Coupled Hidden Markov Models: some computational challenges

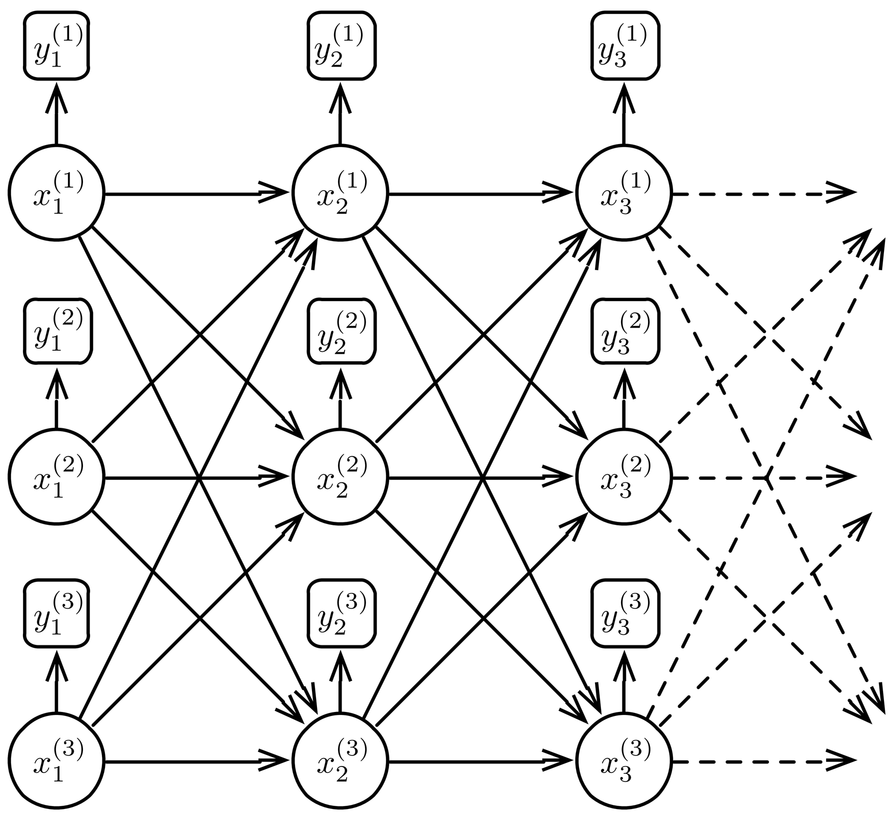
Louis J. M. Aslett (aslett@stats.ox.ac.uk) and Chris C. Holmes (cholmes@stats.ox.ac.uk)
Department of Statistics, University of Oxford

i-like.org.uk

UNIVERSITY OF OXFORD

1. CHMMs

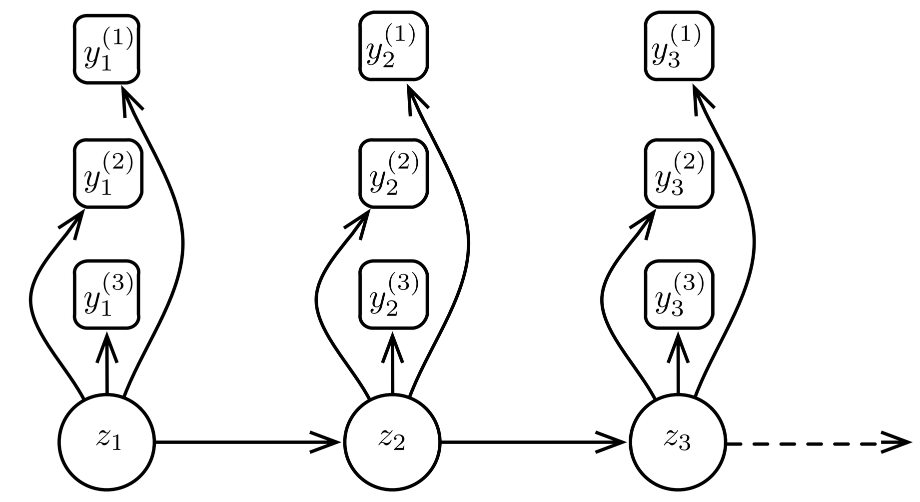
Coupled Hidden Markov Models (CHMMs) are a natural extension of HMMs when there are multiple observation sequences with dependencies:



Notation:
$x_t^{(i)}, y_t^{(i)}$: hidden state/obs at time t in chain i.

