Machine Learning Methods for Neural Data Analysis

Lecture 13: Switching linear dynamical systems

Scott Linderman

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

Agenda

- Switching linear dynamical systems (SLDS)
- Hardness of exact EM for SLDS
- Variational EM
- Coordinate Ascent VI (CAVI)

Recap: Gaussian HMM

Generative Model:

$$
\begin{aligned}
z_1 &\sim \mathrm{Cat}(\pi), \\
z_t \mid z_{t-1} &\sim \mathrm{Cat}(P_{z_{t-1}}), && \text{for } t = 2, \ldots, T. \\
x_t \mid z_t &\sim \mathcal{N}(b_{z_t}, Q_{z_t}) && \text{for } t = 1, \ldots, T
\end{aligned}
$$

Parameters:

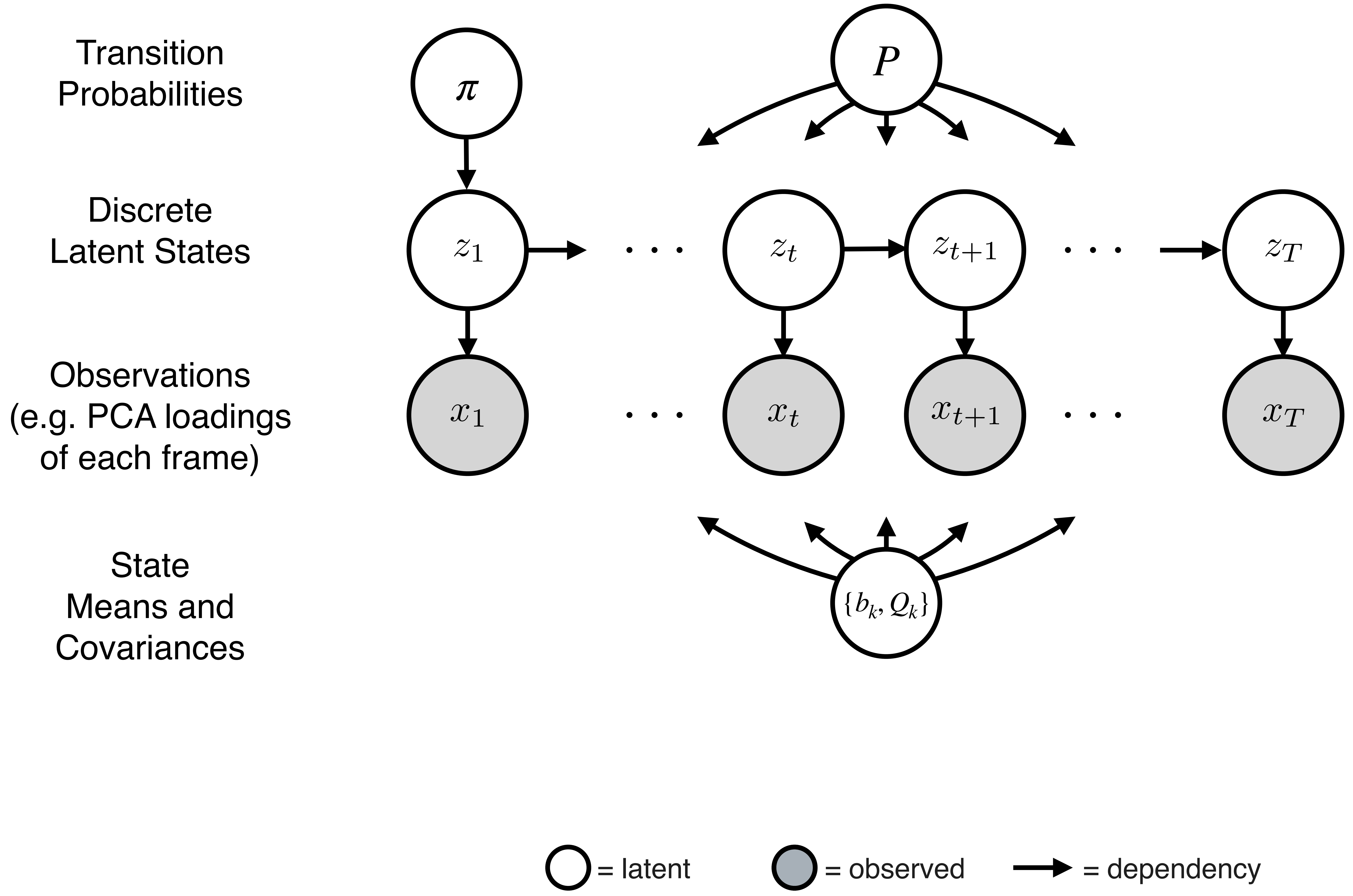
$$\Theta = \pi, P, \{b_k, Q_k\}_{k=1}^{K}$$

Joint probability:

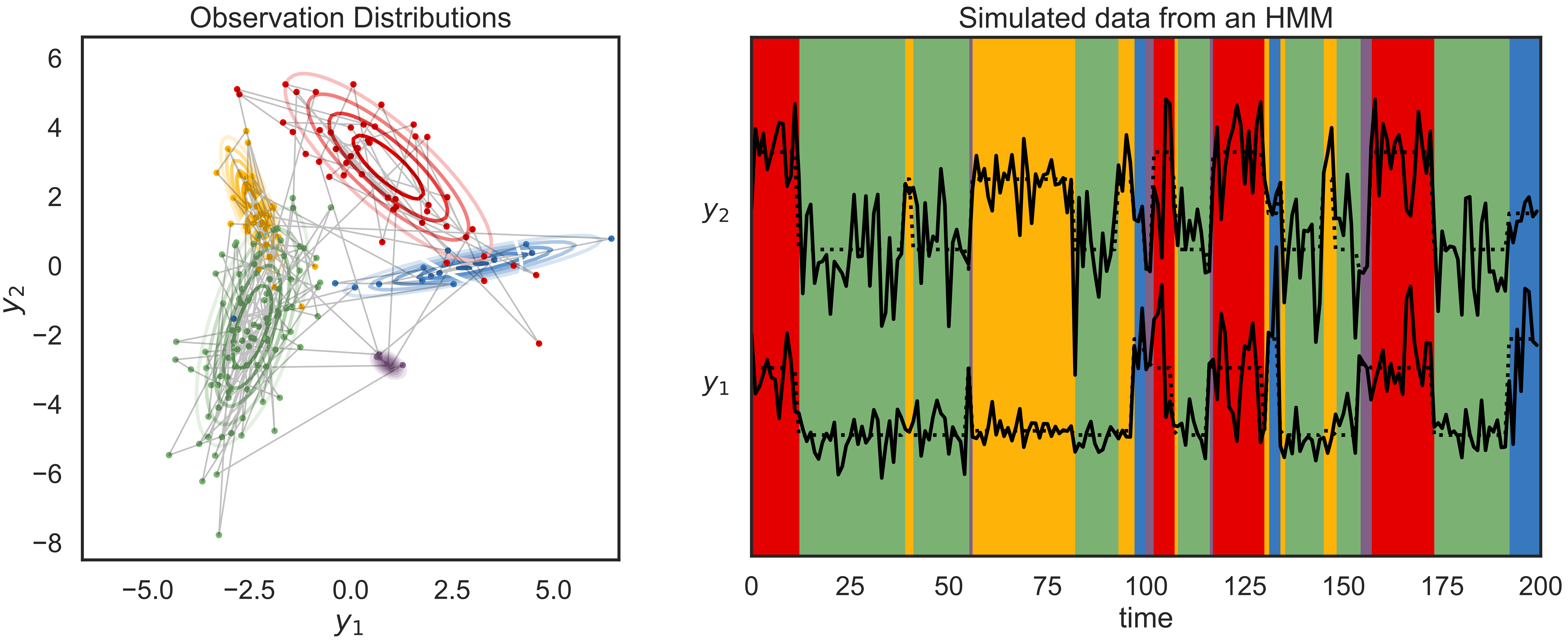
$$p(x, z \mid \Theta) = p(z_1) \prod_{t=1}^{T-1} p(z_{t+1} \mid z_t) \prod_{t=1}^{T} p(x_t \mid z_t)$$

Recap: The Gaussian HMM

Graphical Model



Simulated data from a Gaussian HMM



Autoregressive (AR) HMM

Generative Model:

$$
\begin{aligned}
z_1 &\sim \mathrm{Cat}(\pi), \\
z_t \mid z_{t-1} &\sim \mathrm{Cat}(P_{z_{t-1}}), && \text{for } t = 2,\ldots,T \\
x_1 \mid z_1 &\sim \mathcal{N}(b_{z_1}, Q_{z_1}) && . \\
x_t \mid x_{t-1}, z_t &\sim \mathcal{N}(A_{z_t} x_{t-1} + b_{z_t}, Q_{z_t}) && \text{for } t = 1,\ldots,T
\end{aligned}
$$

Parameters:

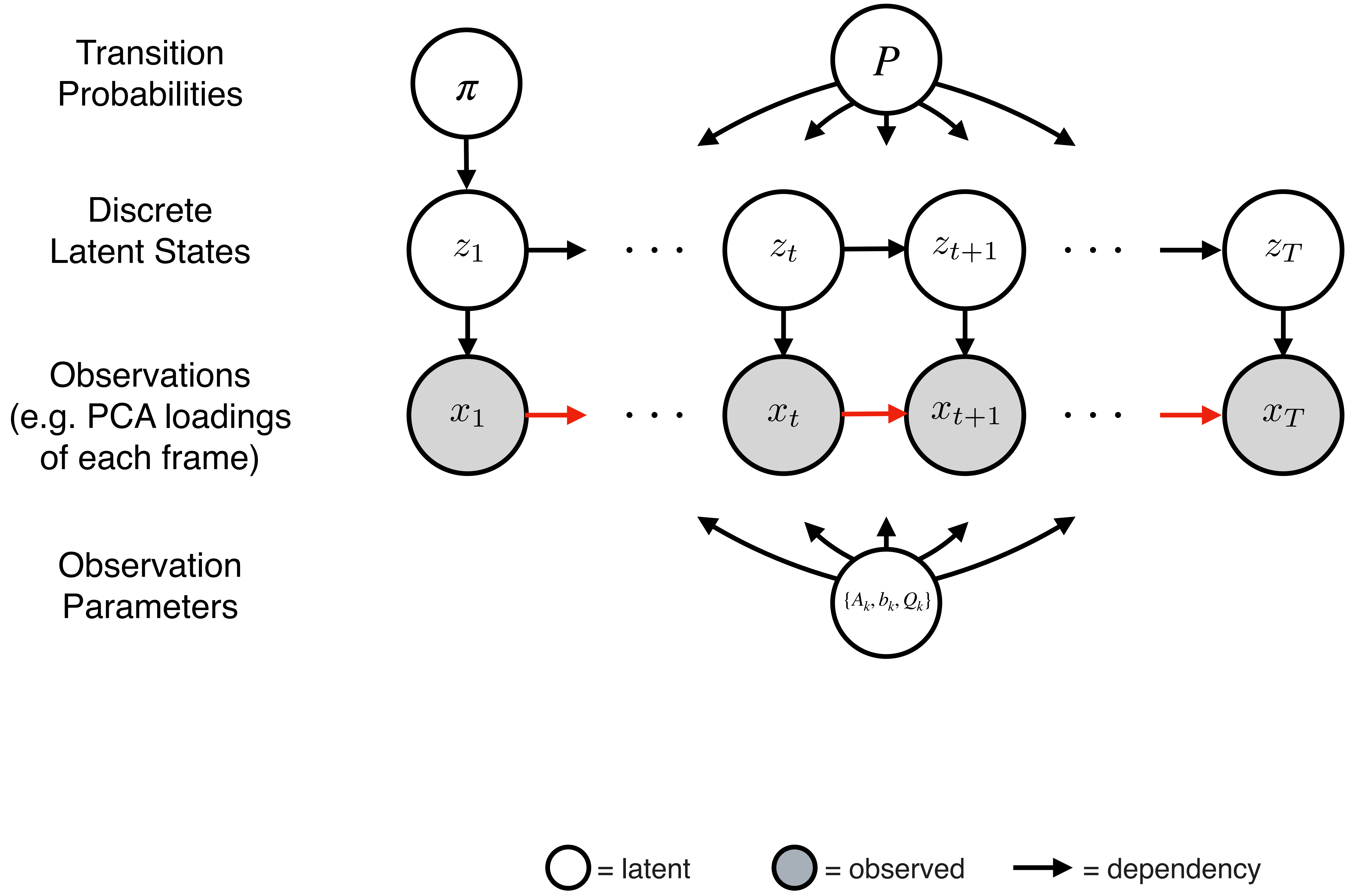
$$\Theta = \pi, P, \{A_k, b_k, Q_k\}_{k=1}^{K}$$

Joint probability:

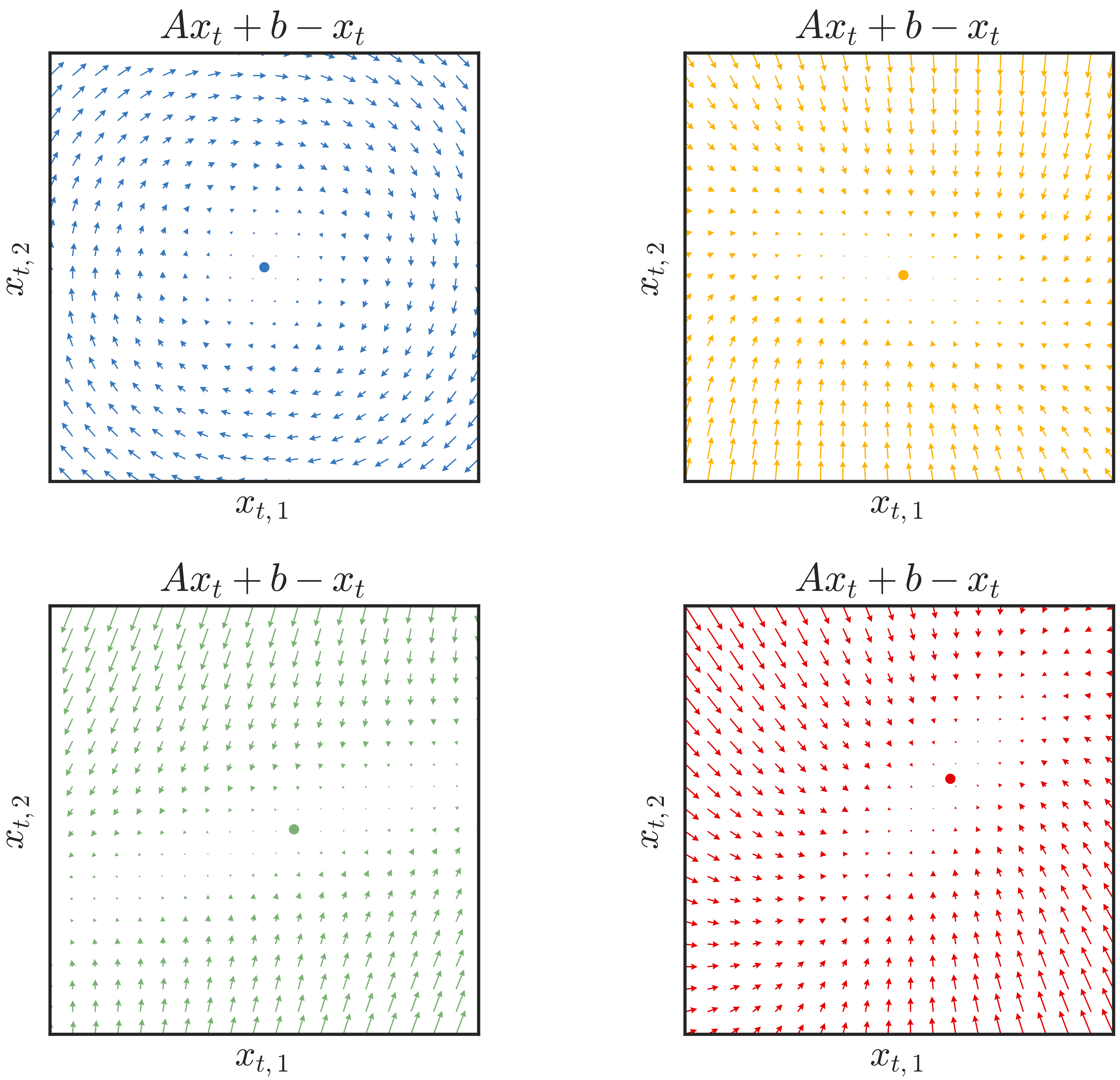
$$p(x, z \mid \Theta) = p(z_1)\, p(x_1 \mid z_1) \prod_{t=1}^{T-1} p(z_{t+1} \mid z_t) \prod_{t=1}^{T} p(x_t \mid x_{t-1}, z_t)$$

Autoregressive (AR) HMM

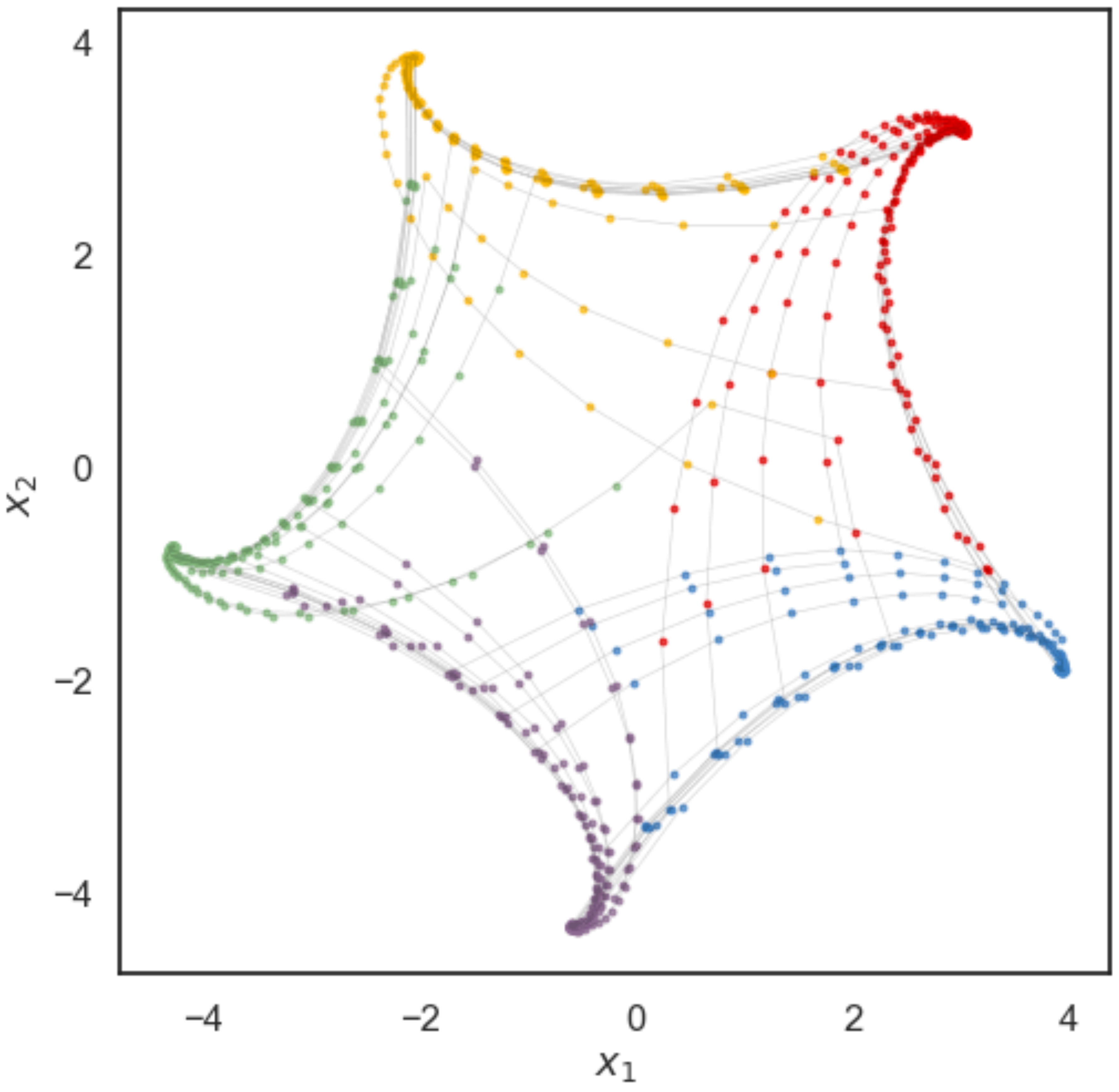
Graphical Model

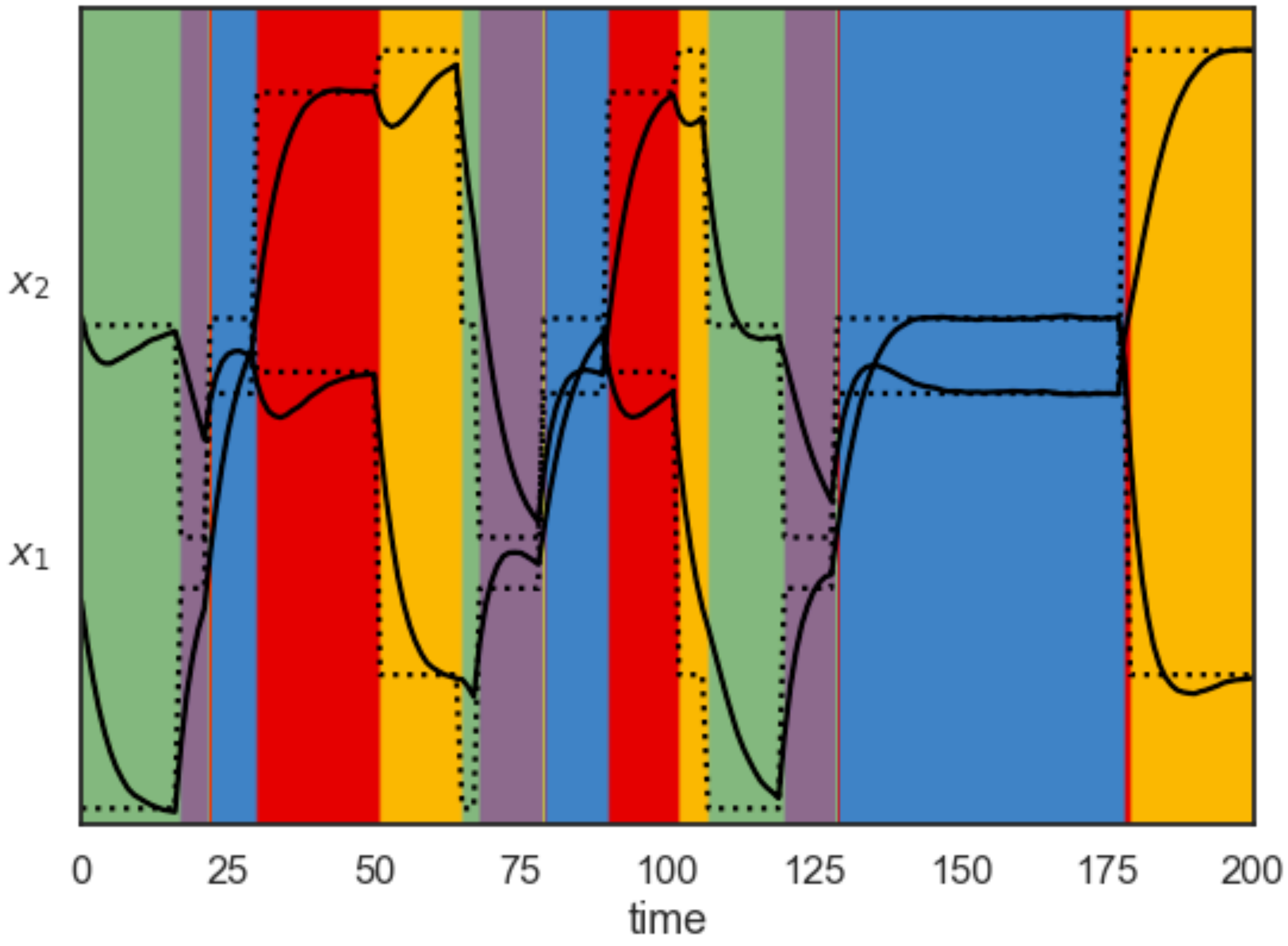


Visualizing Linear Dynamics



Simulated data from an ARHMM





Gaussian Linear Dynamical Systems

Generative Model:

$$\begin{aligned} x_1 &\sim \mathcal{N}(b, Q), & \\ x_t \mid x_{t-1} &\sim \mathcal{N}(Ax_{t-1} + b, Q), & \text{for } t = 2,\ldots,T. \\ y_t \mid x_t &\sim \mathcal{N}(Cx_t + d, R) & \text{for } t = 1,\ldots,T \end{aligned}$$

Parameters:

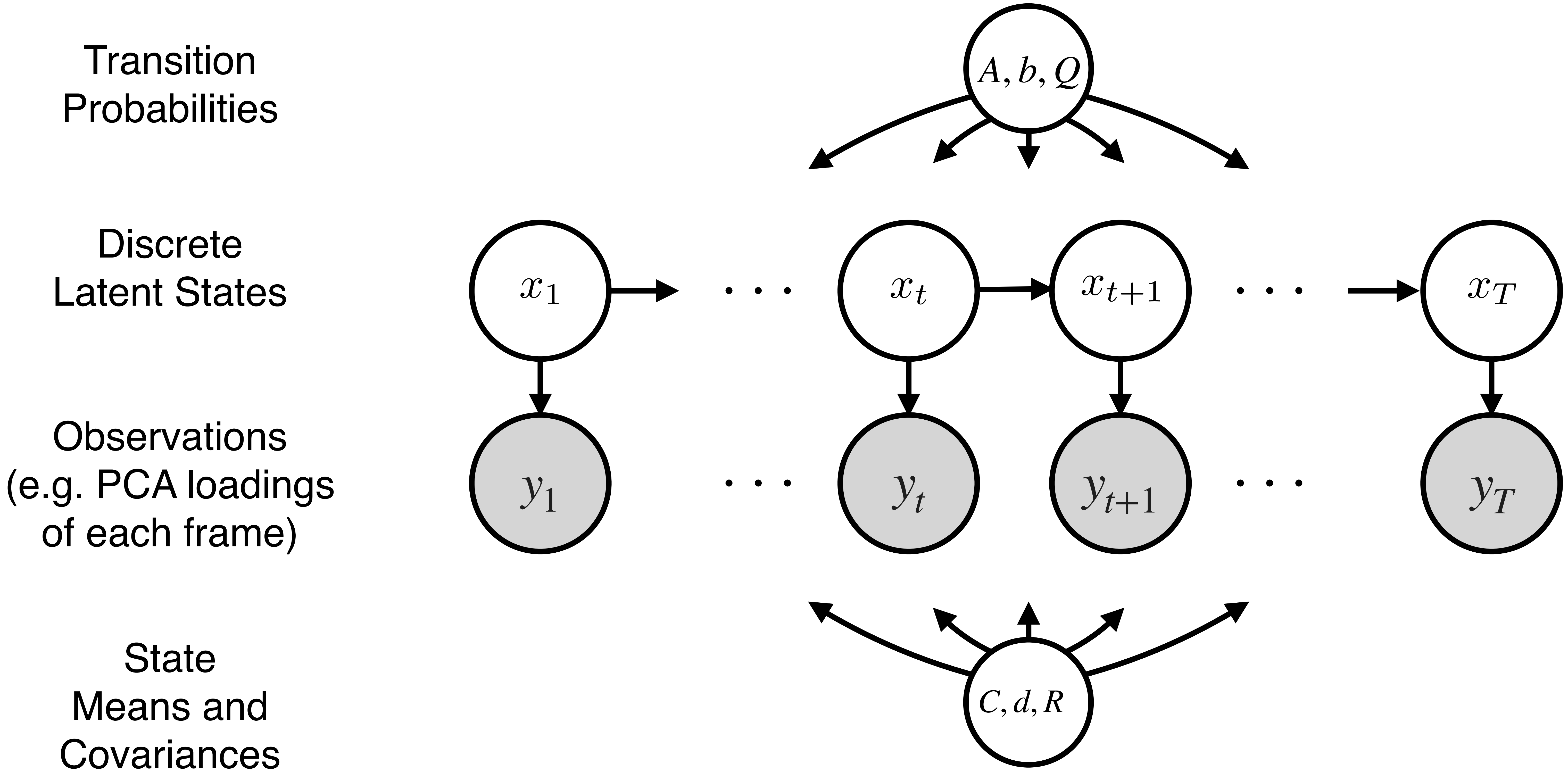
$$\Theta = A, b, Q, C, d, R$$

Joint probability:

$$p(y, x \mid \Theta) = p(x_1) \prod_{t=1}^{T-1} p(x_{t+1} \mid x_t) \prod_{t=1}^{T} p(y_t \mid x_t)$$

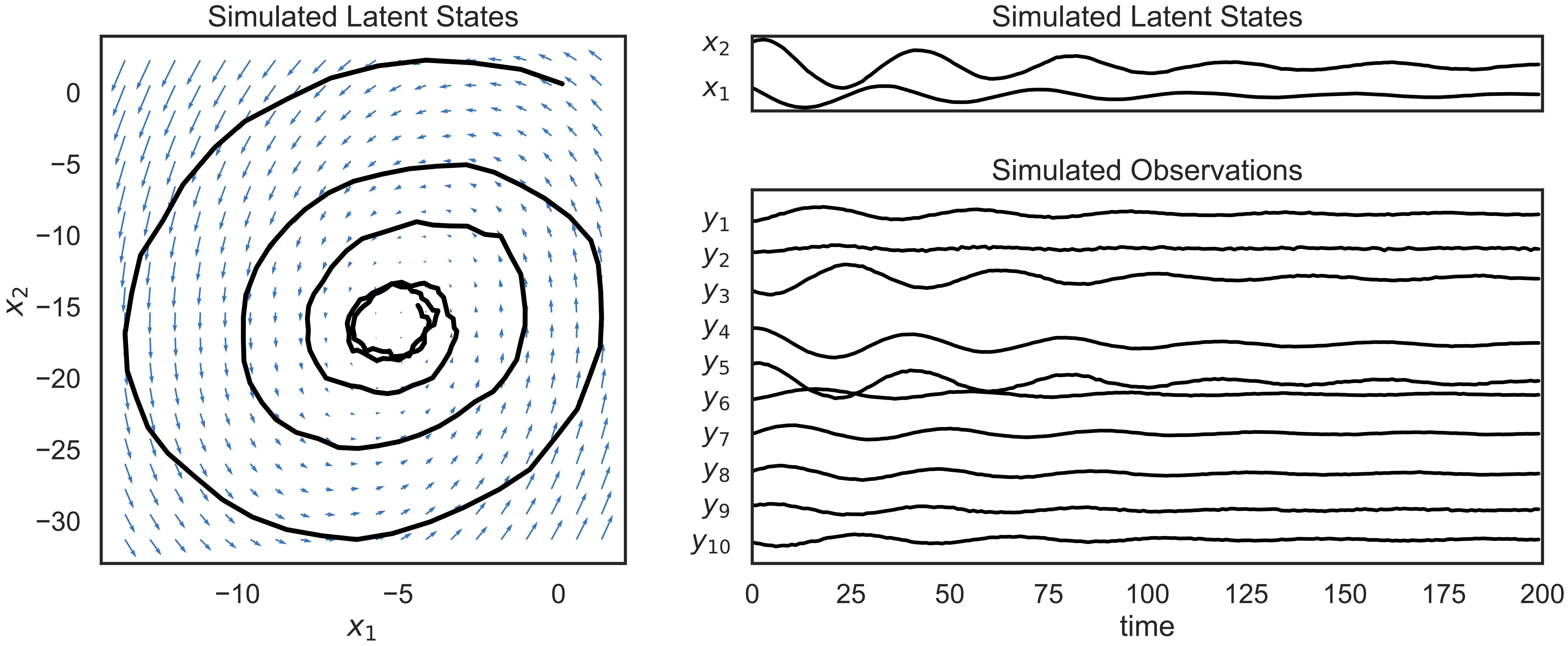
Gaussian Linear Dynamical Systems

Graphical Model

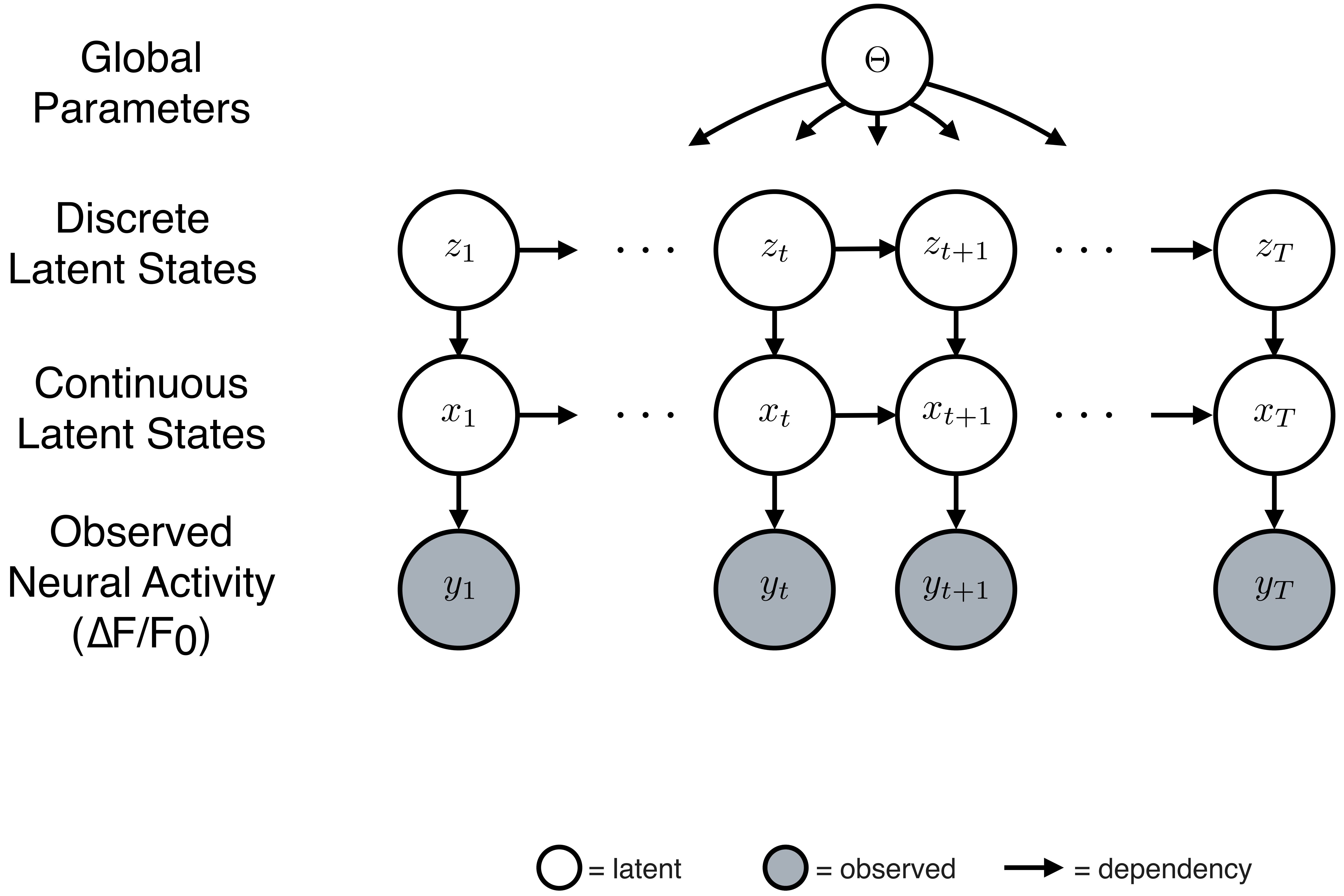


You can do **exact EM** in Gaussian LDS too.
The equivalent **message passing algorithms** are the **Kalman filter** and **Kalman smoother**!

Simulated data from an LDS

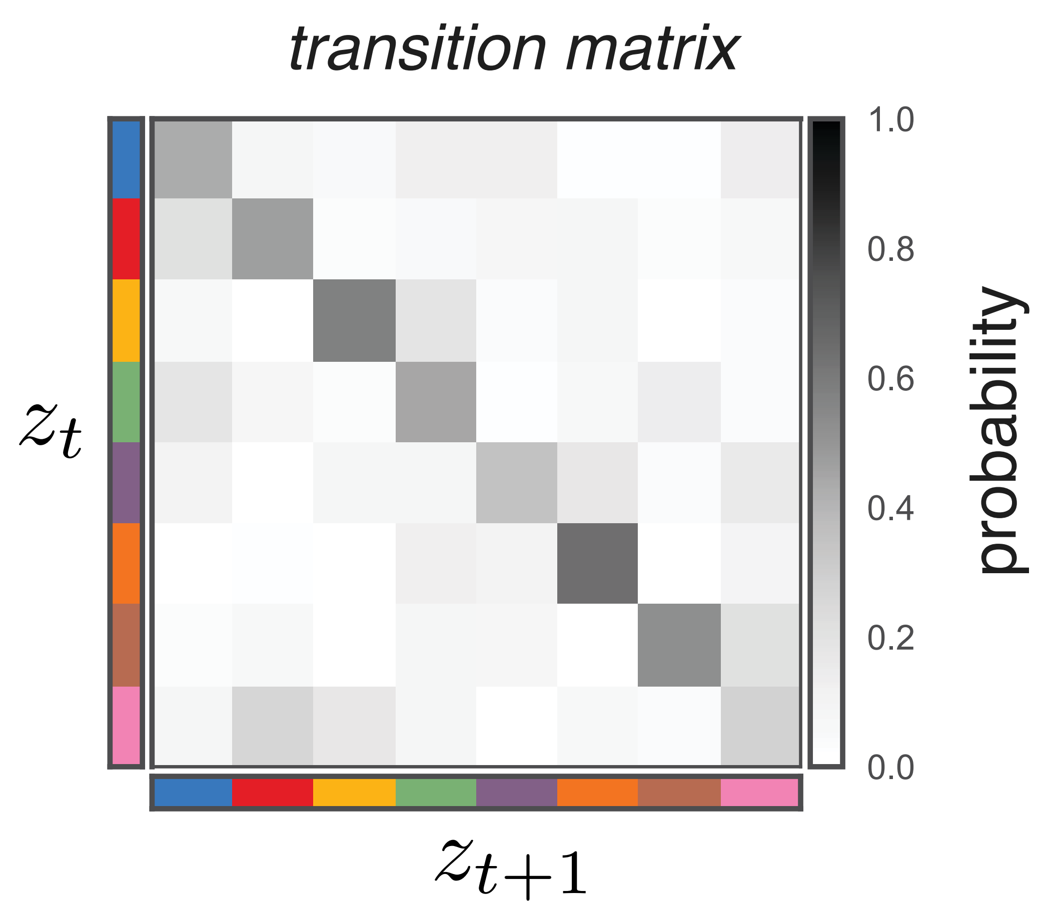


Switching LDS: Best of Both Worlds



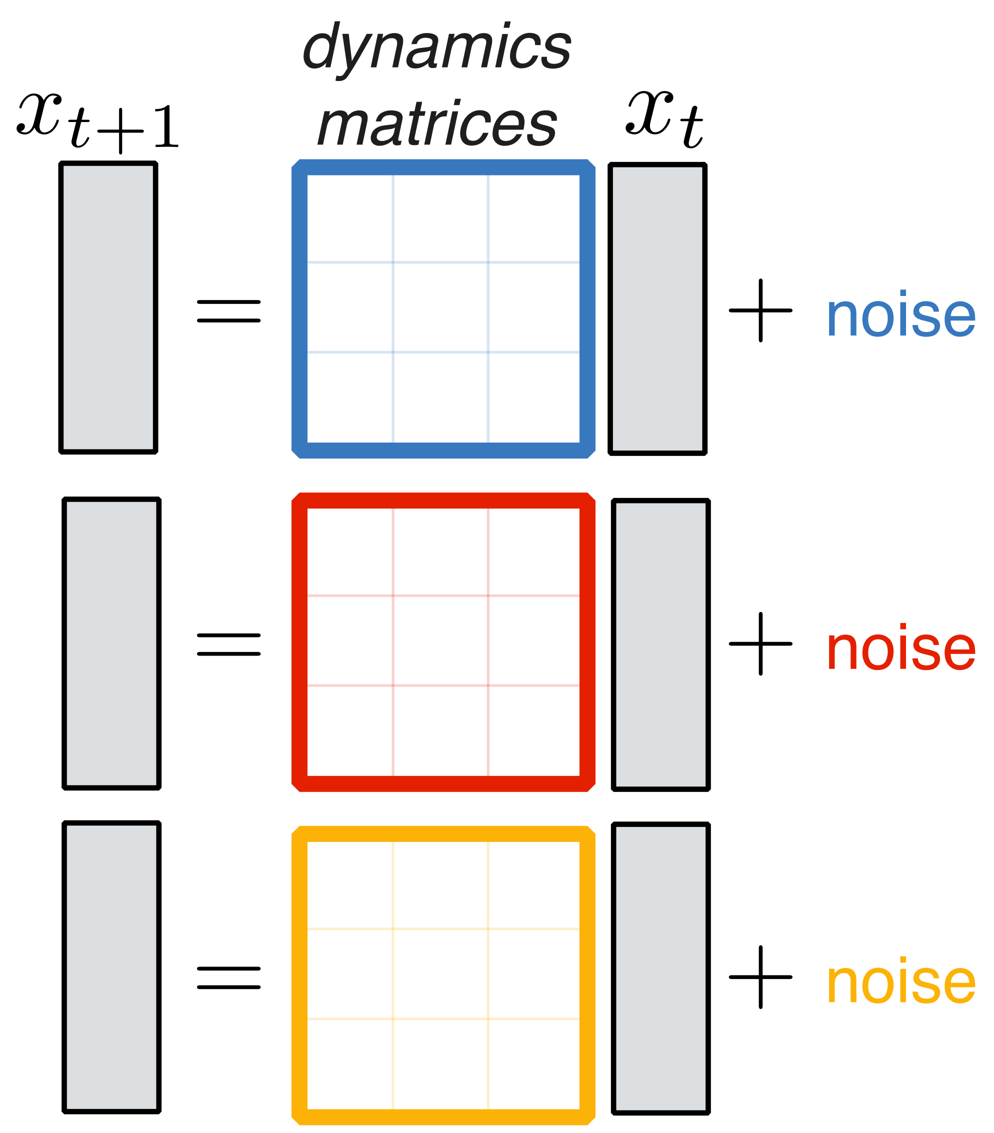
Specifying the form of the dependencies

State-dependent
switching probabilities



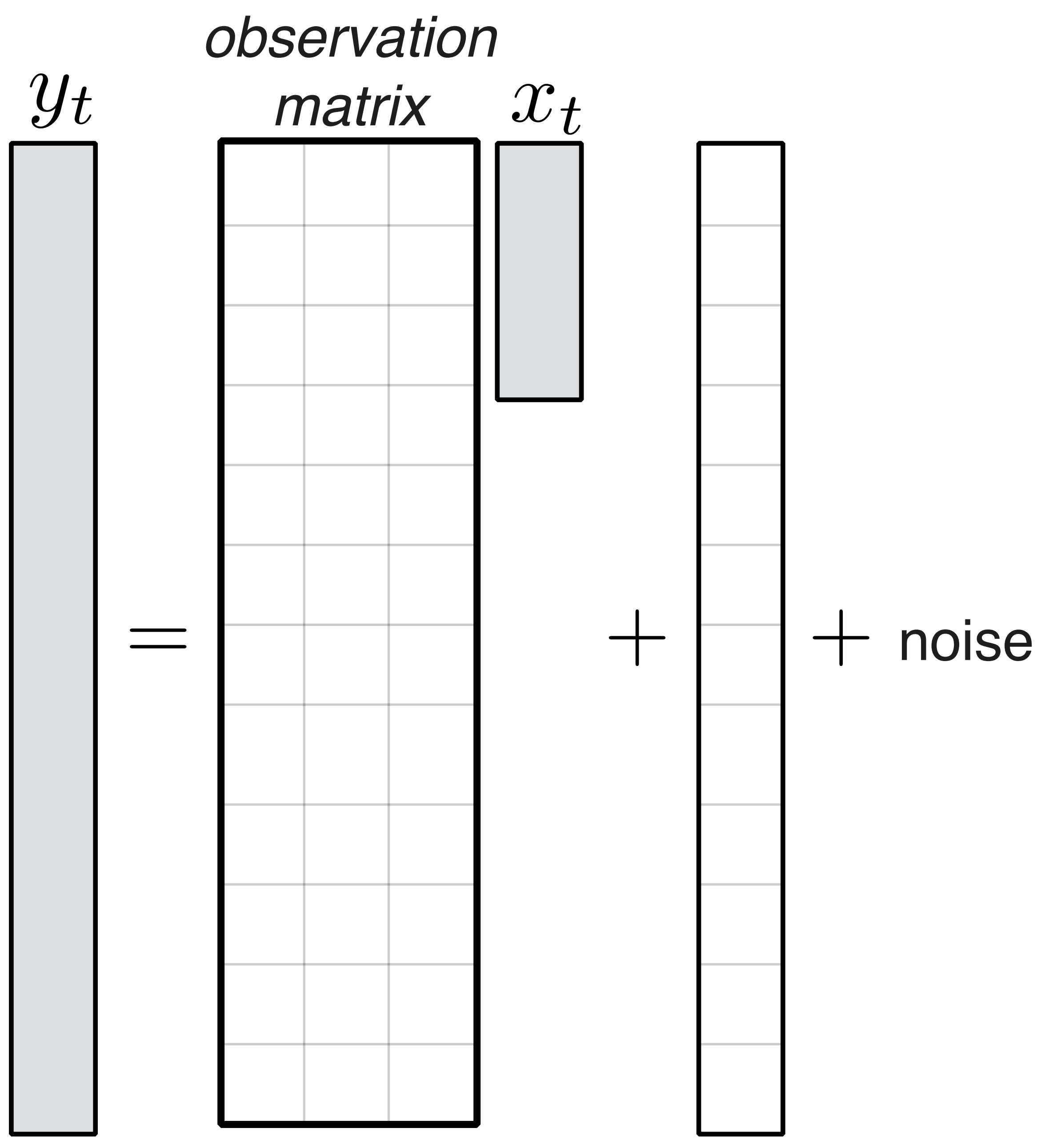
$$\Pr(z_{t+1} = j \mid z_t = i) = P_{ij}$$

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

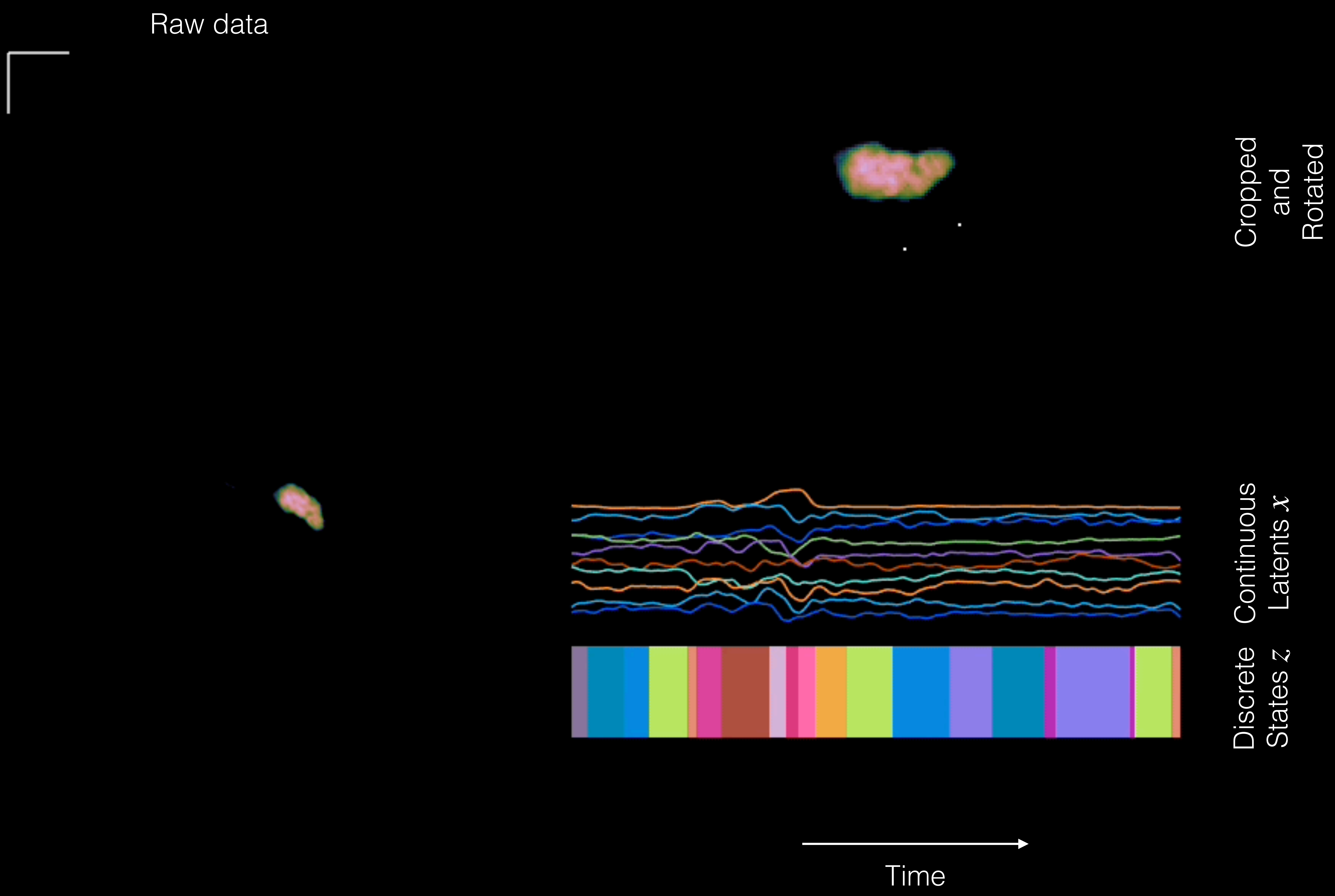


$$y_t = C x_t + d + \delta_t$$

Switching Linear Dynamical System (SLDS)

Combines LDS (Kalman, 1960) and HMMs (Rabiner, 1989).

