Variational Inference

Seungjin Choi

Department of Computer Science and Engineering
Pohang University of Science and Technology
77 Cheongam-ro, Nam-gu, Pohang 37673, Korea
seungjin@postech.ac.kr
http://mlg.postech.ac.kr/∼seungjin

January 24, 2018

Probabilistic Models

Probabilistic Models in Machine Learning

- A **probabilistic model** is a joint distribution,

$$p(\mathbf{x}, \mathbf{z}),$$

over observed variables $\mathbf{x}$ and hidden variables $\mathbf{z}$ (**latent features, latent class**).

- **Inference about unknowns** is carried out by calculating the **posterior** distribution over hidden variables:

$$p(\mathbf{z}|\mathbf{x}) = \frac{p(\mathbf{x}, \mathbf{z})}{p(\mathbf{x})}.$$

- The **evidence** $p(\mathbf{x})$ is not tractable in most of models of interest, we resort to **approximate inference**. Today we focus on **variational inference**.

Latent Class Models (Mixture of Gaussians)

- The joint distribution over $\mathbf{x}$ and $\mathbf{z}$ is given by

$$\begin{aligned} p(\mathbf{x}, \mathbf{z}) &= p(\mathbf{x}|\mathbf{z})p(\mathbf{z}) \\ &= \left[\prod_{k=1}^{K} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k}\right] \left[\prod_{k=1}^{K} \pi_k^{z_k}\right]. \end{aligned}$$

- Then, the likelihood is given by

$$\begin{aligned} p(\mathbf{x}) &= \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z}) \\ &= \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \end{aligned}$$

  and the **inference** $p(\mathbf{z}|\mathbf{x})$ yields which **cluster** the example $\mathbf{x}$ belongs to.

Latent Feature Models (Probabilistic PCA)

- The joint distribution over $\mathbf{x}$ and $\mathbf{z}$ is given by

$$\begin{aligned} p(\mathbf{x}, \mathbf{z}) &= p(\mathbf{x}|\mathbf{z})p(\mathbf{z}) \\ &= \mathcal{N}(\mathbf{x}|\mathbf{A}\mathbf{z} + \boldsymbol{\mu}, \sigma^2\mathbf{I})\mathcal{N}(\mathbf{z}|0, \mathbf{I}) \end{aligned}$$

- Then, the likelihood is given by

$$\begin{aligned} p(\mathbf{x}) &= \int p(\mathbf{x}, \mathbf{z})d\mathbf{z} \\ &= \mathcal{N}(\mathbf{x}|0, \mathbf{A}\mathbf{A}^\top + \sigma^2\mathbf{I}), \end{aligned}$$

  and the **inference** $p(\mathbf{z}|\mathbf{x})$ reveals **features** corresponding to the data $\mathbf{x}$

Hidden Markov Models (HMMs)

Consider the joint distribution over $\mathbf{x}_1, \ldots, \mathbf{x}_N$ and $\mathbf{z}_1, \ldots, \mathbf{z}_N$:

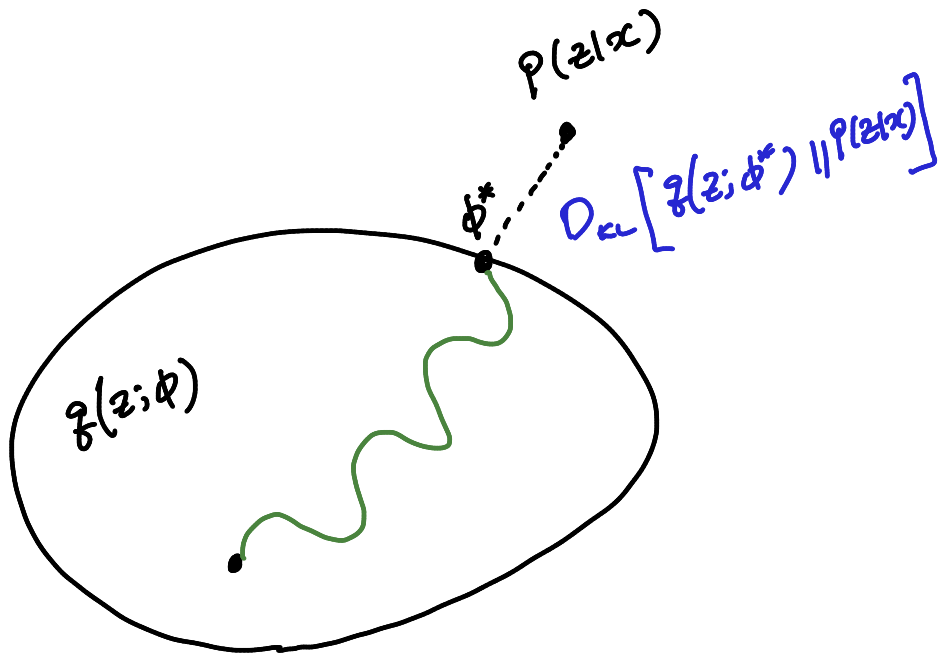
$$p(\mathbf{x}_1, \ldots, \mathbf{x}_N, \mathbf{z}_1, \ldots, \mathbf{z}_N) = \underbrace{p(\mathbf{z}_1)}_{\text{initial state probability}} \underbrace{\left[\prod_{n=2}^{N} p(\mathbf{z}_n|\mathbf{z}_{n-1})\right]}_{\text{state transition probabilities}} \underbrace{\left[\prod_{n=1}^{N} p(\mathbf{x}_n|\mathbf{z}_n)\right]}_{\text{emission probabilities}}.$$

