CSci 8980: Advanced Topics in Graphical Models

Mixture Models, EM, Exponential Families

Instructor: Arindam Banerjee

September 11, 2007

Incremental EM

- Since z_i are independent, optimal $\tilde{p}(Z) = \prod_i \tilde{p}(z_i)$

Incremental EM

- Since z_i are independent, optimal $\tilde{p}(Z) = \prod_i \tilde{p}(z_i)$
- Sufficient to work with such $\tilde{p}$ in $F(\tilde{p}, \theta)$

Incremental EM

- Since z_i are independent, optimal $\tilde{p}(Z) = \prod_i \tilde{p}(z_i)$
- Sufficient to work with such $\tilde{p}$ in $F(\tilde{p}, \theta)$
- Then $F(\tilde{p}, \theta) = \sum_i F_i(\tilde{p}_i, \theta)$ where

$$F_i(\tilde{p}_i, \theta) = E_{\tilde{p}_i}[\log p(x_i, z_i|\theta)] + H(\tilde{p}_i)$$

Incremental EM

- Since z_i are independent, optimal $\tilde{p}(Z) = \prod_i \tilde{p}(z_i)$
- Sufficient to work with such $\tilde{p}$ in $F(\tilde{p}, \theta)$
- Then $F(\tilde{p}, \theta) = \sum_i F_i(\tilde{p}_i, \theta)$ where

$$F_i(\tilde{p}_i, \theta) = E_{\tilde{p}_i}[\log p(x_i, z_i|\theta)] + H(\tilde{p}_i)$$

- Incremental algorithm that works one point at a time

Incremental EM (Contd.)

- Basic Incremental EM

Incremental EM (Contd.)

- Basic Incremental EM
  - E-step: Choose a data item i to be updated
    Set $\tilde{p}_j^{(t)} = \tilde{p}_j^{(t-1)}$ for $j \neq i$
    Set $\tilde{p}_i^{(t)} = p(z_i|x_i, \theta^{(t)})$

Incremental EM (Contd.)

- Basic Incremental EM
  - E-step: Choose a data item i to be updated
    Set $\tilde{p}_j^{(t)} = \tilde{p}_j^{(t-1)}$ for $j \neq i$
    Set $\tilde{p}_i^{(t)} = p(z_i|x_i, \theta^{(t)})$
  - M-step: Set $\theta^{(t)}$ to $\text{argmax}_\theta \; E_{\tilde{p}^{(t)}}[\log p(x, z|\theta)]$

Incremental EM (Contd.)

- Basic Incremental EM
  - E-step: Choose a data item i to be updated
    Set $\tilde{p}_j^{(t)} = \tilde{p}_j^{(t-1)}$ for $j \neq i$
    Set $\tilde{p}_i^{(t)} = p(z_i|x_i, \theta^{(t)})$
  - M-step: Set $\theta^{(t)}$ to $\text{argmax}_\theta \; E_{\tilde{p}^{(t)}}[\log p(x, z|\theta)]$
- M-step needs to look at all components of $\tilde{p}$

Incremental EM (Contd.)

- Basic Incremental EM
  - E-step: Choose a data item i to be updated
    Set $\tilde{p}_j^{(t)} = \tilde{p}_j^{(t-1)}$ for $j \neq i$
    Set $\tilde{p}_i^{(t)} = p(z_i|x_i, \theta^{(t)})$
  - M-step: Set $\theta^{(t)}$ to $\text{argmax}_\theta \;\; E_{\tilde{p}^{(t)}}[\log p(x, z|\theta)]$
- M-step needs to look at all components of $\tilde{p}$
- Can be simplified by using sufficient statistics

Incremental EM (Contd.)

