Determine the Number of States in Hidden Markov Models via Marginal Likelihood

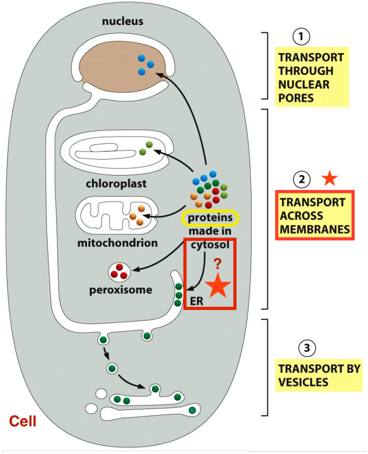
Yang Chen

Department of Statistics & MIDAS
University of Michigan, Ann Arbor

Joint work with C.L.Kao, C.D. Fuh, and S. Kou

Outline

Hidden Markov Models: an example



Ribosome (Manufacturer)

Protein ↓ **How to?**

Endoplasmic Reticulum (Mailroom)

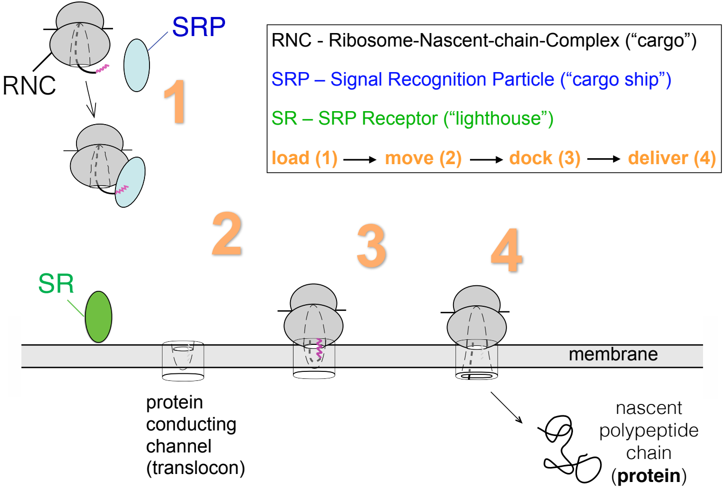
Hidden Markov Models: an example

RNC - Ribosome-Nascent-chain-Complex ("cargo")

SRP – Signal Recognition Particle ("cargo ship")

SR – SRP Receptor ("lighthouse")

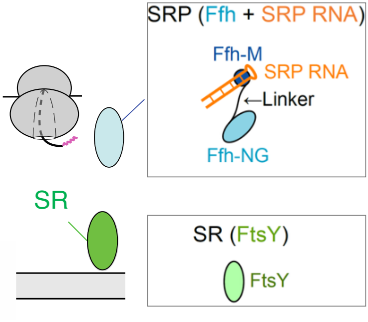
load (1) ⟶ move (2) ⟶ dock (3) ⟶ deliver (4)

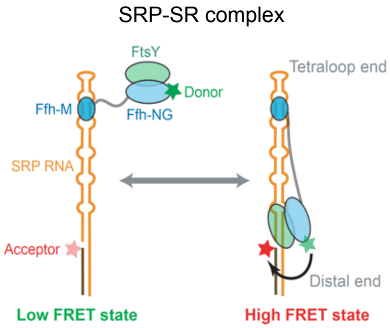


Hidden Markov Models: an example

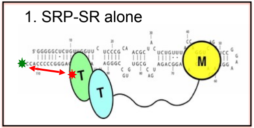
Single-molecule experiments – real time trajectory of FRET (distance).

FRET: energy transfer rate between two light-sensitive molecules.

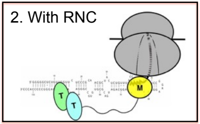




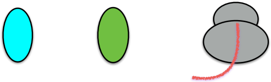
Hidden Markov Models: an example

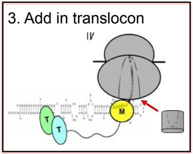


Ship (SRP) + Lighthouse (SR)

