Machine Learning Methods for Neural Data Analysis

Lecture 14: Variational EM for SLDS

Scott Linderman

STATS 220/320 (NBIO220, CS339N). Winter 2021.

Announcements

- **Updated proposals** due this Friday.
  - Looking for progress on downloading and extracting the data.

Agenda

- Variational EM
- Coordinate Ascent VI
- Variational EM for SLDS

Recap

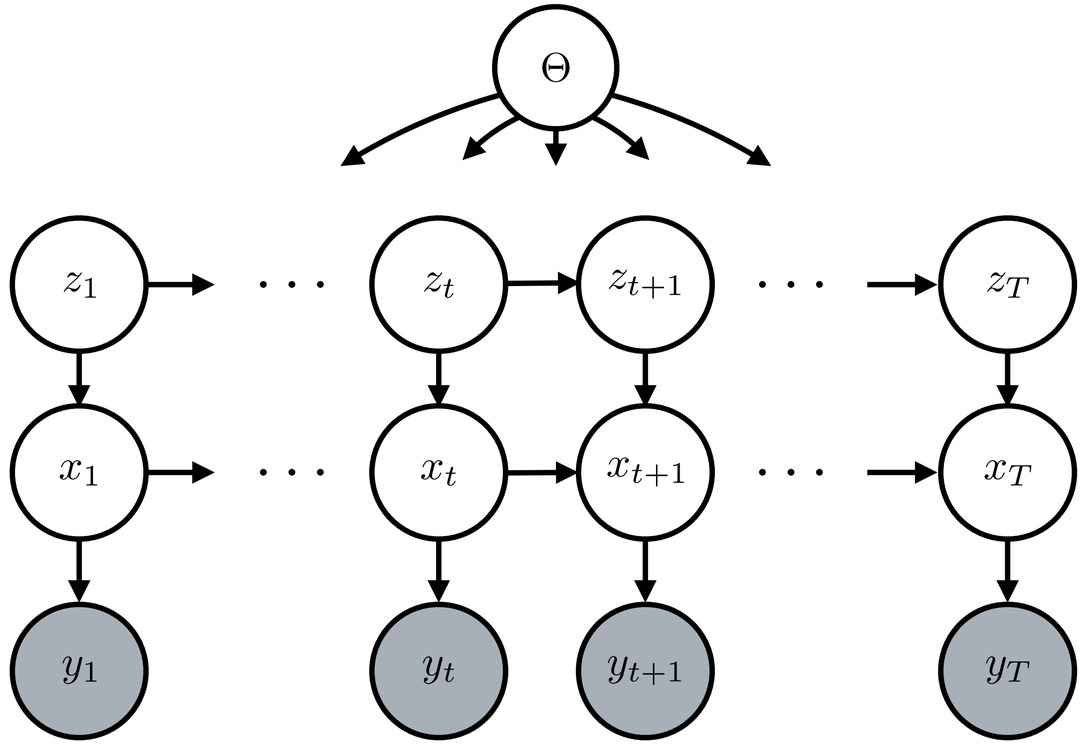
Switching Linear Dynamical Systems

Global Parameters

Discrete Latent States

Continuous Latent States

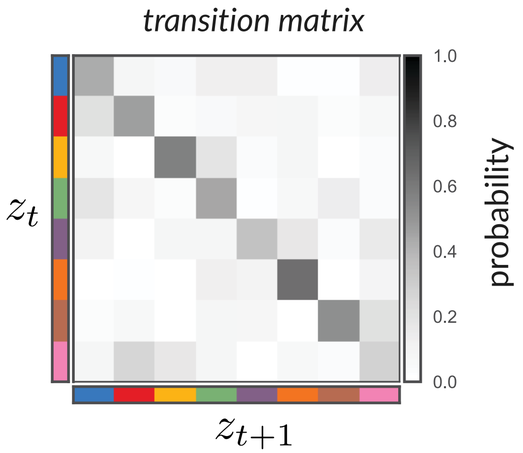
Observed Neural Activity ($\Delta F/F_0$)



= latent = observed = dependency

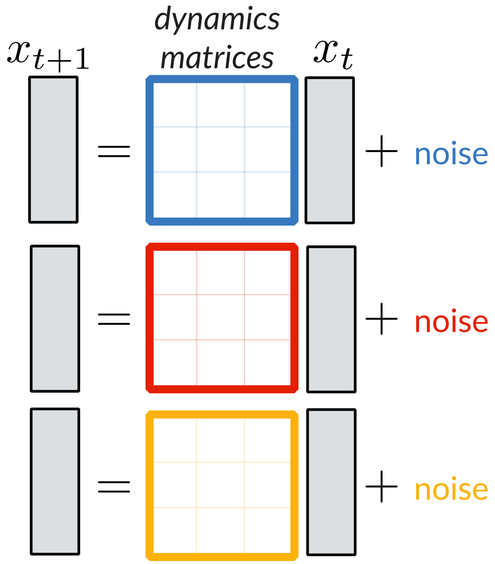
Specifying the form of the dependencies

State-dependent
switching probabilities



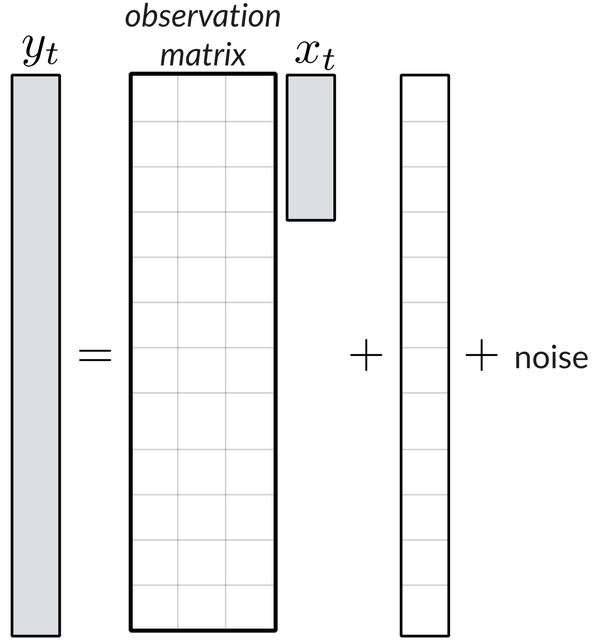
$$\Pr(z_{t+1} = j \mid z_t = i) = P_{ij}$$