Variational Inference

Variational Inference

- Interested in: $p(\mathbf{z}|\mathbf{x})$ where $\mathbf{x}$ is observed variable and $\mathbf{z}$ is hidden variable.
- Find $q_{\phi^*}(\mathbf{z}|\mathbf{x})$ such that

$$\phi^* = \arg\min_{\phi} D_{KL}\left[q_\phi(\mathbf{z}|\mathbf{x}) \| p(\mathbf{z}|\mathbf{x})\right].$$



[Figure source: David Blei]

$$
\begin{aligned}
D_{KL}\left[q(\mathbf{z};\phi)\|p(\mathbf{z}|\mathbf{x})\right] &= \int q(\mathbf{z};\phi)\log\frac{q(\mathbf{z};\phi)}{p(\mathbf{z}|\mathbf{x})}d\mathbf{z} \\
&= \int q(\mathbf{z};\phi)\log\frac{q(\mathbf{z};\phi)p(\mathbf{x})}{p(\mathbf{x},\mathbf{z})}d\mathbf{z} \\
&= \log p(\mathbf{x}) \underbrace{-\mathbb{E}_q \log p(\mathbf{x},\mathbf{z}) - H(q)}_{\textit{variational free energy}}
\end{aligned}
$$

Evidence Lower Bound (ELBO)

Given a set of **observed (visible)** variables $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$ and a set of **unobserved (hidden)** variables $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_N]$, the evidence is given by

$$\begin{aligned}
\log p(\mathbf{X}) &= \log \int p(\mathbf{X}, \mathbf{Z}) d\mathbf{Z} \\
&= \log \int q(\mathbf{Z}) \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} d\mathbf{Z} \\
&\geq \int q(\mathbf{Z}) \log \left[\frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \right] d\mathbf{Z} \quad \text{(Jensen's inequality)} \\
&= \underbrace{\mathcal{F}(q)}_{ELBO}.
\end{aligned}$$

As in the decomposition in EM, one can easily show that

$$\begin{aligned}
\log p(\mathbf{X}) - \mathcal{F}(q) &= D_{KL}[q(\mathbf{Z}, \boldsymbol{\theta}) \| p(\mathbf{Z}, \boldsymbol{\theta} | \mathbf{X})], \\
\arg\max_q \mathcal{F}(q) &\Rightarrow q(\mathbf{Z}, \boldsymbol{\theta}) = p(\mathbf{Z}, \boldsymbol{\theta} | \mathbf{X}).
\end{aligned}$$

ELBO $\mathcal{F}(q)$ is given by

$$\mathcal{F}(q) = \mathbb{E}_q \log p(\mathbf{X}, \mathbf{Z}) - \mathbb{E}_q \log q(\mathbf{Z}).$$

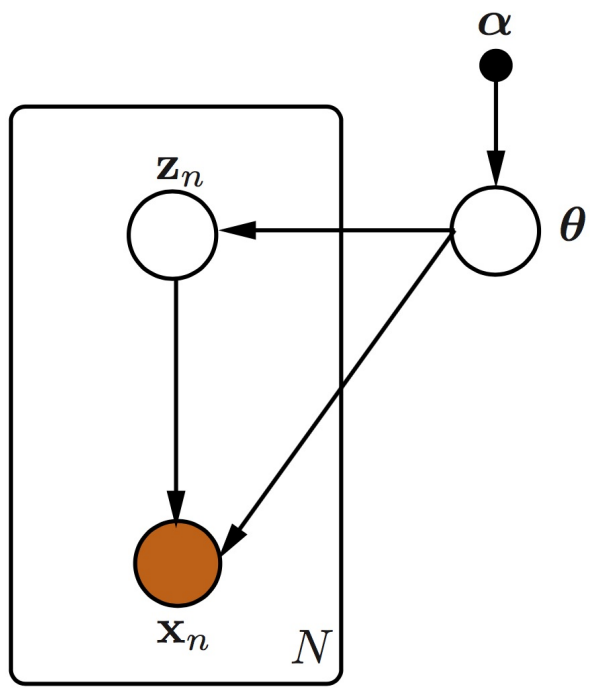
- ELBO is a lower bound on the evidence $p(\mathbf{x})$
- Maximizing the ELBO is equivalent to minimizing the KL divergence between $q(\mathbf{z})$ and $p(\mathbf{z}|\mathbf{x})$.
- The first term prefers $q(\cdot)$ to place its mass on the MAP estimate
- The second term encourages $q(\cdot)$ to diffuse.