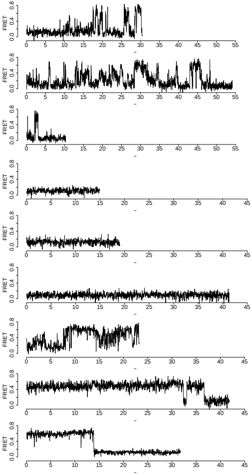


Ship + Lighthouse + Cargo

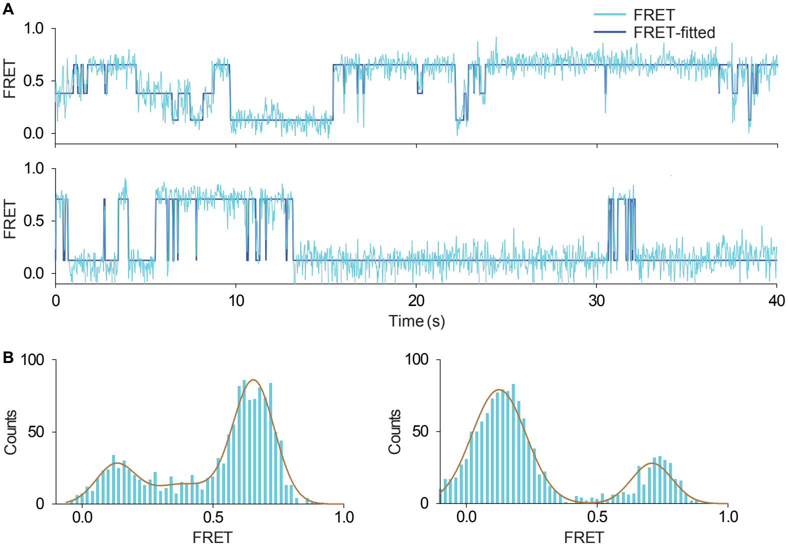




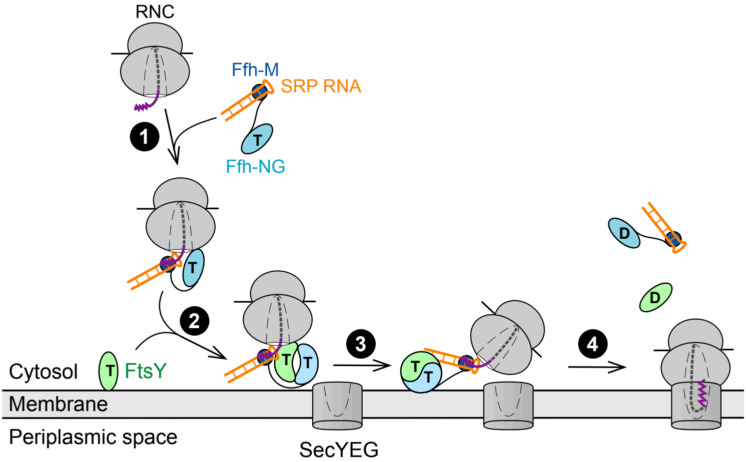
Ship + Lighthouse + Cargo + Dock



Hidden Markov Models: an example



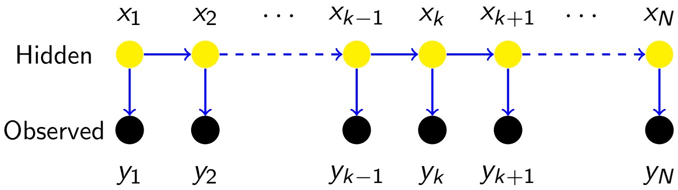
Hidden Markov Models: an example



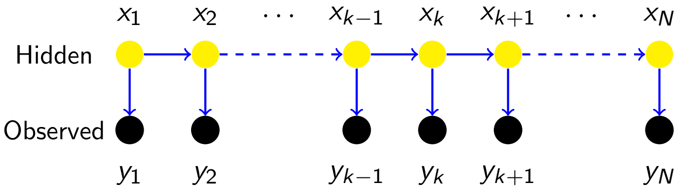
Refer to Chen et al. (2016) for details.

Outline

Introduction: Hidden Markov Models

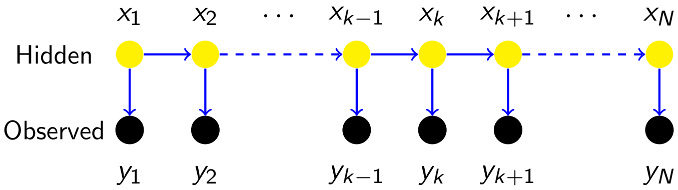


Introduction: Hidden Markov Models



- Observations: $\boldsymbol{y}_{1:N} = (y_1, \ldots, y_N) \in \mathbb{R}^N$.
- Hidden states: $\boldsymbol{x}_{1:N} = (x_1, \ldots, x_N) \in \{1, 2, \ldots, K\}^N$.

Introduction: Hidden Markov Models



- Observations: $\boldsymbol{y}_{1:N} = (y_1, \ldots, y_N) \in \mathbb{R}^N$.
- Hidden states: $\boldsymbol{x}_{1:N} = (x_1, \ldots, x_N) \in \{1, 2, \ldots, K\}^N$.
- Data generating process:

$$P(X_{t+1} = j | X_t = k) = P_{kj}, \;\; Y_t | X_t = k \sim \mathcal{F}(\boldsymbol{\theta}_k).$$

- Parameters: $\boldsymbol{P} = \boldsymbol{P}_{K \times K}$, $\{\boldsymbol{\theta}_k\}_{k=1}^K$.