Different **linear dynamics**
in each discrete state

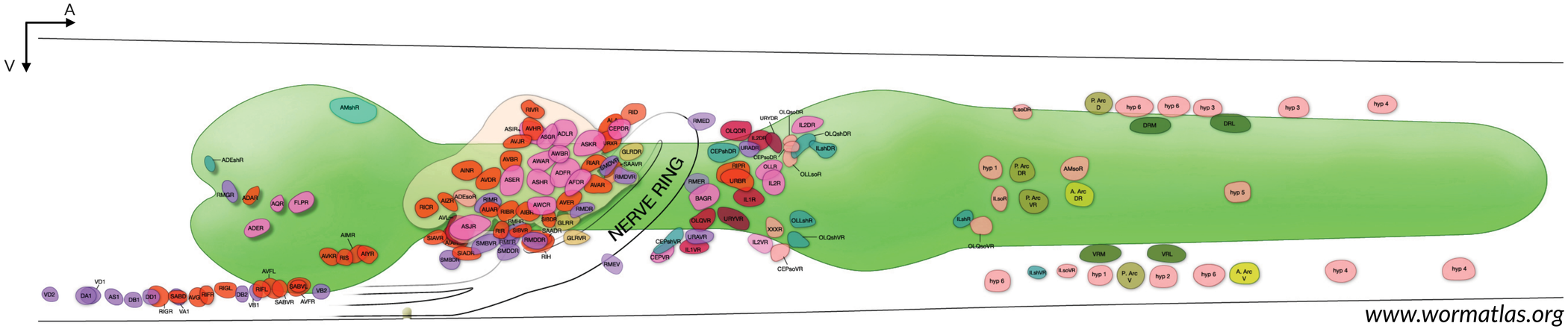


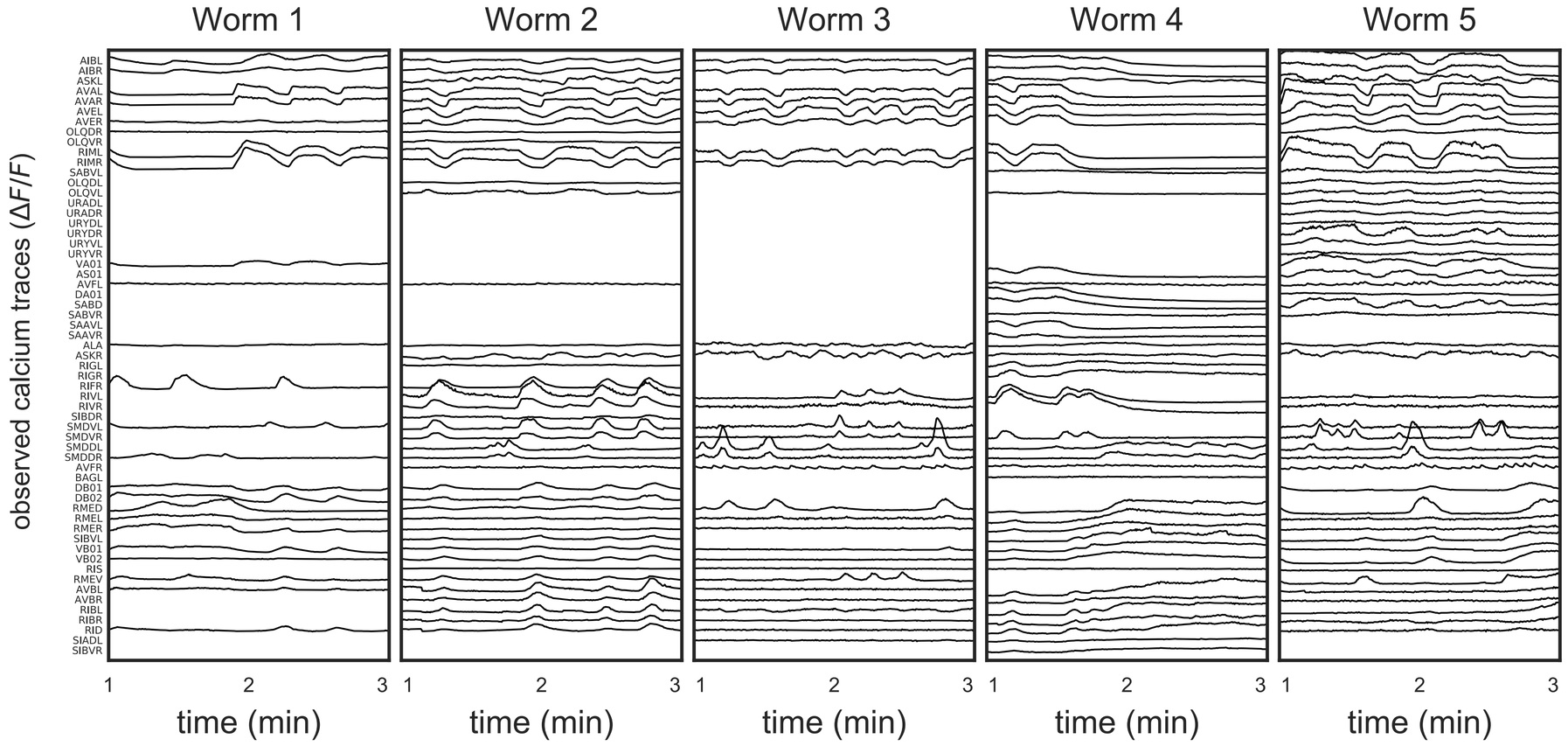
$$x_{t+1} = A_{z_{t+1}} x_t + b_{z_{t+1}} + \epsilon_{t+1}$$

Linear mapping from continuous latent
states to observed neural activity

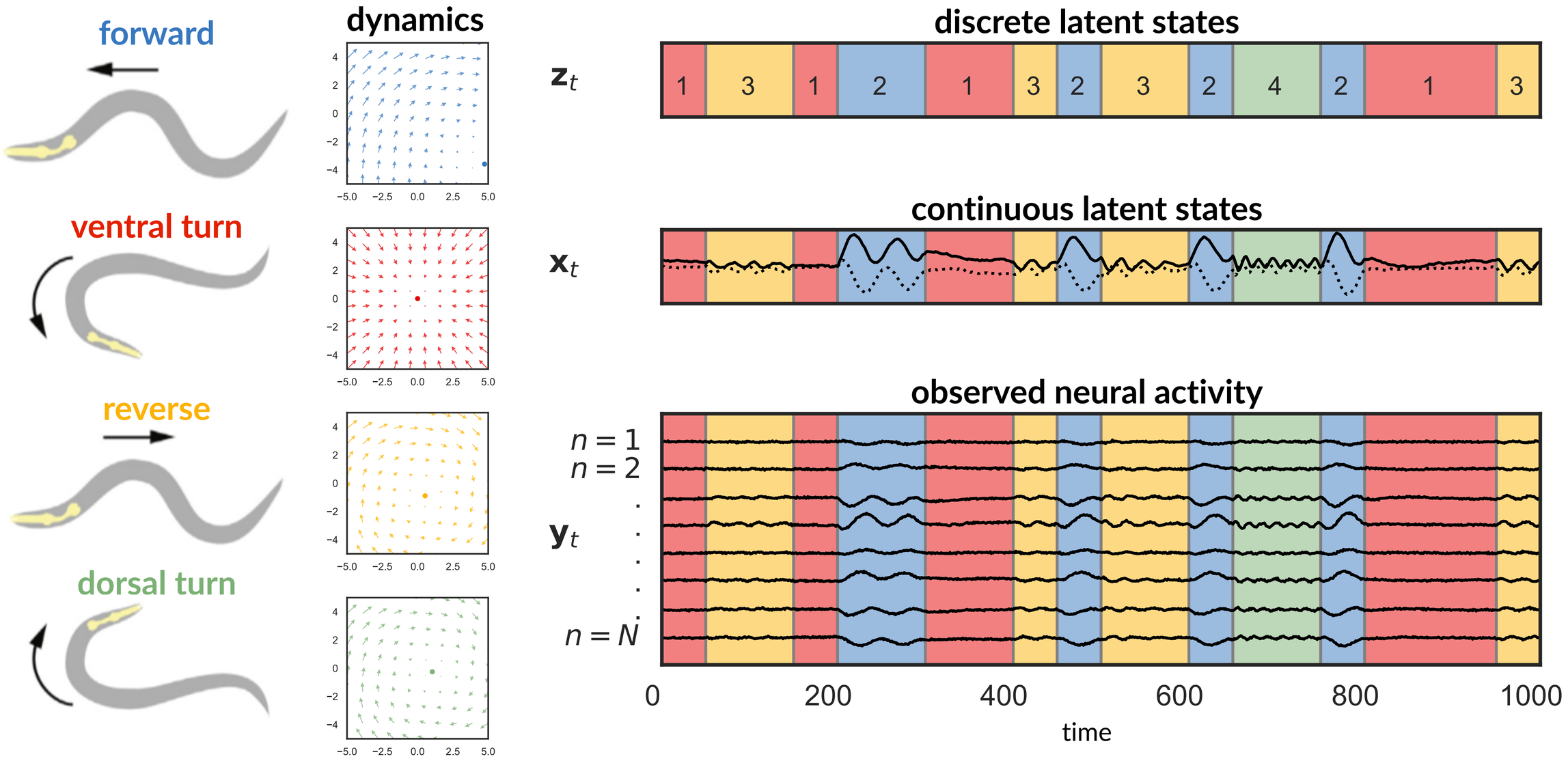


$$y_t = C x_t + d + \delta_t$$


A
V
NERVE RING
www.wormatlas.org


Worm 1
Worm 2
Worm 3
Worm 4
Worm 5
observed calcium traces (ΔF/F)
AIBL
AIBR
ASKL
AVAL
AVAR
AVEL
AVER
OLQDR
OLQVR
RIML
RIMR
SABVL
OLQDL
OLQVL
URADL
URADR
URYDL
URYDR
URYVL
URYVR
VA01
AS01
AVFL
DA01
SABD
SABVR
SAAVL
SAAVR
ALA
ASKR
RIGL
RIGR
RIFR
RIVL
RIVR
SIBDR
SMDVL
SMDVR
SMDDL
SMDDR
AVFR
BAGL
DB01
DB02
RMED
RMEL
RMER
SIBVL
VB01
VB02
RIS
RMEV
AVBL
AVBR
RIBL
RIBR
RID
SIADL
SIBVR
1
2
3
time (min)

Building a probabilistic model of neural data



Exact EM for the SLDS

- **E-step**: Update the posterior over latent variables,

$$q(z, x) \leftarrow p(z, x \mid y, \Theta) = \frac{p(z, x, y \mid \Theta)}{p(y \mid \Theta)}$$

- As before, we only need certain expectations under q,

$$\mathbb{E}_{q(z,x)}\left[\mathbb{I}[z_t = k]\right], \quad \mathbb{E}_{q(z,x)}\left[\mathbb{I}[z_t = k] x_t\right], \quad \mathbb{E}_{q(z,x)}\left[\mathbb{I}[z_t = k] x_t x_t^\top\right], \quad \mathbb{E}_{q(z,x)}\left[\mathbb{I}[z_t = k] x_t x_{t+1}^\top\right],$$

- **M-step**: Update the parameters,

$$\Theta \leftarrow \arg\max \mathbb{E}_{q(z,x)}\left[\log p(z, x, y \mid \Theta)\right]$$

- Unfortunately, computing the necessary expectations is a lot harder now!
- The posterior distribution is a mixture of Gaussians with K^T components.

Variational EM

Bayesian inference in latent variable models

Recall our derivation of the EM algorithm

- Goal: find parameters that maximize the **marginal likelihood** (aka the **model evidence**):

$$\log p(y \mid \Theta) = \log \int p(y, z \mid \Theta) \, \mathrm{d}z$$

Bayesian inference in latent variable models

Recall our derivation of the EM algorithm

- Goal: find parameters that maximize the **marginal likelihood** (aka the **model evidence**):

$$\log p(y \mid \Theta) = \log \int p(y, z \mid \Theta)\, \mathrm{d}z$$

$$= \log \int \frac{q(z)}{q(z)} p(y, z \mid \Theta)\, \mathrm{d}z \qquad \text{for any distribution } q(z)$$

Bayesian inference in latent variable models

Recall our derivation of the EM algorithm

- Goal: find parameters that maximize the **marginal likelihood** (aka the **model evidence**):

$$\begin{aligned}
\log p(y \mid \Theta) &= \log \int p(y, z \mid \Theta)\, \mathrm{d}z \\
&= \log \int \frac{q(z)}{q(z)} p(y, z \mid \Theta)\, \mathrm{d}z \qquad \text{for any distribution } q(z) \\
&= \log \mathbb{E}_{q(z)} \left[\frac{p(y, z \mid \Theta)}{q(z)} \right]
\end{aligned}$$

Bayesian inference in latent variable models

Recall our derivation of the EM algorithm

- Goal: find parameters that maximize the **marginal likelihood** (aka the **model evidence**):

$$
\begin{aligned}
\log p(y \mid \Theta) &= \log \int p(y, z \mid \Theta)\, \mathrm{d}z \\
&= \log \int \frac{q(z)}{q(z)} p(y, z \mid \Theta)\, \mathrm{d}z && \text{for any distribution } q(z) \\
&= \log \mathbb{E}_{q(z)} \left[\frac{p(y, z \mid \Theta)}{q(z)} \right] \\
&\geq \mathbb{E}_{q(z)} \left[\log p(y, z \mid \Theta) - \log q(z) \right] && \text{by Jensen's inequality}
\end{aligned}
$$

Bayesian inference in latent variable models

Recall our derivation of the EM algorithm

- Goal: find parameters that maximize the **marginal likelihood** (aka the **model evidence**):

$$
\begin{aligned}
\log p(y \mid \Theta) &= \log \int p(y, z \mid \Theta)\, \mathrm{d}z \\
&= \log \int \frac{q(z)}{q(z)} p(y, z \mid \Theta)\, \mathrm{d}z && \text{for any distribution } q(z) \\
&= \log \mathbb{E}_{q(z)} \left[\frac{p(y, z \mid \Theta)}{q(z)} \right] \\
&\geq \mathbb{E}_{q(z)} \left[\log p(y, z \mid \Theta) - \log q(z) \right] && \text{by Jensen's inequality} \\
&\triangleq \mathscr{L}[q, \Theta]
\end{aligned}
$$

- $\mathscr{L}$ is called the **evidence lower bound** or the **ELBO** for short.