- Basic Incremental EM
  - E-step: Choose a data item i to be updated
    Set $\tilde{p}_j^{(t)} = \tilde{p}_j^{(t-1)}$ for $j \neq i$
    Set $\tilde{p}_i^{(t)} = p(z_i|x_i, \theta^{(t)})$
  - M-step: Set $\theta^{(t)}$ to $\text{argmax}_\theta \; E_{\tilde{p}^{(t)}}[\log p(x, z|\theta)]$
- M-step needs to look at all components of $\tilde{p}$
- Can be simplified by using sufficient statistics
- For a distribution $p(x|\theta)$, $s(x)$ is a sufficient statistic if

$$p(x|s(x), \theta) = p(x|s(x)) \implies p(x|\theta) = h(x)q(s(x)|\theta)$$

Incremental EM with Sufficient Statistics

- EM with sufficient statistics

Incremental EM with Sufficient Statistics

- EM with sufficient statistics
  - E-step: Set $\tilde{s}^{(t)} = E_{\tilde{p}}[s(x,z)]$ where $\tilde{p}(z) = p(z|x, \theta^{(t-1)})$

Incremental EM with Sufficient Statistics

- EM with sufficient statistics
  - E-step: Set $\tilde{s}^{(t)} = E_{\tilde{p}}[s(x, z)]$ where $\tilde{p}(z) = p(z|x, \theta^{(t-1)})$
  - M-step: Set $\theta^{(t)}$ to θ, the max likelihood given $\tilde{s}^{(t)}$

Incremental EM with Sufficient Statistics

- EM with sufficient statistics
  - E-step: Set $\tilde{s}^{(t)} = E_{\tilde{p}}[s(x, z)]$ where $\tilde{p}(z) = p(z|x, \theta^{(t-1)})$
  - M-step: Set $\theta^{(t)}$ to θ, the max likelihood given $\tilde{s}^{(t)}$
- Incremental EM with sufficient statistics

Incremental EM with Sufficient Statistics

- EM with sufficient statistics
  - E-step: Set $\tilde{s}^{(t)} = E_{\tilde{p}}[s(x,z)]$ where $\tilde{p}(z) = p(z|x, \theta^{(t-1)})$
  - M-step: Set $\theta^{(t)}$ to θ, the max likelihood given $\tilde{s}^{(t)}$
- Incremental EM with sufficient statistics
  - E-step: Choose a data item i to be updated
    Set $\tilde{s}_j^{(t)} = \tilde{s}_j^{(t-1)}$, for $j \neq i$
    Set $\tilde{s}_i^{(t)} = E_{\tilde{p}_i}[s_i(x_i, z_i)]$, where $\tilde{p}_i(z_i) = p(z_i|x_i, \theta^{(t-1)})$
    Set $\tilde{s}^{(t)} = \tilde{s}^{(t-1)} - \tilde{s}_i^{(t-1)} + \tilde{s}_i^{(t)}$

Incremental EM with Sufficient Statistics

- EM with sufficient statistics
  - E-step: Set $\tilde{s}^{(t)} = E_{\tilde{p}}[s(x,z)]$ where $\tilde{p}(z) = p(z|x, \theta^{(t-1)})$
  - M-step: Set $\theta^{(t)}$ to θ, the max likelihood given $\tilde{s}^{(t)}$
- Incremental EM with sufficient statistics
  - E-step: Choose a data item i to be updated
    Set $\tilde{s}_j^{(t)} = \tilde{s}_j^{(t-1)}$, for $j \neq i$
    Set $\tilde{s}_i^{(t)} = E_{\tilde{p}_i}[s_i(x_i, z_i)]$, where $\tilde{p}_i(z_i) = p(z_i|x_i, \theta^{(t-1)})$
    Set $\tilde{s}^{(t)} = \tilde{s}^{(t-1)} - \tilde{s}_i^{(t-1)} + \tilde{s}_i^{(t)}$
  - M-step: Set $\theta^{(t)}$ to θ, the max likelihood given $\tilde{s}^{(t)}$

Example

- Consider a mixture of 2 univariate Gaussians

Example

- Consider a mixture of 2 univariate Gaussians
- Parameter set $\theta = (\alpha, \mu_1, \sigma_1, \mu_2, \sigma_2)$

Example

- Consider a mixture of 2 univariate Gaussians
- Parameter set $\theta = (\alpha, \mu_1, \sigma_1, \mu_2, \sigma_2)$
- Sufficient statistics

$$s_i(x_i, z_i) = [z_i \ (1 - z_i) \ z_i x_i \ (1 - z_i) x_i \ z_i x_i^2 \ (1 - z_i) x_i^2]$$

Example

- Consider a mixture of 2 univariate Gaussians
- Parameter set $\theta = (\alpha, \mu_1, \sigma_1, \mu_2, \sigma_2)$
- Sufficient statistics

$$s_i(x_i, z_i) = [z_i \;\; (1 - z_i) \;\; z_i x_i \;\; (1 - z_i)x_i \;\; z_i x_i^2 \;\; (1 - z_i)x_i^2]$$