2. The naïve approach

One might naïvely reformulate as:



$Z_t = (X_t^{(1)}, \dots, X_t^{(3)})$ and $Y_t^{(i)} \,|\, Z_t = Y_t^{(i)} \,|\, X_t^{(i)}$

$\implies$ for C chains with $X_t^{(i)} \in \{1, \dots, N\}$, $|Z_t| = N^C$.

Natural forward variable becomes:

$$\alpha_t(i_1, \dots, i_C) = \mathbb{P}(Y_{1:t}^{(1:C)} = y_{1:t}^{(1:C)}, X_t^{(1:C)} = i_{1:C})$$

$$= \left(\sum_{j_1=1}^{N} \cdots \sum_{j_C=1}^{N} \alpha_{t-1}(j_1, \dots, j_C) \prod_{k=1}^{C} \mathbb{P}(X_t^{(k)} = i_k \,|\, x_{t-1}^{(1:C)} = j_{1:C}) \right) \prod_{k=1}^{C} f_{Y\,|\,X}(y_t^{(k)} \,|\, i_k)$$

Objective is inference in the minimal setting of $N = 10, C = 100, T = 10^5$. This leads to numerous challenges:

- computing forward variable $\implies$ TN^C additions and TC multiplications; $\geq 10^{105}$ elementary operations
- forward variable requires $8TN^C$ bytes of memory to store; $\geq 7.45 \times 10^{96}$ GB memory
- transition matrix is $N^C \times N^C$. $\geq 9.31 \times 10^{190}$ GB memory

Hence, naïve approach clearly a non-starter.

3. Existing approaches

Saul and Jordan (1999) Mixture Model

$$\mathbb{P}(X_t^{(i)} \,|\, x_{t-1}^{(1:C)}) = \sum_{k=1}^{C} \omega_{ki} \, \mathbb{P}(X_t^{(i)} \,|\, x_{t-1}^{(k)})$$

ω_{ki} can be viewed as mixing weights, or strength of effect of chain k on chain i. Now only NC^2 parameters.

Zhong and Ghosh (2002) Marginal Composite Likelihood

$$\mathbb{P}(Y_{1:T}^{(1:C)}) \approx \prod_{k=1}^{C} \mathbb{P}(Y_{1:T}^{(k)}) = \prod_{k=1}^{C} \sum_{i=1}^{N} \alpha_T^{(k)}(i)$$

with $\alpha_T^{(k)}(i)$ itself a factored approximation of the forward variable. Only $C = 2$ example.

Sherlock *et al.* (2013) Structured Transitions
Uses structured transition matrix for each chain, where probabilities modelled with a logistic regression with others chains (and external factors) as covariates.

Choi *et al.* (2013) Logistic Regression
Similarly, a transition matrix per chain, with logistic regression transition probabilities.
But, for speed, ad-hoc inferential procedure: mixture model EM to infer observation model, Viterbi to select most likely hidden sequence, IRLS on subsample to fit LR with lasso+AIC.
$C = 39, N = 2, T = 15.4 \times 10^6$

4. Our hidden layer model

$$X_t^{(i)} \,|\, X_{t-1}^{(1:C)} \sim \mathcal{M}\left(p_{t1}^{(i)}, \dots, p_{tN}^{(i)}\right), t \in \{2, \dots, T\}$$

$$p_{tj}^{(i)} = \frac{\exp(\tilde{\mathbf{x}}_{t-1}^{(1:C)} \boldsymbol{\beta}_j^{(i)})}{\sum_{n=1}^{N-1} \exp(\tilde{\mathbf{x}}_{t-1}^{(1:C)} \boldsymbol{\beta}_n^{(i)})}$$

$$\beta_{jk}^{(i)} \sim \mathrm{N}(0, v^2)$$

$$X_0^{(i)} \sim \mathcal{M}(N^{-1}, \dots, N^{-1})$$

5. Our main interest and scaling towards $C = 100$ — initial work

There is some interest in inference on model parameters, but our *primary* interest is actually in inferring dependence structure. e.g. in genomics data set this could infer ancestry.
$\therefore$ direct multinomial logistic regression transition model: a blocked spike-and-slab prior for Bayesian variable selection is then equivalent to inferring the hidden layer structure.