Bayesian inference in latent variable models

Coordinate ascent on the ELBO

- **E Step:** Update the posterior distribution on latent variables,

$$q \leftarrow \arg\max_q \mathscr{L}[q, \Theta]$$

Bayesian inference in latent variable models

Coordinate ascent on the ELBO

- **E Step:** Update the posterior distribution on latent variables,

$$
\begin{aligned}
q &\leftarrow \arg\max_q \mathscr{L}[q, \Theta] \\
&= \arg\max_q \mathbb{E}_{q(z)} \left[\log \frac{p(y, z \mid \Theta)}{q(z)} \right]
\end{aligned}
$$

Bayesian inference in latent variable models

Coordinate ascent on the ELBO

- **E Step:** Update the posterior distribution on latent variables,

$$
\begin{aligned}
q &\leftarrow \arg\max_q \mathscr{L}[q, \Theta] \\
&= \arg\max_q \mathbb{E}_{q(z)} \left[\log \frac{p(y, z \mid \Theta)}{q(z)} \right] \\
&= \arg\max_q \mathbb{E}_{q(z)} \left[\log \frac{p(z \mid y \mid \Theta)}{q(z)} \right] + \log p(y \mid \Theta)
\end{aligned}
$$

Bayesian inference in latent variable models

Coordinate ascent on the ELBO

- **E Step:** Update the posterior distribution on latent variables,

$$
\begin{aligned}
q &\leftarrow \arg\max_q \mathscr{L}[q, \Theta] \\
&= \arg\max_q \mathbb{E}_{q(z)}\left[\log \frac{p(y, z \mid \Theta)}{q(z)}\right] \\
&= \arg\max_q \mathbb{E}_{q(z)}\left[\log \frac{p(z \mid y \mid \Theta)}{q(z)}\right] + \log p(y \mid \Theta) \\
&= \arg\max_q \; -\mathrm{KL}\left(q(z) \,\|\, p(z \mid y, \Theta)\right) + \log p(y \mid \Theta)
\end{aligned}
$$

Bayesian inference in latent variable models

Coordinate ascent on the ELBO

- **E Step:** Update the posterior distribution on latent variables,

$$
\begin{aligned}
q &\leftarrow \arg\max_q \mathscr{L}[q, \Theta] \\
&= \arg\max_q \mathbb{E}_{q(z)}\left[\log \frac{p(y, z \mid \Theta)}{q(z)}\right] \\
&= \arg\max_q \mathbb{E}_{q(z)}\left[\log \frac{p(z \mid y \mid \Theta)}{q(z)}\right] + \log p(y \mid \Theta) \\
&= \arg\max_q \; -\mathrm{KL}\left(q(z) \,\|\, p(z \mid y, \Theta)\right) + \log p(y \mid \Theta) \\
&= \arg\min_q \mathrm{KL}\left(q(z) \,\|\, p(z \mid y, \Theta)\right)
\end{aligned}
$$

Bayesian inference in latent variable models

Coordinate ascent on the ELBO

- **E Step:** Update the posterior distribution on latent variables,

$$
\begin{aligned}
q &\leftarrow \arg\max_q \mathscr{L}[q, \Theta] \\
&= \arg\max_q \mathbb{E}_{q(z)} \left[\log \frac{p(y, z \mid \Theta)}{q(z)} \right] \\
&= \arg\max_q \mathbb{E}_{q(z)} \left[\log \frac{p(z \mid y \mid \Theta)}{q(z)} \right] + \log p(y \mid \Theta) \\
&= \arg\max_q \; - \mathrm{KL}\left(q(z) \,\|\, p(z \mid y, \Theta) \right) + \log p(y \mid \Theta) \\
&= \arg\min_q \mathrm{KL}\left(q(z) \,\|\, p(z \mid y, \Theta) \right)
\end{aligned}
$$

- **Maximizing the ELBO** w.r.t. q is equivalent to **minimizing the Kullback-Leibler (KL) divergence**.
- The KL divergence is **non-negative, and it equals zero iff** $q(z) \equiv p(z \mid y, \Theta)$.

Bayesian inference in latent variable models

The Expectation-Maximization (EM) algorithm

- **M-step**: Maximize the expected log probability

$$\Theta \leftarrow = \arg\max_{\Theta} \mathbb{E}_{q(z)}[\log p(y, z, \Theta)]$$

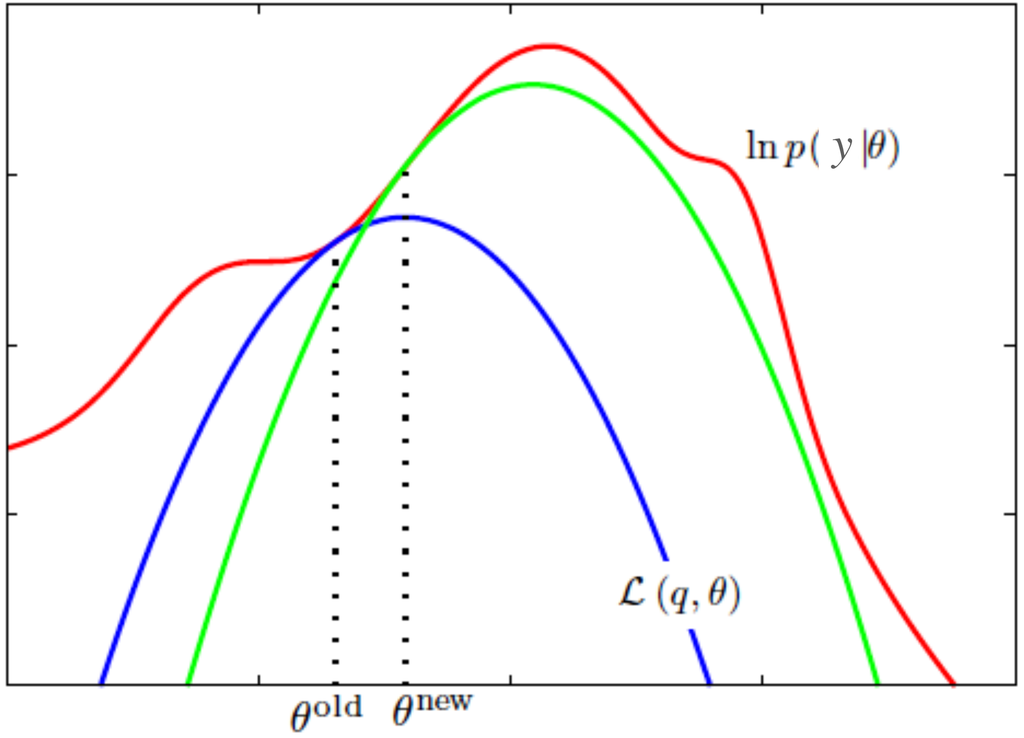
- **E-step**: Update the posterior over latent variables

$$\begin{aligned} q &\leftarrow \arg\max_q \mathscr{L}[q, \Theta] \\ &= \arg\min_q \mathrm{KL}\left(q(z) \,\|\, p(z \mid y, \Theta)\right) \\ &= p(z \mid y, \Theta) \end{aligned}$$

- After each E-step, the **ELBO is tight**:

$$\mathscr{L}[p(z \mid y, \Theta), \Theta] = \log p(y \mid \Theta)$$

- EM converges to **local optima** of the marginal distribution.



Bishop (2006). Pattern Recognition and Machine Learning, Ch 9.4.

Bayesian inference in latent variable models

Variational Expectation-Maximization

- **M-step**: Maximize the expected log probability

$$\Theta \leftarrow = \arg\max_{\Theta} \mathbb{E}_{q(z)}[\log p(y, z, \Theta)]$$

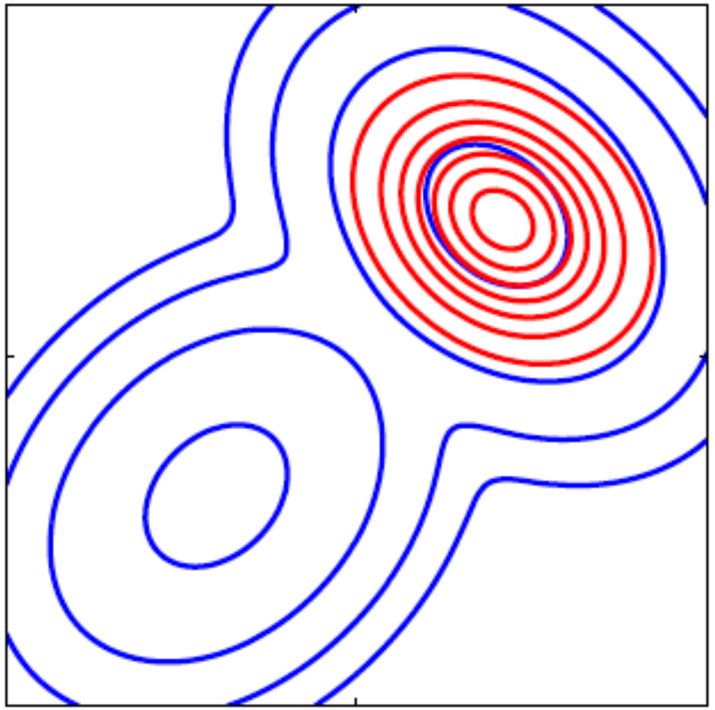
- **Variational E-step**: Update the posterior, subject to $q \in Q$

$$q \leftarrow \arg\max_{q \in \mathbb{Q}} \mathscr{L}[q, \Theta]$$

$$= \arg\min_{q \in \mathbb{Q}} \mathrm{KL}\left(q(z) \,\|\, p(z \mid y, \Theta)\right)$$

  where $\mathbb{Q}$ is a set of **tractable approximate posteriors.**

- For example, $\mathbb{Q}$ could **assume independence** or a **particular functional form.**
- If $\mathbb{Q}$ does not contain the true posterior, the ELBO will be a strict lower bound on the marginal likelihood, $\mathscr{L}[q(z), \Theta] < \log p(y \mid \Theta)$.
- Optimizing over Q to find the best approximation is called **variational inference**, hence the name Variational EM.