- Given $s(x, z) = \sum_i s(x_i, z_i) = (n_1, n_2, m_1, m_2, q_1, q_2)$

$$\alpha = \frac{n_1}{n_1 + n_2}\,, \qquad \mu_h = \frac{m_h}{n_h}\,, \qquad \sigma_h^2 = \frac{q_h}{n_h} - \left(\frac{m_h}{n_h}\right)^2$$

Sparse EM

- Consider a mixture model with many components

Sparse EM

- Consider a mixture model with many components
- Most $p(z|x,\theta)$ will be negligibly small

Sparse EM

- Consider a mixture model with many components
- Most $p(z|x,\theta)$ will be negligibly small
- Computation can be saved by freezing these

Sparse EM

- Consider a mixture model with many components
- Most $p(z|x,\theta)$ will be negligibly small
- Computation can be saved by freezing these
- Only a small set of component posteriors need to be updated

$$\tilde{p}^{(t)}(z) = \begin{cases} q_z^{(t)}, & \text{if } z \notin S_t \\ Q^{(t)} r_z^{(t)} & \text{if } z \in S_t \end{cases}$$

Sparse EM

- Consider a mixture model with many components
- Most $p(z|x,\theta)$ will be negligibly small
- Computation can be saved by freezing these
- Only a small set of component posteriors need to be updated

$$\tilde{p}^{(t)}(z) = \begin{cases} q_z^{(t)}, & \text{if } z \notin S_t \\ Q^{(t)} r_z^{(t)} & \text{if } z \in S_t \end{cases}$$

- S_t = set of plausible values

Sparse EM

- Consider a mixture model with many components
- Most $p(z|x, \theta)$ will be negligibly small
- Computation can be saved by freezing these
- Only a small set of component posteriors need to be updated

$$\tilde{p}^{(t)}(z) = \begin{cases} q_z^{(t)}, & \text{if } z \notin S_t \\ Q^{(t)} r_z^{(t)} & \text{if } z \in S_t \end{cases}$$

- S_t = set of plausible values
  - Can be determined by a reasonable hueristic

Other Variants

- Generalized EM

Other Variants

- Generalized EM
  - M-step finds $\theta^{(t)} = \text{argmax}_\theta \;\; E_{\tilde{p}}[\log p(x, z|\theta)]$

Other Variants

- Generalized EM
  - M-step finds $\theta^{(t)} = \text{argmax}_\theta \; E_{\tilde{p}}[\log p(x, z|\theta)]$
  - Instead find $\theta^{(t)}$ such that

$$E_{\tilde{p}}[\log p(x, z|\theta^{(t)})] \geq E_{\tilde{p}}[\log p(x, z|\theta^{(t-1)})]$$

Other Variants

- Generalized EM
  - M-step finds $\theta^{(t)} = \text{argmax}_\theta \ E_{\tilde{p}}[\log p(x, z|\theta)]$
  - Instead find $\theta^{(t)}$ such that

$$E_{\tilde{p}}[\log p(x, z|\theta^{(t)})] \geq E_{\tilde{p}}[\log p(x, z|\theta^{(t-1)})]$$

- Hard assignments

Other Variants

- Generalized EM
  - M-step finds $\theta^{(t)} = \text{argmax}_\theta \;\; E_{\tilde{p}}[\log p(x, z|\theta)]$
  - Instead find $\theta^{(t)}$ such that

$$E_{\tilde{p}}[\log p(x, z|\theta^{(t)})] \geq E_{\tilde{p}}[\log p(x, z|\theta^{(t-1)})]$$

- Hard assignments
  - Winner-take-all variant of EM

Other Variants

- Generalized EM
  - M-step finds $\theta^{(t)} = \text{argmax}_\theta \; E_{\tilde{p}}[\log p(x, z|\theta)]$
  - Instead find $\theta^{(t)}$ such that

$$E_{\tilde{p}}[\log p(x, z|\theta^{(t)})] \geq E_{\tilde{p}}[\log p(x, z|\theta^{(t-1)})]$$

- Hard assignments
  - Winner-take-all variant of EM
  - Assign 1 to one component, zero to all others