Hidden Markov Models: Order Selection

- Focus: discrete state space hidden Markov models
  - the hidden states X_i have a finite support
  - observed at discrete time points $\{t_1, \ldots, t_n\}$

Hidden Markov Models: Order Selection

- Focus: discrete state space hidden Markov models
  - the hidden states X_i have a finite support
  - observed at discrete time points $\{t_1, \ldots, t_n\}$
- K: size of the support of the hidden states
  - not known beforehand
  - conveys important information of the underlying process

Hidden Markov Models: Order Selection

- Focus: discrete state space hidden Markov models
  - the hidden states X_i have a finite support
  - observed at discrete time points $\{t_1, \ldots, t_n\}$
- K: size of the support of the hidden states
  - not known beforehand
  - conveys important information of the underlying process
- Goal: provide *the marginal likelihood method*
  - to determine K
  - consistent
  - computationally feasible
  - minimal tuning

Outline

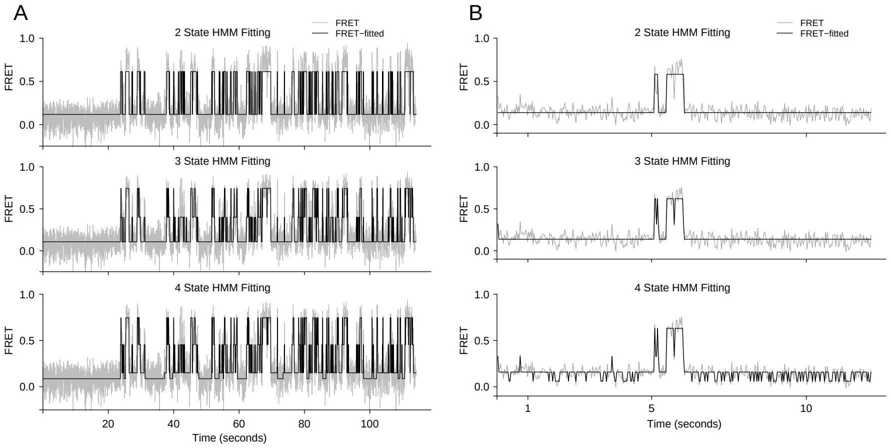
Model Selection

What is **Model Selection**?

Model Selection

What is **Model Selection**?

HMM Model Selection



Outline

Model Selection

Model Selection (General mixture models)

- Penalized likelihood Methods/ Information Criterion
  Chen and Kalbfleisch (1996); Lo et al. (2001); Jeffries (2003); Chen et al. (2008); Chen and Tan (2009); Chen and Li (2009); Chen and Khalili (2009); Huang et al. (2013); Rousseau and Mengersen (2011); Hui et al. (2015).
- Bayes Factors ($\approx$ BIC asymptotically). Kass and Raftery (1995).

Model Selection

Model Selection (General mixture models)

- Penalized likelihood Methods/ Information Criterion
  Chen and Kalbfleisch (1996); Lo et al. (2001); Jeffries (2003); Chen et al. (2008); Chen and Tan (2009); Chen and Li (2009); Chen and Khalili (2009); Huang et al. (2013); Rousseau and Mengersen (2011); Hui et al. (2015).
- Bayes Factors ($\approx$ BIC asymptotically). Kass and Raftery (1995).

Model Selection for HMM

- Existing work on finite-alphabet HMMs.
  Finesso (1990); Ziv and Merhav (1992); Kieffer (1993); Liu and Narayan (1994); Gassiat and Boucheron (2003); Rydén (1995); Ephraim and Merhav (2002).
- Most popular in practice: BIC (Rydén et al., 1998).
- Problem: lack of theoretical justification; unbounded likelihood.

Outline

Proposed Method: Marginal Likelihood

The marginal likelihood of HMM with K hidden states is

$$p_K(\boldsymbol{y}_{1:N}) = \int_\Theta \int_{\mathcal{X}^N} p(\boldsymbol{y}_{1:N}, \boldsymbol{x}_{1:N}|\boldsymbol{\theta}) d\boldsymbol{x}_{1:N} \; p_0(\boldsymbol{\theta}) d\boldsymbol{\theta}.$$

- Posterior samples: $\{\boldsymbol{\theta}_j\}_{j=1}^M \sim p(\boldsymbol{\theta}|\boldsymbol{y}_{1:N})$.

Proposed Method: Marginal Likelihood

The marginal likelihood of HMM with K hidden states is

$$p_K(\boldsymbol{y}_{1:N}) = \int_{\Theta} \int_{\mathcal{X}^N} p(\boldsymbol{y}_{1:N}, \boldsymbol{x}_{1:N} | \boldsymbol{\theta}) d\boldsymbol{x}_{1:N} \; p_0(\boldsymbol{\theta}) d\boldsymbol{\theta}.$$

- Posterior samples: $\{\boldsymbol{\theta}_j\}_{j=1}^M \sim p(\boldsymbol{\theta} | \boldsymbol{y}_{1:N})$.
- $p_K(\boldsymbol{y}_{1:N})$ is the unknown normalizing constant.