Hierarchical Bayesian Models

Variational Bayesian EM

Hierarchical Bayesian Models

We consider a class of models which involves a set of **observed (visible)** variables $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$, a set of **local hidden** variables $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_N]$, **global hidden** variables $\boldsymbol{\theta}$ (mode parameters are treated as random variables as well), and fixed hyperparameters $\boldsymbol{\alpha}$.



The joint distribution factorizes into a global term and a product of local terms,

$$p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\theta}|\boldsymbol{\alpha}) = p(\boldsymbol{\theta}|\boldsymbol{\alpha}) \prod_{n=1}^{N} p(\mathbf{x}_n|\mathbf{z}_n, \boldsymbol{\theta}) p(\mathbf{z}_n|\boldsymbol{\theta}).$$

- **Local variables** can be used for:
  - cluster indicators in mixture models;
  - mixed memberships in topic models;
  - factors (or features) in probabilistic PCA or probabilistic matrix factorization.

- **Global variables** can be used to pool information across data points in the models, serving as model parameters.

Variational Bayesian EM

The log evidence (marginal likelihood) is calculated as:

$$\begin{aligned}
\log p(\mathbf{X}|\boldsymbol{\alpha}) &= \log \int\int p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\theta}|\boldsymbol{\alpha}) d\mathbf{Z} d\boldsymbol{\theta} \\
&= \log \int\int q(\mathbf{Z}, \boldsymbol{\theta}) \frac{p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\theta}|\boldsymbol{\alpha})}{q(\mathbf{Z}, \boldsymbol{\theta})} d\mathbf{Z} d\boldsymbol{\theta} \\
&\geq \int\int q(\mathbf{Z}, \boldsymbol{\theta}) \log \left[\frac{p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\theta}|\boldsymbol{\alpha})}{q(\mathbf{Z}, \boldsymbol{\theta})} \right] d\mathbf{Z} d\boldsymbol{\theta} \\
&= \mathcal{F}(q).
\end{aligned}$$

As in the decomposition in EM, one can easily show that

$$\begin{aligned}
\log p(\mathbf{X}) - \mathcal{F}(q) &= \mathrm{KL}[q(\mathbf{Z}, \boldsymbol{\theta}) \| p(\mathbf{Z}, \boldsymbol{\theta}|\mathbf{X})], \\
\arg\max_q \mathcal{F}(q) &\Rightarrow q(\mathbf{Z}, \boldsymbol{\theta}) = p(\mathbf{Z}, \boldsymbol{\theta}|\mathbf{X}).
\end{aligned}$$

Variational Bayesian EM (cont'd)

We assume that the distribution $q(\mathbf{Z}, \boldsymbol{\theta})$ factorizes as:

$$q(\mathbf{Z}, \boldsymbol{\theta}) = q_z(\mathbf{Z}) q_\theta(\boldsymbol{\theta}),$$

which is known as mean field approximation.

We place no restrictions on the functional forms of $q_z(\mathbf{Z})$ and $q_\theta(\boldsymbol{\theta})$. (free-form optimization)

Then the lower bound to be optimized is given by

$$\mathcal{F}(q_z, q_\theta) = \int \int q_z(\mathbf{Z}) q_\theta(\boldsymbol{\theta}) \log \left[\frac{p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\theta} | \boldsymbol{\alpha})}{q_z(\mathbf{Z}) q_\theta(\boldsymbol{\theta})} \right] d\mathbf{Z} d\boldsymbol{\theta}.$$

Algorithm Outline: VBEM

Optimize $\mathcal{F}(q_z, q_\theta)$ with respect to each of $q_z(\mathbf{Z})$ and $q_\theta(\boldsymbol{\theta})$ in turn, solving

$$\frac{\partial \mathcal{F}(q_z, q_\theta)}{\partial q_z} = 0, \quad \frac{\partial \mathcal{F}(q_z, q_\theta)}{\partial q_\theta} = 0,$$

in turn using calculus of variations (functional differentiation).

