Composing Graphical Models with Neural Networks for Structured Representations and Fast Inference

Matthew James Johnson[1], David Duvenaud[1], Alexander B. Wiltschko[1], Sandeep R. Datta[2], Ryan P. Adams[1]

[1]Harvard University

[2]Harvard Medical School

NIPS,2016

Presenter: Arshdeep Sekhon

Motivation

Combine graphical models with neural networks: Complementary Strengths of Deep Learning and Graphical Models

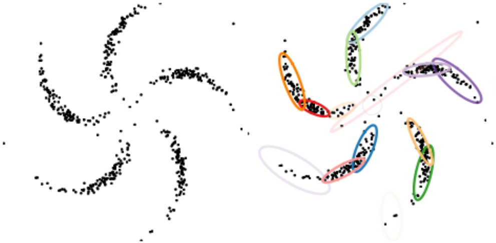
GRAPHICAL MODELS

1. (+) structured representations
2. (+) priors and uncertainty
3. (+) data and computational efficiency:efficient inference procedures
4. (-) assumptions about data
5. (-) feature engineering

DEEP LEARNING

1. (-) hard to understand
2. (-) lot of data
3. (+) flexible: fit anything
4. (+) feature learning

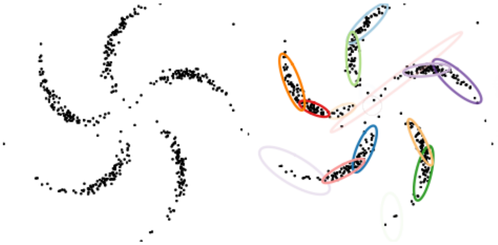
Warped mixtures for arbitrary cluster shapes



(a) Data (b) GMM

$$\pi \sim \mathrm{Dir}(\alpha), \quad (\mu_k, \Sigma_k) \overset{\text{iid}}{\sim} \mathrm{NIW}(\lambda), \quad z_n \,|\, \pi \overset{\text{iid}}{\sim} \pi \quad y_n \,|\, z_n, \{(\mu_k, \Sigma_k)\}_{k=1}^K \overset{\text{iid}}{\sim} \mathcal{N}(\mu_{z_n}, \Sigma_{z_n}).$$

Warped mixtures for arbitrary cluster shapes



(a) Data (b) GMM

1. Cluster data using GMM: Real data does not form nice Gaussian clusters
2. Clusters are there but not explained correctly by GMMs
3. lose the interpretability of the model

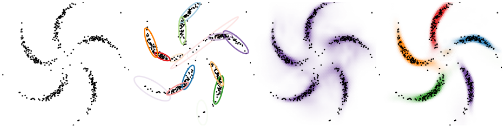
Warped mixtures for arbitrary cluster shapes

(a) Data (b) GMM (c) Density net (VAE)

$$\gamma \sim p(\gamma) \qquad x_n \overset{\text{iid}}{\sim} \mathcal{N}(0, I), \qquad y_n \mid x_n, \gamma \overset{\text{iid}}{\sim} \mathcal{N}(\mu(x_n; \gamma), \Sigma(x_n; \gamma)),$$

1. No structure in data, although captures shape correctly

Warped mixtures for arbitrary cluster shapes



(a) Data (b) GMM (c) Density net (VAE) (d) GMM SVAE

Figure: composing a latent GMM with nonlinear observations

$$\pi \sim \mathrm{Dir}(\alpha), \qquad (\mu_k, \Sigma_k) \overset{\text{iid}}{\sim} \mathrm{NIW}(\lambda), \qquad \gamma \sim p(\gamma)$$

$$z_n \,|\, \pi \overset{\text{iid}}{\sim} \pi \qquad x_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu^{(z_n)}, \Sigma^{(z_n)}), \qquad y_n \,|\, x_n, \gamma \overset{\text{iid}}{\sim} \mathcal{N}(\mu(x_n;\gamma), \Sigma(x_n;\gamma)).$$

1. Flexibility
2. structured

Generatively Modeling a Video

1. Neuroscientists want to do experiments on a mouse and see how it's behavior changes

Generatively Modeling a Video

1. Neuroscientists want to do experiments on a mouse and see how it's behavior changes
2. Experiment involves recording the 'states' of the mouse: sleeping, running, standing, etc.