Proposed Method: Marginal Likelihood

The marginal likelihood of HMM with K hidden states is

$$p_K(\boldsymbol{y}_{1:N}) = \int_{\Theta} \int_{\mathcal{X}^N} p(\boldsymbol{y}_{1:N}, \boldsymbol{x}_{1:N}|\boldsymbol{\theta}) d\boldsymbol{x}_{1:N} \; p_0(\boldsymbol{\theta}) d\boldsymbol{\theta}.$$

- Posterior samples: $\{\boldsymbol{\theta}_j\}_{j=1}^{M} \sim p(\boldsymbol{\theta}|\boldsymbol{y}_{1:N})$.
- $p_K(\boldsymbol{y}_{1:N})$ is the unknown normalizing constant.
- Unnormalized posterior $p(\mathbf{y}_{1:n}|\phi)p_0(\phi)$ can be evaluated at any ϕ: forward algo. (Baum and Petrie, 1966; Baum et al., 1970; Xuan et al., 2001)

Proposed Method: Marginal Likelihood

The marginal likelihood of HMM with K hidden states is

$$p_K(\boldsymbol{y}_{1:N}) = \int_\Theta \int_{\mathcal{X}^N} p(\boldsymbol{y}_{1:N}, \boldsymbol{x}_{1:N}|\boldsymbol{\theta}) d\boldsymbol{x}_{1:N} \; p_0(\boldsymbol{\theta}) d\boldsymbol{\theta}.$$

- Posterior samples: $\{\boldsymbol{\theta}_j\}_{j=1}^M \sim p(\boldsymbol{\theta}|\boldsymbol{y}_{1:N})$.
- $p_K(\boldsymbol{y}_{1:N})$ is the unknown normalizing constant.
- Unnormalized posterior $p(\mathbf{y}_{1:n}|\phi)p_0(\phi)$ can be evaluated at any ϕ: forward algo. (Baum and Petrie, 1966; Baum et al., 1970; Xuan et al., 2001)
- Proposed Procedure:

posterior sampling + estimating normalizing constant.

Outline

Posterior Sampling of HMM

Posterior Sampling of HMM

Data Augmentation (Gibbs Sampling):

- Augment the parameter space with the hidden states (Tanner and Wong, 1987; Rydén, 2008).
- Sample parameters and hidden states iteratively till convergence.

Posterior Sampling of HMM

Data Augmentation (Gibbs Sampling):

- Augment the parameter space with the hidden states (Tanner and Wong, 1987; Rydén, 2008).
- Sample parameters and hidden states iteratively till convergence.
- Pros and cons: Iterative algorithm (slow), full posterior.

Posterior Sampling of HMM

Data Augmentation (Gibbs Sampling):

- Augment the parameter space with the hidden states (Tanner and Wong, 1987; Rydén, 2008).
- Sample parameters and hidden states iteratively till convergence.
- Pros and cons: Iterative algorithm (slow), full posterior.

MCMC + Forward algorithm

- Forward algorithm (Baum and Petrie, 1966; Baum et al., 1970; Xuan et al., 2001): integrate out hidden states in linear time.
- Any MCMC algorithm (Liu, 2001) can be applied here.

Outline

Estimation of Normalizing Constant: Literature

Existing Work

- Laplace approximation & Bartlett adjustment (DiCiccio et al., 1997).

Estimation of Normalizing Constant: Literature

Existing Work

- Laplace approximation & Bartlett adjustment (DiCiccio et al., 1997).
- Methods based on importance sampling and reciprocal importance sampling (Geweke, 1989; Oh and Berger, 1993; Newton and Raftery, 1994; Gelfand and Dey, 1994; Ionides, 2008; Neal, 2005; Steele et al., 2006; Chen and Shao, 1997; DiCiccio et al., 1997).

Estimation of Normalizing Constant: Literature

Existing Work

- Laplace approximation & Bartlett adjustment (DiCiccio et al., 1997).
- Methods based on importance sampling and reciprocal importance sampling (Geweke, 1989; Oh and Berger, 1993; Newton and Raftery, 1994; Gelfand and Dey, 1994; Ionides, 2008; Neal, 2005; Steele et al., 2006; Chen and Shao, 1997; DiCiccio et al., 1997).
- Methods based on Markov chain Monte Carlo (MCMC) output (Chib, 1995; Geyer, 1994; Chib and Jeliazkov, 2001, 2005; de Valpine, 2008; Petris and Tardella, 2007).

Estimation of Normalizing Constant: Literature

Existing Work

- Laplace approximation & Bartlett adjustment (DiCiccio et al., 1997).
- Methods based on importance sampling and reciprocal importance sampling (Geweke, 1989; Oh and Berger, 1993; Newton and Raftery, 1994; Gelfand and Dey, 1994; Ionides, 2008; Neal, 2005; Steele et al., 2006; Chen and Shao, 1997; DiCiccio et al., 1997).
- Methods based on Markov chain Monte Carlo (MCMC) output (Chib, 1995; Geyer, 1994; Chib and Jeliazkov, 2001, 2005; de Valpine, 2008; Petris and Tardella, 2007).
- Estimating ratio of normalizing constants: bridge sampling (Meng and Wong, 1996) and path sampling (Gelman and Meng, 1998).

