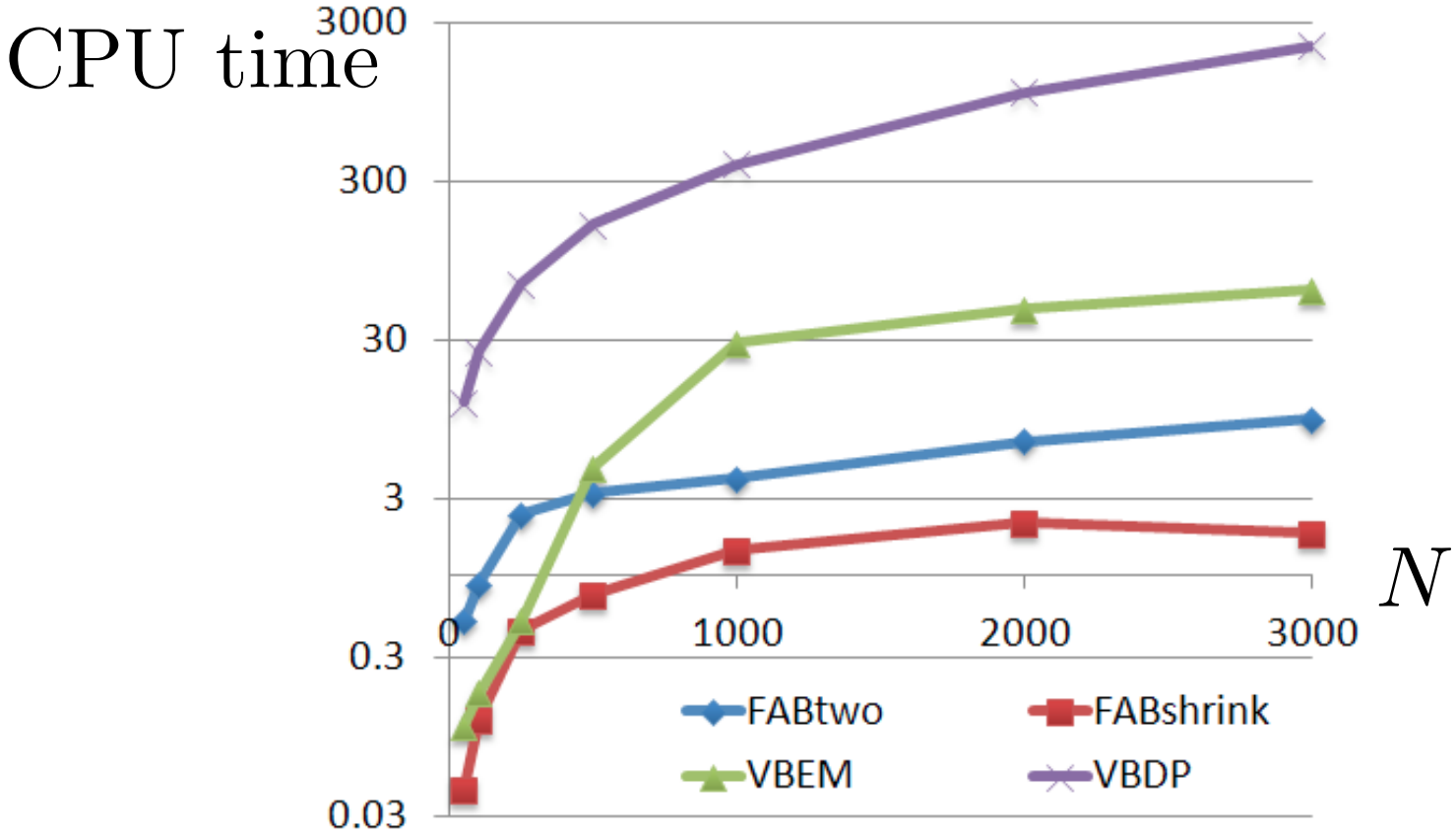
MM: Artificial GMM

- Comparison with VB and Non-parametric Bayes
  - VBEM (Bishop 2005): VBEM algorithm with Dirichlet prior (model selection via variational free energy)
  - VBDP (Blei and Jordan 2006): VB Dirichlet process mixture model (model selection via Dirichlet process prior)