MCMC sampler

- Hidden states: conditional forward/stochastic-backward $\mathbf{X}_{1:T}^{(i)} \,|\, \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{Y}_{1:T}^{(i)}, \mathbf{X}_{1:T}^{(-i)}$ for $i \in \{1, \dots, C\}$
- Multinomial logistic parameters $\boldsymbol{\beta} \,|\, \mathbf{X}_{1:T}^{(1:C)}$
- Observation model parameters $\boldsymbol{\lambda} \,|\, \mathbf{Y}_{1:T}^{(1:C)}, \mathbf{X}_{1:T}^{(1:C)}$

Hidden states
Define conditional forward variable

$$\alpha_{tjk}^{(l)} = \mathbb{P}(y_t^{(l)}, X_{t-1}^{(l)} = j, X_t^{(l)} = k \,|\, \mathbf{y}_{1:t-1}^{(l)}, \mathbf{x}_{1:T}^{(-l)})$$

$$= \left(\sum_{i=1}^{N} \alpha_{(t-1)ij}^{(l)} \right) \frac{\exp(\tilde{\mathbf{x}}_{t-1}^{*_j} \boldsymbol{\beta}_k^{(l)})}{1 + \sum_{n=1}^{N-1} \exp(\tilde{\mathbf{x}}_{t-1}^{*_j} \boldsymbol{\beta}_n^{(l)})}$$

$$\times f_{Y_t^{(i)} \,|\, X_t^{(i)}} \left(y_t^{(i)} \,|\, k \right)$$

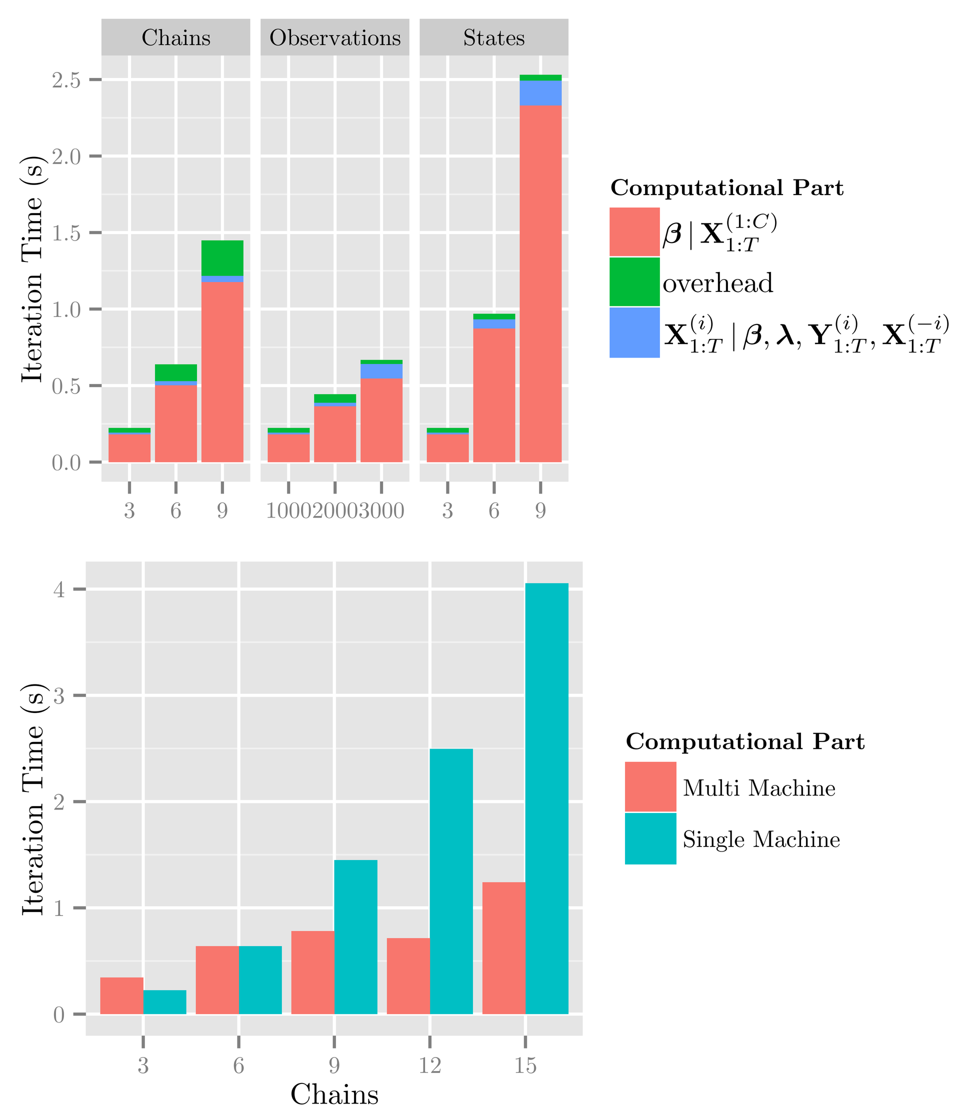
Then sample $\mathbf{X}_{1:T}^{(i)} \,|\, \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{Y}_{1:T}^{(i)}, \mathbf{X}_{1:T}^{(-i)}$ backwards, since:

$$\mathbb{P}(X_T^{(l)} = j \,|\, \mathbf{y}_{1:T}^{(l)}, \mathbf{x}_{1:T}^{(-l)}) = \sum_{i=1}^{N} \alpha_{Tij}^{(l)}$$

$$\mathbb{P}(X_{T-t}^{(l)} = i \,|\, X_{T-t+1}^{(l)} = j, \mathbf{y}_{1:T}^{(l)}, \mathbf{x}_{1:T}^{(-l)}) \propto \alpha_{(n-t+1)ij}^{(l)}$$

Logistic regression
Currently using Holmes and Held (2006).

Results



6. Current work

Probit regression
Adapting Pakman and Paninski (2013), a Hamiltonian Monte Carlo sampler for truncated multivariate Gaussian and binary distributions. Achieved substantial speedup vs author's reference C++ implementation by exploiting problem specific features.
Currently exploring GPU implementation: boundary hit times embarassingly parallel; minimum hit time a reduction operator; entire problem can propagate on GPU.

Hidden states
Also, exploring block sampling hidden states. Need to find an algorithm to partition chains in some sense 'optimally': mixing -vs- compute.

References

Choi, H., Fermin, D., Nesvizhskii, A. I., Ghosh, D. and Qin, Z. S. (2013), 'Sparsely correlated hidden Markov models with application to genome-wide location studies', *Bioinformatics* **29**(5), 533–541.
Holmes, C. C. and Held, L. (2006), 'Bayesian auxiliary variable models for binary and multinomial regression', *Bayesian Analysis* **1**(1), 145–168.
Pakman, A. and Paninski, L. (2013), 'Auxiliary-variable exact Hamiltonian Monte Carlo samplers for binary distributions', *arXiv* (1311.2166).
Saul, L. K. and Jordan, M. I. (1999), 'Mixed memory Markov models: decomposing complex stochastic processes as mixtures of simpler ones', *Machine Learning* **37**(1), 75–86.
Sherlock, C., Xifara, T., Telfer, S. and Begon, M. (2013), 'A coupled hidden Markov model for disease interactions', *Journal of the Royal Statistical Society, Series C* **62**(4), 609–627.
Zhong, S. and Ghosh, J. (2002), HMMs and coupled HMMs for multi-channel EEG classification, *in* 'Proceedings of the International Joint Conference on Neural Networks', Vol. 2, pp. 1154–1159.

Funding

This work is funded under the EPSRC i-like project.
www.i-like.org.uk