Estimation of Normalizing Constant: Literature

- **Importance sampling (IS)** $\hat{C}_I$.
  $q(\cdot)$ should be similar to and have tails no thinner than $h(\cdot)$;
  $q(\cdot) = \phi(\cdot; \hat{\theta}, \hat{\Sigma})$. The locally restricted version $\hat{C}_I^*$.

- **Reciprocal importance sampling (RIS)** $\hat{C}_R$.
  $s(\cdot)$ should be similar to $h(\cdot)$ and has sufficiently thin tails,
  $s(\cdot) = \phi(\cdot; \hat{\theta}, \hat{\Sigma})$. The locally restricted version: $\hat{C}_R^*$.

$$\hat{C}_I = \frac{1}{M}\sum_{i=1}^{M} \frac{h(\tilde{\theta}_i)}{q(\tilde{\theta}_i)}, \quad \hat{C}_R = \left[\frac{1}{m}\sum_i \frac{s(\theta_i)}{h(\theta_i)}\right]^{-1}.$$

$$\hat{C}_I^* = \frac{\frac{1}{M}\sum_i h(\tilde{\theta}_i) Z_B(\tilde{\theta}_i)/q(\tilde{\theta}_i)}{\frac{1}{m}\sum_i Z_B(\theta_i)}, \quad \hat{C}_R^* = \alpha\left[\frac{1}{m}\sum_i \frac{s(\theta_i) Z_B(\theta_i)}{h(\theta_i)}\right]^{-1}.$$

Importance Sampling – Travel with Maps

Outline

Estimation of Normalizing Constant: Procedure I

1. Obtain posterior samples. Sample from $p(\phi_K|\mathbf{y}_{1:n})$ using a preferred MCMC algorithm, and denote the samples by $\{\phi_K^{(i)}\}_{i=1}^N$ (where N is often a few thousand).
2. Find a "good" importance function. Fit a Gaussian/student-t mixture model using the samples $\{\phi_K^{(i)}\}_{i=1}^N$.
3. Choose a finite region. Choose Ω_K to be a bounded subset of the parameter space such that $1/2 < \int_{\Omega_K} g(\cdot) < 1$. This can be achieved through finding an appropriate finite region for each mixing component of $g(\cdot)$, avoiding the tail parts.
4. Estimate $p_K(\boldsymbol{y}_{1:n})$ using either way as follows:

Estimation of Normalizing Constant: Procedure II

- Reciprocal importance sampling. Approximate $p_K(\mathbf{y}_{1:n})$ by

$$\hat{p}_K^{(RIS)}(\mathbf{y}_{1:n}) = \left[\frac{1}{N\int_{\Omega_k} g(\cdot)} \sum_{i=1}^{N} \frac{g(\phi_K^{(i)})}{p(\mathbf{y}_{1:n}, \phi_K^{(i)})} I_{\phi_K^{(i)} \in \Omega_K}\right]^{-1}, \qquad (1)$$

where $I_{\phi_K^{(i)} \in \Omega_K} = 1$ if $\phi_K^{(i)} \in \Omega_K$ and zero otherwise.

- Importance sampling.
  1. Draw M independent samples from $g(\cdot)$, denoted by $\{\psi_K^{(j)}\}_{1 \le j \le M}$.
  2. Approximate $p_K(\mathbf{y}_{1:n})$ by

$$\hat{p}_K^{(IS)}(\mathbf{y}_{1:n}) = \frac{1}{MP_\Omega} \sum_{j=1}^{M} \frac{p(\mathbf{y}_{1:n}, \psi_K^{(j)})}{g(\psi_K^{(j)})} I_{\{\psi_K^{(j)} \in \Omega_K\}}, \qquad (2)$$

where $I_{\{\psi_K^{(j)} \in \Omega_K\}} = 1$ if $\psi_K^{(j)} \in \Omega_K$ and zero otherwise; $P_\Omega = \#\mathcal{S}/N$, where $\mathcal{S} = \{i : \phi_K^{(i)} \in \Omega_K; 1 \le i \le N\}$.