MM: Artificial GMM

- Comparison with VB and Non-parametric Bayes
  - VBEM (Bishop 2005): VBEM algorithm with Dirichlet prior (model selection via variational free energy)
  - VBDP (Blei and Jordan 2006): VB Dirichlet process mixture model (model selection via Dirichlet process prior)



HMMs: E-Book Dataset

- Six books in different topics (http://www.gutenberg.org)
  - Alice's Adventures in Wonderland
  - The Art of War
  - The Metamorphosis
  - The Republic
  - The United States Declaration of Independence
  - The Adventures of Sherlock Holmes
- The first 5000 characters for training and the following 5000 characters for testing.
- Comparison methods
  - VB: variational Bayesian HMMs (Mackay 1997, Beal 2003)
  - iHMM: infinite HMMs with beam sampler (van Gael et al. 2008)

E-book Dataset

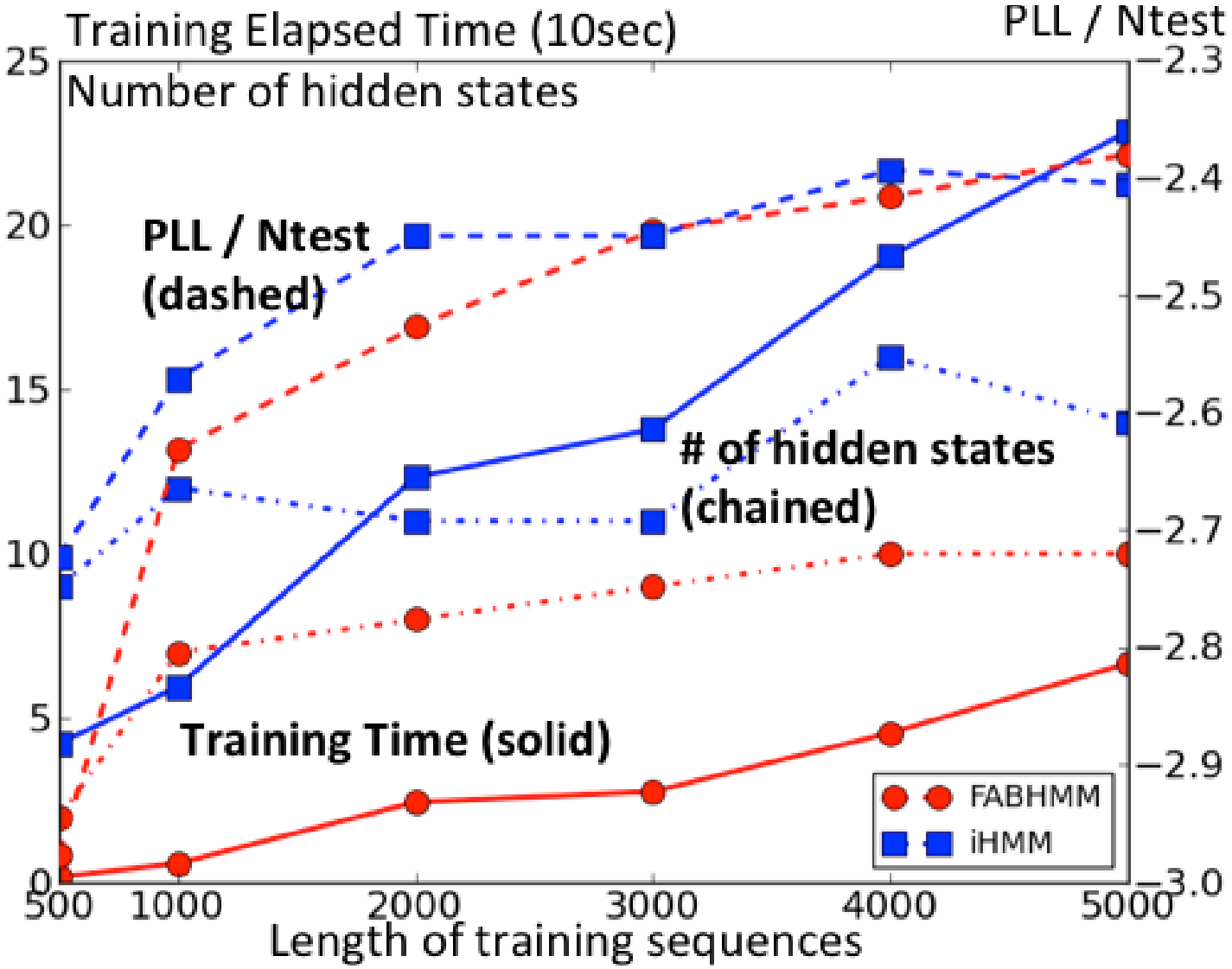
- Predictive Log-likelihood and Elapsed Time

Table 2. Estimated number of hidden states K, training time (sec), and PLLs on the ebook data sets.