Other Variants

- Generalized EM
  - M-step finds $\theta^{(t)} = \text{argmax}_\theta \; E_{\tilde{p}}[\log p(x, z|\theta)]$
  - Instead find $\theta^{(t)}$ such that

$$E_{\tilde{p}}[\log p(x, z|\theta^{(t)})] \geq E_{\tilde{p}}[\log p(x, z|\theta^{(t-1)})]$$

- Hard assignments
  - Winner-take-all variant of EM
  - Assign 1 to one component, zero to all others
  - Hard clustering, equivalent to kmeans

Other Variants

- Generalized EM
  - M-step finds $\theta^{(t)} = \text{argmax}_\theta \; E_{\tilde{p}}[\log p(x, z|\theta)]$
  - Instead find $\theta^{(t)}$ such that

$$E_{\tilde{p}}[\log p(x, z|\theta^{(t)})] \geq E_{\tilde{p}}[\log p(x, z|\theta^{(t-1)})]$$

- Hard assignments
  - Winner-take-all variant of EM
  - Assign 1 to one component, zero to all others
  - Hard clustering, equivalent to kmeans
  - Does not directly optimize $L(\theta)$

Other Variants

- Generalized EM
  - M-step finds $\theta^{(t)} = \text{argmax}_\theta \; E_{\tilde{p}}[\log p(x, z|\theta)]$
  - Instead find $\theta^{(t)}$ such that

$$E_{\tilde{p}}[\log p(x, z|\theta^{(t)})] \geq E_{\tilde{p}}[\log p(x, z|\theta^{(t-1)})]$$

- Hard assignments
  - Winner-take-all variant of EM
  - Assign 1 to one component, zero to all others
  - Hard clustering, equivalent to kmeans
  - Does not directly optimize $L(\theta)$
  - But optimizes a lower bound on $L(\theta)$

Auxiliary Functions

- Consider the problem of minimizing $F(x)$

Auxiliary Functions

- Consider the problem of minimizing $F(x)$
- $G(x, x')$ is an auxiliary function to $F(x)$ if

$$G(x, x') \geq F(x) \qquad G(x, x) = F(x)$$

Auxiliary Functions

- Consider the problem of minimizing $F(x)$
- $G(x, x')$ is an auxiliary function to $F(x)$ if

$$G(x, x') \geq F(x) \qquad G(x, x) = F(x)$$

- F is non-decreasing under the following updates

$$x^t = \text{argmin}_x \; G(x, x^{(t-1)})$$

Auxiliary Functions

- Consider the problem of minimizing $F(x)$
- $G(x, x')$ is an auxiliary function to $F(x)$ if

$$G(x, x') \geq F(x) \qquad G(x, x) = F(x)$$

- F is non-decreasing under the following updates

$$x^t = \text{argmin}_x \; G(x, x^{(t-1)})$$

- By definition

$$F(x^t) \leq G(x^t, x^{(t-1)}) \leq G(x^{(t-1)}, x^{(t-1)}) = F(x^{(t-1)})$$

Auxiliary Functions

- Consider the problem of minimizing $F(x)$
- $G(x, x')$ is an auxiliary function to $F(x)$ if

$$G(x, x') \geq F(x) \qquad G(x, x) = F(x)$$

- F is non-decreasing under the following updates

$$x^t = \text{argmin}_x \ G(x, x^{(t-1)})$$

- By definition

$$F(x^t) \leq G(x^t, x^{(t-1)}) \leq G(x^{(t-1)}, x^{(t-1)}) = F(x^{(t-1)})$$

- The sequence is guaranteed to converge to a local minima

Auxiliary Functions

- Consider the problem of minimizing $F(x)$
- $G(x, x')$ is an auxiliary function to $F(x)$ if

$$G(x, x') \geq F(x) \qquad G(x, x) = F(x)$$

- F is non-decreasing under the following updates

$$x^t = \text{argmin}_x \; G(x, x^{(t-1)})$$

- By definition

$$F(x^t) \leq G(x^t, x^{(t-1)}) \leq G(x^{(t-1)}, x^{(t-1)}) = F(x^{(t-1)})$$

- The sequence is guaranteed to converge to a local minima
- The argument reverses for maximization problems