Bishop (2006). Pattern Recognition and Machine Learning, Ch 10.1

Coordinate Ascent Variational Inference (CAVI)

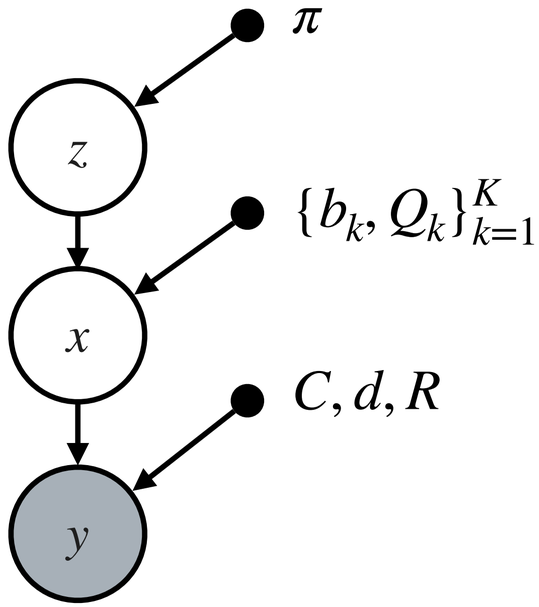
Coordinate ascent VI

Warm-up example

$$z \sim \text{Cat}(\pi)$$

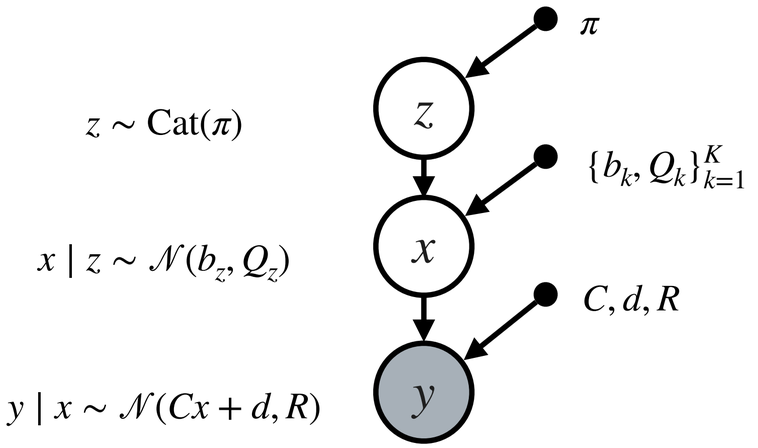
$$x \mid z \sim \mathcal{N}(b_z, Q_z)$$

$$y \mid x \sim \mathcal{N}(Cx + d, R)$$



Coordinate ascent VI

Mean field variational family



- **Assume** $\mathcal{Q}$ is the set of "factored" distributions

$$q(z, x) = q(z)\, q(x)$$

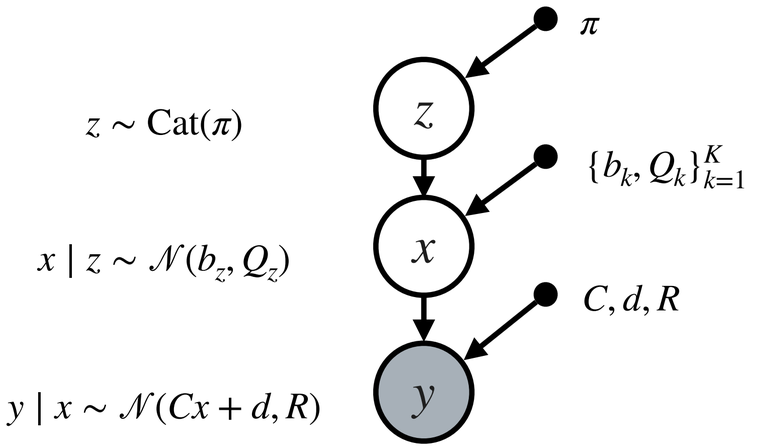
where z and x are independent.

- This is called the **mean field family**.
- We want to find,

$$\arg\max_{q \in \mathcal{Q}} \mathscr{L}[q(z)q(x), \Theta] \equiv \arg\min_{q \in \mathcal{Q}} \text{KL}\left(q(z)q(x) \,\|\, p(z, x \mid y, \Theta)\right)$$

Coordinate ascent VI

Coordinate updates

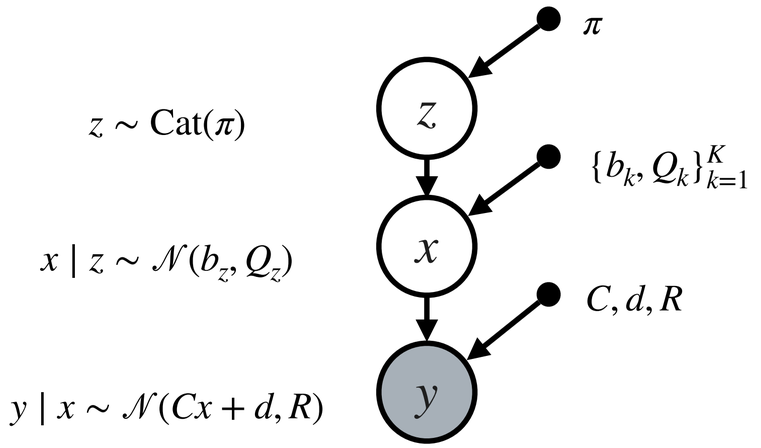


Hold $q(x)$ fixed and optimize w.r.t. $q(z)$:

$$\mathscr{L}[q(z)q(x), \Theta] = \mathbb{E}_{q(z)q(x)}\left[\log p(z, x, y \mid \Theta) - \log q(z) - \log q(x)\right]$$

Coordinate ascent VI

Coordinate updates



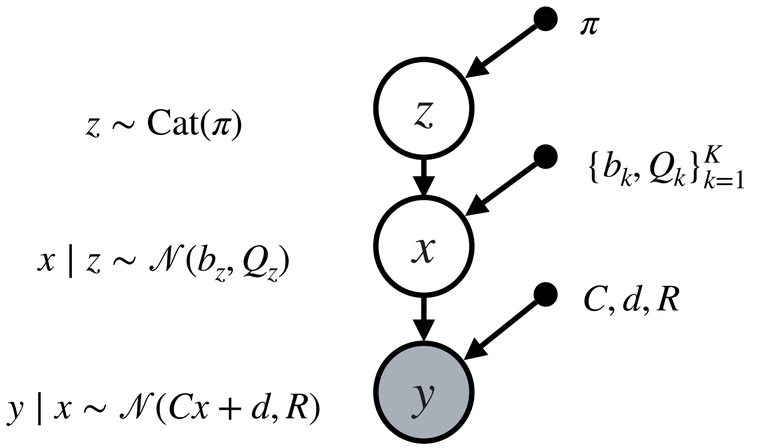
Hold $q(x)$ fixed and optimize w.r.t. $q(z)$:

$$\mathscr{L}[q(z)q(x), \Theta] = \mathbb{E}_{q(z)q(x)}\left[\log p(z, x, y \mid \Theta) - \log q(z) - \log q(x)\right]$$

$$= \mathbb{E}_{q(z)}\left[\mathbb{E}_{q(x)}\left[\log p(z, x, y \mid \Theta)\right] - \log q(z)\right] + \text{c}$$

Coordinate ascent VI

Coordinate updates

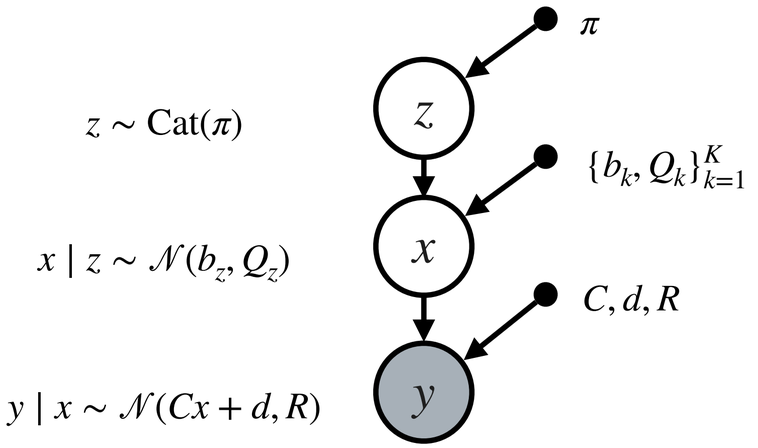


Hold $q(x)$ fixed and optimize w.r.t. $q(z)$:

$$\begin{aligned}
\mathscr{L}[q(z)q(x), \Theta] &= \mathbb{E}_{q(z)q(x)}\left[\log p(z, x, y \mid \Theta) - \log q(z) - \log q(x)\right] \\
&= \mathbb{E}_{q(z)}\left[\mathbb{E}_{q(x)}\left[\log p(z, x, y \mid \Theta)\right] - \log q(z)\right] + \text{c} \\
&= \mathbb{E}_{q(z)}\left[\log \exp\left\{\mathbb{E}_{q(x)}\left[\log p(z, x, y \mid \Theta)\right]\right\} - \log q(z)\right] + \text{c}
\end{aligned}$$