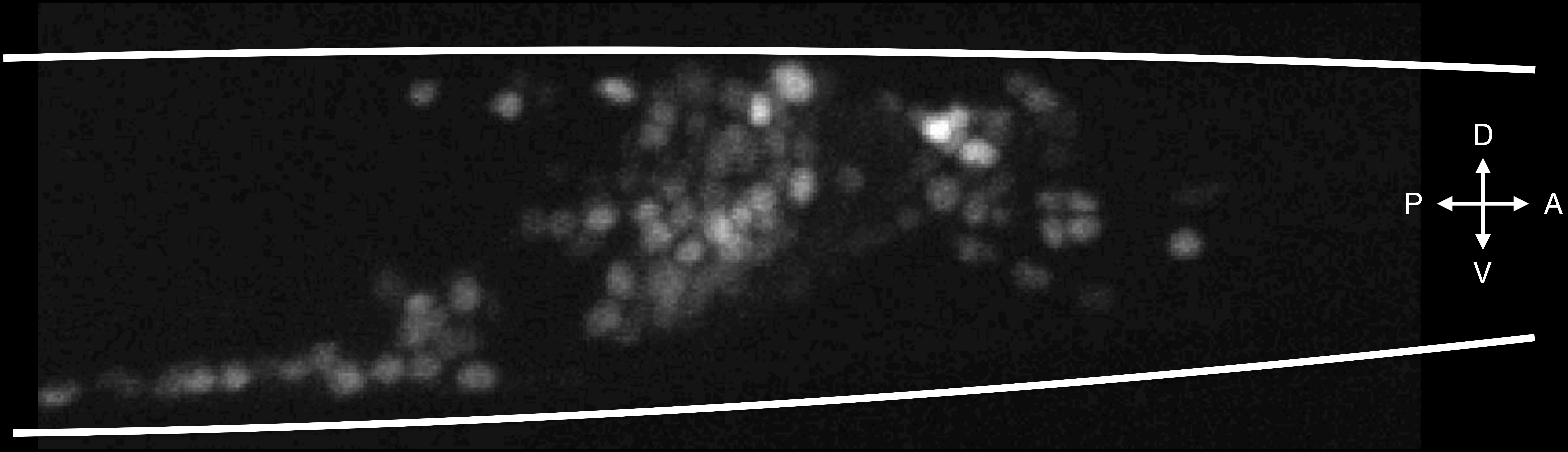
Ghahramani and Hinton (1996); Murphy (199



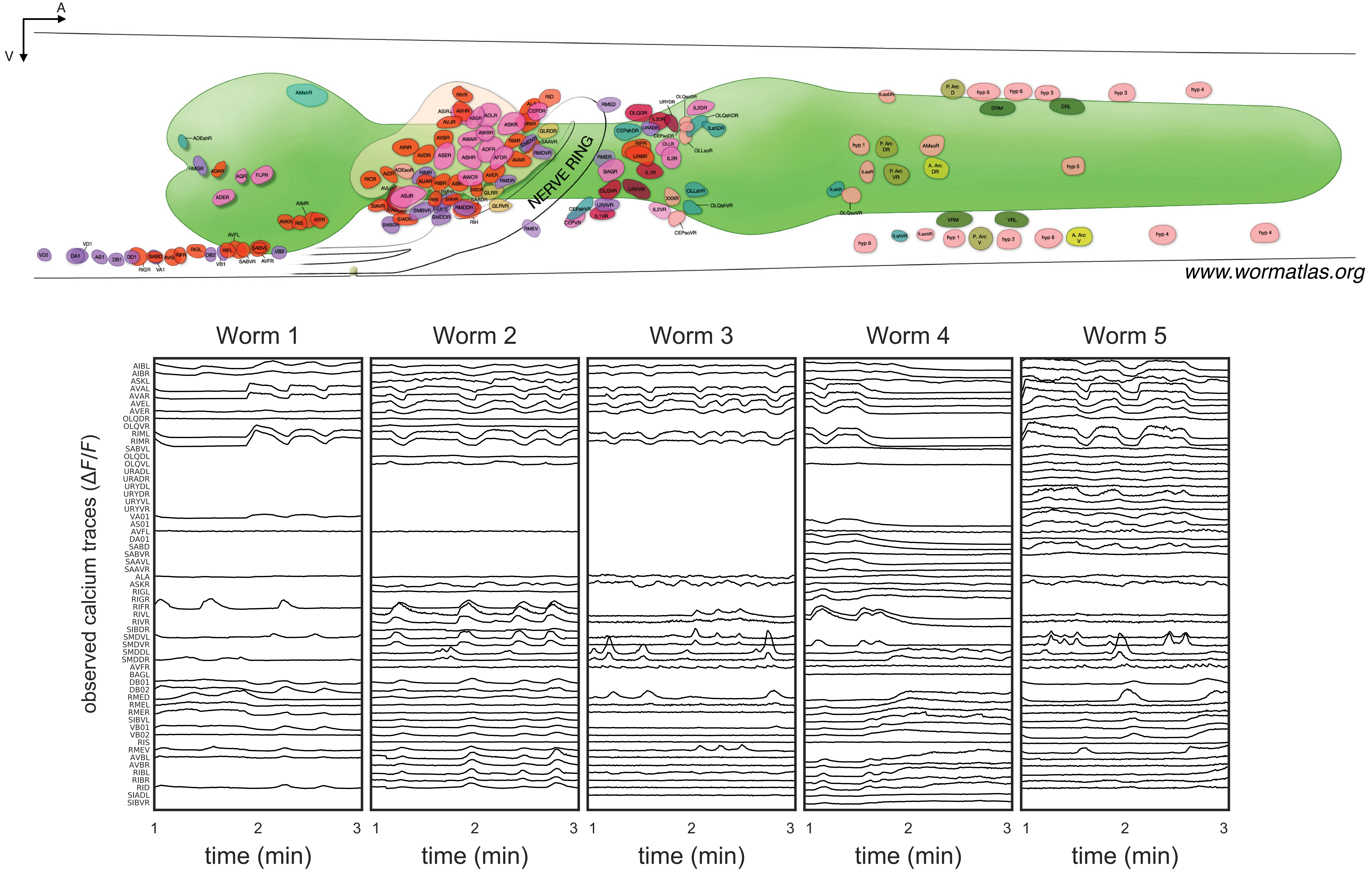
Wiltschko et al. (2015)
Markowitz et al. (2018)

A. E. X. Brown, Imperial College

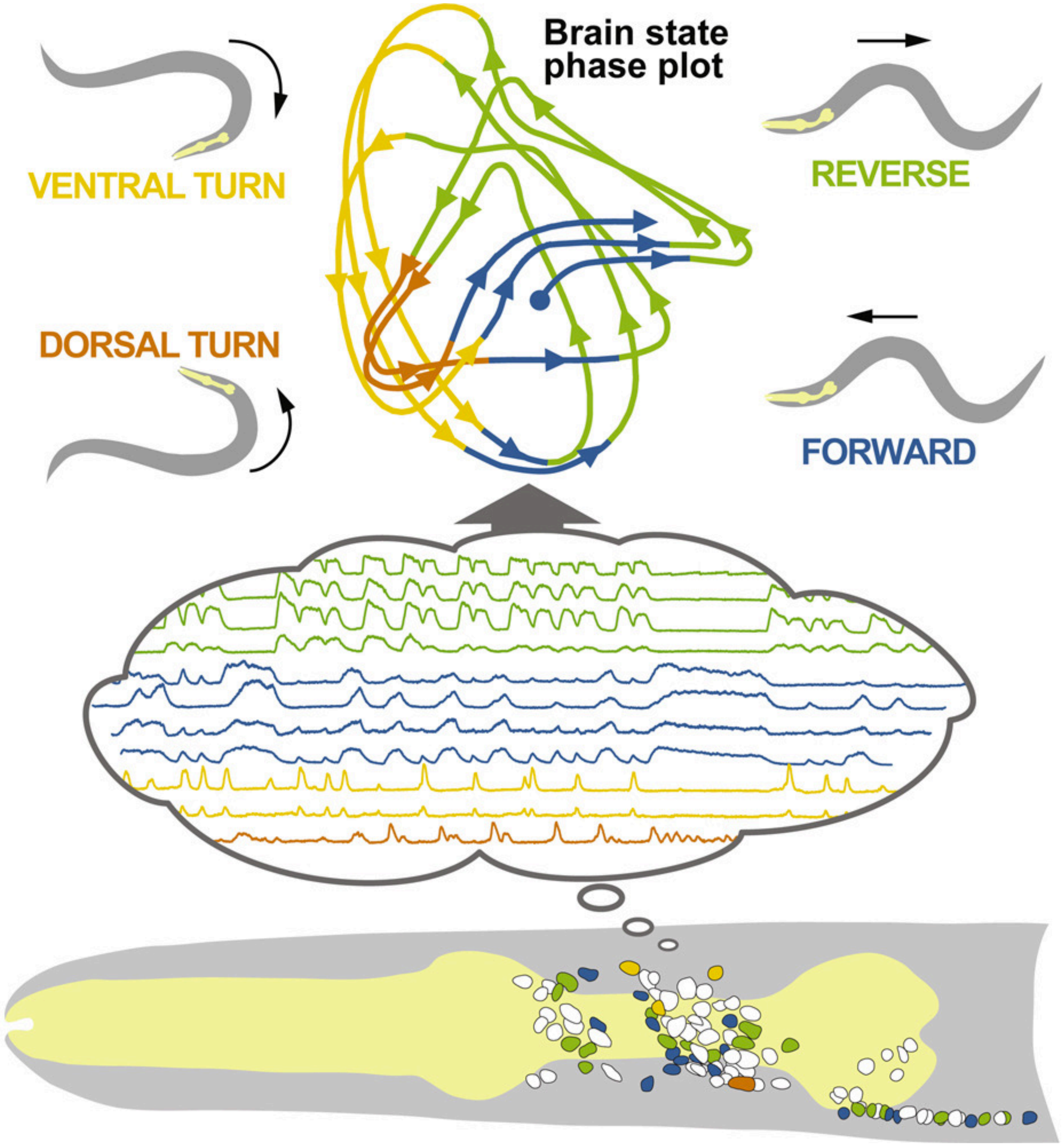
Calcium imaging of ~100 head ganglia neurons in immobilized *C. elegans*

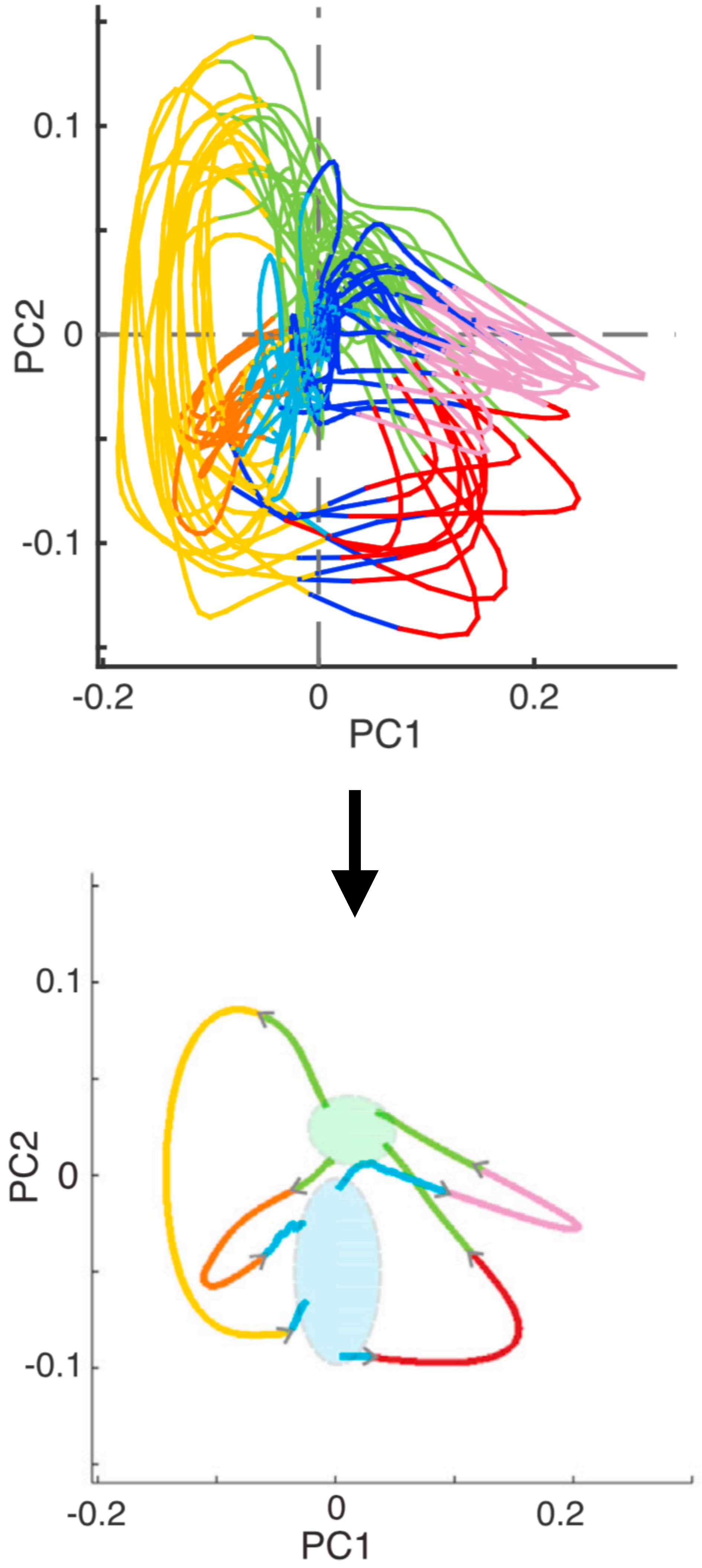


*with **Nichols**, Blei, **Zimmer**, Paninski*


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)

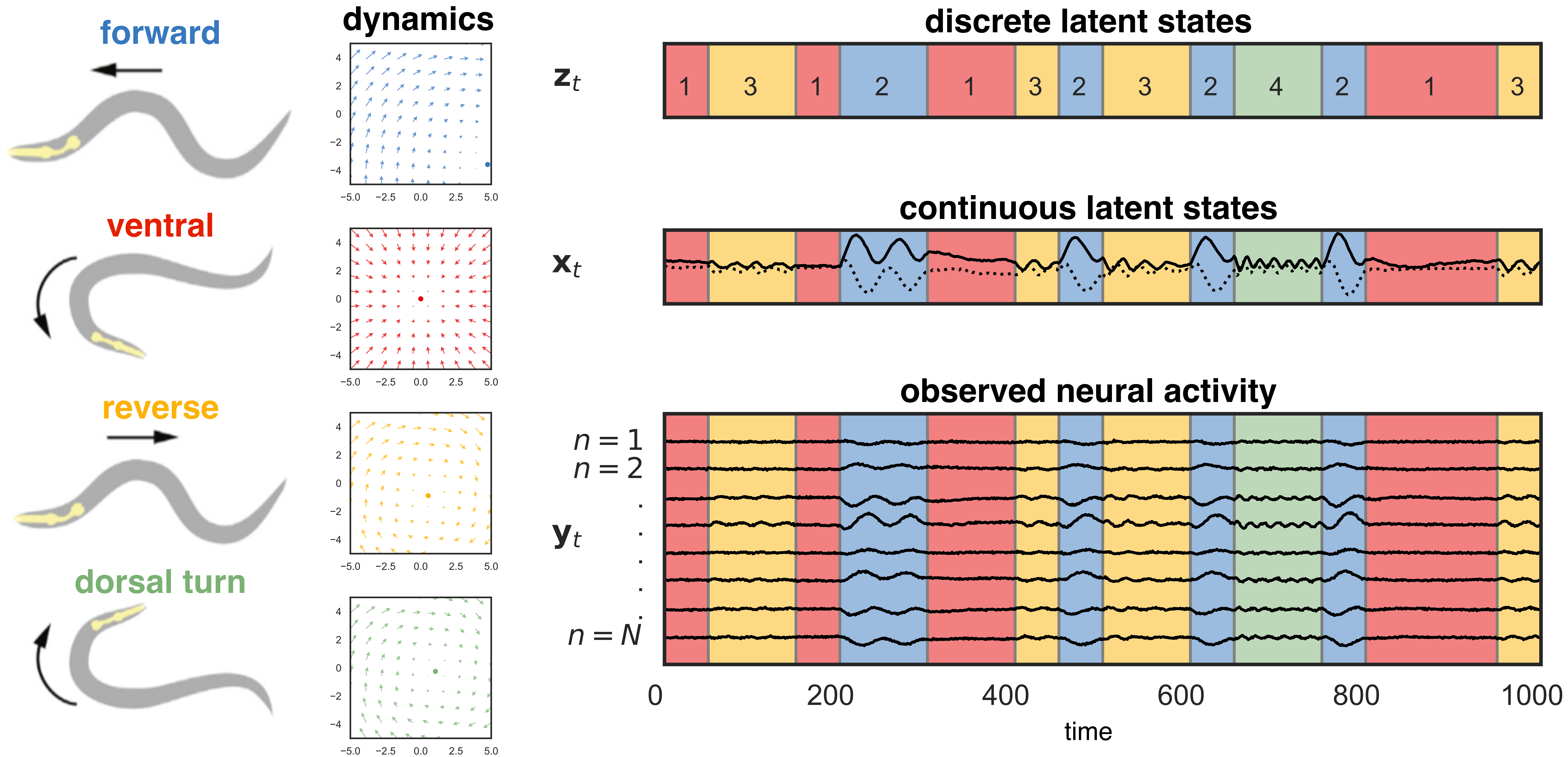
Previous work suggests that this neural activity lies on a low dimensional manifold partitioned by behavior





Kato et al (2015)

Building a probabilistic model of neural data



Hardness of exact EM for SLDS

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,

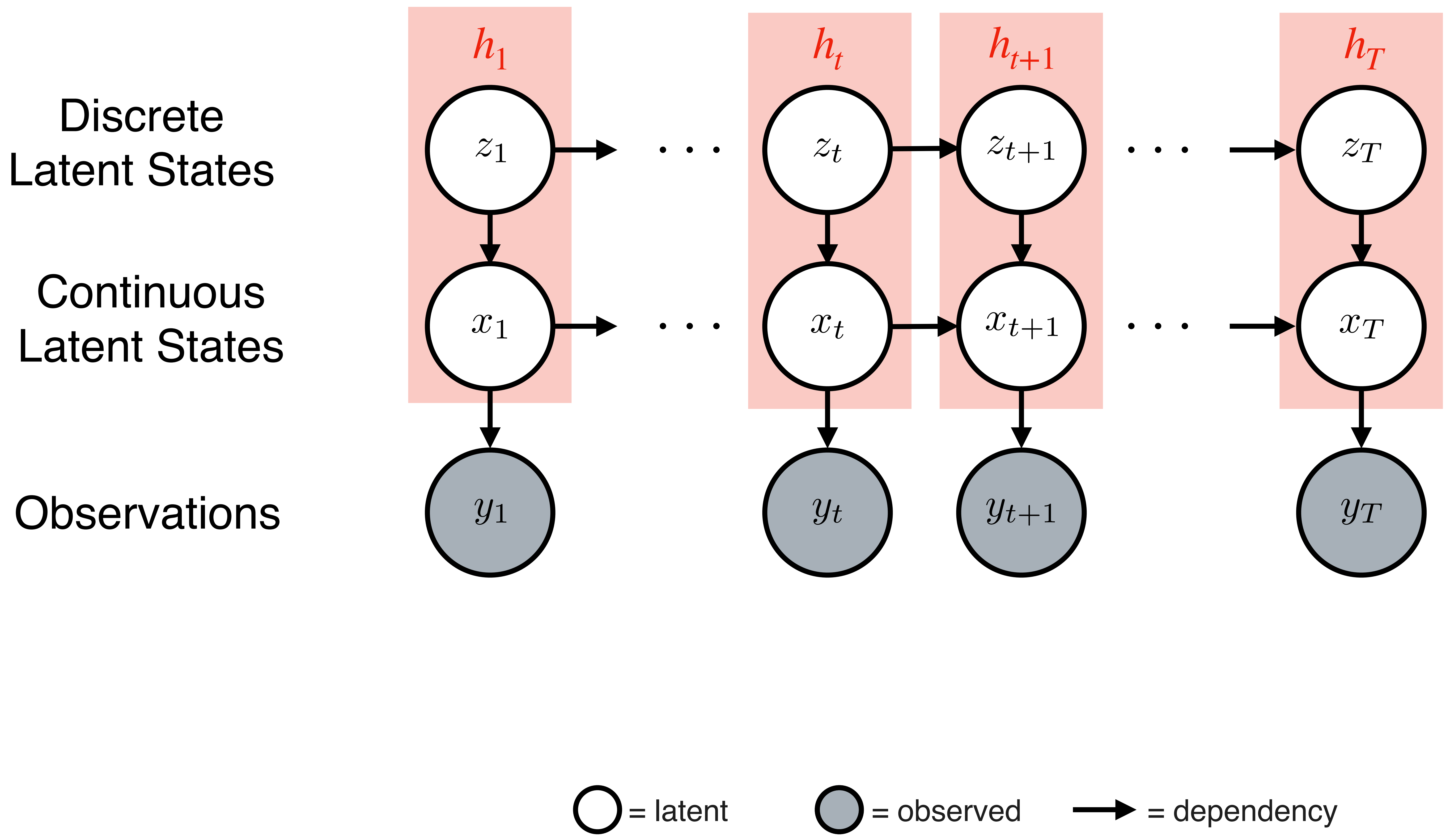
$$\Theta \leftarrow \arg\max \mathbb{E}_{q(z,x)}\left[\log p(z, x, y \mid \Theta)\right]$$

- Unfortunately, computing the necessary expectations is a lot harder now!