- VBE-step: Update $q_z(\mathbf{Z})$

$$q_z^{(k+1)}(\mathbf{Z}) = \frac{1}{Z_z} \exp\left\{\mathbb{E}_{q_\theta^{(k)}} \log p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\theta}|\boldsymbol{\alpha})\right\}.$$

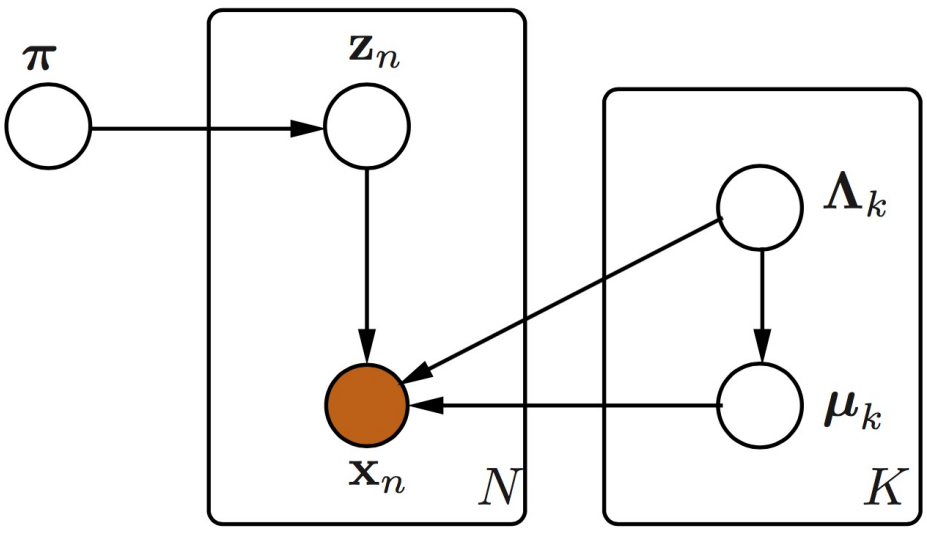
- VBM-step: Update $q_\theta(\boldsymbol{\theta})$

$$q_\theta^{(k+1)}(\boldsymbol{\theta}) = \frac{1}{Z_\theta} \exp\left\{\mathbb{E}_{q_z^{(k+1)}} \log p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\theta}|\boldsymbol{\alpha})\right\}.$$

Example 1:

Variational Mixture of Gaussians

Variational MoG



$$p(\mathbf{x}_n) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}).$$

- Dirichlet prior over $\boldsymbol{\pi}$:

$$p(\boldsymbol{\pi}) = \text{Dir}(\alpha_0, \ldots, \alpha_0).$$

- Gaussian-Wishart prior over $\boldsymbol{\mu}_k$ and $\boldsymbol{\Lambda}_k$:

$$\begin{aligned} p(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) &= p(\boldsymbol{\mu}_k | \boldsymbol{\Lambda}_k) p(\boldsymbol{\Lambda}_k) \\ &= \mathcal{N}\left(\boldsymbol{\mu}_k \,|\, \mathbf{m}_0, (\beta \boldsymbol{\Lambda}_k)^{-1}\right) \\ &\quad \mathcal{W}(\boldsymbol{\Lambda}_k \,|\, \mathbf{W}_0, \nu_0). \end{aligned}$$

- Assume that variational distribution factorizes

$$q(\mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = q(\mathbf{Z}) q(\boldsymbol{\pi}) \prod_{j=1}^{K} q(\boldsymbol{\mu}_k | \boldsymbol{\Lambda}_k) q(\boldsymbol{\Lambda}_k).$$

The log marginal likelihood is given by

$$\begin{aligned}\log p(\mathbf{X}) &= \sum_{\mathbf{Z}} \int\int\int p(\mathbf{X},\mathbf{Z},\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda})\, d\boldsymbol{\pi} d\boldsymbol{\mu} d\boldsymbol{\Lambda} \\ &= \sum_{\mathbf{Z}} \int\int\int p(\mathbf{X}|\mathbf{Z},\boldsymbol{\mu},\boldsymbol{\Lambda}) p(\mathbf{Z}|\boldsymbol{\pi}) p(\boldsymbol{\mu}|\boldsymbol{\Lambda}) p(\boldsymbol{\pi})\, d\boldsymbol{\pi} d\boldsymbol{\mu} d\boldsymbol{\Lambda},\end{aligned}$$

where

$$\begin{aligned}p(\mathbf{X}|\mathbf{Z},\boldsymbol{\mu},\boldsymbol{\Lambda}) &= \prod_{n=1}^{N}\prod_{k=1}^{K} \mathcal{N}\left(\mathbf{x}_n \,|\, \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}\right)^{Z_{k,n}}, \\ p(\mathbf{Z}|\boldsymbol{\pi}) &= \prod_{n=1}^{N}\prod_{k=1}^{K} \pi_k^{Z_{k,n}}, \\ p(\boldsymbol{\mu},\boldsymbol{\Lambda}) &= \prod_{k=1}^{K} \mathcal{N}\left(\boldsymbol{\mu}_k \,|\, \mathbf{m}_0, (\beta\boldsymbol{\Lambda}_k)^{-1}\right) \mathcal{W}(\boldsymbol{\Lambda}_k \,|\, \mathbf{W}_0, \nu_0), \\ p(\boldsymbol{\pi}) &= \frac{1}{Z(\boldsymbol{\alpha}_0)} \prod_{k=1}^{K} \pi_k^{\alpha_0 - 1},\end{aligned}$$