Auxiliary Functions

- Consider the problem of minimizing $F(x)$
- $G(x, x')$ is an auxiliary function to $F(x)$ if

$$G(x, x') \geq F(x) \qquad G(x, x) = F(x)$$

- F is non-decreasing under the following updates

$$x^t = \text{argmin}_x \;\; G(x, x^{(t-1)})$$

- By definition

$$F(x^t) \leq G(x^t, x^{(t-1)}) \leq G(x^{(t-1)}, x^{(t-1)}) = F(x^{(t-1)})$$

- The sequence is guaranteed to converge to a local minima
- The argument reverses for maximization problems
- EM updates are a special case of the general technique

Mixture of Gaussians

- For multi-variate Gaussians, each component

$$p_h(x|\mu_h, \Sigma_h) = \frac{1}{(2\pi)^{d/2}|\Sigma_h|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_h)^T \Sigma_h^{-1}(x-\mu_h)\right)$$

Mixture of Gaussians

- For multi-variate Gaussians, each component

$$p_h(x|\mu_h, \Sigma_h) = \frac{1}{(2\pi)^{d/2}|\Sigma_h|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_h)^T \Sigma_h^{-1}(x-\mu_h)\right)$$

- The Mixture of Gaussians (MoG) model

$$p(x|\alpha, \Theta) = \sum_{h=1}^{k} \alpha_h p_h(x|\mu_h, \Sigma_h)$$

Mixture of Gaussians

- For multi-variate Gaussians, each component

$$p_h(x|\mu_h, \Sigma_h) = \frac{1}{(2\pi)^{d/2}|\Sigma_h|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_h)^T \Sigma_h^{-1}(x-\mu_h)\right)$$

- The Mixture of Gaussians (MoG) model

$$p(x|\alpha, \Theta) = \sum_{h=1}^{k} \alpha_h p_h(x|\mu_h, \Sigma_h)$$

- One of the most widely used mixture models

Mixture of Gaussians

- For multi-variate Gaussians, each component

$$p_h(x|\mu_h, \Sigma_h) = \frac{1}{(2\pi)^{d/2}|\Sigma_h|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_h)^T \Sigma_h^{-1}(x-\mu_h)\right)$$

- The Mixture of Gaussians (MoG) model

$$p(x|\alpha, \Theta) = \sum_{h=1}^{k} \alpha_h p_h(x|\mu_h, \Sigma_h)$$

- One of the most widely used mixture models
- Recent years have seen progress on non-EM algorithm

EM for Mixture of Gaussians: E-step

- E-step is a direct application of Bayes rule

$$p(h|x, \alpha, \Theta) = \frac{\alpha_h p_h(x|\mu_h, \Sigma_h)}{\sum_{h'=1}^{k} \alpha_{h'} p_{h'}(x|\mu_{h'}, \Sigma_{h'})}$$

EM for Mixture of Gaussians: E-step

- E-step is a direct application of Bayes rule

$$p(h|x, \alpha, \Theta) = \frac{\alpha_h p_h(x|\mu_h, \Sigma_h)}{\sum_{h'=1}^{k} \alpha_{h'} p_{h'}(x|\mu_{h'}, \Sigma_{h'})}$$

- Use current parameter values on the r.h.s.

EM for Mixture of Gaussians: E-step

- E-step is a direct application of Bayes rule

$$p(h|x, \alpha, \Theta) = \frac{\alpha_h p_h(x|\mu_h, \Sigma_h)}{\sum_{h'=1}^{k} \alpha_{h'} p_{h'}(x|\mu_{h'}, \Sigma_{h'})}$$

- Use current parameter values on the r.h.s.
- Incremental and sparse variants can be applied in practice

EM for Mixture of Gaussians: M-step

- The auxiliary function

$$Q(\theta, \theta^{(t-1)}) = \sum_i \sum_h \log(\alpha_h) p(h|x_i|\theta^{(t-1)}) + \sum_i \sum_h \log p_h(x|\mu_h, \Sigma_h) p(h|x_i, \theta^{(t-1)})$$

EM for Mixture of Gaussians: M-step

- The auxiliary function

$$Q(\theta, \theta^{(t-1)}) = \sum_i \sum_h \log(\alpha_h) p(h|x_i|\theta^{(t-1)}) + \sum_i \sum_h \log p_h(x|\mu_h, \Sigma_h) p(h|x_i, \theta^{(t-1)})$$