Coordinate ascent VI

Coordinate updates

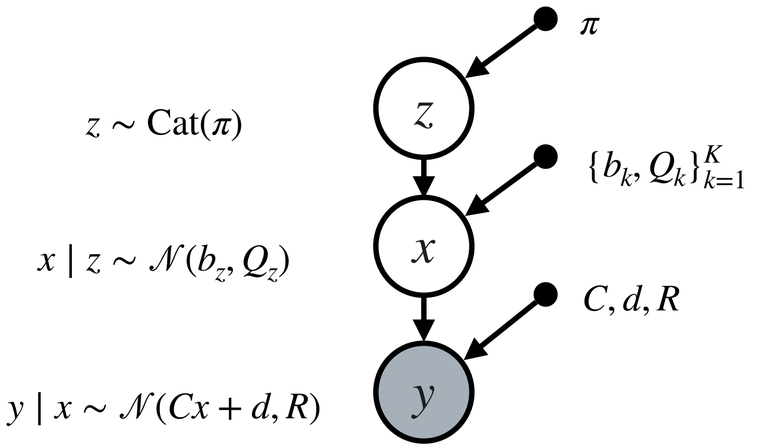


Hold $q(x)$ fixed and optimize w.r.t. $q(z)$:

$$
\begin{aligned}
\mathscr{L}[q(z)q(x), \Theta] &= \mathbb{E}_{q(z)q(x)} \left[\log p(z, x, y \mid \Theta) - \log q(z) - \log q(x) \right] \\
&= \mathbb{E}_{q(z)} \left[\mathbb{E}_{q(x)} \left[\log p(z, x, y \mid \Theta) \right] - \log q(z) \right] + \text{c} \\
&= \mathbb{E}_{q(z)} \left[\log \exp \left\{ \mathbb{E}_{q(x)} \left[\log p(z, x, y \mid \Theta) \right] \right\} - \log q(z) \right] + \text{c} \\
&= \mathbb{E}_{q(z)} \left[\log \tilde{p}(z) - \log q(z) \right] + \text{c}'
\end{aligned}
$$

Coordinate ascent VI

Coordinate updates

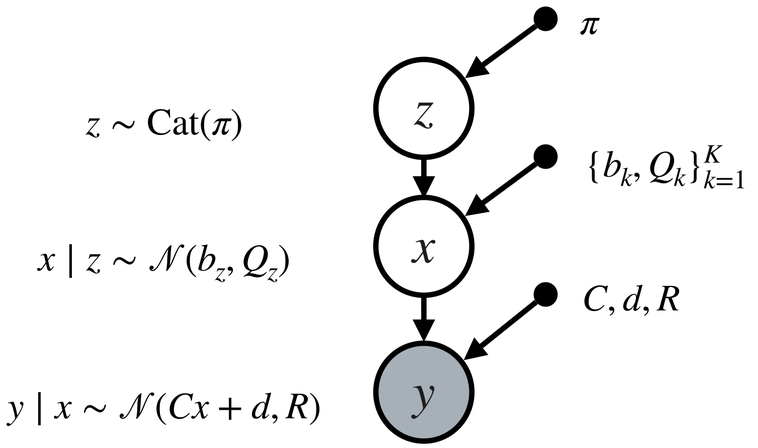


Hold $q(x)$ fixed and optimize w.r.t. $q(z)$:

$$
\begin{aligned}
\mathscr{L}[q(z)q(x), \Theta] &= \mathbb{E}_{q(z)q(x)} \left[\log p(z, x, y \mid \Theta) - \log q(z) - \log q(x) \right] \\
&= \mathbb{E}_{q(z)} \left[\mathbb{E}_{q(x)} \left[\log p(z, x, y \mid \Theta) \right] - \log q(z) \right] + \text{c} \\
&= \mathbb{E}_{q(z)} \left[\log \exp \left\{ \mathbb{E}_{q(x)} \left[\log p(z, x, y \mid \Theta) \right] \right\} - \log q(z) \right] + \text{c} \\
&= \mathbb{E}_{q(z)} \left[\log \tilde{p}(z) - \log q(z) \right] + \text{c}' \\
&= - \text{KL} \left(q(z) \,\|\, \tilde{p}(z) \right] + \text{c}'
\end{aligned}
$$

Coordinate ascent VI

General form



- Thus, as a function of $q(z)$:

$$\mathscr{L}[q(z)q(x), \Theta] = -\,\text{KL}\left(q(z) \,\|\, \tilde{p}(z)\right] + \text{c}'$$

- This is minimized when

$$q(z) = \tilde{p}(z) \propto \exp\left\{ \mathbb{E}_{q(x)}\left[\log p(z, x, y \mid \Theta)\right] \right\}$$

- By symmetry, the optimal update for $q(x)$ is

$$q(x) = \tilde{p}(x) \propto \exp\left\{ \mathbb{E}_{q(z)}\left[\log p(z, x, y \mid \Theta)\right] \right\}$$

- In general, coordinate ascent for mean field variational families takes the form,

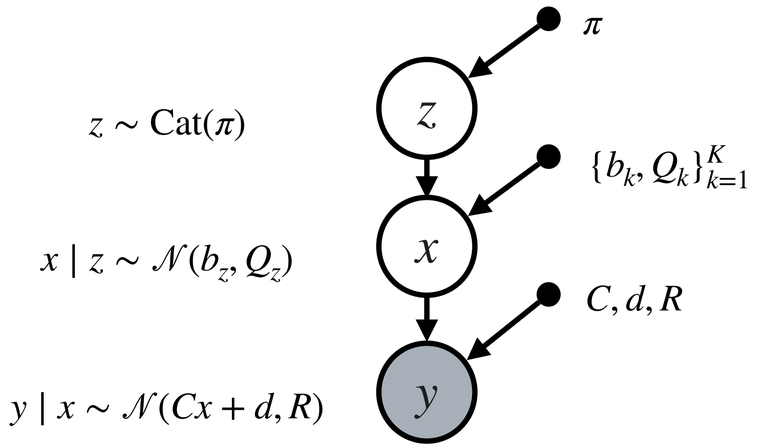
$$q(x_i) \propto \exp\left\{ \mathbb{E}_{q(x_{\neg i})}\left[\log p(x_1, \ldots, x_i, \ldots, x_D, y \mid \Theta)\right] \right\}$$

Coordinate ascent VI

Closed form updates for $q(z)$

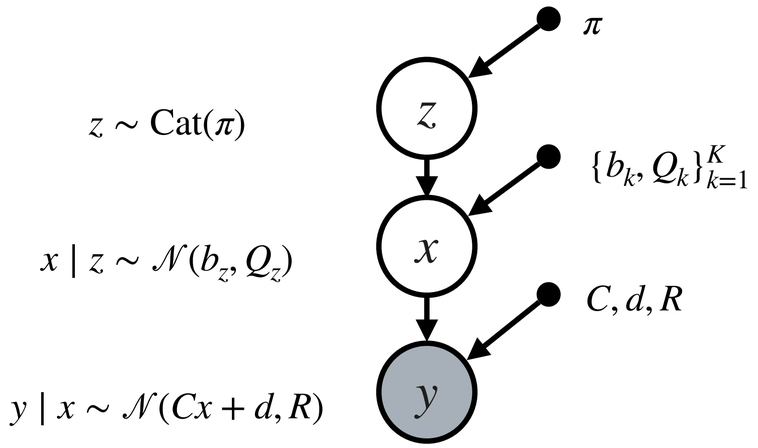
The optimal updates are often available in closed form,

$$\log q(z) = \mathbb{E}_{q(x)}\left[\log p(z, x, y \mid \Theta)\right] + \text{c}$$



Coordinate ascent VI

Closed form updates for $q(z)$

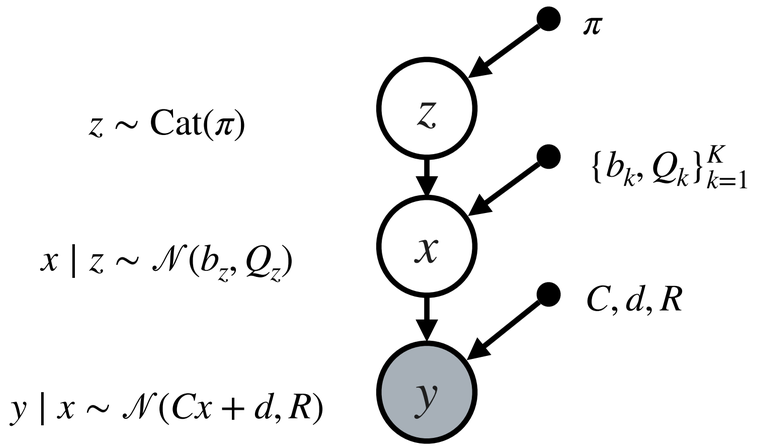


The optimal updates are often available in closed form,

$$\begin{aligned}\log q(z) &= \mathbb{E}_{q(x)}\left[\log p(z, x, y \mid \Theta)\right] + \text{c} \\ &= \mathbb{E}_{q(x)}\left[\log p(z \mid \Theta) + \log p(x \mid z, \Theta) + \log p(y \mid x, \Theta)\right] + \text{c}\end{aligned}$$