Generatively Modeling a Video

1. Neuroscientists want to do experiments on a mouse and see how it's behavior changes
2. Experiment involves recording the 'states' of the mouse: sleeping, running, standing, etc.
3. sequence of actions
4. desired: clustering model of actions

Generatively Modeling a Video

1. Neuroscientists want to do experiments on a mouse and see how it's behavior changes
2. Experiment involves recording the 'states' of the mouse: sleeping, running, standing, etc.
3. sequence of actions
4. desired: clustering model of actions
5. To automate this: Use a Switching Latent Linear Dynamical System

Switching Latent Linear Dynamical System

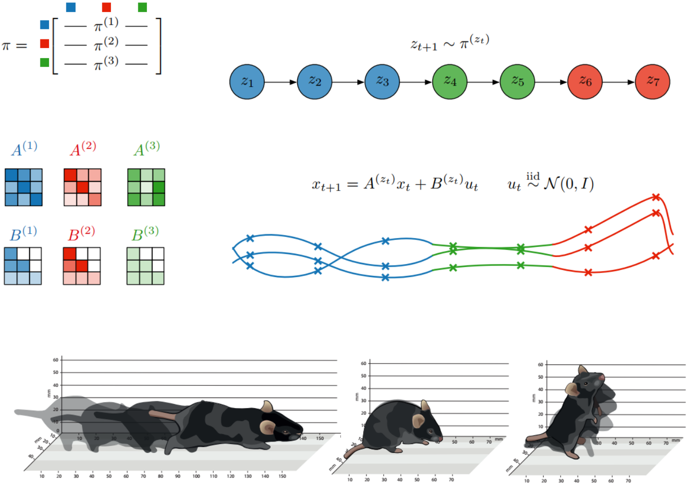


Figure: composing a latent GMM with nonlinear observations

Combining GMs and NNs: SVAE

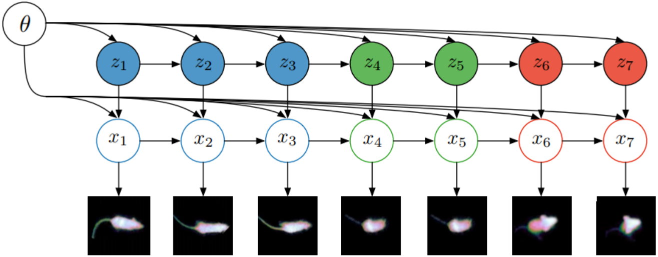
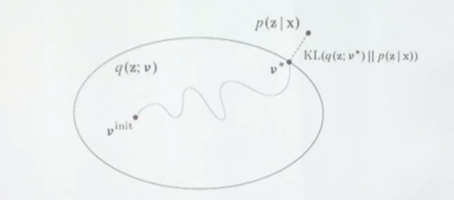


Figure: composing a latent GMM with nonlinear observations

Background: Variational Inference

1. Consider a joint density of latent variables $x = x_{1:m}$ and observations $y = y_{1:m}$
2. Inference in a Bayesian model: conditioning on data and computing the posterior $p(x|y)$
3. $p(x|y) = \dfrac{p(x, y)}{p(y)}$
4. Variational Inference: solve this problem with optimization

Background: Variational Inference

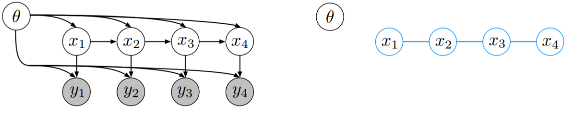
1. posit a variational family $q(z, \nu)$
2. optimize ν to make $q(z, \nu)$ close to $p(x|y)$
3. 

4. Evidence Lower bound(ELBO)

$$\mathbb{E}_q log\{\frac{(p(x, y))}{q(x)}\} \tag{1}$$

5. Solving this maximization problem is equivalent to finding the member of the family that is closest in KL divergence to the posterior