Combining the latent states

SLDS as a "hybrid" state space model

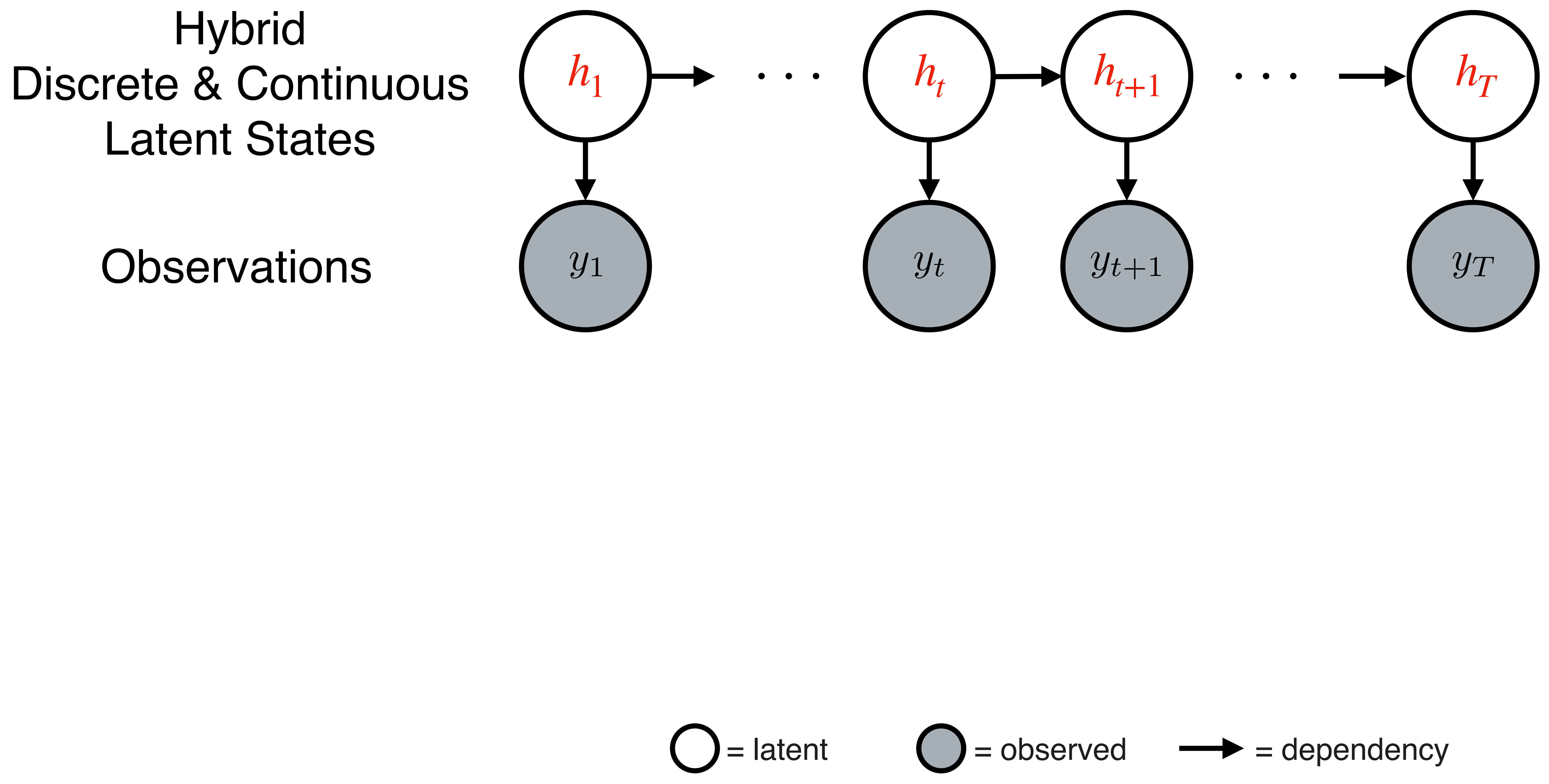
- Let $h_t = (z_t, x_t)$ denote the hybrid discrete & continuous latent state



Combining the latent states

SLDS as a "hybrid" state space model

- Let $h_t = (z_t, x_t)$ denote the hybrid discrete & continuous latent state



Exact EM for SLDS

Computing the marginal distributions

- Consider the marginal probability of the latent states at time t:

$$q(h_t) = \int \mathrm{d}h_1 \cdots \int \mathrm{d}h_{t-1} \int \mathrm{d}h_{t+1} \cdots \int \mathrm{d}h_T \, q(h_1, \ldots, h_{t-1}, h_t, h_{t+1}, \ldots, h_T)$$

Exact EM for SLDS

Computing the marginal distributions

- Consider the marginal probability of the latent states at time t:

$$q(h_t) = \int \mathrm{d}h_1 \cdots \int \mathrm{d}h_{t-1} \int \mathrm{d}h_{t+1} \cdots \int \mathrm{d}h_T \, q(h_1, \ldots, h_{t-1}, h_t, h_{t+1}, \ldots, h_T)$$

$$\propto \left[\int \mathrm{d}h_1 \cdots \int \mathrm{d}h_{t-1} \, p(h_1) \prod_{s=1}^{t-1} p(h_s \mid h_s) \, p(h_{s+1} \mid h_s) \right] \times \left[p(y_t \mid h_t) \right]$$

$$\times \left[\int \mathrm{d}h_{t+1} \cdots \int \mathrm{d}h_T \prod_{u=t+1}^{T} p(h_u \mid h_{u-1}) \, p(y_u \mid h_u) \right]$$

Exact EM for SLDS

Computing the marginal distributions

- Consider the marginal probability of the latent states at time t:

$$
\begin{aligned}
q(h_t) &= \int \mathrm{d}h_1 \cdots \int \mathrm{d}h_{t-1} \int \mathrm{d}h_{t+1} \cdots \int \mathrm{d}h_T \, q(h_1, \ldots, h_{t-1}, h_t, h_{t+1}, \ldots, h_T) \\
&\propto \left[\int \mathrm{d}h_1 \cdots \int \mathrm{d}h_{t-1} \, p(h_1) \prod_{s=1}^{t-1} p(h_s \mid h_s) \, p(h_{s+1} \mid h_s) \right] \times \left[p(y_t \mid h_t) \right] \\
&\qquad \times \left[\int \mathrm{d}h_{t+1} \cdots \int \mathrm{d}h_T \prod_{u=t+1}^{T} p(h_u \mid h_{u-1}) \, p(y_u \mid h_u) \right] \\
&\triangleq \alpha_t(h_t) \times p(y_t \mid h_t) \times \beta_t(h_t)
\end{aligned}
$$

Exact EM for SLDS

Computing the forward messages $\alpha_t(h_t)$

- Consider the "forward messages":

$$\alpha_t(h_t) \triangleq \int \mathrm{d}h_1 \cdots \int \mathrm{d}h_{t-1}\, p(h_1) \prod_{s=1}^{t-1} p(h_s \mid h_s)\, p(h_{s+1} \mid h_s)$$

Exact EM for SLDS

Computing the forward messages $\alpha_t(h_t)$

- Consider the "forward messages":

$$\alpha_t(h_t) \triangleq \int \mathrm{d}h_1 \cdots \int \mathrm{d}h_{t-1}\, p(h_1) \prod_{s=1}^{t-1} p(h_s \mid h_s)\, p(h_{s+1} \mid h_s)$$

$$= \int \mathrm{d}h_{t-1} \left[\left(\int \mathrm{d}h_1 \cdots \int \mathrm{d}h_{t-2}\, p(h_1) \prod_{s=1}^{t-2} p(y_s \mid h_s) p(h_{s+1} \mid h_s) \right) p(y_{t-1} \mid h_{t-1})\, p(h_t \mid h_{t-1}) \right]$$

Exact EM for SLDS

Computing the forward messages $\alpha_t(h_t)$

- Consider the “forward messages”:

$$\alpha_t(h_t) \triangleq \int \mathrm{d}h_1 \cdots \int \mathrm{d}h_{t-1}\, p(h_1) \prod_{s=1}^{t-1} p(h_s \mid h_s)\, p(h_{s+1} \mid h_s)$$

$$= \int \mathrm{d}h_{t-1} \left[\left(\int \mathrm{d}h_1 \cdots \int \mathrm{d}h_{t-2}\, p(h_1) \prod_{s=1}^{t-2} p(y_s \mid h_s) p(h_{s+1} \mid h_s) \right) p(y_{t-1} \mid h_{t-1})\, p(h_t \mid h_{t-1}) \right]$$

$$= \int \mathrm{d}h_{t-1}\, \alpha_{t-1}(h_{t-1})\, p(y_{t-1} \mid h_{t-1})\, p(h_t \mid h_{t-1})$$

- We can compute these messages **recursively!**

Hardness of Exact EM for SLDS

Computing the forward messages $\alpha_t(h_t)$

- Now substitute $h_t = (z_t, x_t)$.
- Base case:

$$
\begin{aligned}
\alpha_1(z_1, x_1) &= p(z_1)\, p(x_1 \mid z_1) \\
&= \sum_{k=1}^{K} \left[\mathrm{Cat}(z_1 \mid \pi)\, \mathcal{N}(x_1 \mid b_k, Q_k)\right]^{\mathbb{I}[z_1=k]}
\end{aligned}
$$

Hardness of Exact EM for SLDS

Computing the forward messages $\alpha_t(h_t)$

- Now substitute $h_t = (z_t, x_t)$.
- Base case:

$$\begin{aligned}\alpha_1(z_1, x_1) &= p(z_1)\, p(x_1 \mid z_1) \\ &= \sum_{k=1}^{K} \left[\mathrm{Cat}(z_1 \mid \pi)\, \mathcal{N}(x_1 \mid b_k, Q_k)\right]^{\mathbb{I}[z_1=k]}\end{aligned}$$

- Second time step:

$$\begin{aligned}\alpha_2(z_2, x_2) &= \sum_{z_1=1}^{K} \int \mathrm{d}x_1\, \alpha_1(z_1, x_1)\, p(y_1 \mid x_1)\, p(z_2 \mid z_1)\, p(x_2 \mid x_1, z_2) \\ &= \sum_{z_1=1}^{K} \sum_{k=1}^{K} \left[\mathrm{Cat}\left(z_2 \mid \rho(z_1, y_1)\right) \mathcal{N}\left(x_2 \mid \mu(z_1, z_2, y_1), \Sigma(z_1, z_2, y_1)\right)\right]^{\mathbb{I}[z_2=k]}\end{aligned}$$

Hardness of Exact EM for SLDS

Computing the forward messages $\alpha_t(h_t)$

- By the t-th time step,

$$\alpha_t(z_t, x_t) = \sum_{z_1=1}^{K} \cdots \sum_{z_{t-1}=1}^{K} \sum_{k=1}^{K} \left[\mathrm{Cat}\left(z_t \mid \rho(z_{1:t-1}, y_{1:t-1})\right) \mathcal{N}\left(x_t \mid \mu(z_t, z_{1:t-1}, y_{1:t-1}), \Sigma(z_t, z_{1:t-1}, y_{1:t-1})\right) \right]^{\mathbb{I}[z_t=k]}$$

- This is still a mixture of Gaussians.

- **Question:** How many components does it have?

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**):

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

- $\mathscr{L}$ 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

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

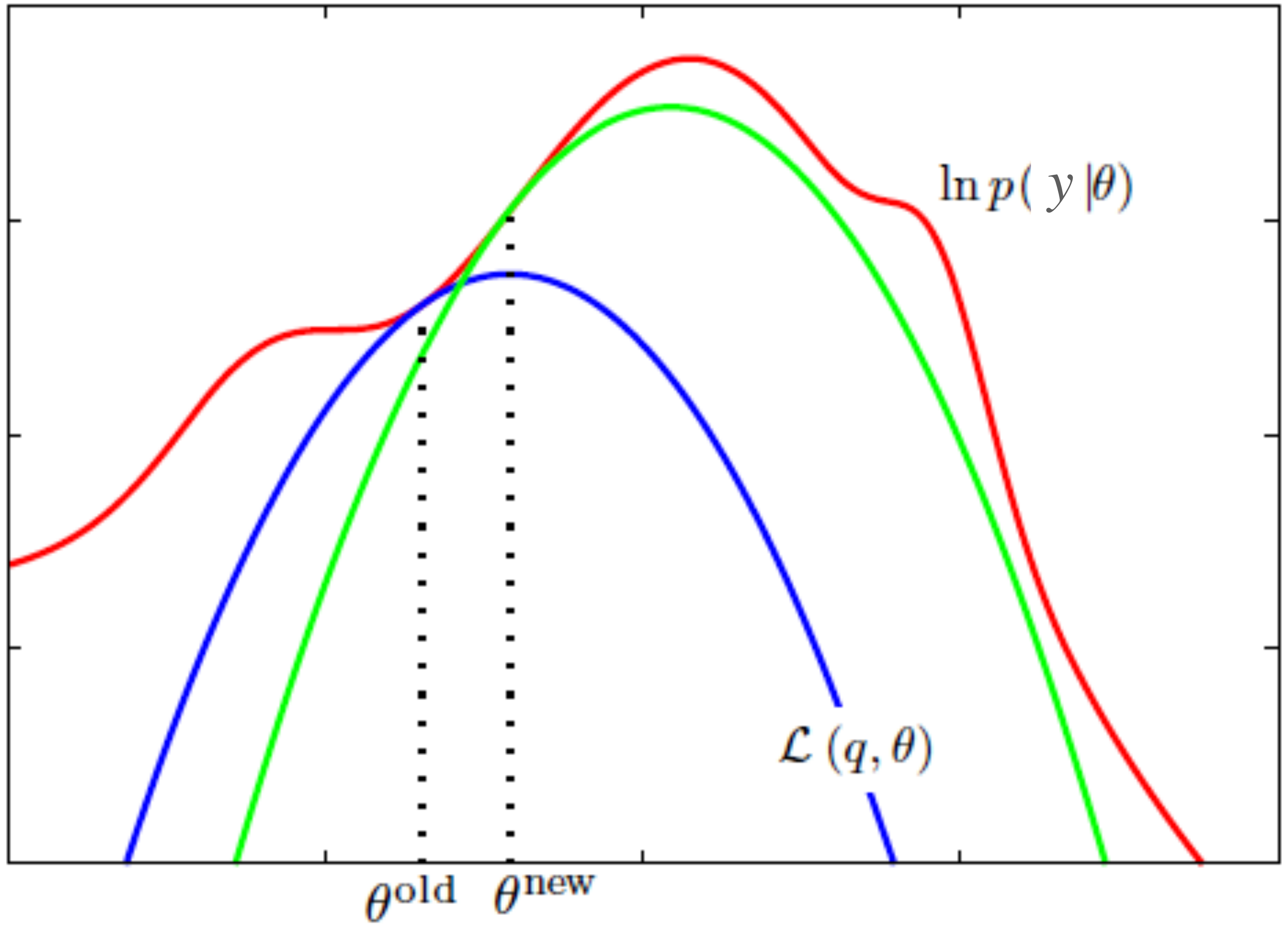
$$= \arg\min_q \mathrm{KL}\left(q(z) \,\|\, p(z \mid y, \Theta)\right)$$

$$= p(z \mid y, \Theta)$$

- 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 $q \in Q$

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

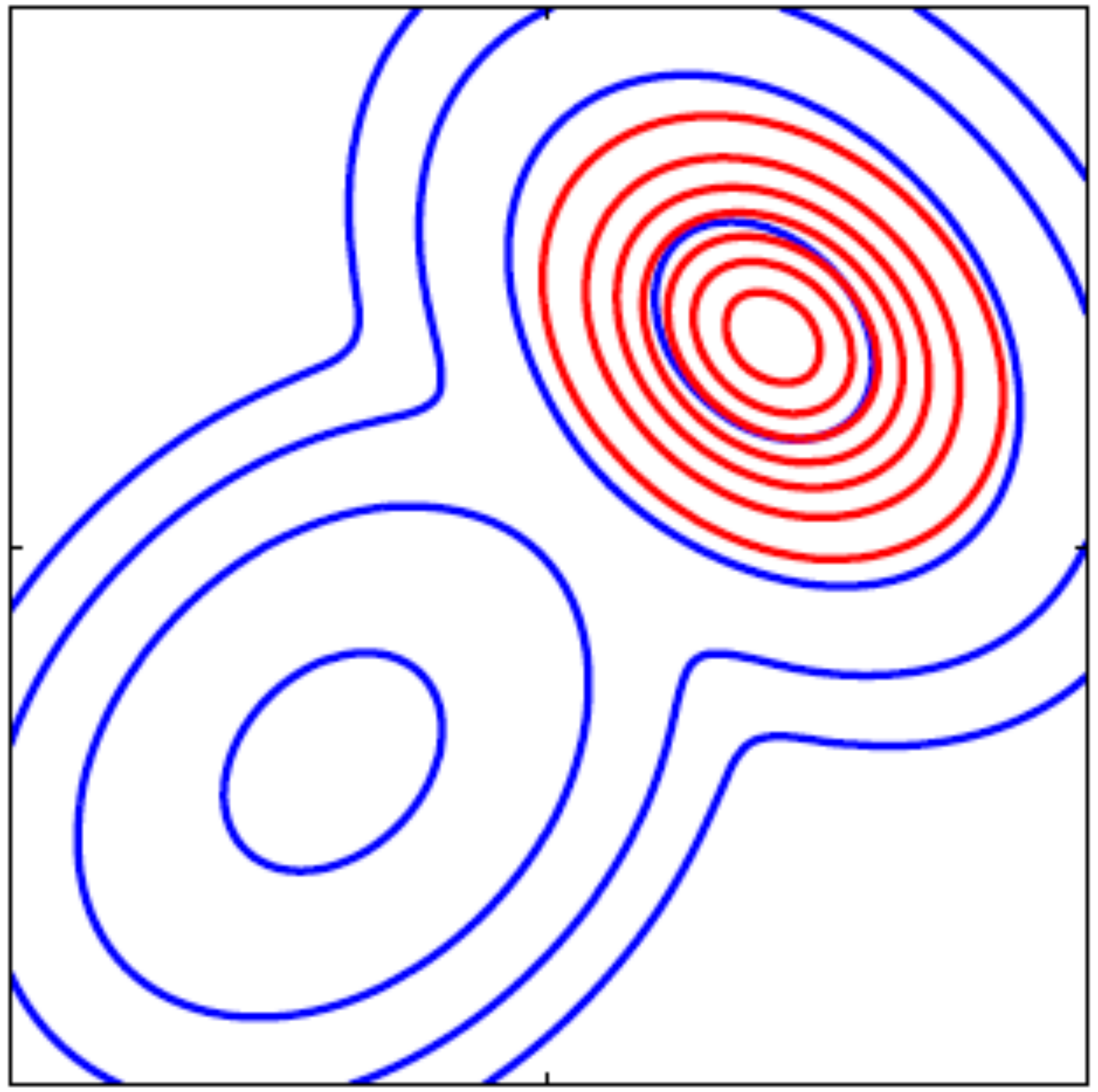
$$= \arg\min_{q \in \mathcal{Q}} \mathrm{KL}\left(q(z) \,\|\, p(z \mid y, \Theta)\right)$$

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

- For example, $\mathcal{Q}$ could **assume independence** or a **particular functional form.**

- If $\mathcal{Q}$ does not contain the true posterior, the ELBO will be a strict lower bound on the marginal likelihood, $\mathcal{L}[q(z), \Theta] < \log p(y \mid \Theta)$.

- 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)

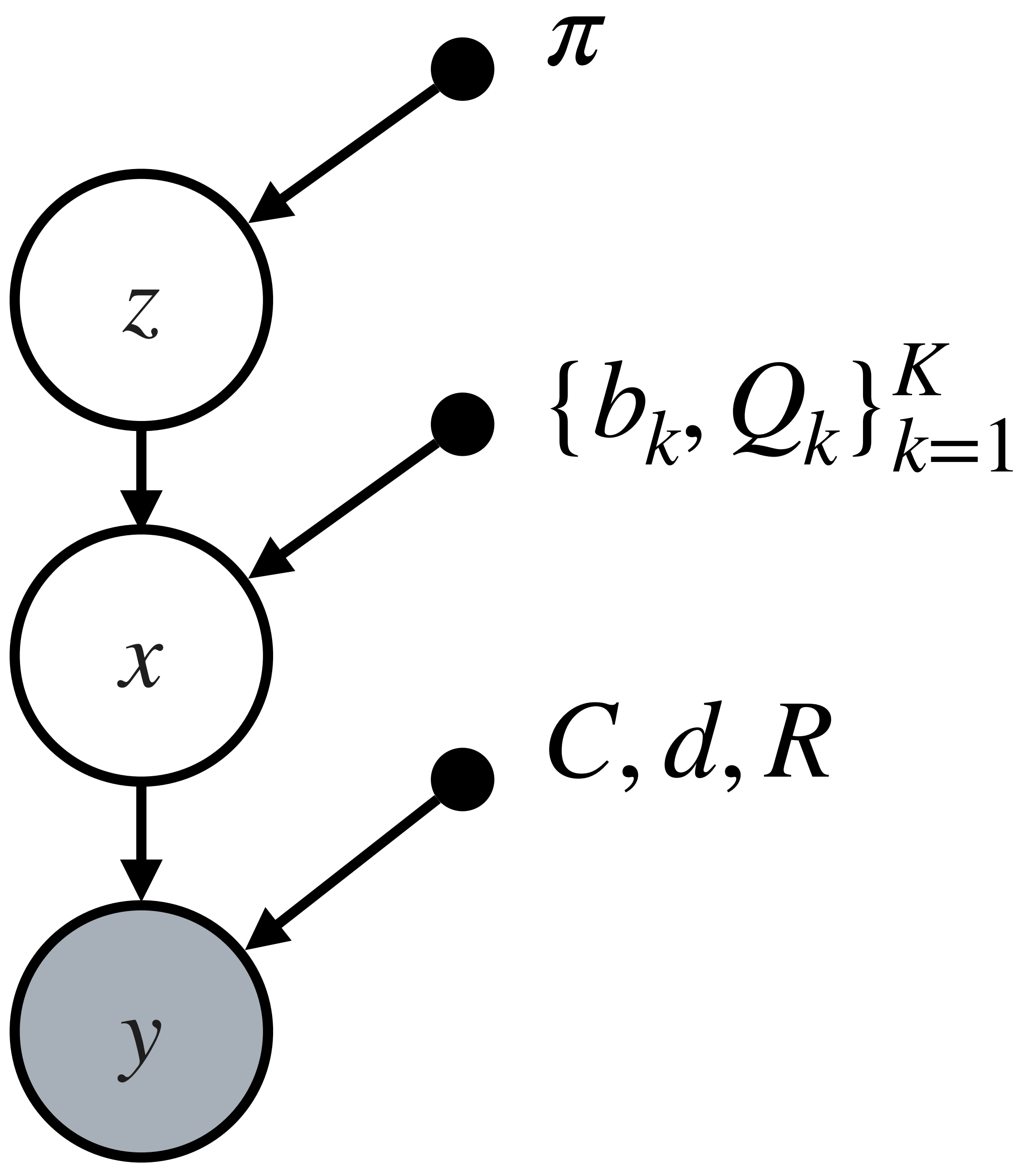
Coordinate ascent VI

Warm-up example

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

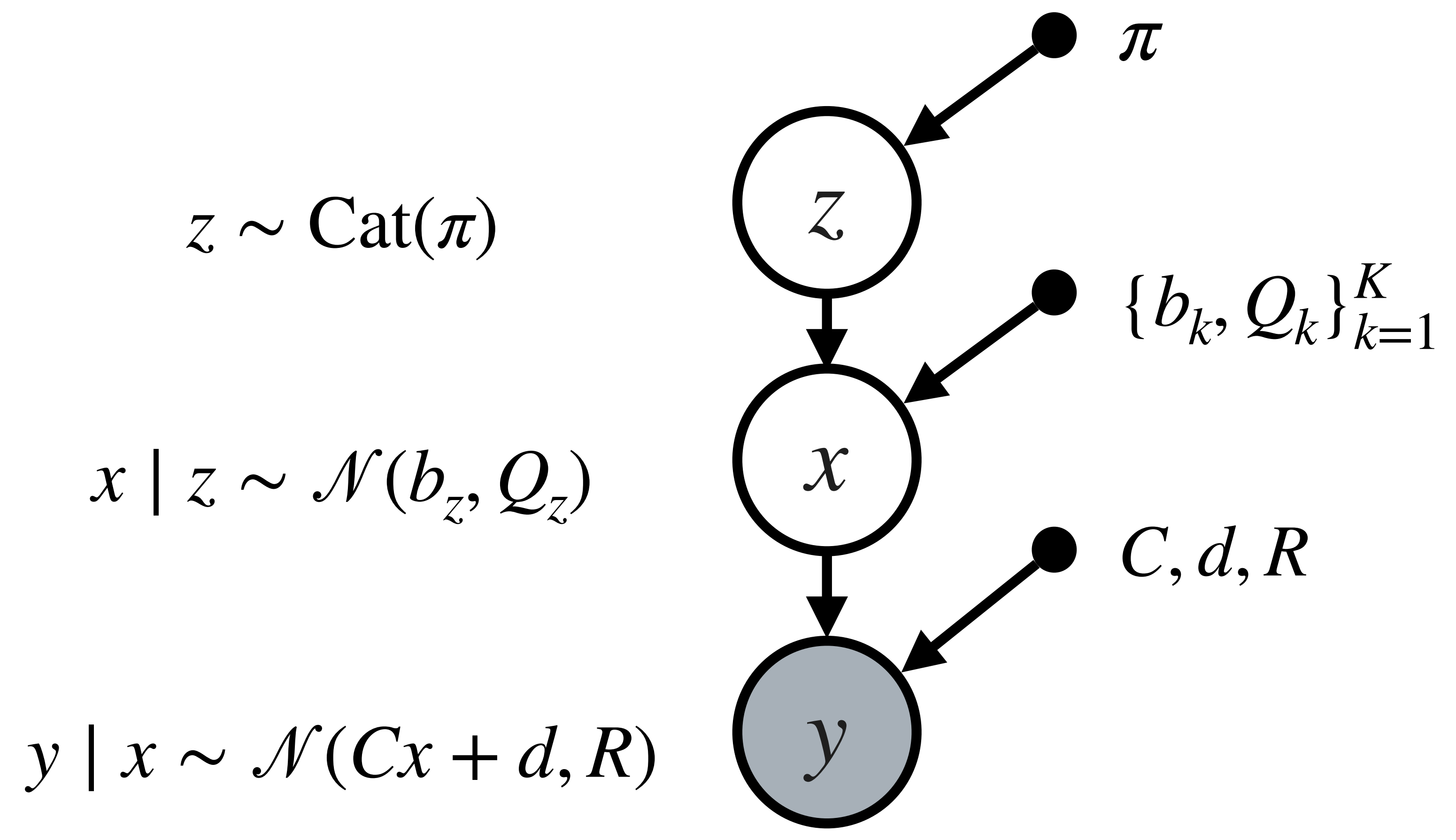
$$x \mid z \sim \mathcal{N}(b_z, Q_z)$$

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



Coordinate ascent VI

Mean field variational family



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

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

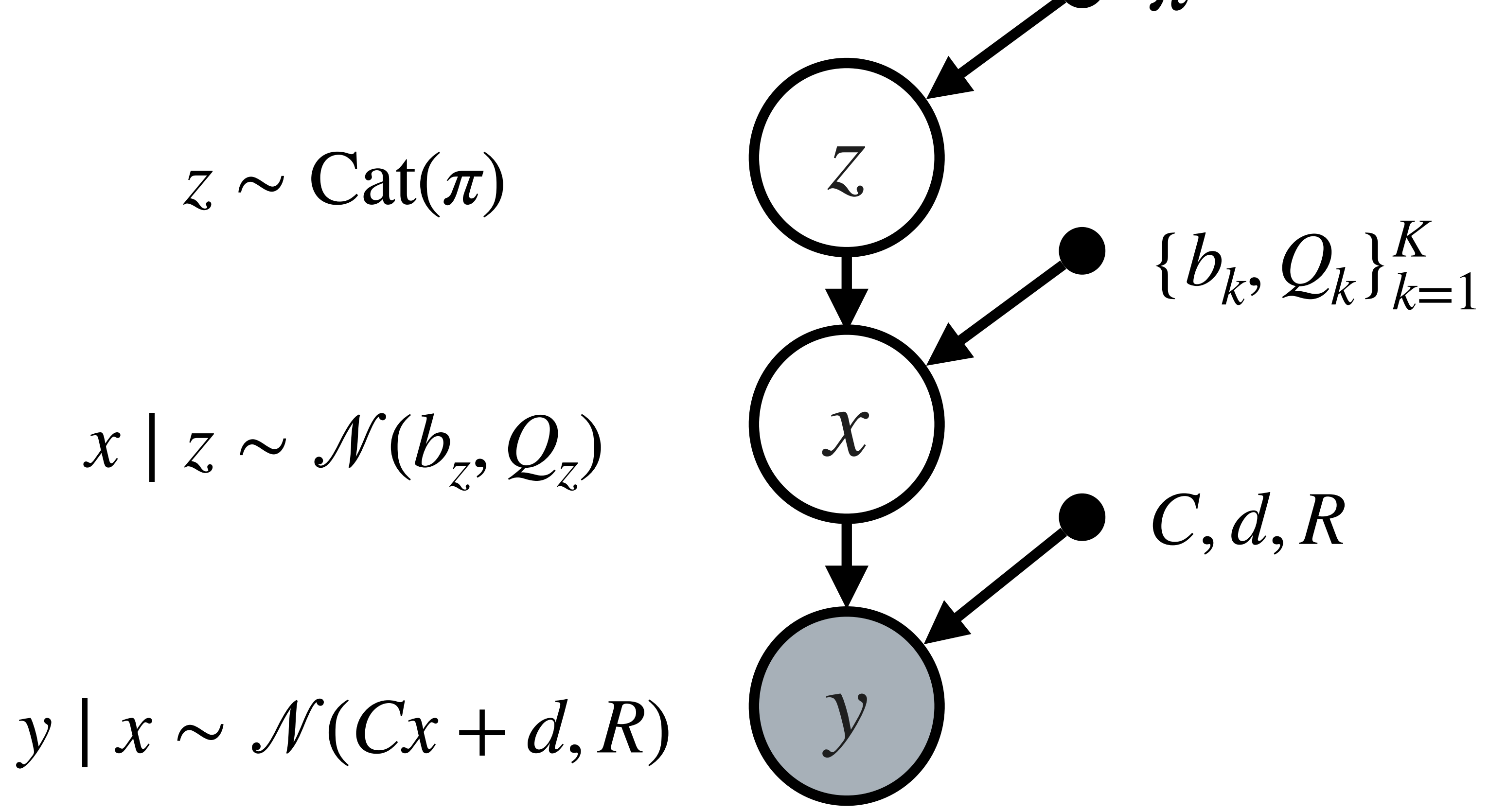
where z and x are independent.