	FABHMM			iHMM			VBHMM		
Data	K	Time	PLL	K	Time	PLL	K	Time	PLL
Alice	9	64.2	-2.57	13	234	-2.54	15	122.3	-2.73
Kafka	10	47.0	-2.36	11	218	-2.40	12	104.2	-2.62
Plato	10	66.6	-2.38	14	228	-2.41	8	144.9	-2.63
Sherl	11	72.3	-2.58	12	227	-2.52	19	98.0	-2.75
Sunzi	10	63.2	-2.56	14	228	-2.52	14	110.8	-2.72
DOI	10	73.8	-2.98	12	232	-2.75	11	94.6	-2.76

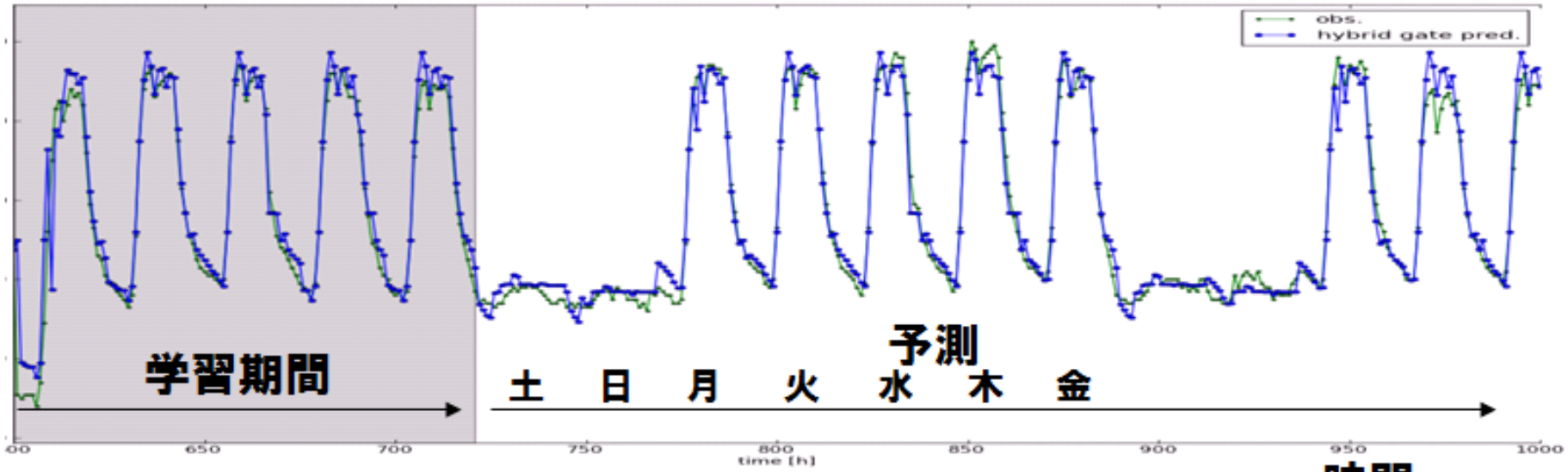
E-book Dataset