Variational inference in Linear Dynamical Systems



$p(x \mid \theta)$ is linear dynamical system
$p(y \mid x, \theta)$ is linear-Gaussian
$p(\theta)$ is conjugate prior

$$q(\theta)q(x) \approx p(\theta, x \mid y)$$

$$\mathcal{L}[\, q(\theta)q(x)\,] \triangleq \mathbb{E}_{q(\theta)q(x)}\left[\log \frac{p(\theta,x,y)}{q(\theta)q(x)}\right]$$

$$q(\theta) \leftrightarrow \eta_\theta \qquad q(x) \leftrightarrow \eta_x$$

Figure: Efficient Inference for Conjugate Family distributions

If the posterior distributions $p(\theta|x)$ are in the same family as the prior probability distribution $p(\theta)$, the prior and posterior are then called conjugate distributions, and the prior is called a conjugate prior for the likelihood function. Makes it easier to calculate posterior

Variational inference in Linear Dynamical Systems

$$\mathcal{L}(\eta_\theta, \eta_x) \triangleq \mathbb{E}_{q(\theta)q(x)}\left[\log \frac{p(\theta,x,y)}{q(\theta)q(x)}\right]$$

$$\eta_x^*(\eta_\theta) \triangleq \arg\max_{\eta_x} \mathcal{L}(\eta_\theta, \eta_x) \qquad \mathcal{L}_{\text{SVI}}(\eta_\theta) \triangleq \mathcal{L}(\eta_\theta, \eta_x^*(\eta_\theta))$$

Proposition (natural gradient SVI of Hoffman et al. 2013)

$$\widetilde{\nabla}\mathcal{L}_{\text{SVI}}(\eta_\theta) = \eta_\theta^0 + \mathbb{E}_{q^*(x)}(t_{xy}(x,y), 1) - \eta_\theta$$

Figure: Efficient Inference for Exponential Family distributions

Because the observation model $p(y|x,\theta)$ is conjugate to the latent variable model $p(x|\theta)$, for any fixed $q(\theta)$ the optimal factor $q^*(x)$, $argmax_{q(x)} L[q(\theta)q(x)]$ is itself a Gaussian linear dynamical system with parameters that are simple functions of the expected statistics of $q(\theta)$ and the data y.

Combine Both: Structured Variational Autoencoder

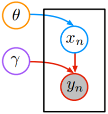
Basic Idea

Keep graphical models for latent variables (the clusters), connect these to data that doesn't fit our assumptions

1. Similar to Supervised Learning: transform data into a latent space, which separates the data

Inference in SVAE

1. The main difficulty with combining rich latent variable structure and flexible likelihoods is inference.
2. The most efficient inference algorithms used in graphical models, like structured mean field and message passing, depend on conjugate exponential family likelihoods to preserve tractable structure.



$p(\theta)$	conjugate prior on global variables
$p(x \mid \theta)$	exponential family on local variables
$p(\gamma)$	any prior on observation parameters
$p(y \mid x, \gamma)$	neural network observation model

3.

SVAE Model Class

1. a conjugate pair of exponential family densities on global latent variables θ and local latent variables x
2. Let $p(x|\theta)$ be an exponential family and let $p(\theta)$ be its corresponding natural exponential family conjugate prior

$$p(\theta) = \exp\left\{\langle \eta_\theta^0, t_\theta(\theta)\rangle - \log Z_\theta(\eta_\theta^0)\right\},$$

$$p(x\,|\,\theta) = \exp\left\{\langle \eta_x^0(\theta), t_x(x)\rangle - \log Z_x(\eta_x^0(\theta))\right\} = \exp\left\{\langle t_\theta(\theta), (t_x(x), 1)\rangle\right\},$$

New ELBO: SVAE

$$\mathcal{L}[\,q(\theta)q(\gamma)q(x)\,] \triangleq \mathbb{E}_{q(\theta)q(\gamma)q(x)}\left[\log \frac{p(\theta)p(\gamma)p(x\,|\,\theta)p(y\,|\,x,\gamma)}{q(\theta)q(\gamma)q(x)}\right].$$

without conjugacy structure finding a local partial optimizer may be computationally expensive for general densities $p(y|x,\lambda)$,

1. general observation model means that conjugate updates and natural gradient SVI cannot be directly applied
2. choose η_x by optimizing over a surrogate objective L with conjugacy structure