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

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

Coordinate ascent VI

Coordinate updates

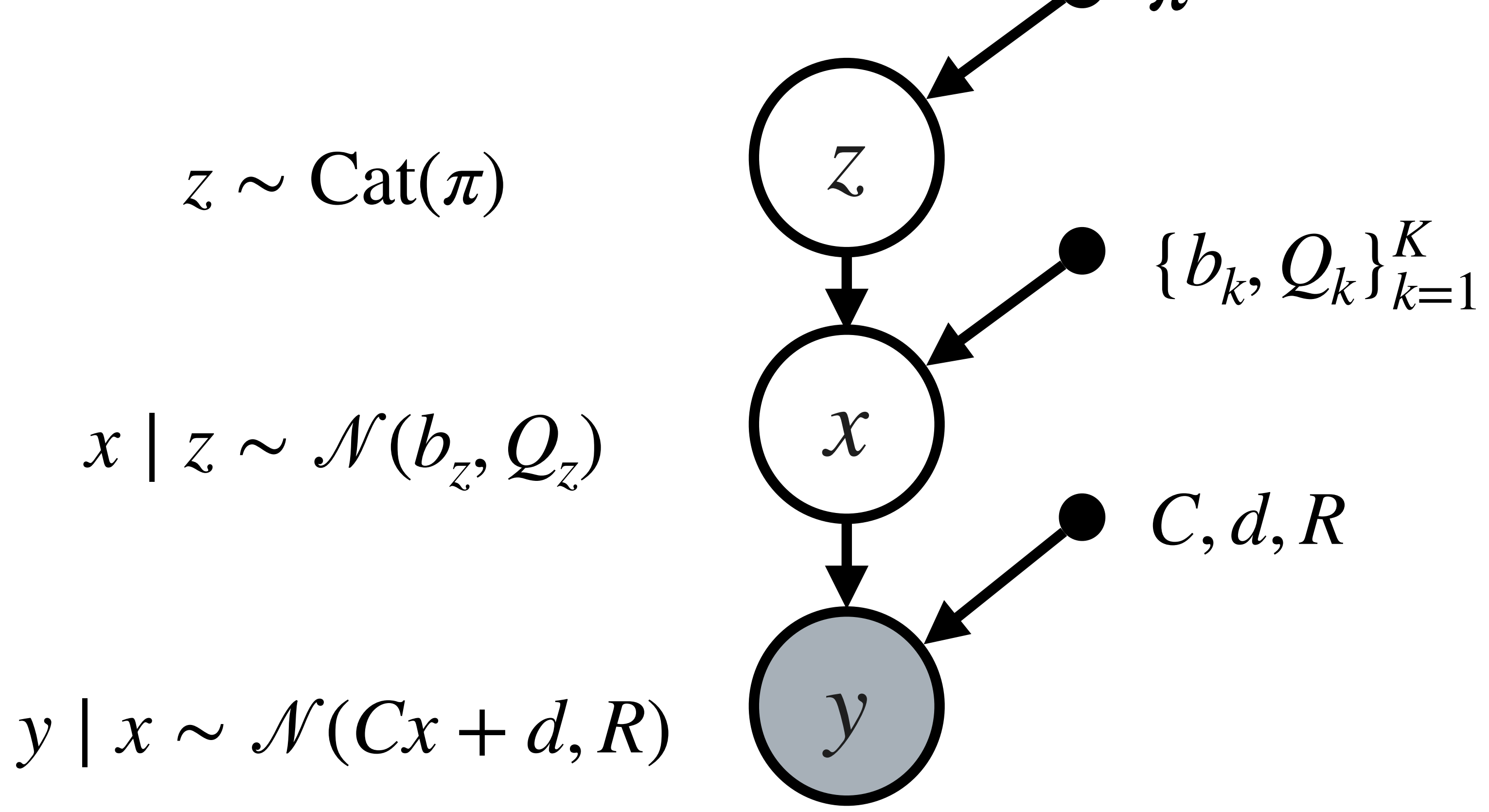


Hold $q(x)$ fixed and optimize w.r.t. $q(z)$:

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

Coordinate ascent VI

Coordinate updates



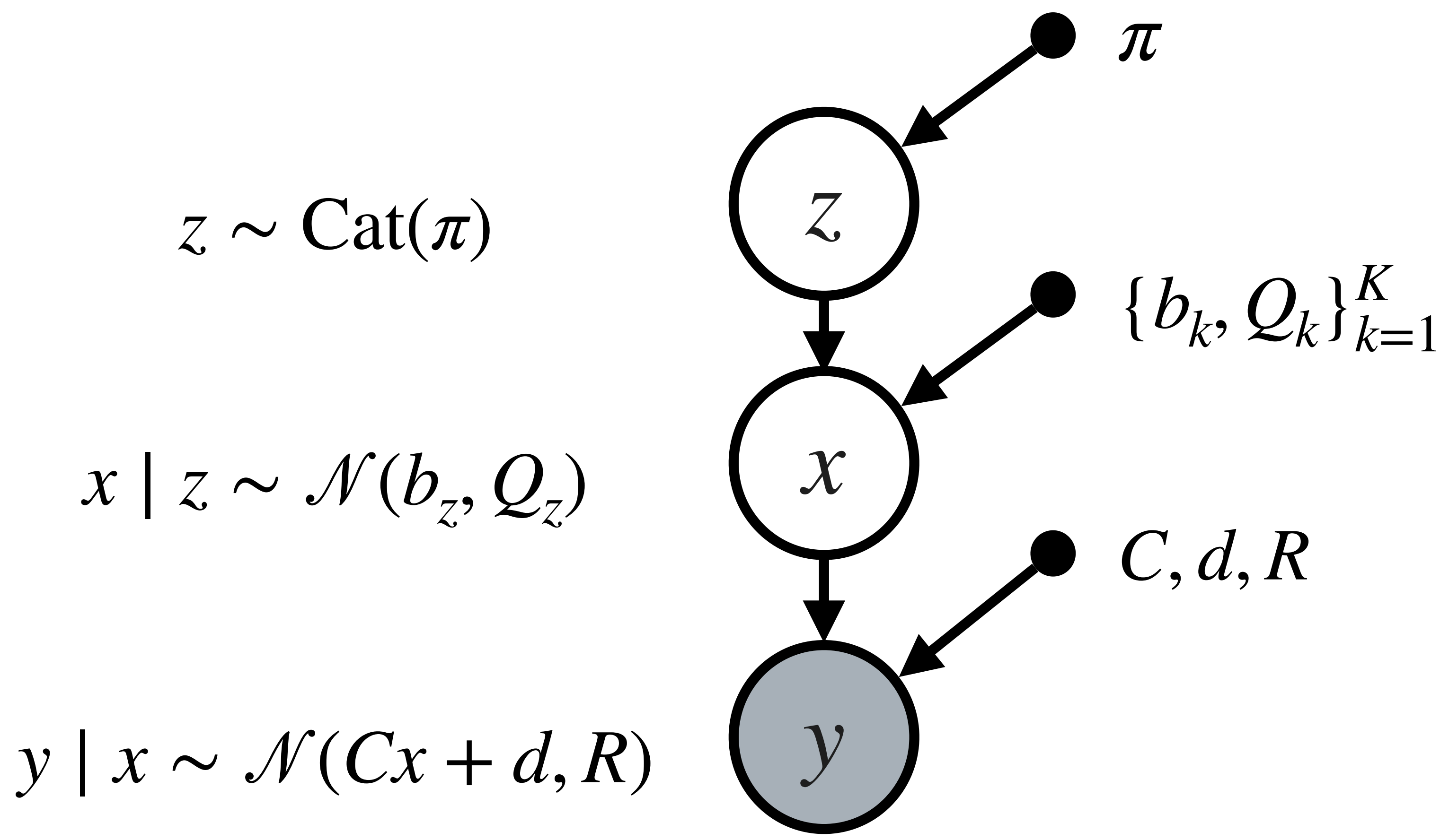
Hold $q(x)$ fixed and optimize w.r.t. $q(z)$:

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

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

Coordinate ascent VI

Coordinate updates

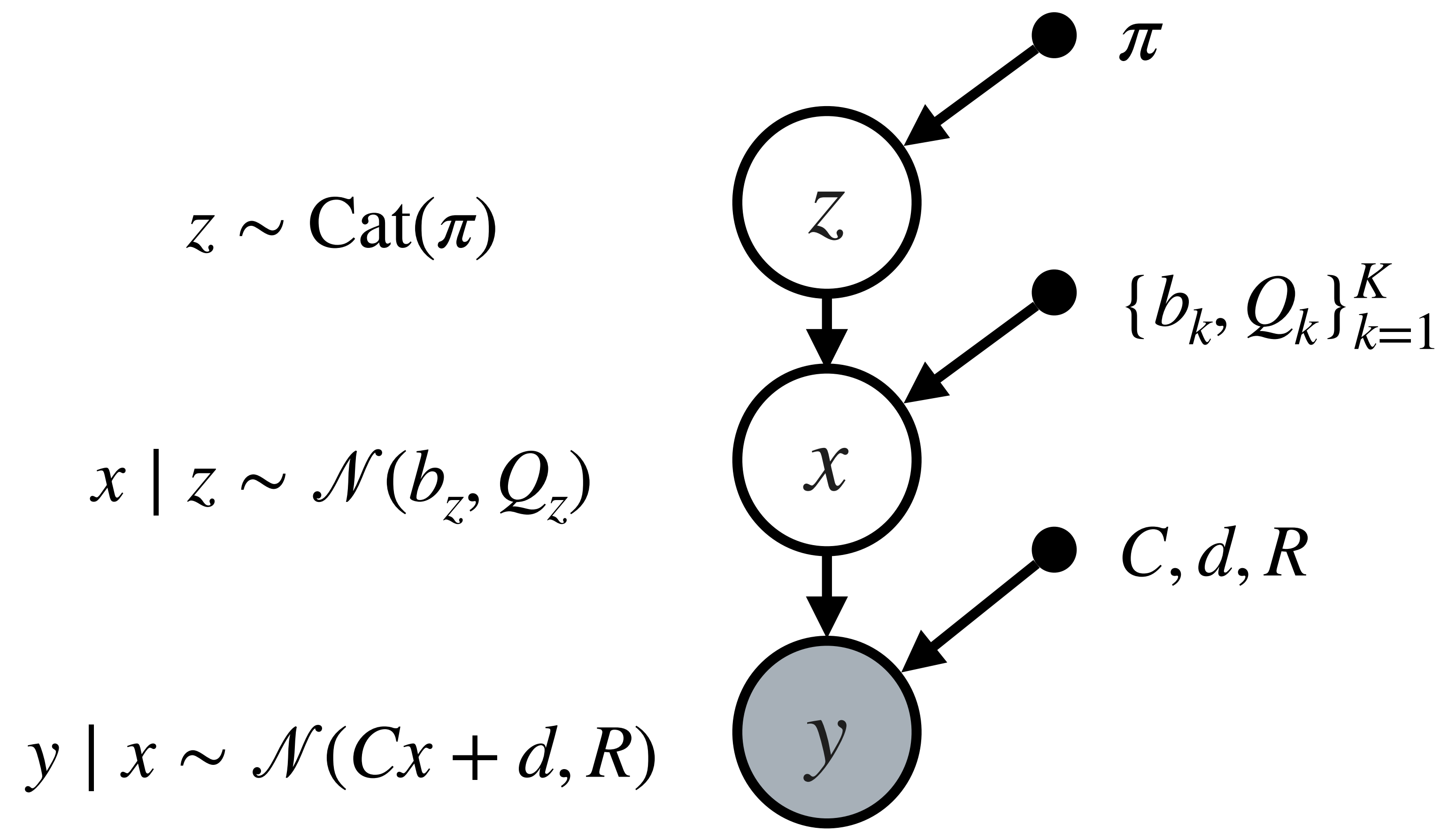


Hold $q(x)$ fixed and optimize w.r.t. $q(z)$:

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

Coordinate ascent VI

Coordinate updates

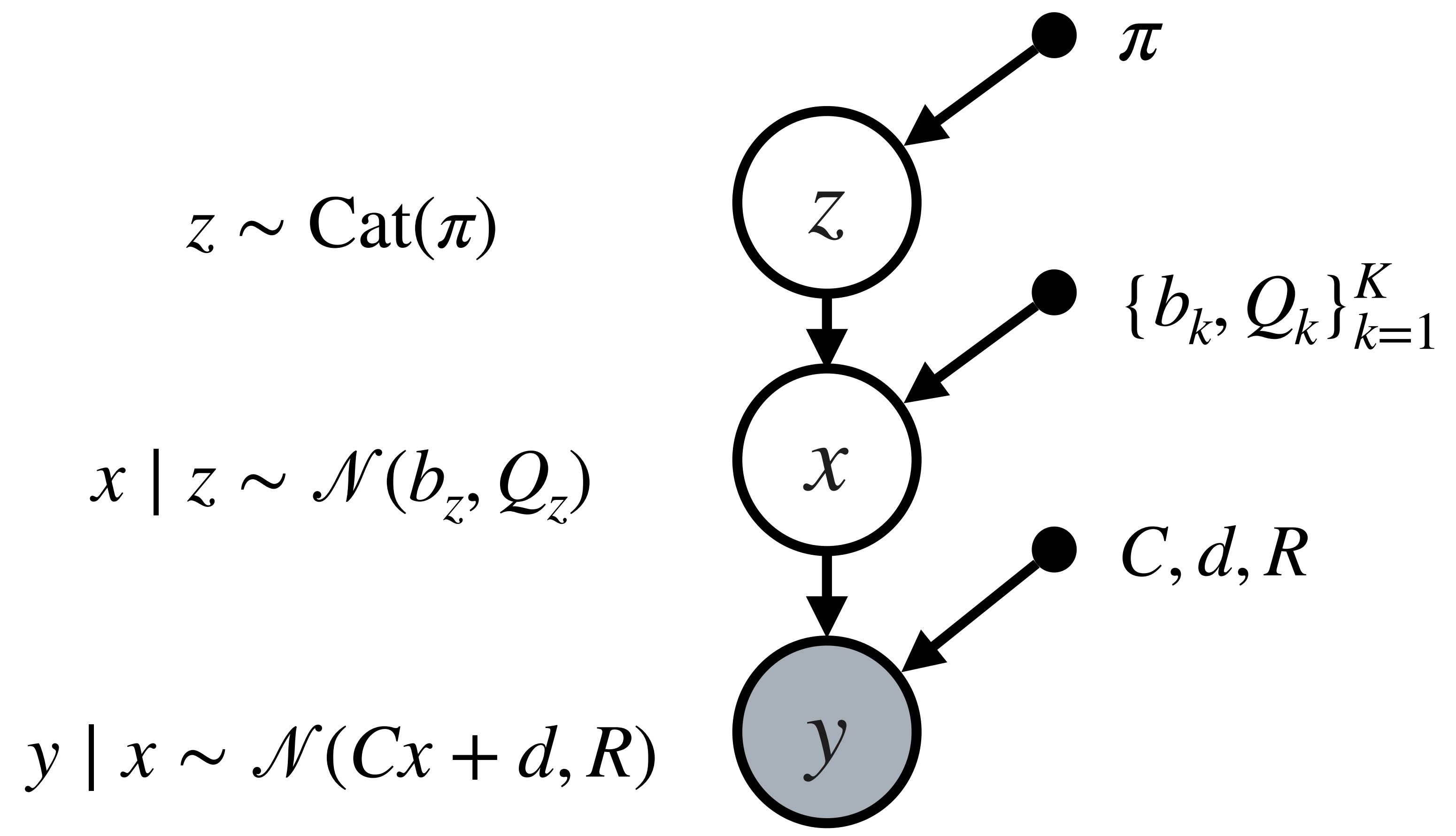


Hold $q(x)$ fixed and optimize w.r.t. $q(z)$:

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

Coordinate ascent VI

Coordinate updates

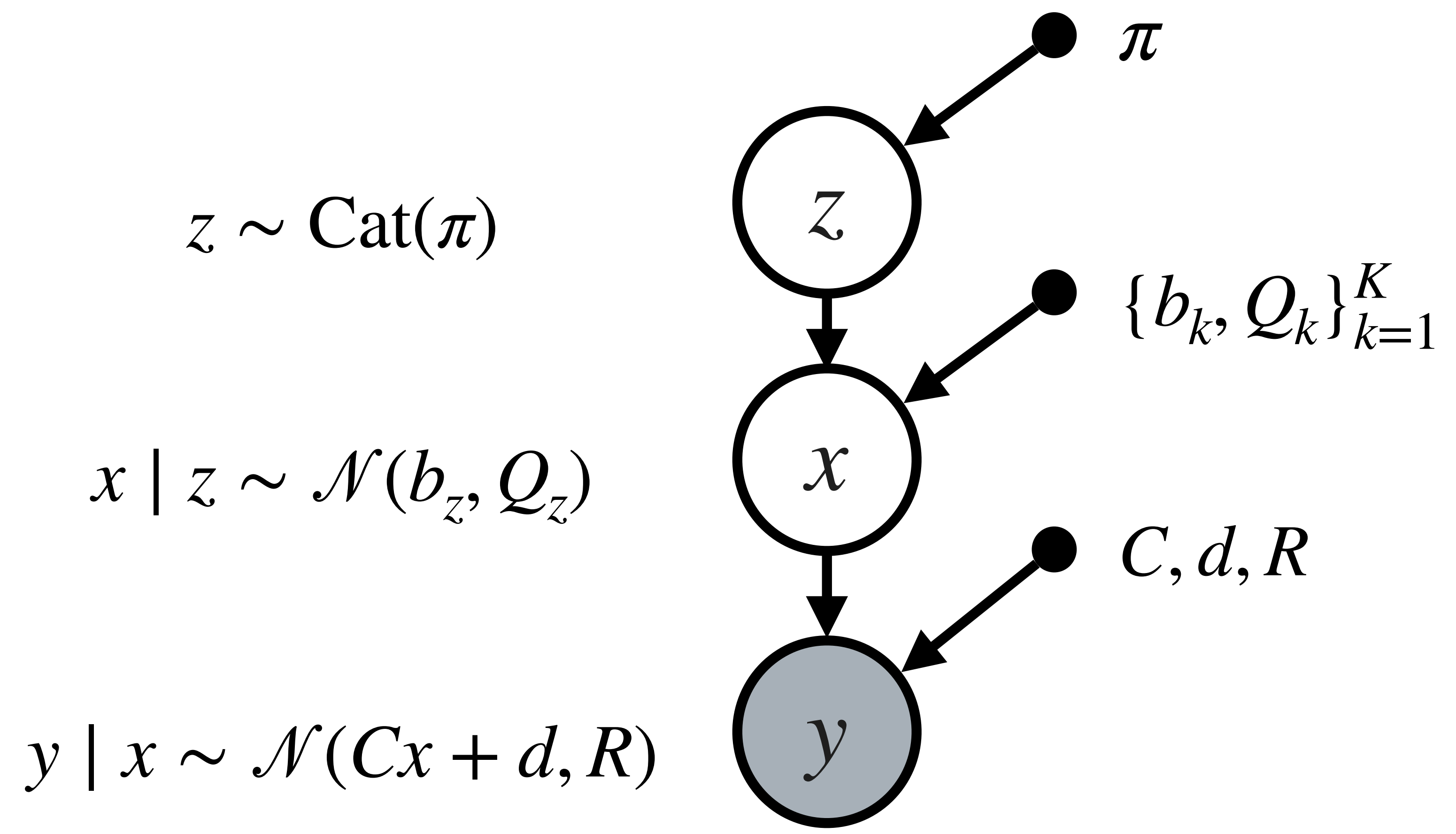


Hold $q(x)$ fixed and optimize w.r.t. $q(z)$:

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

Coordinate ascent VI

General form



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

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

- 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,

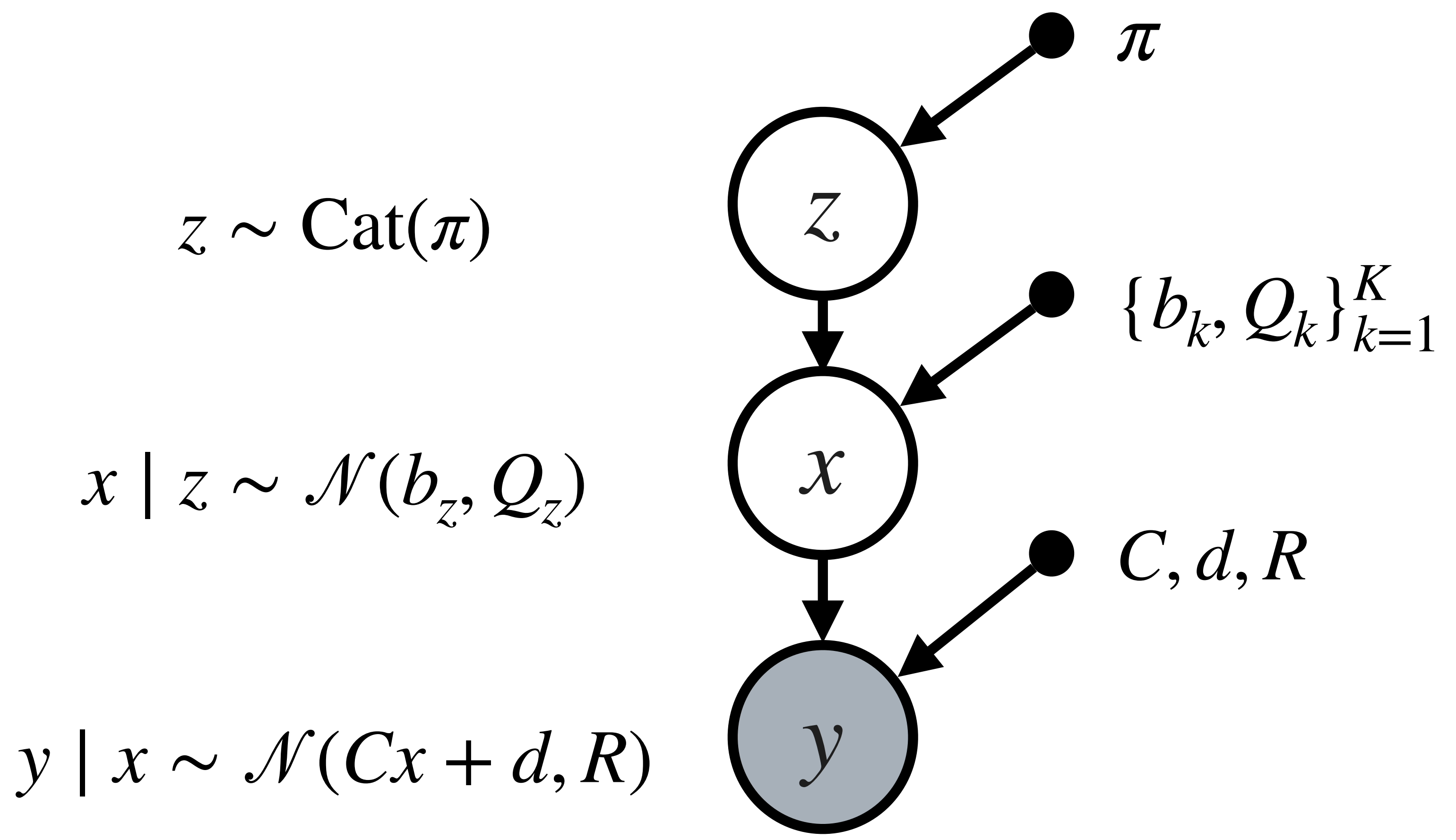
$$q(x_i) \propto \exp\left\{\mathbb{E}_{q(x_{\neg i})}\left[\log p(x_1, \ldots, x_i, \ldots, x_D, y \mid \Theta)\right]\right\}$$