Assume that variational distribution factorizes

$$q(\mathbf{Z}, \pi, \mu, \mathbf{\Lambda}) = \quad q(\mathbf{Z})q(\pi)\prod_{j=1}^{K} q(\mu_k|\mathbf{\Lambda}_k)q(\mathbf{\Lambda}_k).$$

The variational lower-bound is

$$\begin{aligned}
\log p(\mathbf{X}) \quad &\geq \quad \sum_{\mathbf{Z}} \int\int\int q(\mathbf{Z}, \pi, \mu, \mathbf{\Lambda}) \log\left(\frac{p(\mathbf{X}, \mathbf{Z}, \pi, \mu, \mathbf{\Lambda})}{q(\mathbf{Z}, \pi, \mu, \mathbf{\Lambda})}\right) d\pi d\mu d\mathbf{\Lambda} \\
&= \quad \mathbb{E}\log p(\mathbf{X}|\mathbf{Z}, \mu, \mathbf{\Lambda}) + \mathbb{E}\log p(\mathbf{Z}|\pi) + \mathbb{E}\log p(\pi) + \mathbb{E}\log p(\mu, \mathbf{\Lambda}) \\
&\quad -\mathbb{E}\log q(\mathbf{Z}) - \mathbb{E}\log q(\pi) - \mathbb{E}\log q(\mu, \mathbf{\Lambda}).
\end{aligned}$$

Algorithm Outline: Variational E-Step

Compute the optimized variational distributions over latent variables using the current distributions over model parameters:

$$
\begin{aligned}
q^*(\mathbf{Z}) &= \prod_{n=1}^{N}\prod_{j=1}^{K} R_{k,n}^{Z_{k,n}}, \qquad R_{k,n} = \frac{\rho_{k,n}}{\sum_{j=1}^{K}\rho_{j,n}}, \\
\log \rho_{k,n} &= \mathbb{E}_{\boldsymbol{\pi}}\left[\log \pi_k\right] + \frac{1}{2}\mathbb{E}_{\boldsymbol{\Lambda}}\left[\log|\boldsymbol{\Lambda}_k|\right] - \frac{D}{2}\log 2\pi \\
&- \frac{1}{2}\mathbb{E}_{\boldsymbol{\mu},\boldsymbol{\Lambda}}\left[(\mathbf{x}_n - \boldsymbol{\mu}_k)^\top \boldsymbol{\Lambda}_k(\mathbf{x}_n - \boldsymbol{\mu}_k)\right], \\
R_{k,n} &\propto \pi_k|\boldsymbol{\Lambda}_k|^{\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu}_k)^\top \boldsymbol{\Lambda}_k(\mathbf{x}_n - \boldsymbol{\mu}_k)\right\},
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{\pi}}\left\{\log \pi_k\right\} &= \psi(\alpha_k) - \psi(\alpha_1 + \cdots + \alpha_K), \\
\mathbb{E}_{\boldsymbol{\Lambda}}\left\{\log|\boldsymbol{\Lambda}_k|\right\} &= \sum_{i=1}^{D}\psi\left(\frac{\nu_k + 1 - i}{2}\right) + D\log 2 + \log|\mathbf{W}_k|, \\
\mathbb{E}_{\boldsymbol{\mu},\boldsymbol{\Lambda}}\left\{(\mathbf{x}_n - \boldsymbol{\mu}_k)^\top \boldsymbol{\Lambda}_k(\mathbf{x}_n - \boldsymbol{\mu}_k)\right\} &= D\beta_k^{-1} + \nu_k(\mathbf{x}_n - \mathbf{m}_k)^\top \mathbf{W}_k(\mathbf{x}_n - \mathbf{m}_k).
\end{aligned}
$$

Algorithm Outline: Variational M-Step

Compute the optimized variational distributions over model parameters using the current distributions over latent variables:

$$q^*(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \prod_{k=1}^{K} \mathcal{N}\left(\boldsymbol{\mu}_k \,|\, \mathbf{m}_k, (\beta_k \boldsymbol{\Lambda}_k)^{-1}\right) \mathcal{W}\left(\boldsymbol{\Lambda}_k \,|\, \nu_k\right),$$

$$\begin{aligned} N_k &= \sum_{n=1}^{N} R_{k,n}, \quad \bar{\mathbf{x}}_k = \frac{1}{N_k} \sum_{n=1}^{N} R_{k,n} \mathbf{x}_n, \\ \mathbf{S}_k &= \frac{1}{N_k} \sum_{n=1}^{N} R_{k,n} (\mathbf{x}_n - \bar{\mathbf{x}}_k)(\mathbf{x}_n - \bar{\mathbf{x}}_k)^\top. \end{aligned}$$