- Optimize over $(\alpha_h, \mu_h, \Sigma_h), [h]_1^k$

EM for Mixture of Gaussians: M-step

- The auxiliary function

$$Q(\theta, \theta^{(t-1)}) = \sum_i \sum_h \log(\alpha_h) p(h|x_i|\theta^{(t-1)}) + \sum_i \sum_h \log p_h(x|\mu_h, \Sigma_h) p(h|x_i, \theta^{(t-1)})$$

- Optimize over $(\alpha_h, \mu_h, \Sigma_h), [h]_1^k$
- α is a discrete distribution, forms additional constraint

EM for Mixture of Gaussians: M-step

- The auxiliary function

$$Q(\theta, \theta^{(t-1)}) = \sum_i \sum_h \log(\alpha_h) p(h|x_i|\theta^{(t-1)}) + \sum_i \sum_h \log p_h(x|\mu_h, \Sigma_h) p(h|x_i, \theta^{(t-1)})$$

- Optimize over $(\alpha_h, \mu_h, \Sigma_h), [h]_1^k$
- α is a discrete distribution, forms additional constraint
- Focus on first term for α_h, true for all mixtures

EM for Mixture of Gaussians: M-step

- The auxiliary function

$$Q(\theta, \theta^{(t-1)}) = \sum_i \sum_h \log(\alpha_h) p(h|x_i|\theta^{(t-1)}) + \sum_i \sum_h \log p_h(x|\mu_h, \Sigma_h) p(h|x_i, \theta^{(t-1)})$$

- Optimize over $(\alpha_h, \mu_h, \Sigma_h), [h]_1^k$
- α is a discrete distribution, forms additional constraint
- Focus on first term for α_h, true for all mixtures
- Focus on second term for (μ_h, Σ_h)

EM for Mixture of Gaussians: M-step (Contd.)

- For any finite mixture model

$$\alpha_h = \frac{1}{N} \sum_{i=1}^{N} p(h|x_i, \theta^{(t-1)})$$

EM for Mixture of Gaussians: M-step (Contd.)

- For any finite mixture model

$$\alpha_h = \frac{1}{N} \sum_{i=1}^{N} p(h|x_i, \theta^{(t-1)})$$

- For Mixture of Gaussians

$$\mu_h = \frac{\sum_i x_i p(h|x_i, \theta^{(t-1)})}{\sum_i p(h|x_i, \theta^{(t-1)})}$$

$$\Sigma_h = \frac{\sum_i p(h|x_i, \theta_n)(x_i - \mu_h)(x_i - \mu_h)^T}{\sum_i p(h|x_i, \theta_n)}$$

Exponential Family Distributions

- Multi-variate parametric distributions of the form

$$p_\psi(x|\theta) = \exp(x^T\theta - \psi(\theta))p_0(x)$$

Exponential Family Distributions

- Multi-variate parametric distributions of the form

$$p_\psi(x|\theta) = \exp(x^T\theta - \psi(\theta))p_0(x)$$

- x is the sufficient statistic

Exponential Family Distributions

- Multi-variate parametric distributions of the form

$$p_\psi(x|\theta) = \exp(x^T\theta - \psi(\theta))p_0(x)$$

- x is the sufficient statistic
- θ is the natural parameter

Exponential Family Distributions

- Multi-variate parametric distributions of the form

$$p_\psi(x|\theta) = \exp(x^T\theta - \psi(\theta))p_0(x)$$

- x is the sufficient statistic
- θ is the natural parameter
- $\psi(\cdot)$ is the cumulant or log-partition function

Exponential Family Distributions

- Multi-variate parametric distributions of the form

$$p_\psi(x|\theta) = \exp(x^T\theta - \psi(\theta))p_0(x)$$

- x is the sufficient statistic
- θ is the natural parameter
- $\psi(\cdot)$ is the cumulant or log-partition function
- Expectation parameter

$$\mu = E[X] = \nabla\psi(\theta)$$

Exponential Family Distributions

- Multi-variate parametric distributions of the form

$$p_\psi(x|\theta) = \exp(x^T\theta - \psi(\theta))p_0(x)$$

- x is the sufficient statistic
- θ is the natural parameter
- $\psi(\cdot)$ is the cumulant or log-partition function
- Expectation parameter

$$\mu = E[X] = \nabla\psi(\theta)$$

- Examples: Gaussian, Bernoulli, Poisson, Multinomial, Dirichlet

The Cumulant Function

- The Laplace transform viewpoint

$$L(\theta) = \exp(\psi(\theta)) = \int_x \exp(\mathbf{x}^T\theta)p_0(x)\ dx = E_{p_0}[\exp(\mathbf{x}^T\theta)]$$