New ELBO: SVAE

$$\mathcal{L}(\eta_\theta, \eta_\gamma, \eta_x) \triangleq \mathbb{E}_{q(\theta)q(\gamma)q(x)}\left[\log \frac{p(\theta,\gamma,x)p(y\,|\,x,\gamma)}{q(\theta)q(\gamma)q(x)}\right]$$

$$\widehat{\mathcal{L}}(\eta_\theta, \eta_x, \phi) \triangleq \mathbb{E}_{q(\theta)q(\gamma)q(x)}\left[\log \frac{p(\theta,\gamma,x)\exp\{\psi(x;y,\phi)\}}{q(\theta)q(\gamma)q(x)}\right]$$

$$\psi(x;y,\phi) \triangleq \langle r(y;\phi),\ t_x(x)\rangle,$$

$$\eta_x^*(\eta_\theta, \phi) \triangleq \arg\max_{\eta_x} \widehat{\mathcal{L}}(\eta_\theta, \eta_x, \phi) \qquad \mathcal{L}_{\mathrm{SVAE}}(\eta_\theta, \eta_\gamma, \phi) \triangleq \mathcal{L}(\eta_\theta, \eta_\gamma, \eta_x^*(\eta_\theta, \phi))$$

the potentials have a form conjugate to the exponential family $p(x|\theta)$.

Computing SVAE gradients

Algorithm 1 Estimate SVAE lower bound and its gradients

Input: Variational parameters $(\eta_\theta, \eta_\gamma, \phi)$, data sample y

  function SVAEGRADIENTS$(\eta_\theta, \eta_\gamma, \phi, y)$

    $\psi \leftarrow r(y_n; \phi)$ ▷ Get evidence potentials

    $(\hat{x},\ \bar{t}_x,\ \mathrm{KL}^{\mathrm{local}}) \leftarrow$ PGMINFERENCE(η_θ, ψ) ▷ Combine evidence with prior

    $\hat{\gamma} \sim q(\gamma)$ ▷ Sample observation parameters

    $\mathcal{L} \leftarrow N \log p(y \,|\, \hat{x}, \hat{\gamma}) - N\,\mathrm{KL}^{\mathrm{local}} - \mathrm{KL}(q(\theta)q(\gamma) \| p(\theta)p(\gamma))$ ▷ Estimate variational bound

    $\widetilde{\nabla}_{\eta_\theta}\mathcal{L} \leftarrow \eta_\theta^0 - \eta_\theta + N(\bar{t}_x, 1) + N(\nabla_{\eta_x} \log p(y \,|\, \hat{x}, \hat{\gamma}), 0)$ ▷ Compute natural gradient

    return lower bound $\mathcal{L}$, natural gradient $\widetilde{\nabla}_{\eta_\theta}\mathcal{L}$, gradients $\nabla_{\eta_\gamma, \phi}\mathcal{L}$

  function PGMINFERENCE(η_θ, ψ)

    $q^*(x) \leftarrow$ OPTIMIZELOCALFACTORS(η_θ, ψ) ▷ Fast message-passing inference

    return sample $\hat{x} \sim q^*(x)$, statistics $\mathbb{E}_{q^*(x)} t_x(x)$, divergence $\mathbb{E}_{q(\theta)}\,\mathrm{KL}(q^*(x) \| p(x \,|\, \theta))$

SVAE objective and natural gradient

The SVAE objective lower-bounds the mean field objective

The SVAE objective function $\mathcal{L}_{\text{SVAE}}$ *lower-bounds the mean field objective* $\mathcal{L}$ *in the sense that*

$$\max_{q(x)} \mathcal{L}[\, q(\theta)q(\gamma)q(x) \,] \geq \max_{\eta_x} \mathcal{L}(\eta_\theta, \eta_\gamma, \eta_x) \geq \mathcal{L}_{\text{SVAE}}(\eta_\theta, \eta_\gamma, \phi) \quad \forall \phi \in \mathbb{R}^m,$$

for any parameterized function class $\{r(y;\phi)\}_{\phi\in\mathbb{R}^m}$. *Furthermore, if there is some* $\phi^* \in \mathbb{R}^m$ *such that* $\psi(x; y, \phi^*) = \mathbb{E}_{q(\gamma)} \log p(y \,|\, x, \gamma)$, *then the bound can be made tight in the sense that*

$$\max_{q(x)} \mathcal{L}[\, q(\theta)q(\gamma)q(x) \,] = \max_{\eta_x} \mathcal{L}(\eta_\theta, \eta_\gamma, \eta_x) = \max_{\phi} \mathcal{L}_{\text{SVAE}}(\eta_\theta, \eta_\gamma, \phi).$$