Outline

Design of Simulation Experiments

- Parmeters: $\boldsymbol{\mu} = (1, 2, \ldots, K)$, $\boldsymbol{\sigma}^2 = (\sigma^2, \ldots, \sigma^2)$.
- Four kinds of transition matrices: flat ($P_K^{(1)}$), moderate and strongly diagonal ($P_K^{(2)}, P_K^{(3)}$) and strongly off-diagonal ($P_K^{(4)}$).
- For example, if $K = 4$, the four matrices are:

$$P_4^{(1)} = \begin{pmatrix} 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \end{pmatrix}, \quad P_4^{(2)} = \begin{pmatrix} 0.8 & 1/15 & 1/15 & 1/15 \\ 1/15 & 0.8 & 1/15 & 1/15 \\ 1/15 & 1/15 & 0.8 & 1/15 \\ 1/15 & 1/15 & 1/15 & 0.8 \end{pmatrix},$$

$$P_4^{(3)} = \begin{pmatrix} 0.95 & 1/60 & 1/60 & 1/60 \\ 1/60 & 0.95 & 1/60 & 1/60 \\ 1/60 & 1/60 & 0.95 & 1/60 \\ 1/60 & 1/60 & 1/60 & 0.95 \end{pmatrix}, \quad P_4^{(4)} = \begin{pmatrix} 0.1 & 0.3 & 0.3 & 0.3 \\ 0.3 & 0.1 & 0.3 & 0.3 \\ 0.3 & 0.3 & 0.1 & 0.3 \\ 0.3 & 0.3 & 0.3 & 0.1 \end{pmatrix}.$$

HMM State Selection Correct Frequency

K	σ	n	$Q_K = P_K^{(1)}$		$Q_K = P_K^{(2)}$		$Q_K = P_K^{(3)}$		$Q_K = P_K^{(4)}$	
			ML	BIC	ML	BIC	ML	BIC	ML	BIC
2	0.2	200	100	100	100	100	100	100	100	100
2	0.3	200	100	100	100	100	100	100	100	100
2	0.4	200	100	100	100	100	100	100	100	100
3	0.2	200	100	100	100	100	95.0	96.0	100	100
3	0.3	200	62.5	22.5	100	99.5	96.0	94.5	99.0	92.5
3	0.4	200	1.50	0.00	91.0	77.0	88.5	88.0	25.0	10.5
4	0.2	200	100	90.0	100	100	81.0	76.0	100	97.5
4	0.3	200	4.00	0.00	97.0	85.0	65.0	60.0	22.0	0.50
4	0.4	200	0.00	0.00	45.0	21.0	37.5	37.0	0.00	0.00
5	0.2	200	99.0	15.5	99.5	95.0	55.0	44.0	99.5	29.0
5	0.3	200	0.50	0.00	82.0	37.0	24.0	19.0	1.00	0.00
5	0.4	200	0.00	0.00	10.5	1.00	7.00	4.50	0.00	0.00

HMM State Selection Correct Frequency

K	σ	n	$Q_K = P_K^{(1)}$ ML	$Q_K = P_K^{(1)}$ BIC	$Q_K = P_K^{(2)}$ ML	$Q_K = P_K^{(2)}$ BIC	$Q_K = P_K^{(3)}$ ML	$Q_K = P_K^{(3)}$ BIC	$Q_K = P_K^{(4)}$ ML	$Q_K = P_K^{(4)}$ BIC
2	0.2	2000	100	100	100	100	100	100	100	100
2	0.3	2000	100	100	100	100	100	100	100	100
2	0.4	2000	100	100	100	100	100	100	100	100
3	0.2	2000	100	100	100	100	100	100	100	100
3	0.3	2000	100	100	100	100	100	100	100	100
3	0.4	2000	98.5	72.0	100	100	100	100	100	100
4	0.2	2000	100	100	100	100	100	100	100	100
4	0.3	2000	99.5	98.5	100	100	100	100	100	100
4	0.4	2000	4.50	0.00	100	100	100	100	84.0	20.5
5	0.2	2000	100	100	100	100	100	100	100	100
5	0.3	2000	95.0	23.5	100	100	100	100	99.0	87.0
5	0.4	2000	0.00	0.00	100	100	100	100	2.00	0.00

Outline

Consistency of Marginal Likelihood Method: HMM

Theorem

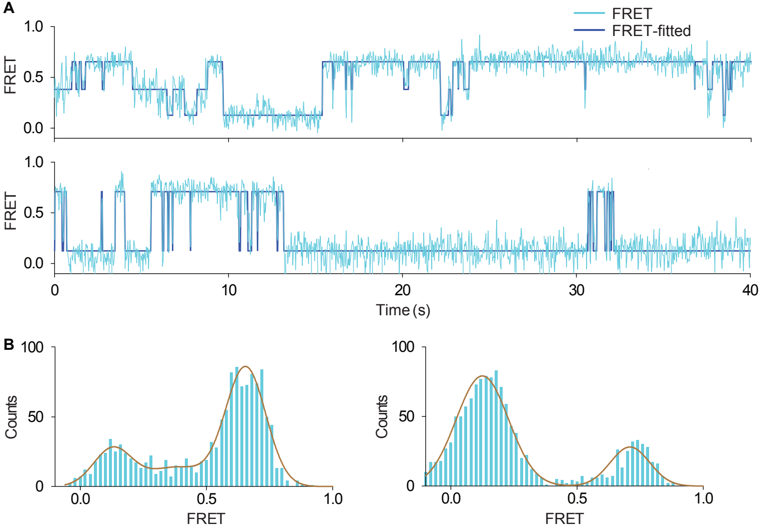
Assume regularity conditions 1)-5). Then for any $K \neq K^*$, as $n \to \infty$,

$$\frac{p_K(\mathbf{y}_{1:n})}{p_{K^*}(\mathbf{y}_{1:n})} = o_P(n^{-1/2} \log n). \tag{3}$$