Coordinate ascent VI

Closed form updates for $q(z)$

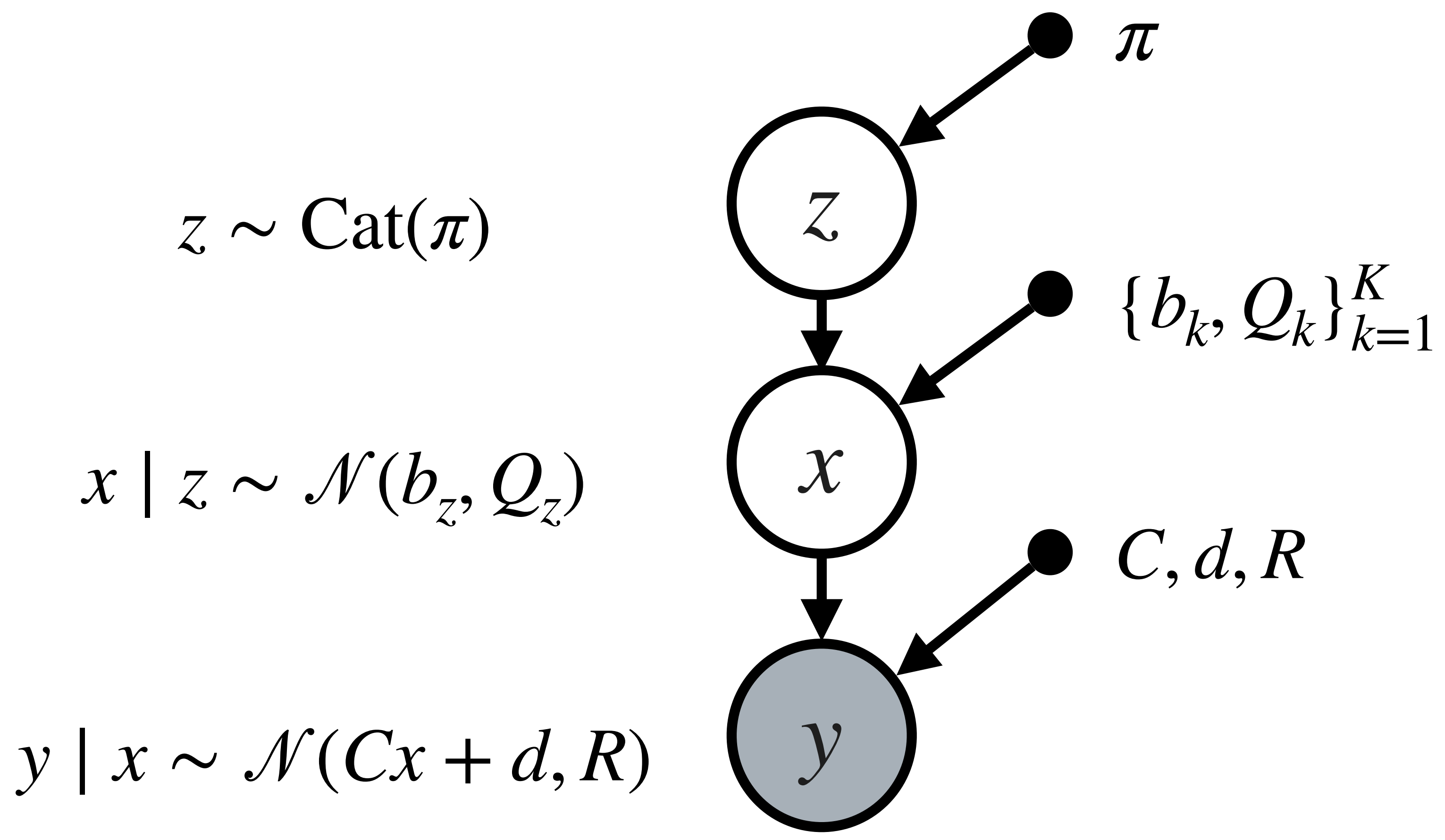
The optimal updates are often available in closed form,

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



Coordinate ascent VI

Closed form updates for $q(z)$

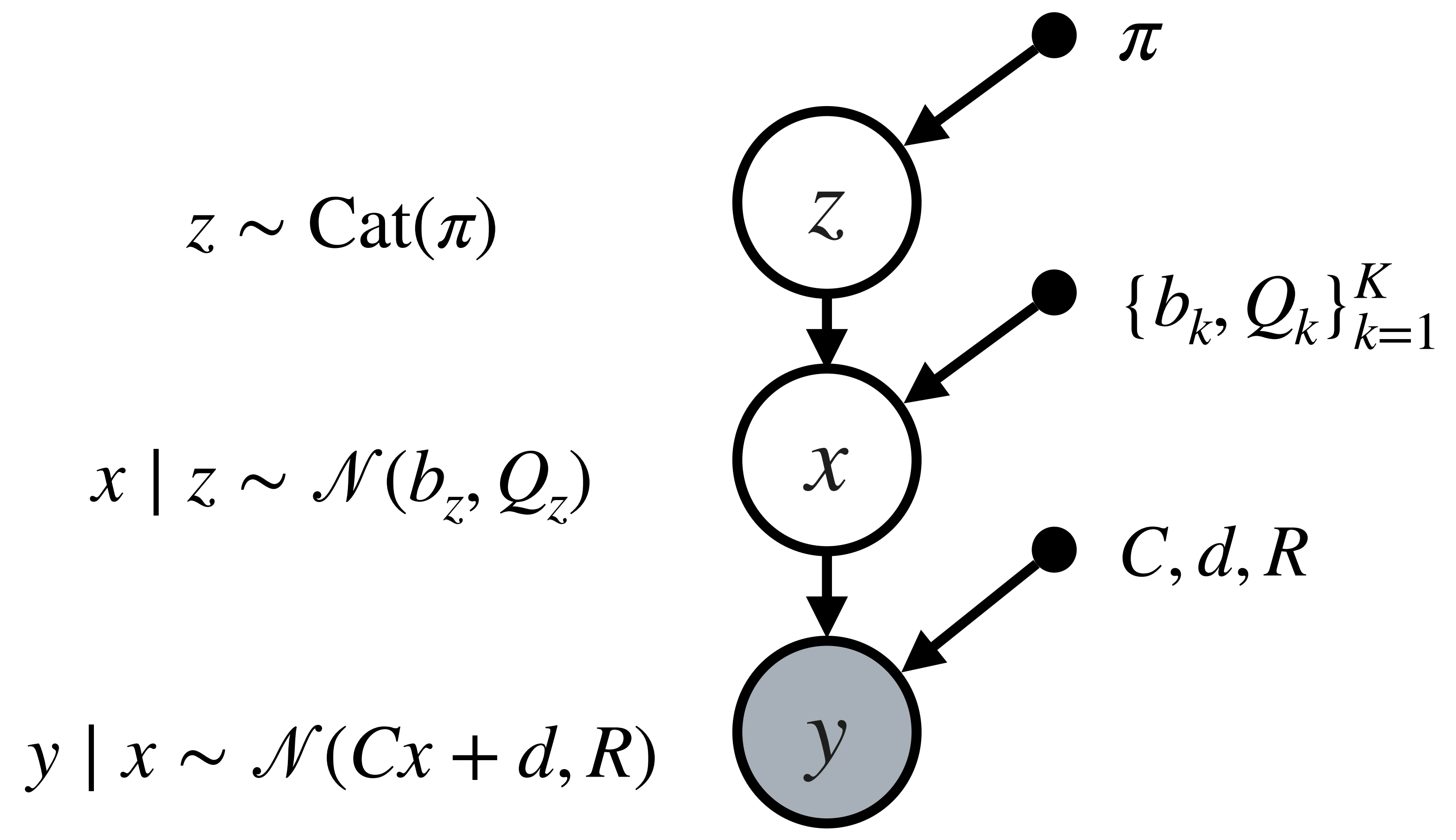


The optimal updates are often available in closed form,

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

Coordinate ascent VI

Closed form updates for $q(z)$

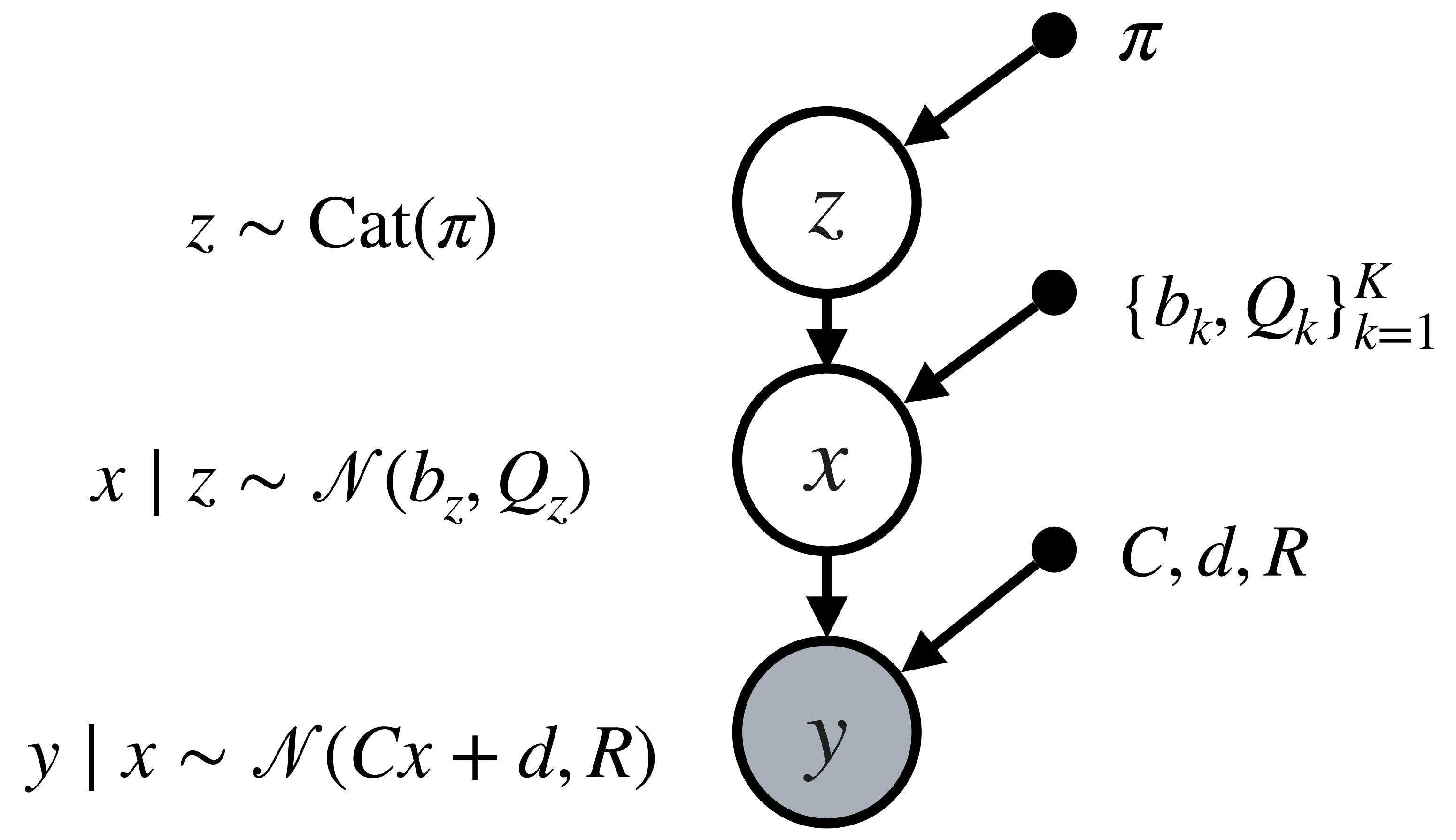


The optimal updates are often available in closed form,

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

Coordinate ascent VI

Closed form updates for $q(z)$



The optimal updates are often available in closed form,

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

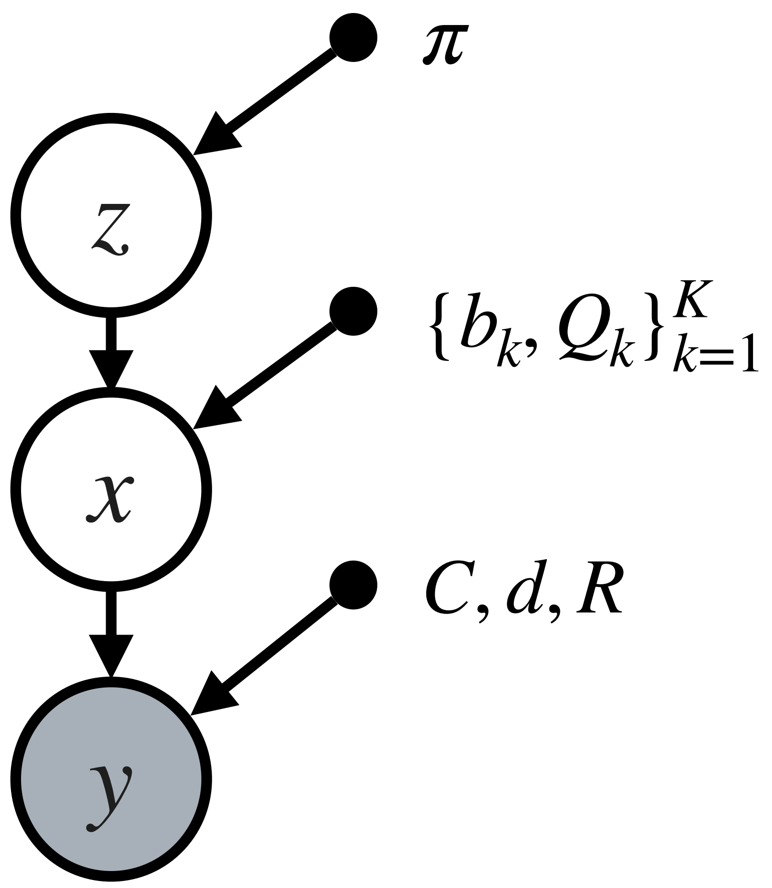
Coordinate ascent VI

Closed form updates for $q(z)$

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

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

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



The optimal updates are often available in closed form,

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

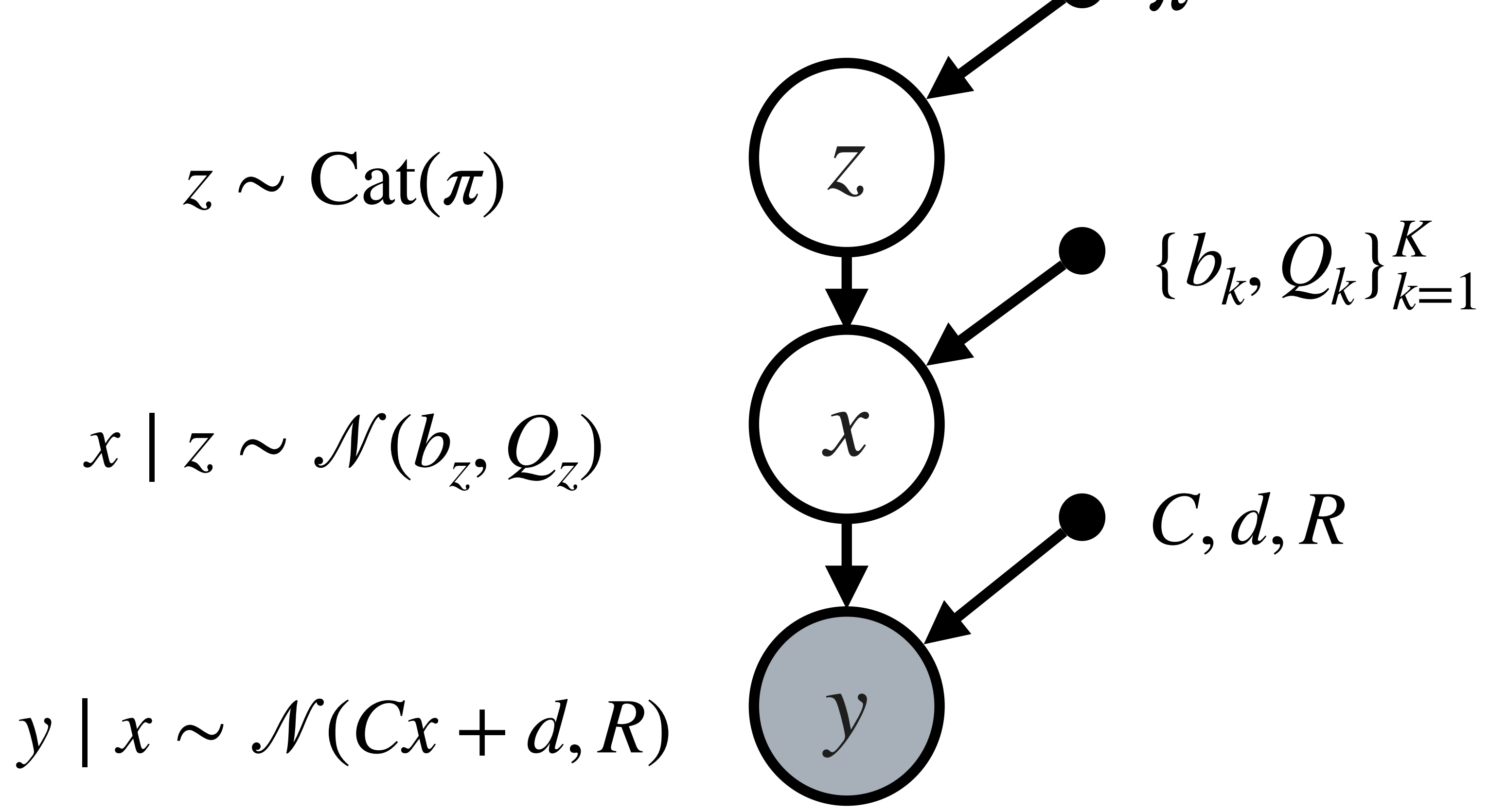
where

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

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 $q(x)$

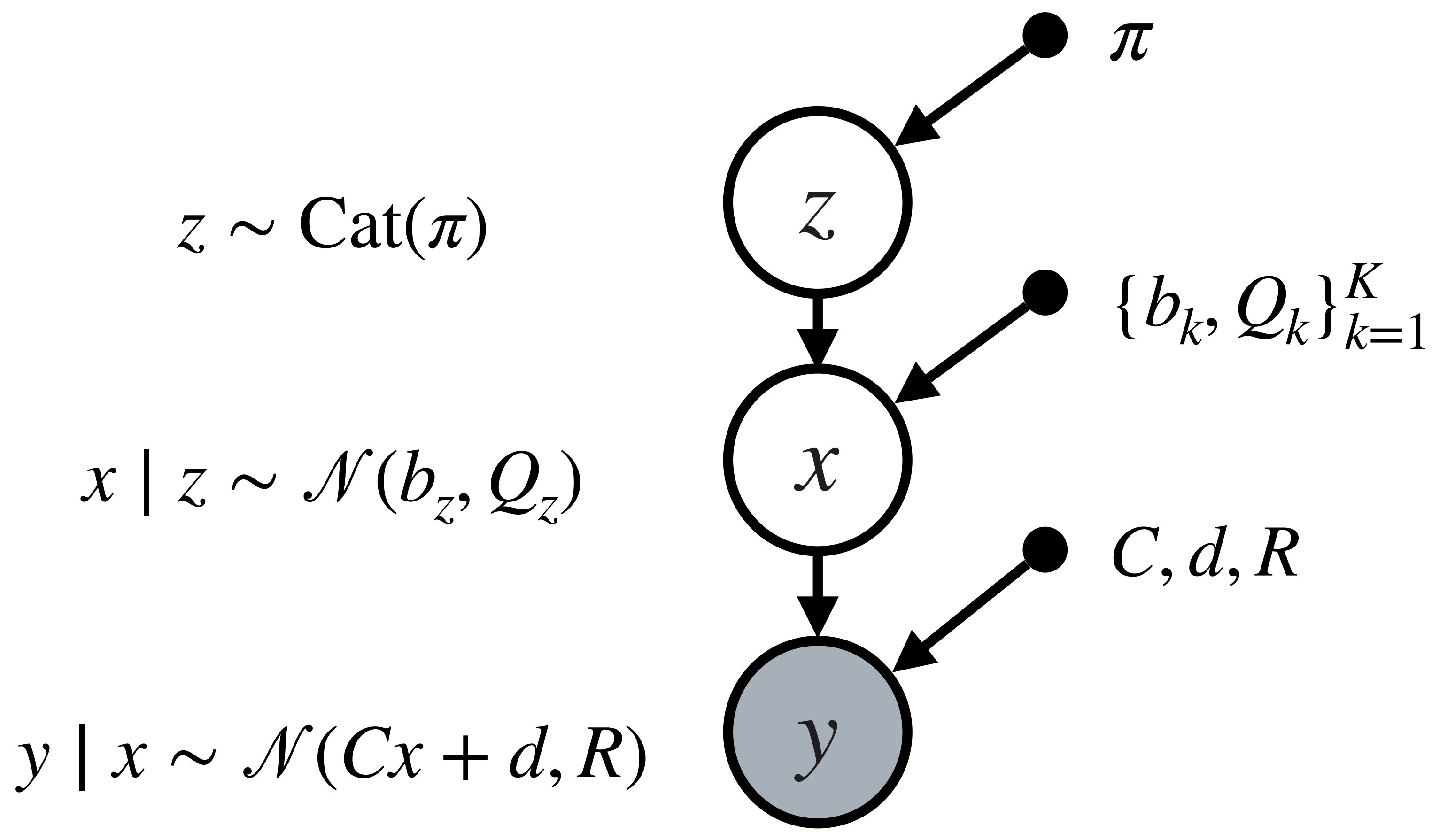


Now do the same for $q(x)$:

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

Coordinate ascent VI

Closed form updates for $q(x)$

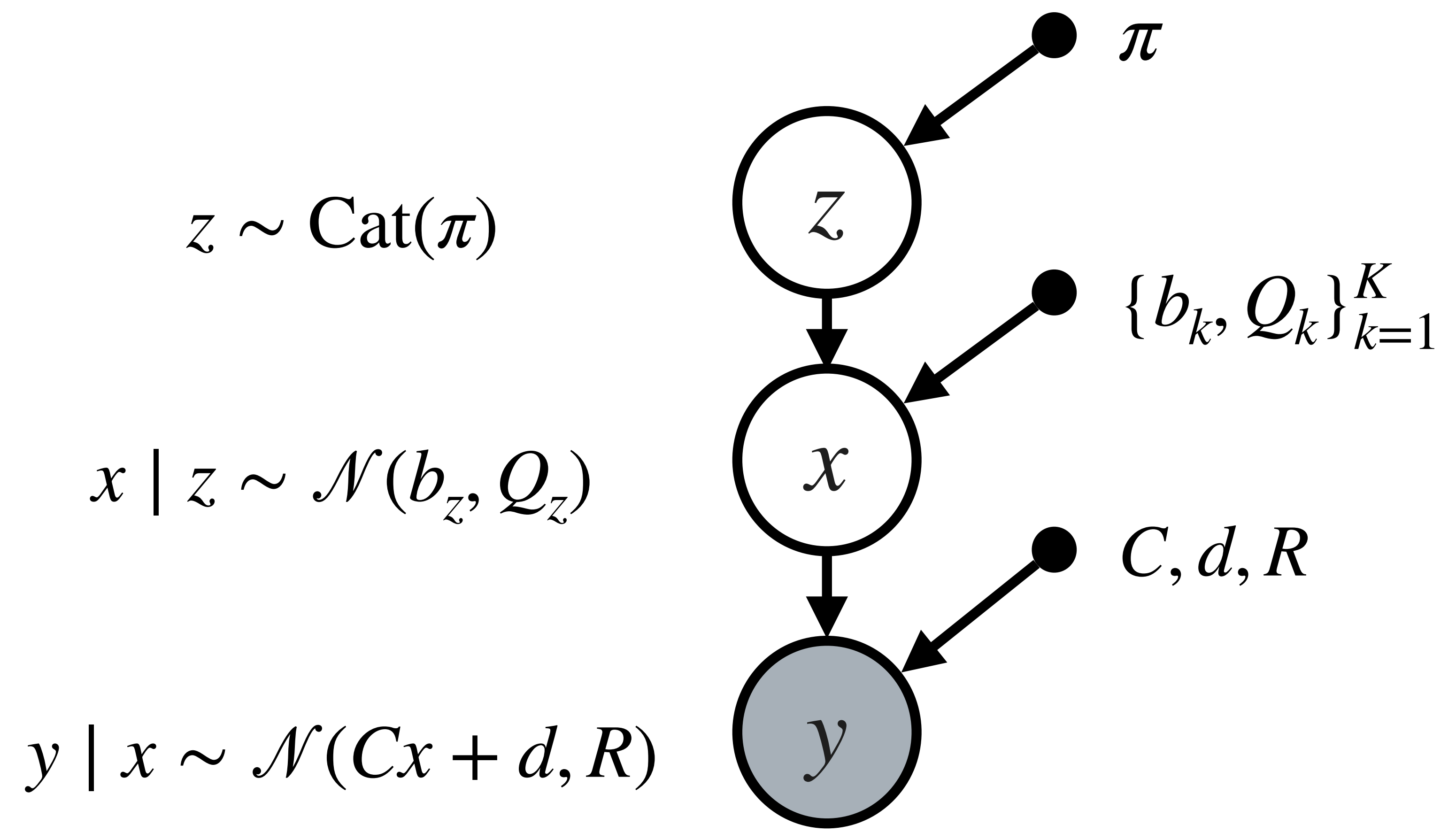


Now do the same for $q(x)$:

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

Coordinate ascent VI

Closed form updates for $q(x)$



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

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

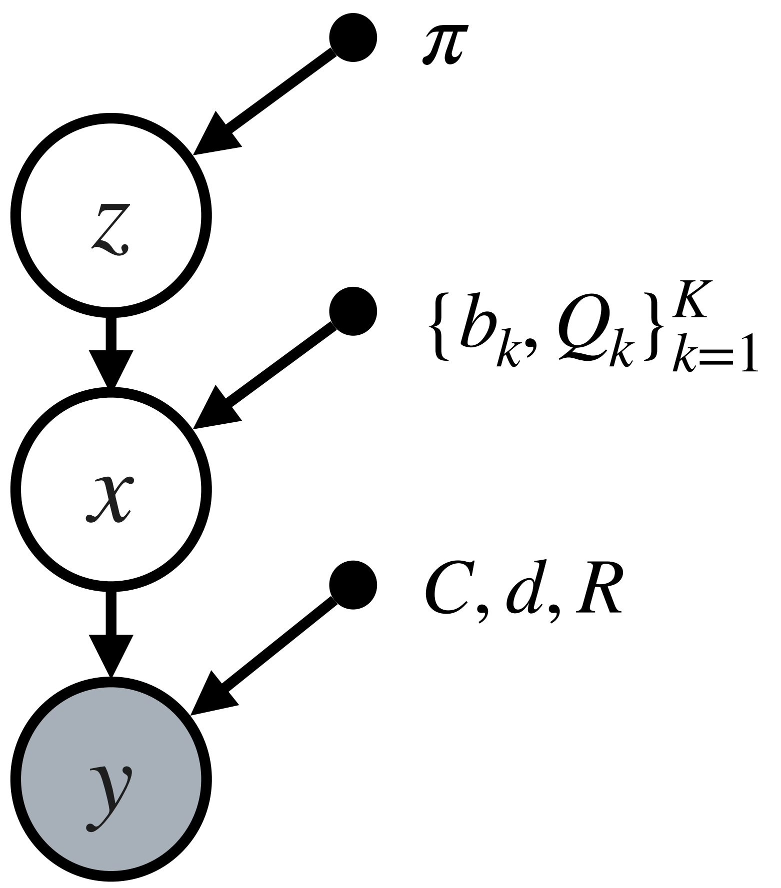
Coordinate ascent VI

Closed form updates for $q(x)$

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

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

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

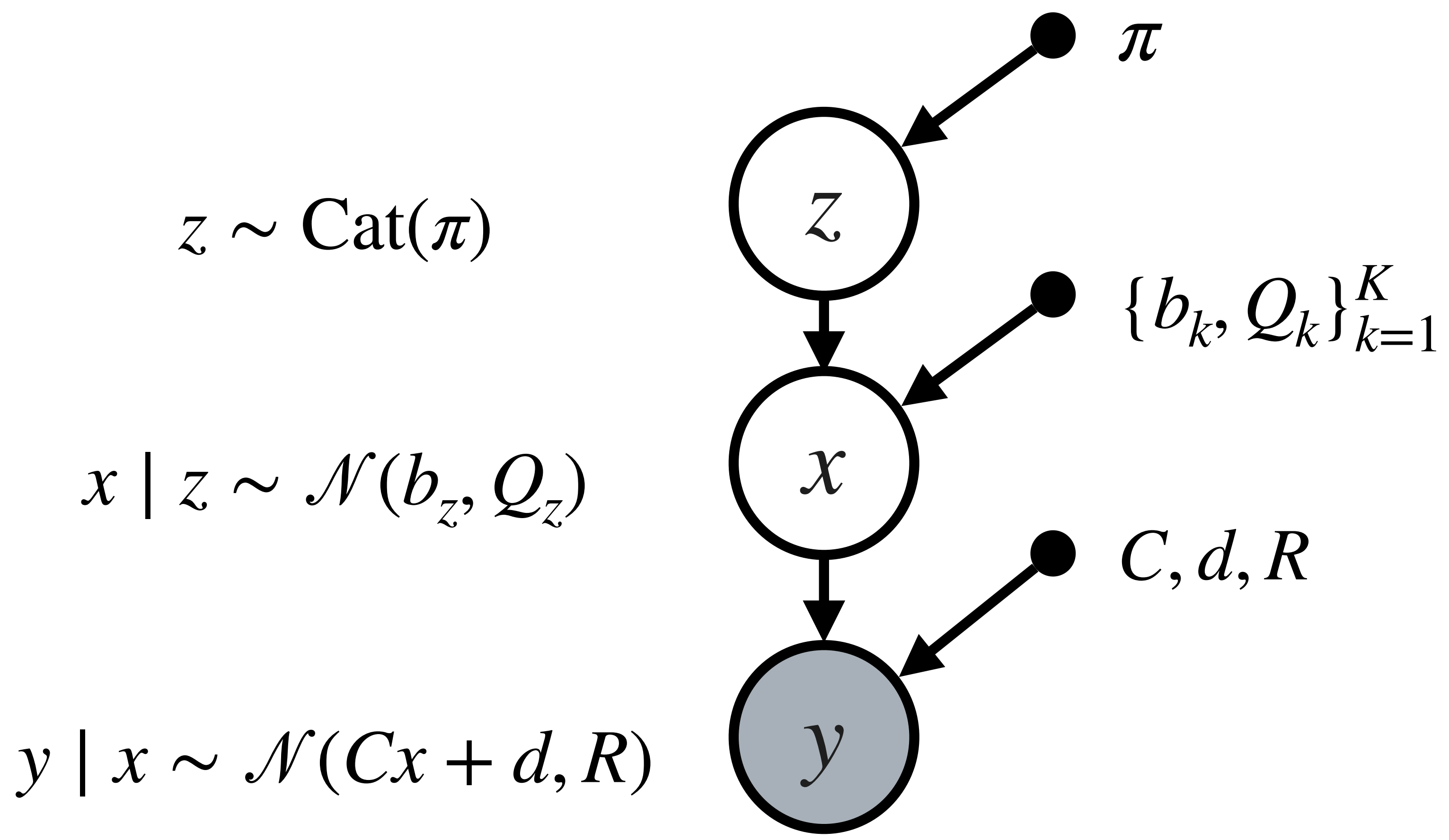


Now do the same for $q(x)$:

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

Coordinate ascent VI

Closed form updates for $q(x)$

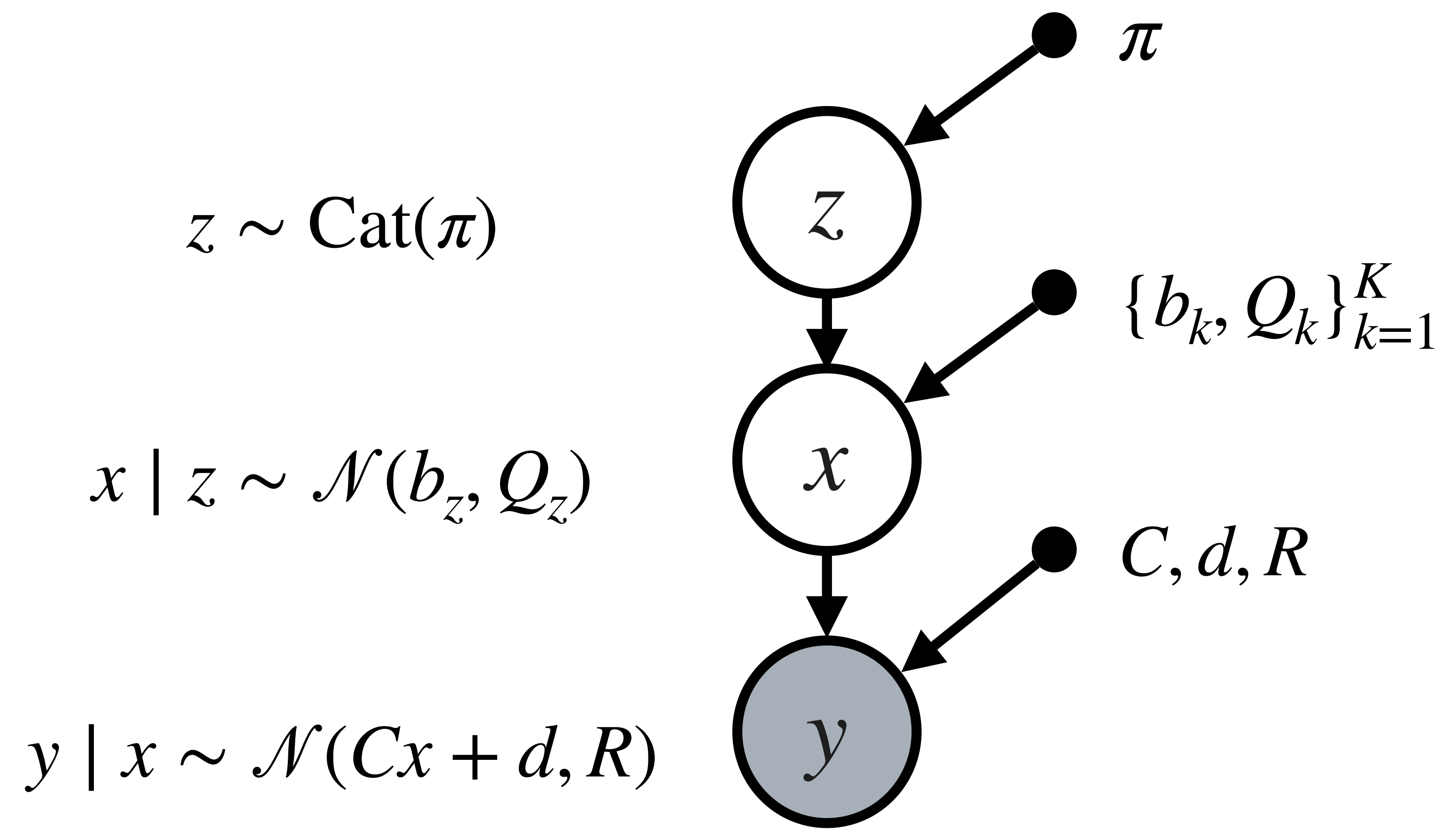


Now do the same for $q(x)$:

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

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

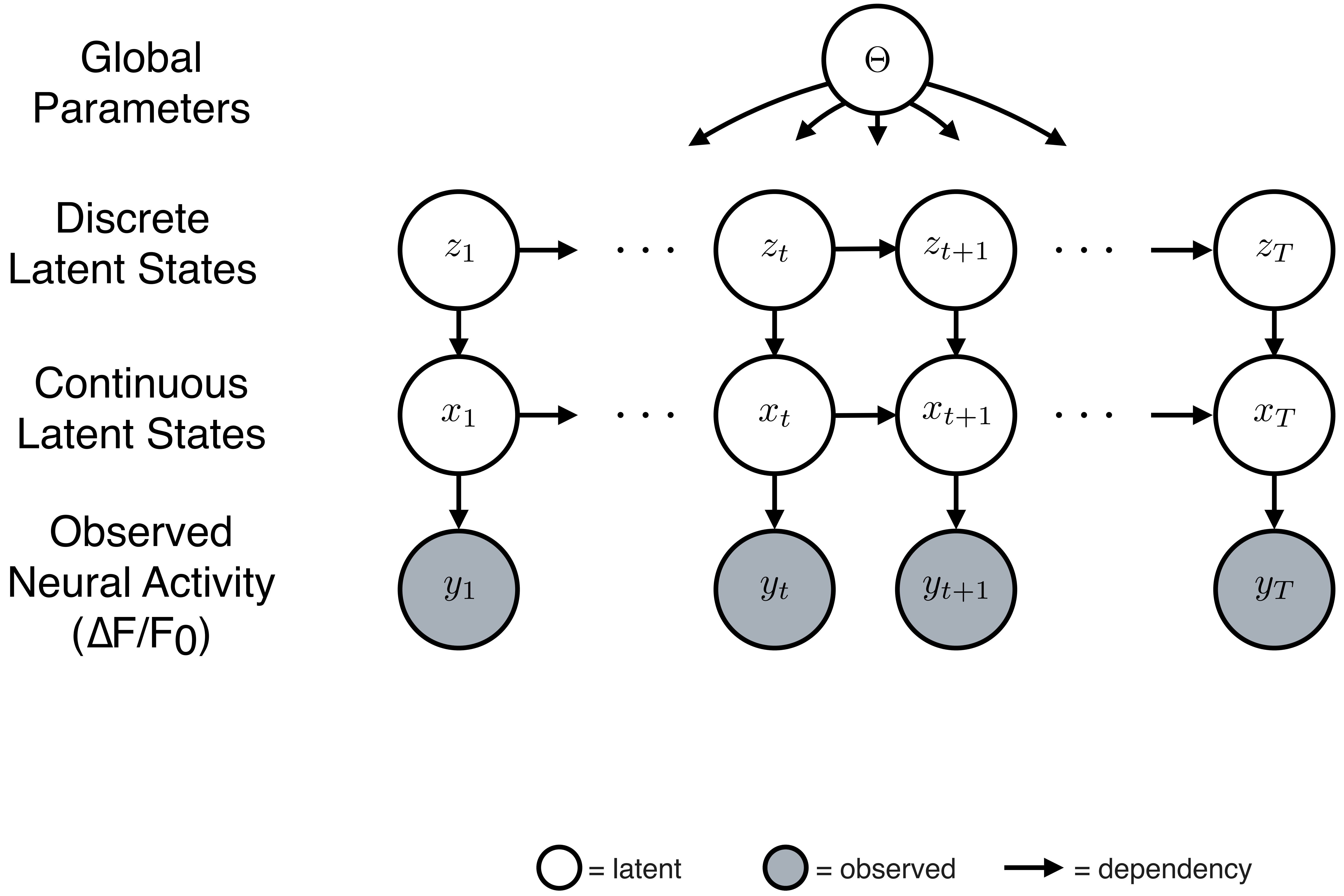
$$\tilde{\mu} = \tilde{J}^{-1}\tilde{h} \qquad \tilde{\Sigma} = \tilde{J}^{-1}$$

$$\begin{aligned} \tilde{h} &= \mathbb{E}_{q(z)}[Q_z^{-1} b_z] + C^\top R^{-1}(y - d) \\ &= \sum_{k=1}^K \left[q(z = k) Q_k^{-1} b_k\right] + C^\top R^{-1}(y - d) \end{aligned} \qquad \begin{aligned} \tilde{J} &= \mathbb{E}_{q(z)}[Q_z^{-1}] + C^\top R^{-1} C \\ &= \sum_{k=1}^K \left[q(z = k) Q_k^{-1}\right] + C^\top R^{-1} C \end{aligned}$$

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

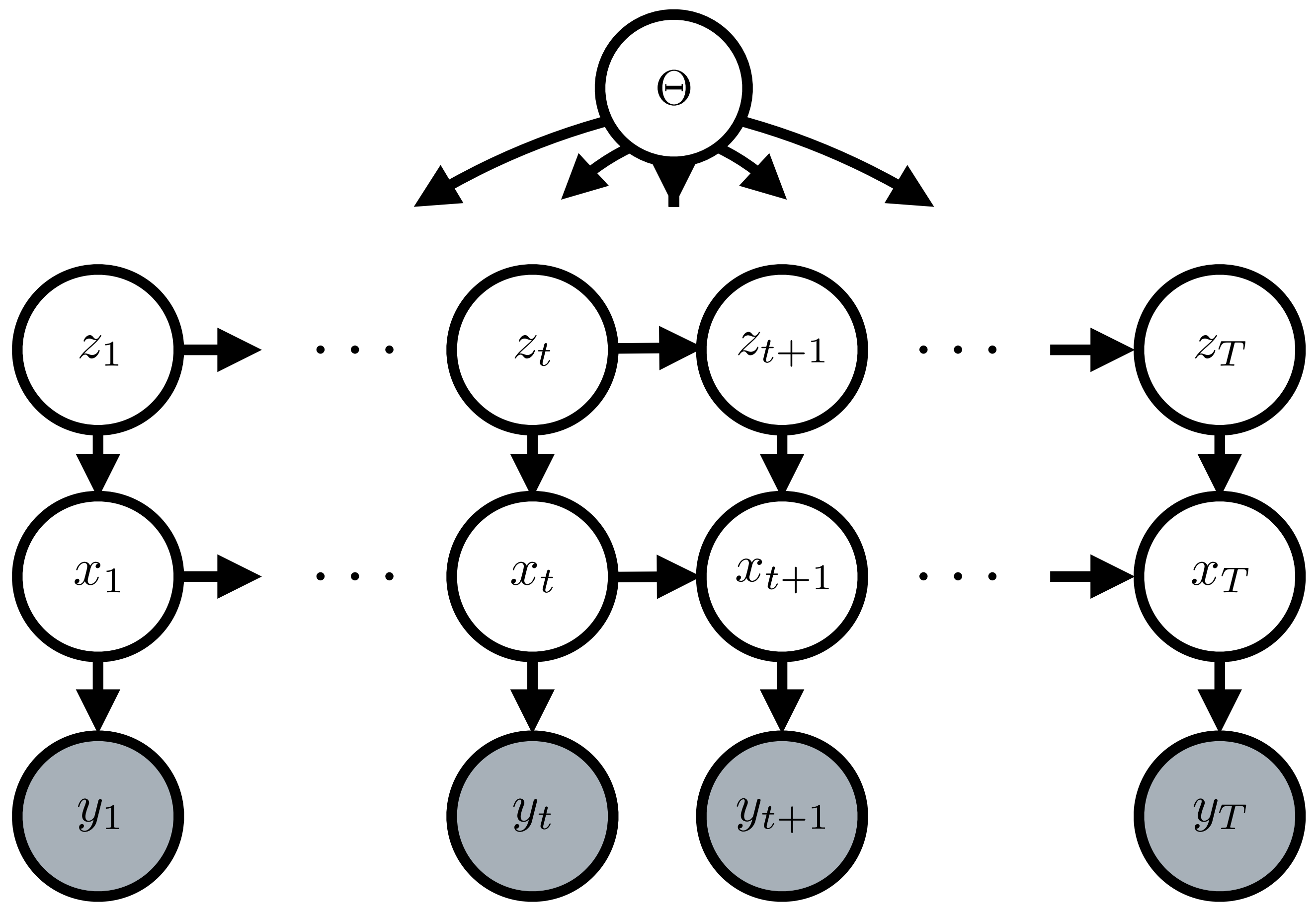
Variational EM for SLDS

Variational EM for SLDS



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.

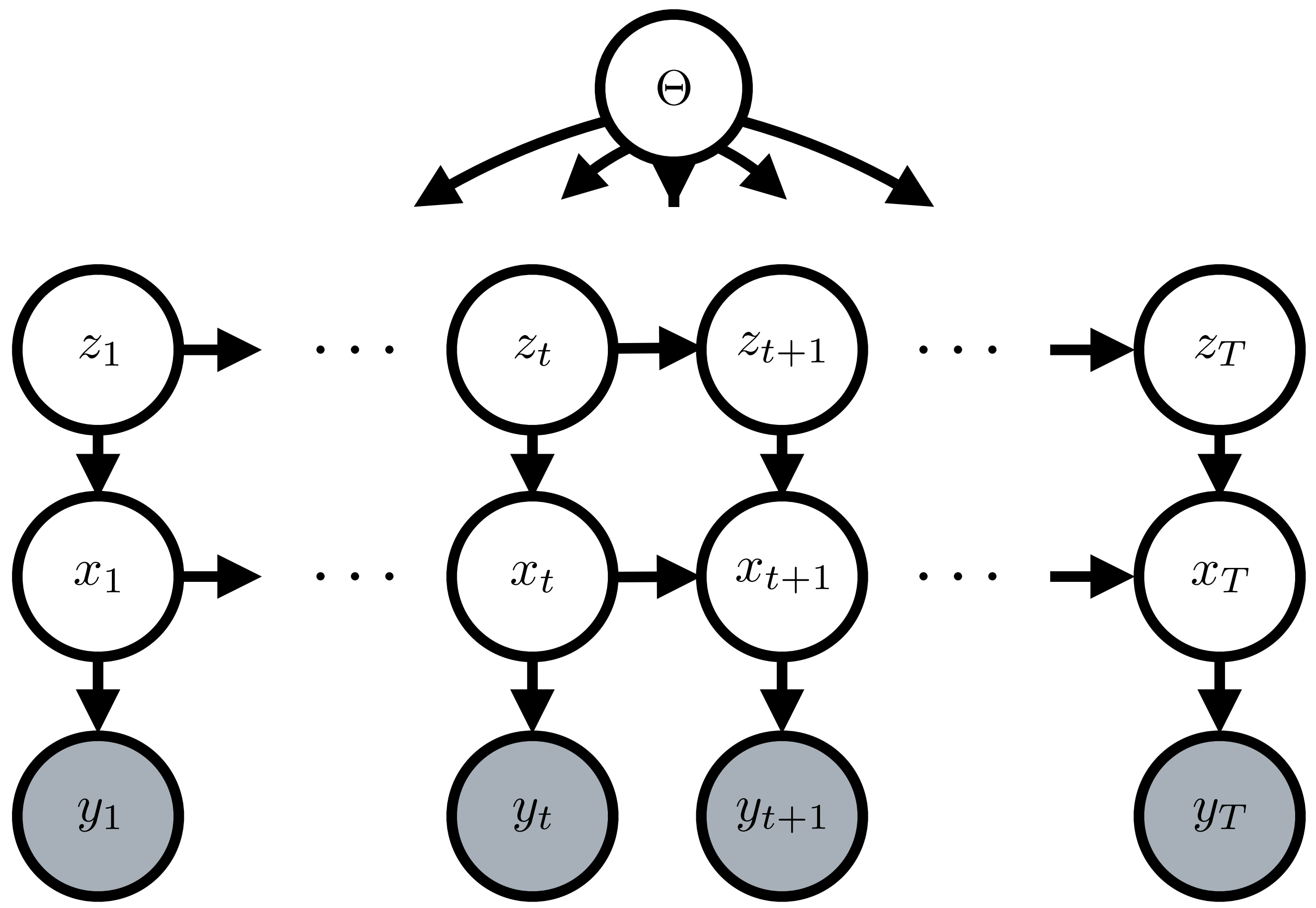
- There can still be dependencies within each factor though!
- This is called a **structured mean field family**.
- We want to find,

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

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

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

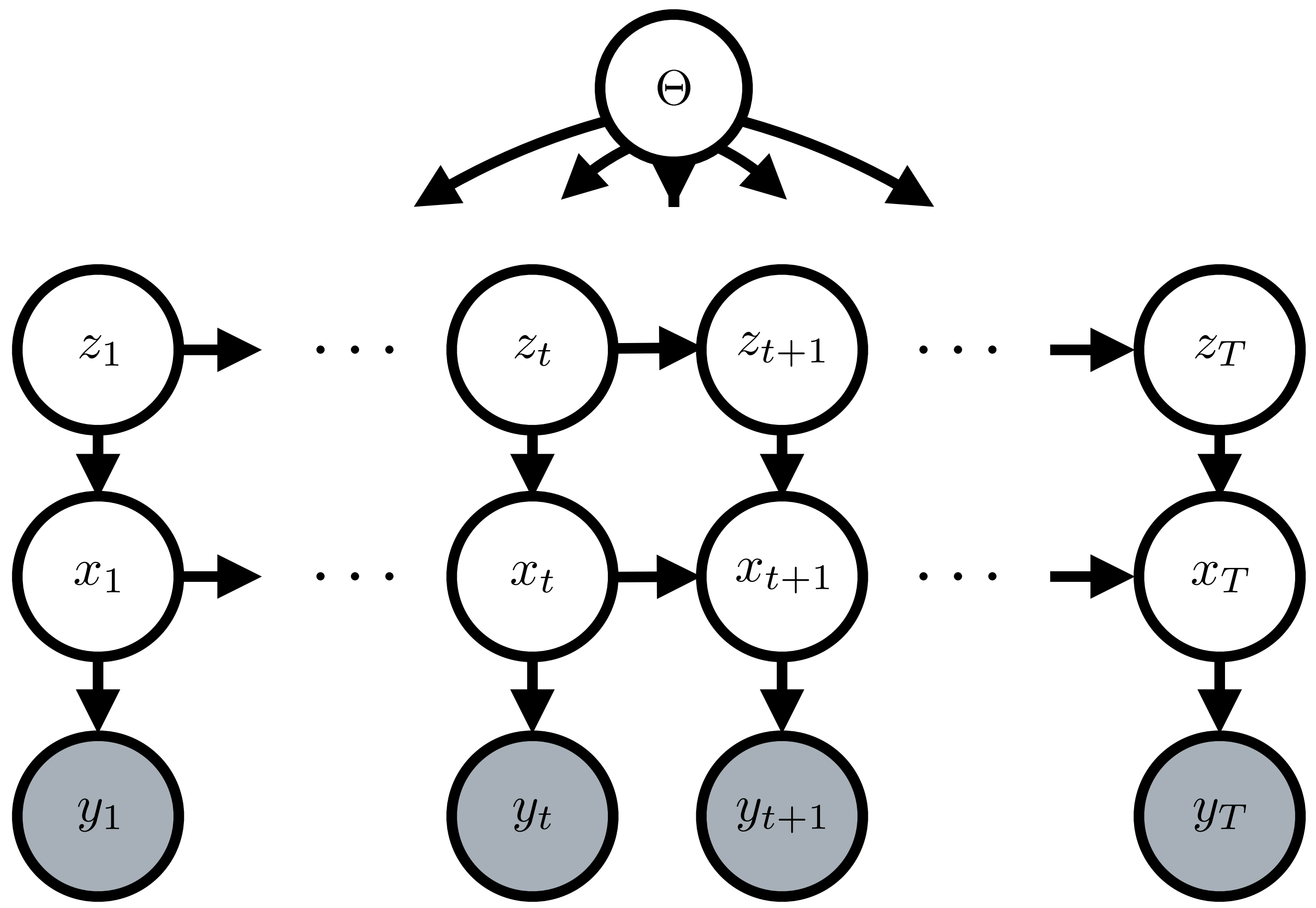
$$\log \tilde{\ell}_{tk} = \mathbb{E}_{q(x)} \left[\log \mathcal{N}(x_t \mid A_k x_{t-1} + b_k, Q_k) \right]$$