The Cumulant Function

- The Laplace transform viewpoint

$$L(\theta) = \exp(\psi(\theta)) = \int_x \exp(\mathbf{x}^T\theta)p_0(x)\ dx = E_{p_0}[\exp(\mathbf{x}^T\theta)]$$

- Holder's inequality implies: For $1 \leq p, q \leq \infty, 1/p + 1/q = 1$,

$$E[|X|^p]^{1/p}E[|Y|^q]^{1/q} \geq E[|XY|]$$

The Cumulant Function

- The Laplace transform viewpoint

$$L(\theta) = \exp(\psi(\theta)) = \int_x \exp(\mathbf{x}^T\theta)p_0(x)\; dx = E_{p_0}[\exp(\mathbf{x}^T\theta)]$$

- Holder's inequality implies: For $1 \le p, q \le \infty, 1/p + 1/q = 1$,

$$E[|X|^p]^{1/p} E[|Y|^q]^{1/q} \ge E[|XY|]$$

- Hence

$$\begin{aligned}
&\lambda\psi(\theta_1) + (1-\lambda)\psi(\theta_2) \\
&\quad = \log\left(E_{p_0}[\exp(x^T\theta_1)]^\lambda E_{p_0}[\exp(x^T\theta_2)]^{1-\lambda}\right) \\
&\quad \ge \log\left(E_{p_0}[\exp(x^T(\lambda\theta_1 + (1-\lambda)\theta_2))]\right) \\
&\quad = \psi(\lambda\theta_1 + (1-\lambda)\theta_2)
\end{aligned}$$

The Cumulant Function

- The Laplace transform viewpoint

$$L(\theta) = \exp(\psi(\theta)) = \int_x \exp(\mathbf{x}^T\theta) p_0(x)\ dx = E_{p_0}[\exp(\mathbf{x}^T\theta)]$$

- Holder's inequality implies: For $1 \leq p, q \leq \infty, 1/p + 1/q = 1$,

$$E[|X|^p]^{1/p} E[|Y|^q]^{1/q} \geq E[|XY|]$$

- Hence

$$\begin{aligned}
&\lambda\psi(\theta_1) + (1-\lambda)\psi(\theta_2) \\
&\quad = \log\left(E_{p_0}[\exp(x^T\theta_1)]^\lambda E_{p_0}[\exp(x^T\theta_2)]^{1-\lambda}\right) \\
&\quad \geq \log\left(E_{p_0}[\exp(x^T(\lambda\theta_1 + (1-\lambda)\theta_2))]\right) \\
&\quad = \psi(\lambda\theta_1 + (1-\lambda)\theta_2)
\end{aligned}$$

- The cumulant $\psi(\theta)$ is a convex function

Maximum Likelihood Estimation, Conjugate

- Let $s = s(x)$ be the sufficient statistic for a set of points x

Maximum Likelihood Estimation, Conjugate

- Let $s = s(x)$ be the sufficient statistic for a set of points x
- Then maximizing log-likelihood is

$$\phi(s) = \max_{\theta} \left(s^T\theta - \psi(\theta)\right)$$

Maximum Likelihood Estimation, Conjugate

- Let $s = s(x)$ be the sufficient statistic for a set of points x
- Then maximizing log-likelihood is

$$\phi(s) = \max_{\theta} \left(s^T\theta - \psi(\theta) \right)$$

- Has a unique maximizer since $\psi(\theta)$ is convex

Maximum Likelihood Estimation, Conjugate

- Let $s = s(x)$ be the sufficient statistic for a set of points x
- Then maximizing log-likelihood is