- Predictive Log-likelihood and Elapsed Time



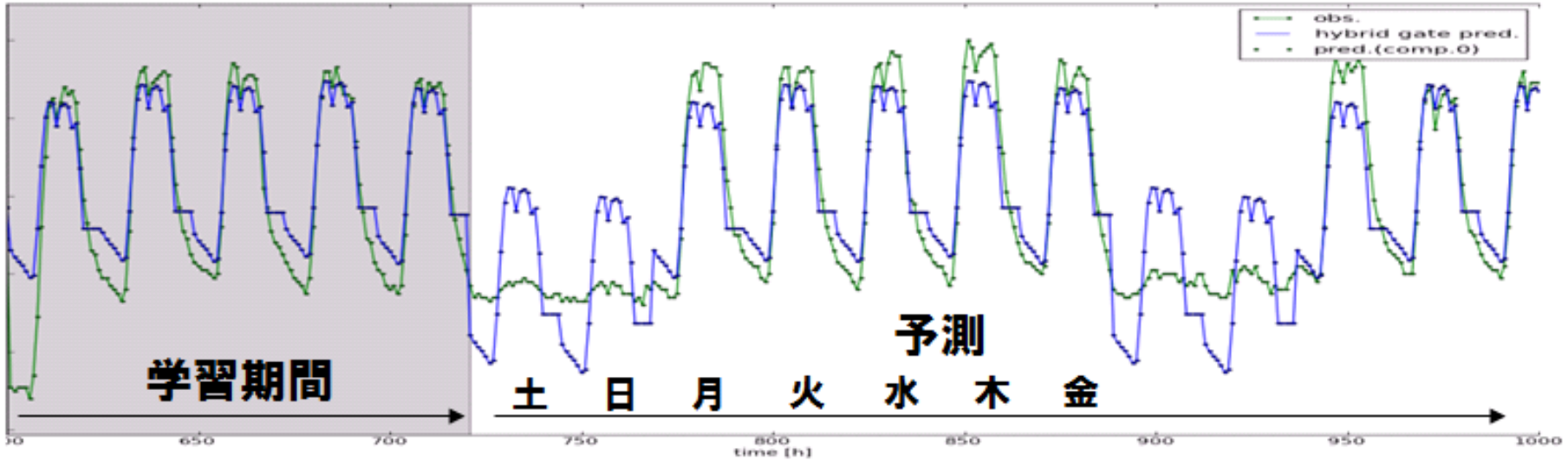
Application to Building Power Prediction

NEC Laboratories America
Relentless passion for innovation



自動分割(3つのパターンの切り替えで予測)

時間

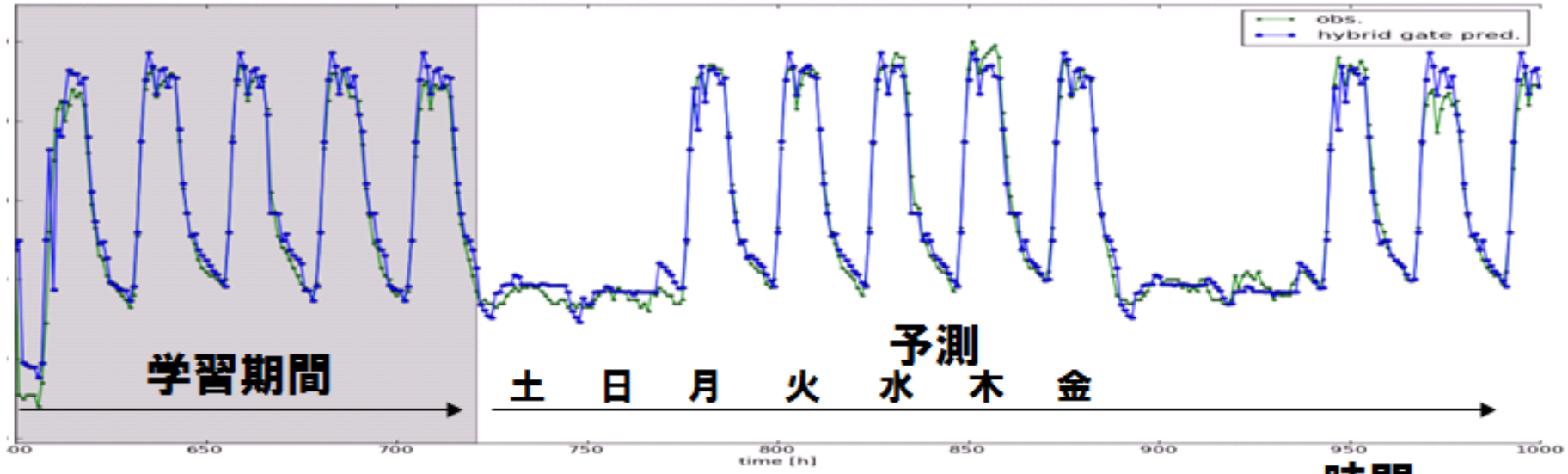


単純予測(切り替え無で、1つのパターンを使って予測)

時間

Application to Building Power Prediction

NEC Laboratories America
Relentless passion for innovation



自動分割(3つのパターンの切り替えで予測)