Furthermore, if K^* is bounded from above, i.e. there exists a finite positive constant $\overline{K}$ such that $K^* \leq \overline{K}$, then as $n \to \infty$,

$$\widehat{K}_n := arg\ max_{1 \leq K \leq \overline{K}}\ p_K(\mathbf{y}_{1:n}) \xrightarrow{P} K^*. \tag{4}$$

Connections of HMM and GM



Consistency of Marginal Likelihood Method: GM

Theorem

Assume that all the conditions in Theorem 1 hold, except that condition (C1) is replaced by (C1') in Appendix, which restricts $\nu_K(\cdot|\beta_K)$ *to be supported on* $\tilde{\mathcal{Q}}_K = \{Q : q_{1k} = q_{2k} = \cdots = q_{Kk}$ *for all* $1 \leq k \leq K\}$*, i.e., assuming a prior for a mixture model without state dependency. Then the consistency of* $\widehat{K}_n$ *defined in (4) still holds.*

Consistency of Marginal Likelihood Method: GM

Theorem

Assume that all the conditions in Theorem 1 hold, except that condition (C1) is replaced by (C1') in Appendix, which restricts $\nu_K(\cdot|\beta_K)$ to be supported on $\tilde{\mathcal{Q}}_K = \{Q : q_{1k} = q_{2k} = \cdots = q_{Kk} \text{ for all } 1 \le k \le K\}$, i.e., assuming a prior for a mixture model without state dependency. Then the consistency of $\widehat{K}_n$ defined in (4) still holds.

- Computational cost: Theorem 1 (HMM) $>$ Theorem 2 (GM).

Consistency of Marginal Likelihood Method: GM

Theorem

Assume that all the conditions in Theorem 1 hold, except that condition (C1) is replaced by (C1') in Appendix, which restricts $\nu_K(\cdot|\beta_K)$ *to be supported on* $\tilde{\mathcal{Q}}_K = \{Q : q_{1k} = q_{2k} = \cdots = q_{Kk}$ *for all* $1 \leq k \leq K\}$*, i.e., assuming a prior for a mixture model without state dependency. Then the consistency of* $\widehat{K}_n$ *defined in (4) still holds.*

- Computational cost: Theorem 1 (HMM) $>$ Theorem 2 (GM).
- Theorem 2 requires n to be large so that $\boldsymbol{y}_{1:n}$ shows a "mixture model" behaviour through *stability convergence* $\Rightarrow$ a larger constant term in front of the common rate $n^{-1/2} \log n$.

Outline

Questions

Email: ychenang@umich.edu

References I

L. E. Baum and T. Petrie. Statistical inference for probabilistic functions of finite state Markov chains. *Ann. Math. Stat.*, 37(6):1554–1563, 1966.

L. E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Stat.*, 41(1):164–171, 1970.

J. Chen and J. D. Kalbfleisch. Penalized minimum-distance estimates in finite mixture models. *Can. J. Stat.*, 24(2):167–175, 1996.

J. Chen and A. Khalili. Order selection in finite mixture models with a nonsmooth penalty. *J. Amer. Statist. Assoc.*, 104(485):187–196, 2009.

J. Chen and P. Li. Hypothesis test for normal mixture models: the EM approach. *Ann. Statist.*, 37:2523–2542, 2009.

References II

J. Chen and X. Tan. Inference for multivariate normal mixtures. *J. Multivar. Anal.*, 100(7):1367–1383, 2009.

J. Chen, X. Tan, and R. Zhang. Inference for normal mixtures in mean and variance. *Stat. Sin.*, pages 443–465, 2008.

M.-H. Chen and Q.-M. Shao. On Monte Carlo methods for estimating ratios of normalizing constants. *Ann. Statist.*, 25(4):1563–1594, 1997.

Yang Chen, Kuang Shen, Shu-Ou Shan, and SC Kou. Analyzing single-molecule protein transportation experiments via hierarchical hidden Markov models. *Journal of the American Statistical Association*, 111(515):951–966, 2016.

S. Chib. Marginal likelihood from the Gibbs output. *J. Amer. Statist. Assoc.*, 90(432):1313–1321, 1995.

References III

S. Chib and I. Jeliazkov. Marginal likelihood from the Metropolis-Hastings output. *J. Amer. Statist. Assoc.*, 96(453):270–281, 2001.

S. Chib and I. Jeliazkov. Accept-reject Metropolis-Hastings sampling and marginal likelihood estimation. *Statist. Neerlandica*, 59(1):30–44, 2005.

P. de Valpine. Improved estimation of normalizing constants from Markov chain Monte Carlo output. *J. Comput. Graph. Statist.*, 17(2):333–351, 2008.