$$\phi(s) = \max_{\theta} \left(s^T\theta - \psi(\theta)\right)$$

- Has a unique maximizer since $\psi(\theta)$ is convex
- The conjugate of ψ is

$$\phi(s) = \sup_{\theta} \left(s^T\theta - \psi(\theta)\right)$$

Maximum Likelihood Estimation, Conjugate

- Let $s = s(x)$ be the sufficient statistic for a set of points x
- Then maximizing log-likelihood is

$$\phi(s) = \max_{\theta} \left(s^T\theta - \psi(\theta)\right)$$

- Has a unique maximizer since $\psi(\theta)$ is convex
- The conjugate of ψ is

$$\phi(s) = \sup_{\theta} \left(s^T\theta - \psi(\theta)\right)$$

- ϕ is a convex function of s

Maximum Likelihood Estimation, Conjugate

- Let $s = s(x)$ be the sufficient statistic for a set of points x
- Then maximizing log-likelihood is

$$\phi(s) = \max_{\theta} \left(s^T\theta - \psi(\theta)\right)$$

- Has a unique maximizer since $\psi(\theta)$ is convex
- The conjugate of ψ is

$$\phi(s) = \sup_{\theta} \left(s^T\theta - \psi(\theta)\right)$$

- ϕ is a convex function of s
- Technically, ψ, ϕ are "Legendre" functions

Mixtures of Exponential Family Distributions

- A finite mixture model

$$p(x|\alpha, \Theta) = \sum_{h=1}^{k} \alpha_h p_\psi(x|\theta_h)$$

Mixtures of Exponential Family Distributions

- A finite mixture model

$$p(x|\alpha, \Theta) = \sum_{h=1}^{k} \alpha_h p_\psi(x|\theta_h)$$

- ψ determines the family

Mixtures of Exponential Family Distributions

- A finite mixture model

$$p(x|\alpha, \Theta) = \sum_{h=1}^{k} \alpha_h p_\psi(x|\theta_h)$$

- ψ determines the family
- All mixture components are of the same family

Mixtures of Exponential Family Distributions

- A finite mixture model

$$p(x|\alpha,\Theta) = \sum_{h=1}^{k} \alpha_h p_\psi(x|\theta_h)$$

- ψ determines the family
- All mixture components are of the same family
- θ determines the distribution in the family

Mixtures of Exponential Family Distributions

- A finite mixture model

$$p(x|\alpha, \Theta) = \sum_{h=1}^{k} \alpha_h p_\psi(x|\theta_h)$$

- ψ determines the family
- All mixture components are of the same family
- θ determines the distribution in the family
- Each component has different parameters

Mixtures of Exponential Family Distributions (Contd.)

- E-step: Exactly same as before

$$\alpha_h = \frac{1}{N}\sum_{i=1}^{N} p(h|x_i, \theta^{(t-1)})$$

Mixtures of Exponential Family Distributions (Contd.)

- E-step: Exactly same as before

$$\alpha_h = \frac{1}{N} \sum_{i=1}^{N} p(h|x_i, \theta^{(t-1)})$$

- M-step: Taking gradient w.r.t. θ_h

$$\nabla \psi(\theta_h) = \frac{\sum_i x_i p(h|x_i, \theta^{(t-1)})}{\sum_i p(h|x_i, \theta^{(t-1)})}$$

Mixtures of Exponential Family Distributions (Contd.)

- E-step: Exactly same as before

$$\alpha_h = \frac{1}{N} \sum_{i=1}^{N} p(h|x_i, \theta^{(t-1)})$$

- M-step: Taking gradient w.r.t. θ_h

$$\nabla\psi(\theta_h) = \frac{\sum_i x_i p(h|x_i, \theta^{(t-1)})}{\sum_i p(h|x_i, \theta^{(t-1)})}$$

- $\nabla\psi$ is monotonic increasing, inverse is well defined

Mixtures of Exponential Family Distributions (Contd.)

- E-step: Exactly same as before

$$\alpha_h = \frac{1}{N} \sum_{i=1}^{N} p(h|x_i, \theta^{(t-1)})$$

- M-step: Taking gradient w.r.t. θ_h

$$\nabla \psi(\theta_h) = \frac{\sum_i x_i p(h|x_i, \theta^{(t-1)})}{\sum_i p(h|x_i, \theta^{(t-1)})}$$

- $\nabla \psi$ is monotonic increasing, inverse is well defined
- Recall the expression for μ_h for Gaussian mixtures

Mixture Models as a Bayes Net