Coordinate ascent VI

Closed form updates for $q(z)$

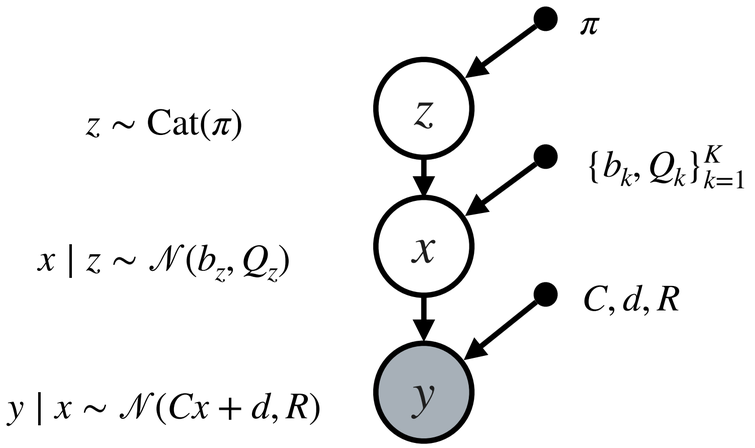


The optimal updates are often available in closed form,

$$\begin{aligned}
\log q(z) &= \mathbb{E}_{q(x)}\left[\log p(z, x, y \mid \Theta)\right] + \mathrm{c} \\
&= \mathbb{E}_{q(x)}\left[\log p(z \mid \Theta) + \log p(x \mid z, \Theta) + \log p(y \mid x, \Theta)\right] + \mathrm{c} \\
&= \log \mathrm{Cat}(z \mid \pi) + \mathbb{E}_{q(x)}\left[\log \mathcal{N}(x \mid b_z, Q_z)\right] + \mathrm{c}
\end{aligned}$$

Coordinate ascent VI

Closed form updates for $q(z)$

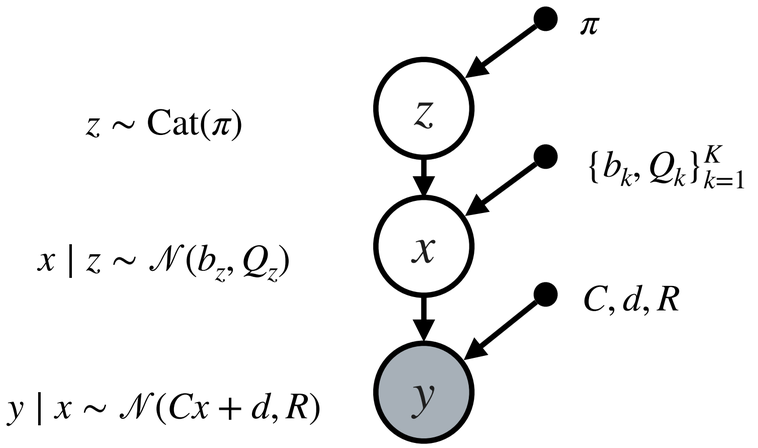


The optimal updates are often available in closed form,

$$
\begin{aligned}
\log q(z) &= \mathbb{E}_{q(x)}\left[\log p(z, x, y \mid \Theta)\right] + \mathrm{c} \\
&= \mathbb{E}_{q(x)}\left[\log p(z \mid \Theta) + \log p(x \mid z, \Theta) + \log p(y \mid x, \Theta)\right] + \mathrm{c} \\
&= \log \mathrm{Cat}(z \mid \pi) + \mathbb{E}_{q(x)}\left[\log \mathcal{N}(x \mid b_z, Q_z)\right] + \mathrm{c} \\
&= \sum_{k=1}^{K} \mathbb{I}[z = k]\left(\log \pi_k + \mathbb{E}_{q(x)}\left[\log \mathcal{N}(x \mid b_k, Q_k)\right]\right) + \mathrm{c}
\end{aligned}
$$

Coordinate ascent VI

Closed form updates for $q(z)$



The optimal updates are often available in closed form,

$$\begin{aligned}
\log q(z) &= \mathbb{E}_{q(x)}\left[\log p(z, x, y \mid \Theta)\right] + \text{c} \\
&= \mathbb{E}_{q(x)}\left[\log p(z \mid \Theta) + \log p(x \mid z, \Theta) + \log p(y \mid x, \Theta)\right] + \text{c} \\
&= \log \text{Cat}(z \mid \pi) + \mathbb{E}_{q(x)}\left[\log \mathcal{N}(x \mid b_z, Q_z)\right] + \text{c} \\
&= \sum_{k=1}^{K} \mathbb{I}[z = k]\left(\log \pi_k + \mathbb{E}_{q(x)}\left[\log \mathcal{N}(x \mid b_k, Q_k)\right]\right) + \text{c} \\
&= \log \text{Cat}(z \mid \tilde{\pi})
\end{aligned}$$

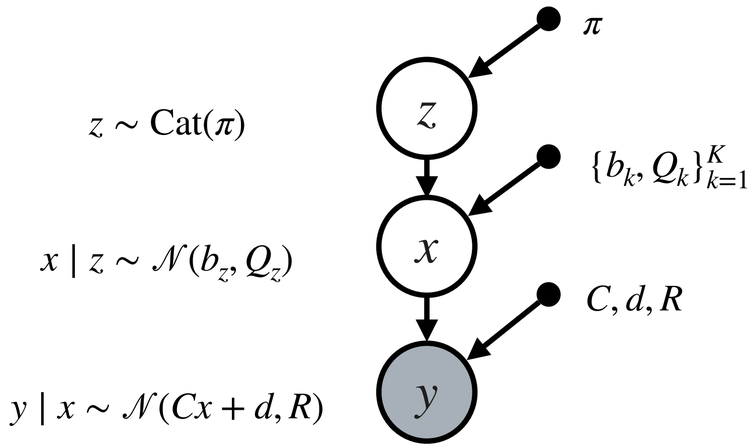
where

$$\log \tilde{\pi}_k = \log \pi_k + \underbrace{\mathbb{E}_{q(x)}\left[\log \mathcal{N}(x \mid b_k, Q_k)\right]}_{\text{"expected log likelihood"}} + c$$

The **expected log likelihood** is also called the **cross entropy** between $q(x)$ and $p(x \mid z)$.

Coordinate ascent VI

Closed form updates for $q(x)$

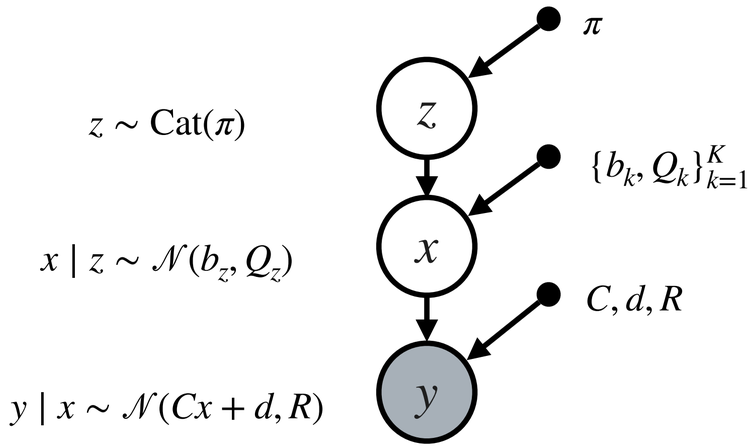


Now do the same for $q(x)$:

$$\log q(x) = \mathbb{E}_{q(z)} \left[\log p(z, x, y \mid \Theta) \right] + \text{c}$$

Coordinate ascent VI

Closed form updates for $q(x)$

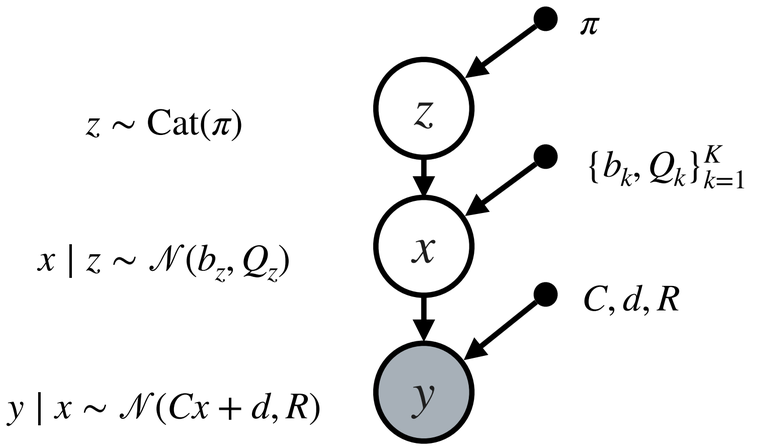


Now do the same for $q(x)$:

$$\log q(x) = \mathbb{E}_{q(z)}\left[\log p(z, x, y \mid \Theta)\right] + \text{c}$$
$$= \mathbb{E}_{q(z)}\left[\log p(x \mid z, \Theta)\right] + \log p(y \mid x, \Theta) + \text{c}$$

Coordinate ascent VI

Closed form updates for $q(x)$

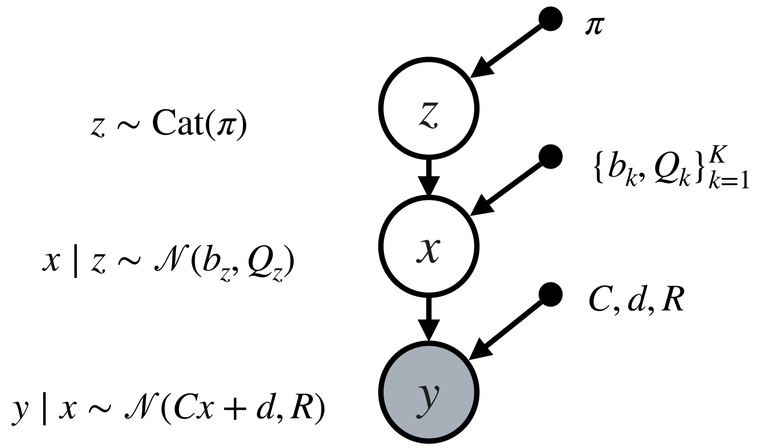


Now do the same for $q(x)$:

$$\begin{aligned}
\log q(x) &= \mathbb{E}_{q(z)}\left[\log p(z, x, y \mid \Theta)\right] + \text{c} \\
&= \mathbb{E}_{q(z)}\left[\log p(x \mid z, \Theta)\right] + \log p(y \mid x, \Theta) + \text{c} \\
&= \mathbb{E}_{q(z)}\left[\log \mathcal{N}(x \mid b_z, Q_z)\right] + \log p(y \mid x, \Theta) + \text{c}
\end{aligned}$$