Natural gradient of the SVAE objective

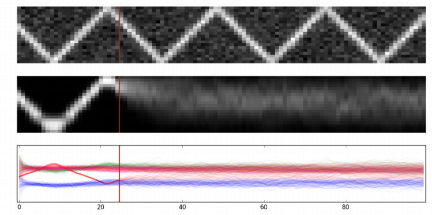
The natural gradient of the SVAE objective $\mathcal{L}_{\text{SVAE}}$ *with respect to* η_θ *can be estimated as*

$$\widetilde{\nabla}_{\eta_\theta} \mathcal{L}_{\text{SVAE}}(\eta_\theta, \eta_\gamma, \phi) = \left(\eta_\theta^0 + \mathbb{E}_{q^*(x)}\left[(t_x(x), 1)\right] - \eta_\theta\right) + \left(\nabla^2 \log Z_\theta(\eta_\theta)\right)^{-1} \nabla F(\eta_\theta),$$

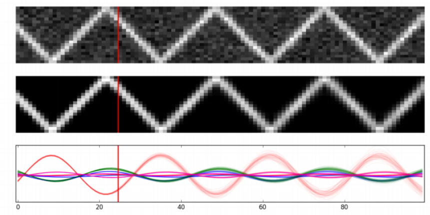
where $F(\eta_\theta') = \mathcal{L}(\eta_\theta, \eta_\gamma, \eta_x^*(\eta_\theta', \phi))$. *When there is only one local variational factor* $q(x)$, *then can simplify the estimator to*

$$\widetilde{\nabla}_{\eta_\theta} \mathcal{L}_{\text{SVAE}}(\eta_\theta, \eta_\gamma, \phi) = \left(\eta_\theta^0 + \mathbb{E}_{q^*(x)}\left[(t_x(x), 1)\right] - \eta_\theta\right) + \left(\nabla_{\eta_x} \mathcal{L}(\eta_\theta, \eta_\gamma, \eta_x^*(\eta_\theta, \phi)), 0\right).$$

Experiments and Results: Synthetic Data

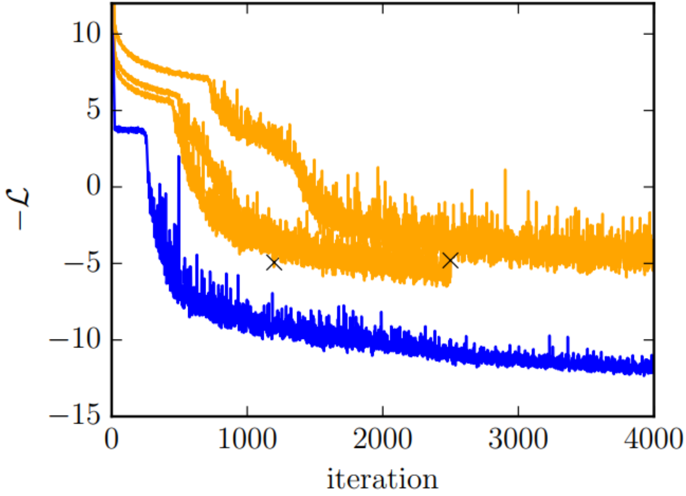


(a) Predictions after 200 training steps.



(b) Predictions after 1100 training steps.

Experiments and Results: Synthetic Data



(a) Natural (blue) and standard (orange) gradient updates.

Experiments and Results: Mouse Video

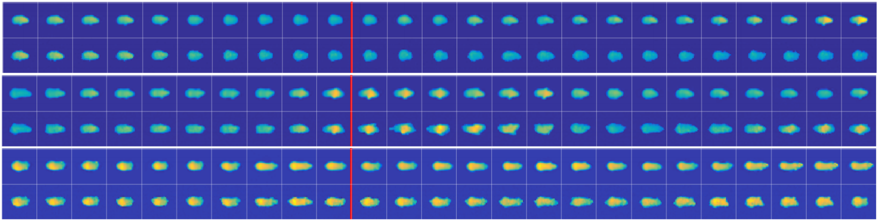


Figure 6: Predictions from an LDS SVAE fit to depth video. In each panel, the top is a sampled prediction and the bottom is real data. The model is conditioned on observations to the left of the line.