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 $q(x_{1:T})$



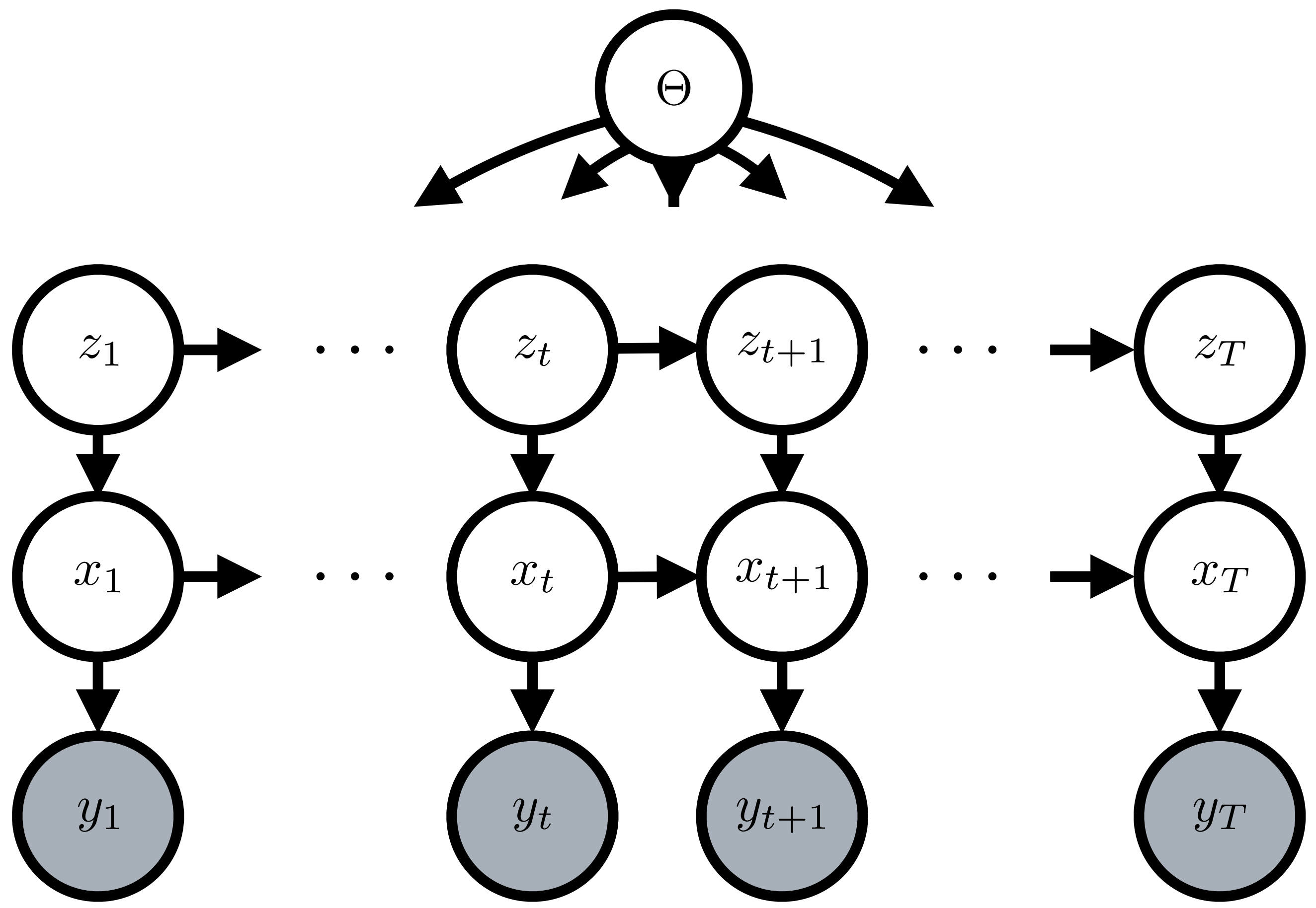
The optimal update for the continuous states takes the form

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

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

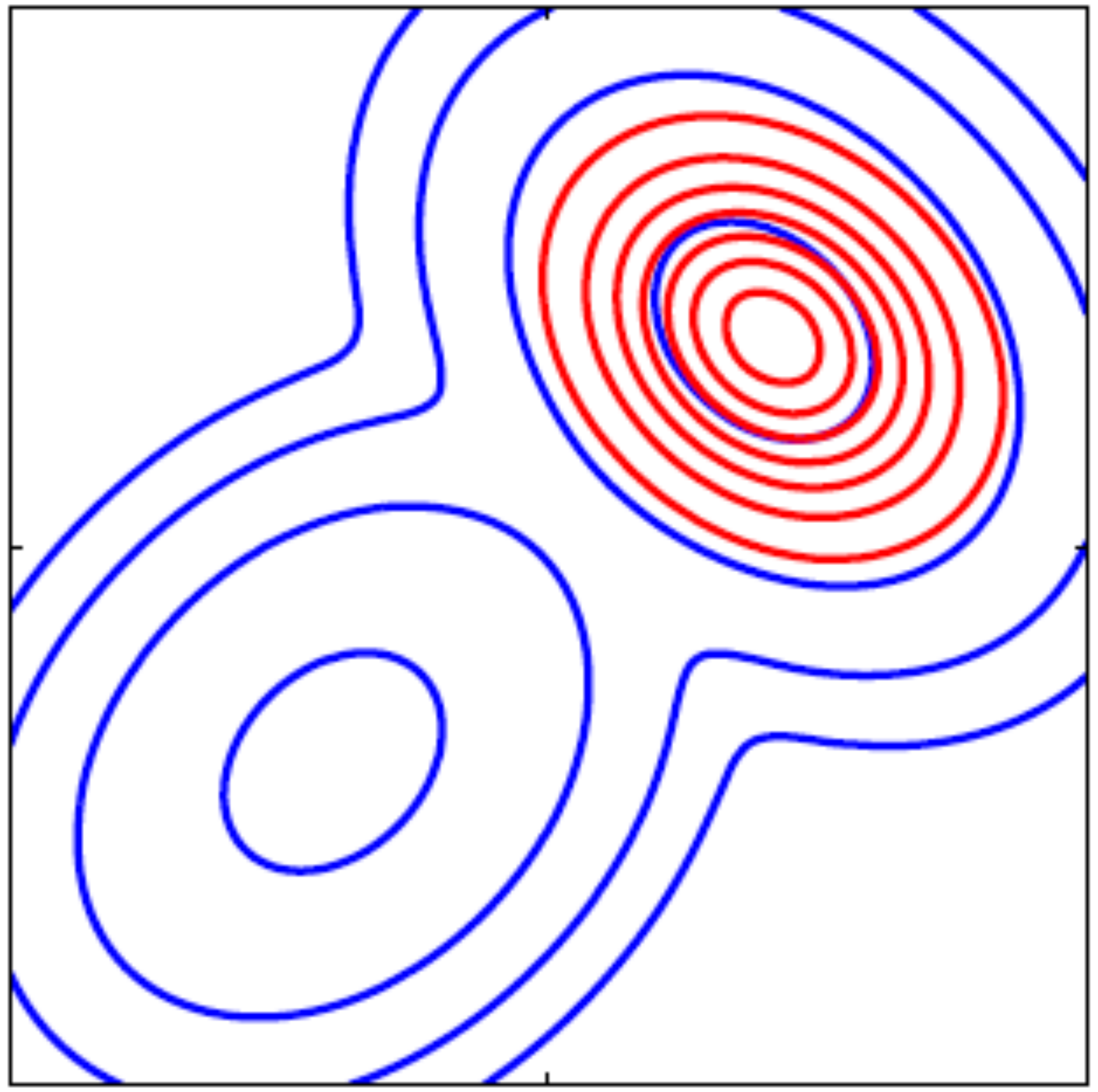
- **M-step**: Maximize the expected log probability

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

  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

- Switching LDS combine ARHMMs and LDS to get the best of both worlds.
- They approximate nonlinear dynamical systems by switching between linear dynamical states.
- However, posterior inference is much harder! The posterior has K^T modes.
- 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.**
- Linderman, Scott W., Matthew J. Johnson, Andrew C. Miller, Ryan P. Adams, David M. Blei, and Liam Paninski. 2017. “Bayesian Learning and Inference in Recurrent Switching Linear Dynamical Systems.” In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS).

Approximate Message Passing

Approximating the forward messages
Assumed density filtering (ADF)

- **Idea:** Approximate the forward messages with a tractable family of distributions $\mathscr{A}$.
  - For example, assume $\alpha_t(h_t) \approx \tilde{\alpha}_t(h_t) \in \mathscr{A}$ where $\mathscr{A}$ is the set of Gaussian mixtures with at most M components.
- Suppose we have $\tilde{\alpha}_{t-1}(h_{t-1}) \in \mathscr{A}$. Our **target for** $\alpha_t(h_t)$ is,

$$\hat{\alpha}_t(h_t) \triangleq \int \mathrm{d}h_{t-1}\, \tilde{\alpha}_{t-1}(h_{t-1})\, p(y_{t-1} \mid h_{t-1})\, p(h_t \mid h_{t-1})$$

  This may not be in $\mathscr{A}$!

- Find the member of $\mathscr{A}$ that best approximates $\hat{\alpha}_t(h_t)$:

$$\tilde{\alpha}_t(h_t) \leftarrow \arg\min_{\mathscr{A}} D\Big(\tilde{\alpha}_t(h_t) \,\|\, \hat{\alpha}_t(h_t)\Big)$$

  where $D(\,\cdot \,\|\, \cdot\,)$ is a measure of divergence between two densities; e.g. reverse Kullback-Leibler (KL) divergence.

- This is called **assumed density filtering (ADF)**, and it is closely related to **expectation-propagation (EP)** and the **unscented/ extended Kalman filter.**

Approximating the forward messages

Assumed density filtering (ADF) with GMM messages

- For example, let $\mathscr{A}$ be the set of Gaussian mixtures with parameters $\rho_t, \{\mu_{t,k}, \Sigma_{t,k}\}_{k=1}^{K}$:

$$\tilde{\alpha}_t(z_t, x_t) = \tilde{\alpha}_t(z_t)\, \tilde{\alpha}_t(x_t \mid z_t) = \mathrm{Cat}(z_t \mid \rho_t)\, \mathcal{N}(x_t \mid \mu_{t,z_t}, \Sigma_{t,z_t}).$$

Approximating the forward messages

Assumed density filtering (ADF) with GMM messages

- For example, let $\mathscr{A}$ be the set of Gaussian mixtures with parameters $\rho_t, \{\mu_{t,k}, \Sigma_{t,k}\}_{k=1}^K$:

$$\tilde{\alpha}_t(z_t, x_t) = \tilde{\alpha}_t(z_t)\, \tilde{\alpha}_t(x_t \mid z_t) = \mathrm{Cat}(z_t \mid \rho_t)\, \mathcal{N}(x_t \mid \mu_{t,z_t}, \Sigma_{t,z_t}).$$

- The target is

$$\hat{\alpha}_t(z_t, x_t) \triangleq \sum_{z_{t-1}=1}^{K} \int \mathrm{d}x_{t-1}\, \tilde{\alpha}_{t-1}(z_{t-1}, x_{t-1})\, p(y_{t-1} \mid x_{t-1})\, p(z_t \mid z_{t-1})\, p(x_t \mid x_{t-1}, z_{t-1})$$

Approximating the forward messages

Assumed density filtering (ADF) with GMM messages

- For example, let $\mathscr{A}$ be the set of Gaussian mixtures with parameters $\rho_t, \{\mu_{t,k}, \Sigma_{t,k}\}_{k=1}^K$:

$$\tilde{\alpha}_t(z_t, x_t) = \tilde{\alpha}_t(z_t)\, \tilde{\alpha}_t(x_t \mid z_t) = \mathrm{Cat}(z_t \mid \rho_t)\, \mathcal{N}(x_t \mid \mu_{t,z_t}, \Sigma_{t,z_t}).$$

- The target is

$$\begin{aligned}
\hat{\alpha}_t(z_t, x_t) &\triangleq \sum_{z_{t-1}=1}^{K} \int \mathrm{d}x_{t-1}\, \tilde{\alpha}_{t-1}(z_{t-1}, x_{t-1})\, p(y_{t-1} \mid x_{t-1})\, p(z_t \mid z_{t-1})\, p(x_t \mid x_{t-1}, z_{t-1}) \\
&= \sum_{z_{t-1}=1}^{K} \int \mathrm{d}x_{t-1}\, \rho_{t-1,z_{t-1}} \mathcal{N}(x_{t-1} \mid \mu_{t-1,z_{t-1}}, \Sigma_{t-1,z_{t-1}})\, \mathcal{N}(y_{t-1} \mid Cx_{t-1} + d, R)\, P_{z_{t-1},z_t}\, \mathcal{N}(x_t \mid A_{z_t} x_{t-1} + b_{z_t}, Q_{z_t})
\end{aligned}$$

Approximating the forward messages
Assumed density filtering (ADF) with GMM messages

- For example, let $\mathscr{A}$ be the set of Gaussian mixtures with parameters $\rho_t, \{\mu_{t,k}, \Sigma_{t,k}\}_{k=1}^K$:

$$\tilde{\alpha}_t(z_t, x_t) = \tilde{\alpha}_t(z_t)\,\tilde{\alpha}_t(x_t \mid z_t) = \mathrm{Cat}(z_t \mid \rho_t)\,\mathcal{N}(x_t \mid \mu_{t,z_t}, \Sigma_{t,z_t}).$$

- The target is

$$\begin{aligned}
\hat{\alpha}_t(z_t, x_t) &\triangleq \sum_{z_{t-1}=1}^K \int \mathrm{d}x_{t-1}\,\tilde{\alpha}_{t-1}(z_{t-1}, x_{t-1})\,p(y_{t-1} \mid x_{t-1})\,p(z_t \mid z_{t-1})\,p(x_t \mid x_{t-1}, z_{t-1}) \\
&= \sum_{z_{t-1}=1}^K \int \mathrm{d}x_{t-1}\,\rho_{t-1,z_{t-1}}\mathcal{N}(x_{t-1} \mid \mu_{t-1,z_{t-1}}, \Sigma_{t-1,z_{t-1}})\,\mathcal{N}(y_{t-1} \mid Cx_{t-1} + d, R)\,P_{z_{t-1},z_t}\,\mathcal{N}(x_t \mid A_{z_t}x_{t-1} + b_{z_t}, Q_{z_t}) \\
&\propto \sum_{z_{t-1}=1}^K \rho(z_t, z_{t-1})\mathcal{N}\left(x_t \mid \mu(z_t, z_{t-1}), \Sigma(z_t, z_{t-1})\right)
\end{aligned}$$

- This is a GMM with K components for each assignment of z_t. We want to approximate it a single Gaussian for each z_t.

Approximating the forward messages
Assumed density filtering (ADF) with GMM messages

- **More generally**, let $\mathscr{A}$ be the set of Gaussian mixtures with at most M components for each assignment of z_t,

$$\tilde{\alpha}_t(z_t, x_t) = \tilde{\alpha}_t(z_t)\,\tilde{\alpha}_t(x_t \mid z_t) = \mathrm{Cat}(z_t \mid \rho_t) \left(\sum_{m=1}^{M} \omega_{t,z_t,m} \mathcal{N}(x_t \mid \mu_{t,z_t,m}, \Sigma_{t,z_t,k}) \right).$$

- The target is

$$\begin{aligned}
\hat{\alpha}_t(z_t, x_t) &\triangleq \sum_{z_{t-1}=1}^{K} \int \mathrm{d}x_{t-1}\, \tilde{\alpha}_{t-1}(z_{t-1}, x_{t-1})\, p(y_{t-1} \mid x_{t-1})\, p(z_t \mid z_{t-1})\, p(x_t \mid x_{t-1}, z_{t-1}) \\
&= \sum_{z_{t-1}=1}^{K} \sum_{m=1}^{M} \int \mathrm{d}x_{t-1}\, \rho_{t-1,z_{t-1}}\, \omega_{t,z_{t-1},m} \mathcal{N}(x_{t-1} \mid \mu_{t-1,z_{t-1},m}, \Sigma_{t-1,z_{t-1},m})\, \mathcal{N}(y_{t-1} \mid Cx_{t-1} + d, R)\, P_{z_{t-1},z_t}\, \mathcal{N}(x_t \mid A_{z_t} x_{t-1} + b_{z_t}, Q_{z_t}) \\
&\propto \sum_{z_{t-1}=1}^{K} \sum_{m=1}^{M} \rho(z_t, z_{t-1}, m) \mathcal{N}\big(x_t \mid \mu(z_t, z_{t-1}, m), \Sigma(z_t, z_{t-1}, m)\big)
\end{aligned}$$

- This is a GMM with KM components for each assignment of z_t. We want to approximate it with a GMM with only M components.

Approximating the forward messages

Projecting onto the set $\mathscr{A}$

- Find the member of $\mathscr{A}$ that best approximates $\hat{\alpha}_t(z_t, h_t)$ by minimizing the reverse Kullback-Leibler divergence,

$$\mathrm{KL}\Big(\hat{\alpha}_t(z_t, x_t) \parallel \tilde{\alpha}_t(z_t, x_t)\Big) = \mathbb{E}_{\hat{\alpha}_t(z_t, x_t)}\left[\log \hat{\alpha}_t(z_t, x_t) - \log \tilde{\alpha}_t(z_t, x_t)\right]$$

Approximating the forward messages

Projecting onto the set $\mathscr{A}$

- Find the member of $\mathscr{A}$ that best approximates $\hat{\alpha}_t(z_t, h_t)$ by minimizing the reverse Kullback-Leibler divergence,

$$\mathrm{KL}\Big(\hat{\alpha}_t(z_t, x_t) \parallel \tilde{\alpha}_t(z_t, x_t)\Big) = \mathbb{E}_{\hat{\alpha}_t(z_t, x_t)}\left[\log \hat{\alpha}_t(z_t, x_t) - \log \tilde{\alpha}_t(z_t, x_t)\right]$$

$$= -\,\mathbb{E}_{\hat{\alpha}_t(z_t, x_t)}\left[\log \tilde{\alpha}_t(z_t, x_t)\right] + \mathrm{c}$$

Approximating the forward messages

Projecting onto the set $\mathscr{A}$

- Find the member of $\mathscr{A}$ that best approximates $\hat{\alpha}_t(z_t, h_t)$ by minimizing the reverse Kullback-Leibler divergence,

$$\mathrm{KL}\left(\hat{\alpha}_t(z_t, x_t) \parallel \tilde{\alpha}_t(z_t, x_t)\right) = \mathbb{E}_{\hat{\alpha}_t(z_t, x_t)}\left[\log \hat{\alpha}_t(z_t, x_t) - \log \tilde{\alpha}_t(z_t, x_t)\right]$$

$$= -\,\mathbb{E}_{\hat{\alpha}_t(z_t, x_t)}\left[\log \tilde{\alpha}_t(z_t, x_t)\right] + \mathrm{c}$$

$$= -\,\mathbb{E}_{\hat{\alpha}_t(z_t, x_t)}\left[\log \mathrm{Cat}(z_t \mid \rho_t) + \log \mathscr{N}(x_t \mid \mu_{t,z_t}, \Sigma_{t,z_t})\right] + \mathrm{c}$$

Approximating the forward messages

Projecting onto the set $\mathscr{A}$

- Find the member of $\mathscr{A}$ that best approximates $\hat{\alpha}_t(z_t, h_t)$ by minimizing the reverse Kullback-Leibler divergence,

$$
\begin{aligned}
\mathrm{KL}\Big(\hat{\alpha}_t(z_t, x_t) \parallel \tilde{\alpha}_t(z_t, x_t)\Big) &= \mathbb{E}_{\hat{\alpha}_t(z_t,x_t)}\left[\log \hat{\alpha}_t(z_t, x_t) - \log \tilde{\alpha}_t(z_t, x_t)\right] \\
&= -\mathbb{E}_{\hat{\alpha}_t(z_t,x_t)}\left[\log \tilde{\alpha}_t(z_t, x_t)\right] + \mathrm{c} \\
&= -\mathbb{E}_{\hat{\alpha}_t(z_t,x_t)}\left[\log \mathrm{Cat}(z_t \mid \rho_t) + \log \mathscr{N}(x_t \mid \mu_{t,z_t}, \Sigma_{t,z_t})\right] + \mathrm{c} \\
&= -\mathbb{E}_{\hat{\alpha}_t(z_t,x_t)}\left[\sum_{k=1}^{K} \mathbb{I}[z_t = k]\Big(\rho_{tk} - \tfrac{1}{2}\log|\Sigma_{tk}| - \tfrac{1}{2}x_t^\top \Sigma_{tk}^{-1} x_t + \mu_{tk}^\top \Sigma_{tk}^{-1} x_t - \tfrac{1}{2}\mu_{tk}^\top \Sigma_{tk}^{-1}\mu_{tk}\Big)\right] + \mathrm{c}
\end{aligned}
$$

Approximating the forward messages

Projecting onto the set $\mathscr{A}$

- Find the member of $\mathscr{A}$ that best approximates $\hat{\alpha}_t(z_t, h_t)$ by minimizing the reverse Kullback-Leibler divergence,

$$
\begin{aligned}
\mathrm{KL}\Big(\hat{\alpha}_t(z_t, x_t) \,\|\, \tilde{\alpha}_t(z_t, x_t)\Big) &= \mathbb{E}_{\hat{\alpha}_t(z_t,x_t)}\left[\log \hat{\alpha}_t(z_t, x_t) - \log \tilde{\alpha}_t(z_t, x_t)\right] \\
&= -\,\mathbb{E}_{\hat{\alpha}_t(z_t,x_t)}\left[\log \tilde{\alpha}_t(z_t, x_t)\right] + \mathrm{c} \\
&= -\,\mathbb{E}_{\hat{\alpha}_t(z_t,x_t)}\left[\log \mathrm{Cat}(z_t \mid \rho_t) + \log \mathcal{N}(x_t \mid \mu_{t,z_t}, \Sigma_{t,z_t})\right] + \mathrm{c} \\
&= -\,\mathbb{E}_{\hat{\alpha}_t(z_t,x_t)}\left[\sum_{k=1}^{K} \mathbb{I}[z_t = k]\Big(\rho_{tk} - \tfrac{1}{2}\log|\Sigma_{tk}| - \tfrac{1}{2}x_t^\top \Sigma_{tk}^{-1} x_t + \mu_{tk}^\top \Sigma_{tk}^{-1} x_t - \tfrac{1}{2}\mu_{tk}^\top \Sigma_{tk}^{-1}\mu_{tk}\Big)\right] + \mathrm{c} \\
&= -\sum_{k=1}^{K} \langle \rho_{tk} - \tfrac{1}{2}\log|\Sigma_{tk}|, \bar{N}_k\rangle + \langle -\tfrac{1}{2}\Sigma_{tk}^{-1}, \bar{\psi}_{k,1}\rangle + \langle \Sigma_{tk}^{-1}\mu_{tk}, \bar{\psi}_{k,2}\rangle + \langle -\tfrac{1}{2}\mu_{tk}^\top \Sigma_{tk}^{-1}\mu_{tk}, \bar{\psi}_{k,3}\rangle + \mathrm{c}
\end{aligned}
$$

Where $\bar{N}_k = \mathbb{E}_{\hat{\alpha}_t(z_t,x_t)}[\mathbb{I}[z_t = k]]$, $\bar{\psi}_{k,1} = \mathbb{E}_{\hat{\alpha}_t(z_t,x_t)}[\mathbb{I}[z_t = k]\, x_t x_t^\top]$, $\bar{\psi}_{k,2} = \mathbb{E}_{\hat{\alpha}_t(z_t,x_t)}[\mathbb{I}[z_t = k]\, x_t]$, $\bar{\psi}_{k,3} = \bar{N}_k$.