T. J. DiCiccio, R. E. Kass, A. E. Raftery, and L. Wasserman. Computing Bayes factors by combining simulation and asymptotic approximations. *J. Amer. Statist. Assoc.*, 92(439):903–915, 1997.

Y. Ephraim and N. Merhav. Hidden Markov processes. *IEEE Trans. Info. Theory*, 48(6):1518–1569, 2002.

References IV

L. Finesso. Consistent estimation of the order for Markov and hidden Markov chains. Technical report, MARYLAND UNIV COLLEGE PARK INST FOR SYSTEMS RESEARCH, 1990.

E. Gassiat and S. Boucheron. Optimal error exponents in hidden Markov model order estimation. *IEEE Trans. Info. Theory*, 48(4):964–980, 2003.

A. E. Gelfand and D. K. Dey. Bayesian model choice: Asymptotics and exact calculations. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 56(3): 501–514, 1994.

A. Gelman and X.-L. Meng. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Stat. Sci.*, 13 (2):163–185, 1998.

References V

J. Geweke. Bayesian inference in econometric models using Monte Carlo integration. *Econometrica*, 57:1317–1339, 1989.

C. J. Geyer. Estimating normalizing constants and reweighting mixtures in Markov chain Monte Carlo. *Technical Report No. 568, School of Statistics, Univ. Minnesota*, 1994.

T. Huang, H. Peng, and K. Zhang. Model selection for Gaussian mixture models. *arXiv preprint arXiv:1301.3558*, 2013.

F. K. Hui, D. I. Warton, and S. D. Foster. Order selection in finite mixture models: complete or observed likelihood information criteria? *Biometrika*, 102(3):724–730, 2015.

E. L. Ionides. Truncated importance sampling. *J. Comput. Graph. Statist.*, 17(2):295–311, 2008.

References VI

N. O. Jeffries. A note on 'Testing the number of components in a normal mixture'. *Biometrika*, 90(4):991–994, 2003.

R. E. Kass and A. E. Raftery. Bayes factors. *J. Amer. Statist. Assoc.*, 90 (430):773–795, 1995.

J. C. Kieffer. Strongly consistent code-based identification and order estimation for constrained finite-state model classes. *IEEE Trans. Info. Theory*, 39(3):893–902, 1993.

C.-C. Liu and P. Narayan. Order estimation and sequential universal data compression of a hidden Markov source by the method of mixtures. *IEEE Trans. Info. Theory*, 40(4):1167–1180, 1994.

J. S. Liu. *Monte Carlo Strategies in Scientific Computing*. Springer-Verlag New York, Inc., 2001.

References VII

Y. Lo, N. R. Mendell, and D. B. Rubin. Testing the number of components in a normal mixture. *Biometrika*, 88(3):767–778, 2001.

X.-L. Meng and W. H. Wong. Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Stat. Sin.*, 6:831–860, 1996.

R. M. Neal. Estimating ratios of normalizing constants using linked importance sampling. *Technical Report No. 0511, Department of Statistics, Univ. Toronto*, 2005.

M. A. Newton and A. E. Raftery. Approximate Bayesian inference with the weighted likelihood bootstrap. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 56(1):3–48, 1994.

References VIII

M.-S. Oh and J. O. Berger. Integration of multimodal functions by Monte Carlo importance sampling. *J. Amer. Statist. Assoc.*, 88(422):450–456, 1993.

G. Petris and L. Tardella. New perspectives for estimating normalizing constants via posterior simulation. *Technical Report, DSPSA, Sapienza Universitá di Roma*, 2007.

J. Rousseau and K. Mengersen. Asymptotic behaviour of the posterior distribution in overfitted mixture models. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 73(5):689–710, 2011.

T. Rydén. Estimating the order of hidden Markov models. *Statistics*, 26 (4):345–354, 1995.

References IX

T. Rydén. EM versus Markov chain Monte Carlo for estimation of hidden Markov models: A computational perspective. *Bayesian Anal.*, 3(4): 659–688, 2008.

T. Rydén, T. Teräsvirta, and S. Asbrink. Stylized facts of daily returns series and the hidden Markov model. *J. Appl. Econ.*, 13:217–244, 1998.

R. J. Steele, A. E. Raftery, and M. J. Emond. Computing normalizing constants for finite mixture models via incremental mixture importance sampling. *J. Comput. Graph. Statist.*, 15(3):712–734, 2006.

M. A. Tanner and W. H. Wong. The calculation of posterior distributions by data augmentation. *J. Amer. Statist. Assoc.*, 82:528–540, 1987.

References X

G. Xuan, W. Zhang, and P. Chai. EM algorithms of gaussian mixture model and hidden markov model. In *Int. Conf. Image. Proc.*, volume 1, pages 145–148. IEEE, 2001.

J. Ziv and N. Merhav. Estimating the number of states of a finite-state source. *IEEE Trans. Info. Theory*, 38(1):61–65, 1992.