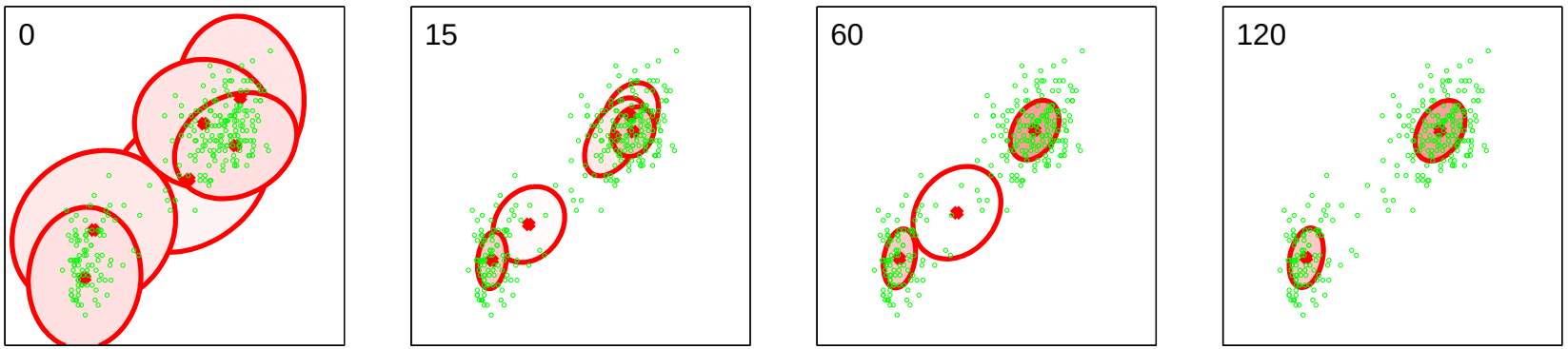
$$\begin{aligned} \beta_k &= \beta_0 + N_k, \quad \mathbf{m}_k = \frac{1}{\beta_k} \left(\beta_0 \mathbf{m}_0 + N_k \bar{\mathbf{x}}_k\right), \quad \nu_j = \nu_0 + N_j, \\ \mathbf{W}_k^{-1} &= \mathbf{W}_0^{-1} + N_k \mathbf{S}_j + \frac{\beta_0 N_k}{\beta_0 + N_k} (\bar{\mathbf{x}}_k - \mathbf{m}_0)(\bar{\mathbf{x}}_k - \mathbf{m}_0)^\top, \end{aligned}$$

The optimized $q(\boldsymbol{\pi})$ is given by

$$q^*(\boldsymbol{\pi}) = \mathsf{Dir}(\boldsymbol{\pi} \mid \boldsymbol{\alpha}) \propto \prod_{k=1}^{K} \pi_k^{\alpha_k - 1},$$

where $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_K]^\top$ and $\alpha_k = \alpha_0 + N_k$.

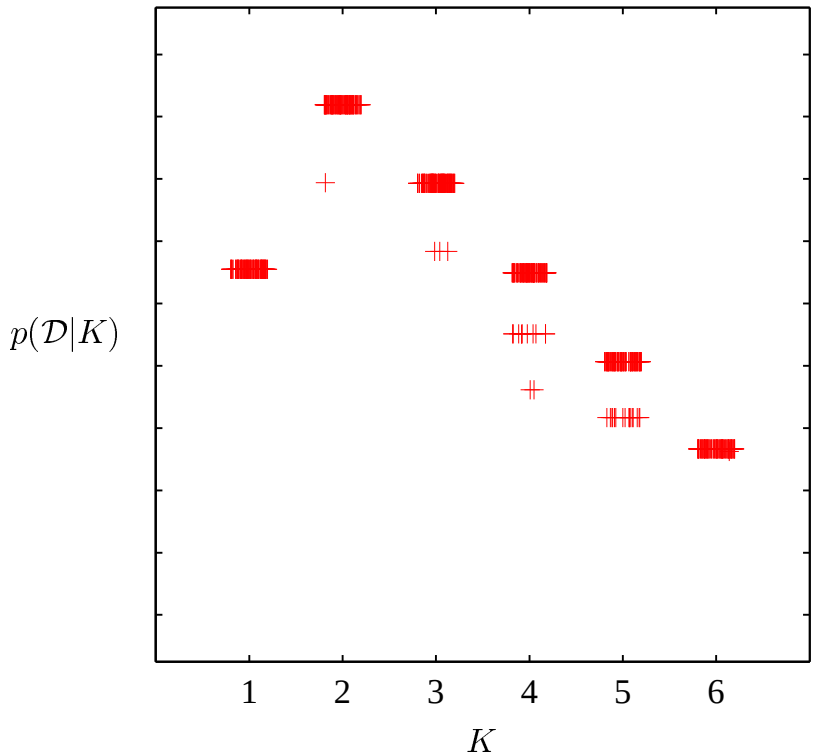
Example



Variational MoG with $K = 6$. Components whose expected mixing coefficients are numerically indistinguishable from zero are not plotted.