Coordinate ascent VI

Closed form updates for $q(x)$

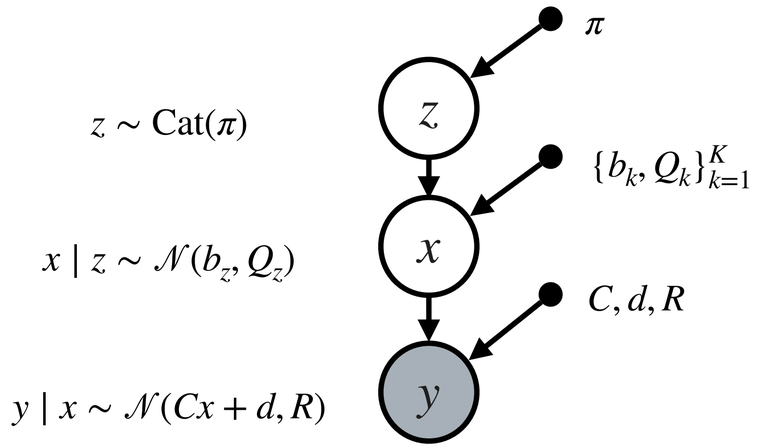


Now do the same for $q(x)$:

$$
\begin{aligned}
\log q(x) &= \mathbb{E}_{q(z)}\left[\log p(z, x, y \mid \Theta)\right] + \text{c} \\
&= \mathbb{E}_{q(z)}\left[\log p(x \mid z, \Theta)\right] + \log p(y \mid x, \Theta) + \text{c} \\
&= \mathbb{E}_{q(z)}\left[\log \mathcal{N}(x \mid b_z, Q_z)\right] + \log p(y \mid x, \Theta) + \text{c} \\
&= -\frac{1}{2} x^\top \mathbb{E}_{q(z)}[Q_z^{-1}] x + x^\top \mathbb{E}_{q(z)}[Q_z^{-1} b_z] - \frac{1}{2} x^\top C^\top R^{-1} C x + x^\top C^\top R^{-1}(y - d) + \text{c}
\end{aligned}
$$

Coordinate ascent VI

Closed form updates for $q(x)$

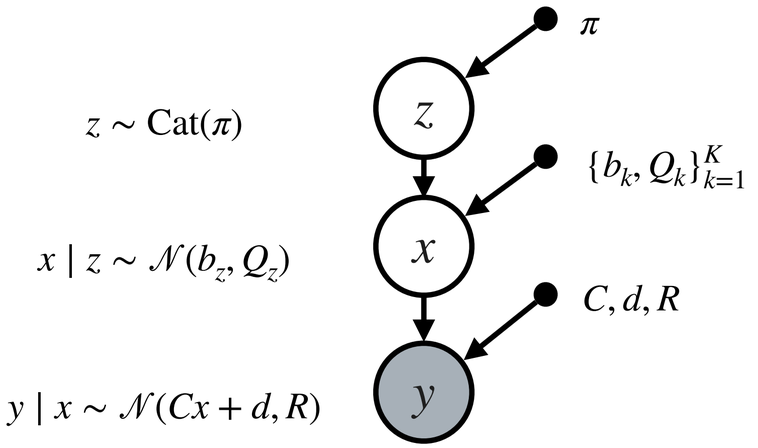


Now do the same for $q(x)$:

$$
\begin{aligned}
\log q(x) &= \mathbb{E}_{q(z)}\left[\log p(z, x, y \mid \Theta)\right] + \text{c} \\
&= \mathbb{E}_{q(z)}\left[\log p(x \mid z, \Theta)\right] + \log p(y \mid x, \Theta) + \text{c} \\
&= \mathbb{E}_{q(z)}\left[\log \mathcal{N}(x \mid b_z, Q_z)\right] + \log p(y \mid x, \Theta) + \text{c} \\
&= -\frac{1}{2} x^\top \mathbb{E}_{q(z)}[Q_z^{-1}] x + x^\top \mathbb{E}_{q(z)}[Q_z^{-1} b_z] - \frac{1}{2} x^\top C^\top R^{-1} C x + x^\top C^\top R^{-1}(y - d) + \text{c} \\
&= \log \mathcal{N}(x \mid \tilde{\mu}, \tilde{\Sigma})
\end{aligned}
$$

Coordinate ascent VI

Closed form updates for $q(x)$



The final result is,

$$\log q(x) = \log \mathcal{N}(x \mid \tilde{\mu}, \tilde{\Sigma})$$

where

$$\tilde{\mu} = \tilde{J}^{-1}\tilde{h} \qquad \tilde{\Sigma} = \tilde{J}^{-1}$$

$$\tilde{h} = \mathbb{E}_{q(z)}[Q_z^{-1}b_z] + C^\top R^{-1}(y-d) \qquad \tilde{J} = \mathbb{E}_{q(z)}[Q_z^{-1}] + C^\top R^{-1}C$$

$$= \sum_{k=1}^K \left[q(z=k)Q_k^{-1}b_k\right] + C^\top R^{-1}(y-d) \qquad = \sum_{k=1}^K \left[q(z=k)Q_k^{-1}\right] + C^\top R^{-1}C$$

In other words, the natural parameters are the **expected natural parameters** under $q(z)$.

Variational EM for SLDS

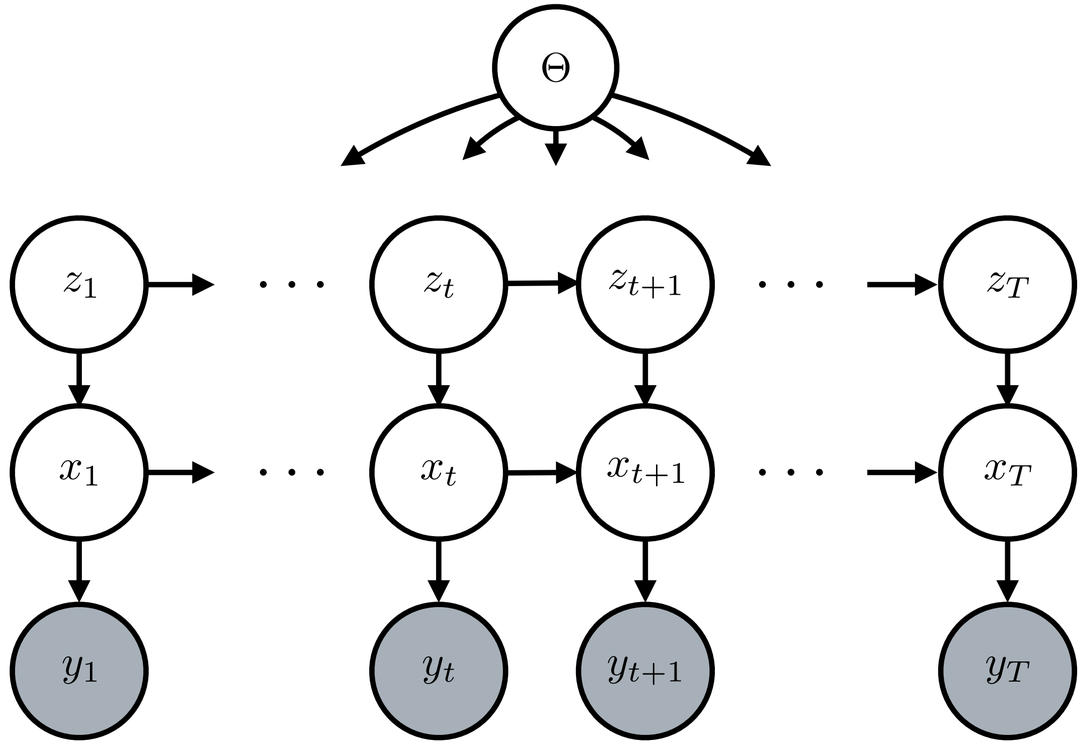
Variational EM for SLDS

Global Parameters

Discrete Latent States

Continuous Latent States

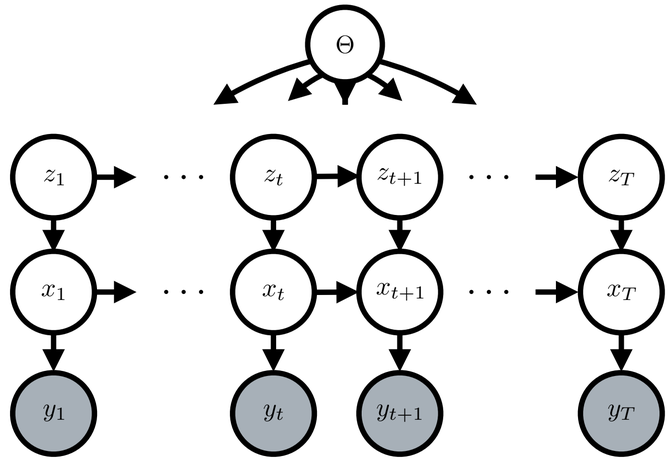
Observed Neural Activity ($\Delta F/F_0$)



= latent = observed = dependency

Variational EM for SLDS

Structured mean field variational family



- Assume the variational posterior factors over discrete and continuous states.

$$q(z_{1:T}, x_{1:T}) = q(z_{1:T})\, q(x_{1:T})$$

where $z_{1:T}$ and $x_{1:T}$ are independent.

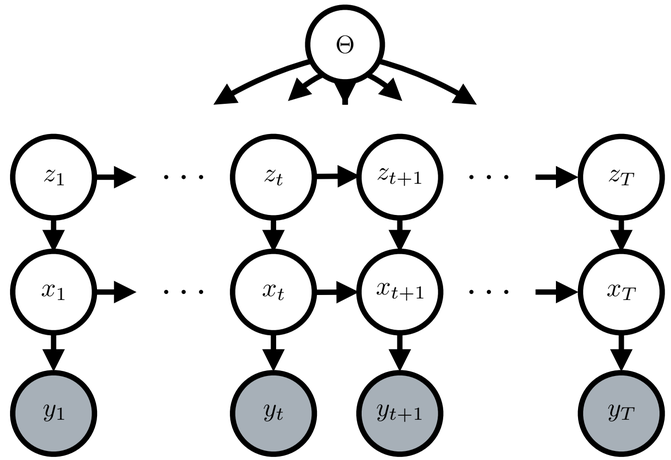
- There can still be dependencies within each factor though!
- This is called a **structured mean field family**.
- We want to find,

$$\arg\max_{q \in \mathcal{Q}} \mathscr{L}[q(z_{1:T})q(x_{1:T}), \Theta]$$

$$\equiv \arg\min_{q \in \mathcal{Q}} \mathrm{KL}\left(q(z_{1:T})q(x_{1:T}) \,\|\, p(z_{1:T}, x_{1:T} \mid y_{1:T}, \Theta)\right)$$

Variational EM for SLDS

Updating the discrete state posterior $q(z_{1:T})$



The optimal update for the discrete states takes the form

$$\log q(z_{1:T}) = \log \mathrm{Cat}(z_1 \mid \pi) + \sum_{t=2}^{T} \log \mathrm{Cat}(z_t \mid P_{z_{t-1}})$$

$$+ \sum_{t=1}^{T} \sum_{k=1}^{K} \mathbb{I}[z_t = k] \log \tilde{\ell}_{tk} + \mathrm{c}$$

where

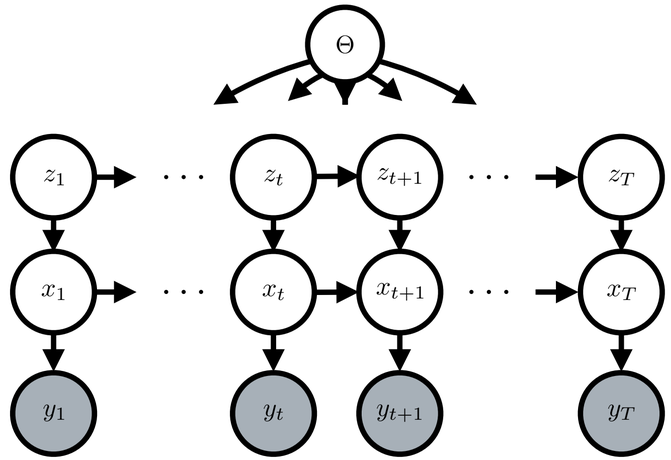
$$\log \tilde{\ell}_{tk} = \mathbb{E}_{q(x)} \left[\log \mathcal{N}(x_t \mid A_k x_{t-1} + b_k, Q_k) \right]$$

This is the **same form as the posterior in a hidden Markov model!**

But here, the log likelihoods are replaced with **expected log likelihoods** under $q(x)$.

Variational EM for SLDS

Updating the continuous state posterior $q(x_{1:T})$



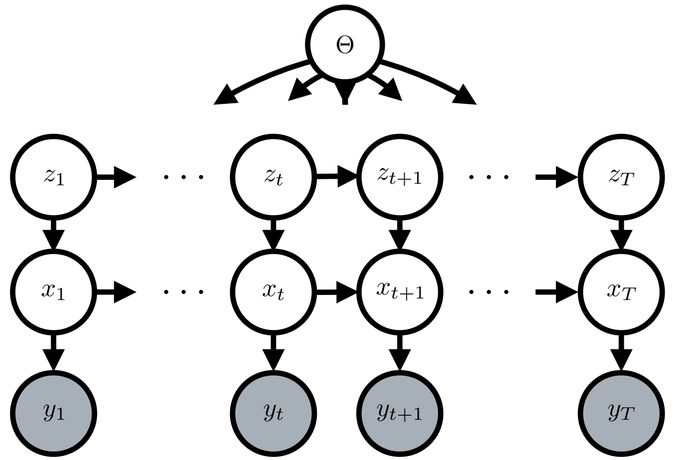
The optimal update for the continuous states takes the form

$$\log q(x_{1:T}) = \mathbb{E}_{q(z)}\left[\log \mathcal{N}(x_1 \mid b_{z_1}, Q_{z_1})\right] + \sum_{t=2}^{T} \mathbb{E}_{q(z)}\left[\log \mathcal{N}(x_t \mid A_{z_t} x_{t-1} + b_{z_t}, Q_{z_t})\right] + \sum_{t=1}^{T} \log \mathcal{N}(y_t \mid C x_t + d, R) + \text{c}$$

Question: can you see what form this will take?

Variational EM for SLDS

Updating the continuous state posterior $q(x_{1:T})$



The optimal update for the continuous states is the same form as the posterior in **linear dynamical system.**

$$\log q(x_{1:T}) = \mathcal{N}(\mathrm{vec}(x_{1:T}) \mid \tilde{J}^{-1}\tilde{h}, \tilde{J}^{-1})$$

where

$$\tilde{J}_{tt} = \mathbb{E}_{q(z)}[Q_{z_t}^{-1}] + \mathbb{E}_{q(z)}[A_{z_{t+1}} Q_{z_{t+1}}^{-1} A_{z_{t+1}}] + C^\top R^{-1} C$$

$$\tilde{J}_{t,t-1} = \mathbb{E}_{q(z)}[Q_{z_t}^{-1} A_{z_t}]$$

$$\tilde{h}_t = \mathbb{E}_{q(z)}[Q_{z_t}^{-1} b_{z_t}] - \mathbb{E}_{q(z)}[A_{z_{t+1}} Q_{z_{t+1}}^{-1} b_{z_{t+1}}] + C^\top R^{-1}(y_t - d)$$

But here, the **natural parameters are expectations** under $q(z)$.

Variational EM for SLDS

Putting it all together

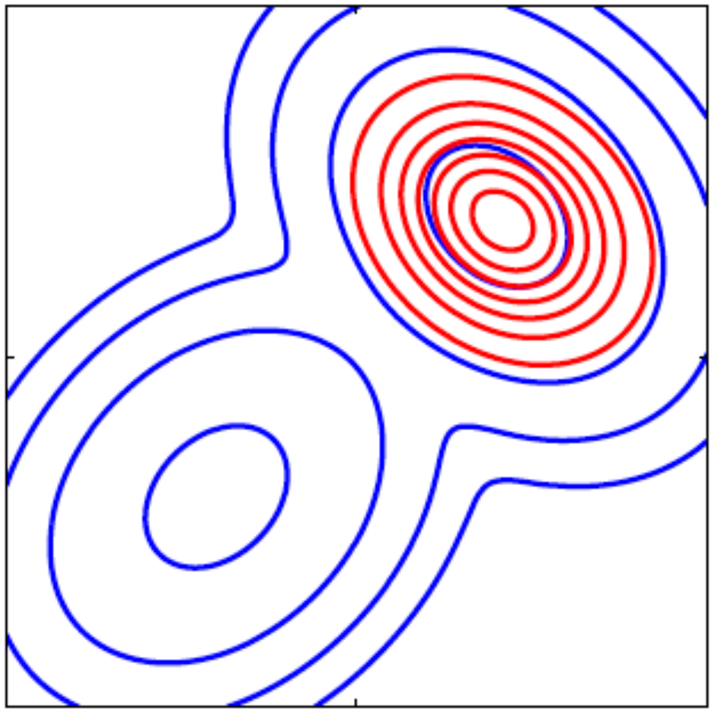
- **M-step**: Maximize the expected log probability

$$\Theta \leftarrow = \arg\max_{\Theta} \mathbb{E}_{q(z)q(x)}[\log p(y, x, z, \Theta)]$$

  using expected sufficient statistics.

- **Variational E-step**:
  - Repeat until convergence:
    1. $q(z) \leftarrow \mathrm{HMM}(\pi, P, \tilde{\ell})$
       where $\tilde{\ell}$ are the expected log likelihoods under $q(x)$
    2. $q(x) \leftarrow \mathrm{LDS}(\tilde{J}, \tilde{h})$
       where $\tilde{J}$ and $\tilde{h}$ are the expected natural parameters under $q(z)$
  - Compute the ELBO

$$\mathscr{L}[q(z)\, q(x), \Theta] \leq \log p(y \mid \Theta)$$



Bishop (2006). Pattern Recognition and Machine Learning, Ch 10.1

Conclusion

- In many latent variable models, including SLDS, the posterior distribution is intractable.
- Variational EM replaces the E step with a tractable variational approximation that minimizes the divergence to the true but intractable posterior.
- Mean field approximations are commonly used, as they often admit simple coordinate updates.
- In the SLDS, we can use a structured mean field approximation that retains dependencies across time while assuming independence of the discrete and continuous states.

Further Reading

- Barber, David. 2012. Bayesian Reasoning and Machine Learning. Cambridge University Press. **Chapter 25.**

- Blei, David M., Alp Kucukelbir, and Jon D. McAuliffe. 2017. “Variational Inference: A Review for Statisticians.” Journal of the American Statistical Association 112 (518): 859–77.

- Ghahramani, Z., and G. E. Hinton. 2000. “Variational Learning for Switching State-Space Models.” Neural Computation 12 (4): 831–64.

- Zoltowski, David, Jonathan Pillow, and Scott Linderman. 2020. “A General Recurrent State Space Framework for Modeling Neural Dynamics during Decision-Making.” In Proceedings of the 37th International Conference on Machine Learning (ICML).