[Figure source: Bishop PRML]

Determining the Number of Components



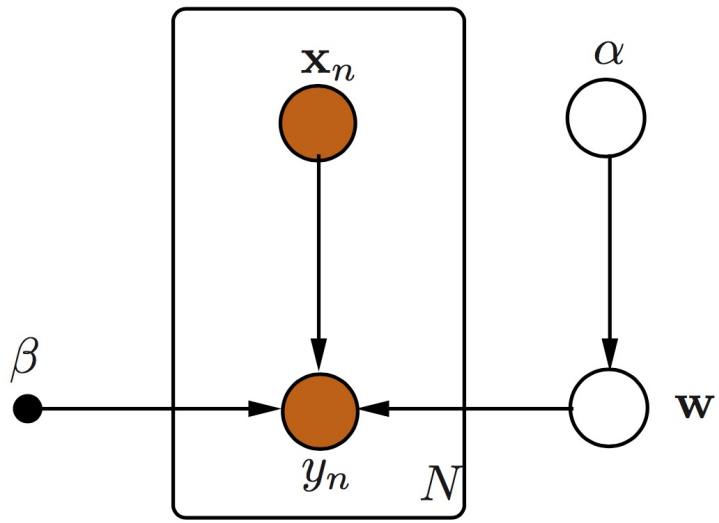
- Plot the variational lower bound on the marginal likelihood versus the number K of components in variational MoG, for the Old Faithful data.
- Peak at $K = 2$.
- For each value of K, the model is trained from 100 different random starts.

[Figure source: Bishop PRML]

Example 2:

Variational Linear Regression

Graphical Model for Variational Linear Regression



- The joint distribution of all the variables is given by

$$\begin{aligned} & p(\mathbf{y}, \mathbf{X}, \mathbf{w}, \alpha) \\ & \quad = \quad p(\mathbf{y}|\mathbf{X}, \mathbf{w}) p(\mathbf{w}|\alpha) p(\alpha). \end{aligned}$$

- The likelihood for $\mathbf{w}$ is given by

$$p(\mathbf{y}|\mathbf{X}, \mathbf{w}) = \prod_{n=1}^{N} \mathcal{N}\left(y_n|\mathbf{w}^\top \mathbf{x}_n, \beta^{-1}\right).$$

- The prior over $\mathbf{w}$ is given by

$$p(\mathbf{w}|\alpha) = \mathcal{N}\left(\mathbf{w}|0, \alpha^{-1}\mathbf{I}\right).$$

- The prior over α is given by

$$p(\alpha) = \mathsf{Gam}(\alpha|a_0, b_0).$$

Algorithm Outline

- Re-estimate $q^*(\alpha)$:

$$\begin{aligned} q^*(\alpha) &= \mathrm{Gam}(\alpha \,|\, a_N, b_N), \\ a_N &= a_0 + \frac{D}{2}, \\ b_N &= b_0 + \frac{1}{2}\mathbb{E}_{\mathbf{w}}\left[\mathbf{w}^\top \mathbf{w}\right], \\ \mathbb{E}_{\mathbf{w}}\left[\mathbf{w}^\top \mathbf{w}\right] &= \mathbf{m}_N^\top \mathbf{m}_N + \mathrm{tr}\left\{\mathbf{S}_N\right\}. \end{aligned}$$

- Re-estimate $q^*(\mathbf{w})$:

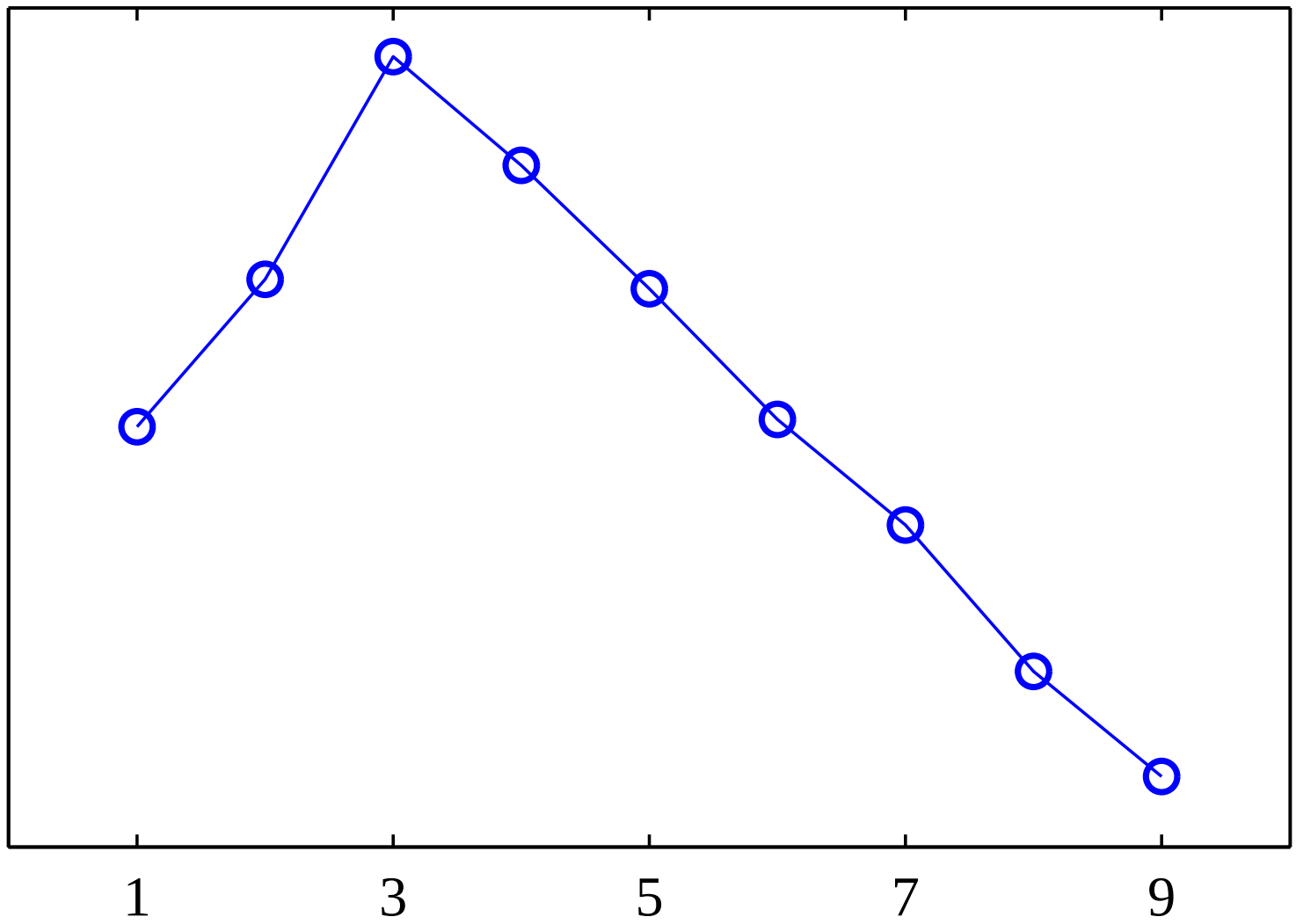
$$\begin{aligned} q^*(\mathbf{w}) &= \mathcal{N}(\mathbf{w} \,|\, \mathbf{m}_N, \mathbf{S}_N), \\ \mathbf{m}_N &= \beta \mathbf{S}_N \mathbf{X}^\top \mathbf{y}, \\ \mathbf{S}_N &= \left(\mathbb{E}_\alpha\left[\alpha\right]\mathbf{I} + \beta \mathbf{X}^\top \mathbf{X}\right)^{-1}, \\ \mathbb{E}_\alpha\left\{\alpha\right\} &= \frac{a_N}{b_N}. \end{aligned}$$

Predictive Distribution

The predictive distribution over y_*, given a new input $\mathbf{x}_*$, is evaluated using the Gaussian variational posterior:

$$
\begin{aligned}
p(y_*|\mathbf{x}_*, \mathbf{X}, \mathbf{y}) &= \int p(y_*, \mathbf{w}|\mathbf{x}_*, \mathbf{X}, \mathbf{y}) d\mathbf{w} \\
&= \int p(y_*|\mathbf{w}, \mathbf{x}_*) p(\mathbf{w}|\mathbf{X}, \mathbf{y}) d\mathbf{w} \\
&\approx \int p(y_*|\mathbf{w}, \mathbf{x}_*) q(\mathbf{w}) d\mathbf{w} \\
&= \int \mathcal{N}\left(y_* \,|\, \mathbf{w}^\top \mathbf{x}_*, \beta^{-1}\right) \mathcal{N}\left(\mathbf{w} \,|\, \mathbf{m}_N, \mathbf{S}_N\right) d\mathbf{w} \\
&= \mathcal{N}\left(y_* \,|\, \mathbf{m}_N^\top \mathbf{x}_*, \beta^{-1} + \mathbf{x}_*^\top \mathbf{S}_N \mathbf{x}_*\right).
\end{aligned}
$$

Lower Bound vs. Model Order in Polynomial Curve Fitting



Summary

- Variational method and inference
- Bayesian hierarchical models and variational EM
- Examples: variational MoG and variational linear regression

- Expressiveness of variational distributions
  - Mean-field approximation where a factorized form of distribution is assumed
  - Structured mean-field approximations that incorporate some basic form of dependency within the approximate posterior
  - Approximate posterior as a mixture model
  - Normalizing flows
  - Hierarchical variational models
  - Implicit models