Instructor: Sael Lee

CS549 Spring – Computational Biology

LECTURE 9: MIXTURE MODELS

Reference:
1. “Pattern Recognition and Machine Learning” Chapter 10: Approximate Inference
2. Performance of Bayesian Model Selection Criteria for Gaussian Mixture Models–R.J. Steele and A.E Raftery.

MODEL SELECTION OUTLINE

Selecting number of components

1. Selecting best model among ones that use different number of components

   For k1 … kG

   Computer score of the model that use kg components

   End for

   Use the best one or use weighted combination.

2. Learn in the model

   Indirectly model the number of components

CRITERIA FOR CHOOSING THE NUMBER OF MIXTURE COMPONENTS

- Consider univariate Gaussian mixture model with G components

$$p(y_i|\mu, \sigma^2, \lambda) = \sum_{g=1}^{G} \lambda_g f(y_i|\mu_g, \sigma_g^2),$$

- Maximum a postriori (MAP) estimate of G
- Bayesian information criteria (BIC) estimate of G
  - -> High data count, less complexity
- Deviance information criterion (DIC) estimate of G
- ICL estimate of G
- Akaike information criterion (AIC) estimate of G
  - -> Low data count, strives for less complexity

MAXIMUM A POSTERIORI (MAP) ESTIMATE OF G

a **point estimate** of an unobserved quantity on the basis of empirical data

-> Estimate an unobserved population parameter θ on the basis of observations x.

Maximum Likelihood

$$\theta \mapsto f(x|\theta)$$

$$\hat{\theta}_{\mathrm{ML}}(x) = \arg\max_{\theta} f(x|\theta)$$

Maximum a Posteriori

Prior

$$\theta \mapsto f(\theta|x) = \frac{f(x|\theta)\, g(\theta)}{\int_{\vartheta\in\Theta} f(x|\vartheta)\, g(\vartheta)\, d\vartheta}$$

$$\hat{\theta}_{\mathrm{MAP}}(x) = \arg\max_{\theta} \frac{f(x|\theta)\, g(\theta)}{\int_{\vartheta} f(x|\vartheta)\, g(\vartheta)\, d\vartheta}$$

$$= \arg\max_{\theta} f(x|\theta)\, g(\theta).$$

COMPUTING MAP ESTIMATES

1. When conjugate priors are used, MAP can be computed **analytically**
   - mode(s) of the posterior distribution can be given in closed form.
2. Via **numerical optimization methods**
   - Ex> conjugate gradient method; Newton's method.
   - Usually requires first or second derivatives, which have to be evaluated analytically or numerically.
3. Via a modification of an **expectation-maximization** algorithm.
   - This does not require derivatives of the posterior density.
4. Via a Monte Carlo method using **simulated annealing**

MODEL SELECTION: AKAIKE INFORMATION CRITERION (AIC)

- A measure of the relative goodness of fit of a statistical model
  - Offering a **relative measure of the information lost** when a given model is used to describe real distribution
  - Tradeoff between bias (accuracy) and variance (complexity) in model construction,
- Best known (but not the best performing) of the information criteria used for determining the number of components

$$AIC(G) = -2 \ln p(y|\hat{\tau}, G) + 2d$$

d: number of free parameters in the mixture

The **penalty term** is larger in BIC than in AIC.

p: maximized value of the **likelihood function** for the estimated model

Assumes a Gaussian prior for each of the μ_j with prior mean ξ_j and prior variance $\sigma_j^2 \tau_j$.

MODEL SELECTION: AIC CONT.

- Justification:
  - Choosing the minimum value of the AIC asymptotically minimizes the mean K-L divergence for discrimination between the proposed distribution (q) and the true distribution (p),
- Reality:
  - Overestimates the number of components for mixtures,
    - Most likely due to violations of the "regularity conditions" required for the approximation to hold
  - The estimate is only valid asymptotically;
    - If the number of data points is small, then some correction is often necessary
- How to apply AIC in practice
  - Start with a set of candidate models
  - Find corresponding AIC values
  - Find relative probability that the *i*th model minimizes the (estimated) information loss by computing (AIC_{min} is the min AIC value calculated)

$$\exp((AIC_{min} - AIC_i)/2)$$

MODEL SELECTION: BAYESIAN INFORMATION CRITERION (BIC)

- BIC is a criterion for model selection among a finite set of models

d: number of free parameters in the mixture

$$BIC(G) = -2p(y|\hat{\tau}, G) + d\ln(n)$$

The **penalty term** is larger in BIC than in AIC.

- BIC is consistent for choosing the number of components in a mixture model

Schwarz, G. (1978). Estimating the dimension of a model. The Annals of Statistics 6, 461–464

MODEL SELECTION: DEVIANCE INFORMATION CRITERION

- **Number of effective model parameters** is used for likelihood penalization criterion
  - Unlike the actual number of free parameters in the model as in AIC

$$DIC(G) = -2\ln p(y|\hat{\tau}, G) + 2p_d$$

$$p_d = \mathrm{E}_{(\tau|\mathrm{y})}\big(\ln p(y|\hat{\tau})\big) - \ln p(y|\hat{\tau})$$

$$\hat{p}_d = \frac{1}{\mathrm{T}}\sum_{t=1}^{T}\big(\ln p(y|\tau_t)\big) - \ln p(y|\hat{\tau})$$

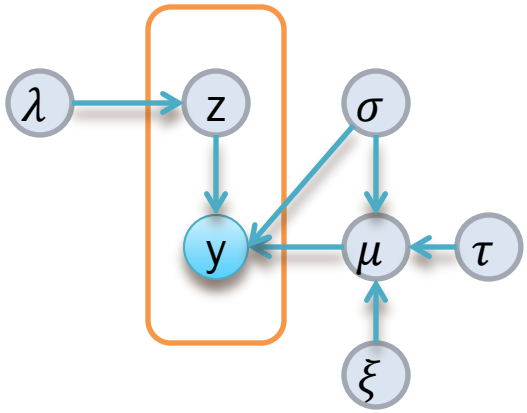
To estimate τ , find the largest posterior mode: $\hat{\tau}$

INFORMATION CRITERION BASED OF COMPLETE DATA LIKELIHOOD: ICL

- Information criterion based on the complete data likelihood

$$p(y,z|G) = \int p(y,z|\tau,G)p(\tau)d\tau$$

$$p(y|z,G)p(z|G) = \int p(y|z,\tau,G)p(z|G,\tau)p(\tau)d\tau$$



Approximate

$$\ln p(y|z,G) \approx \ln(p(y|z,\hat{\tau},\mathrm{G}))) - \mathrm{d}/2\ln(2)$$

$$\mathrm{ICL} = -2 * \log(p(y|\hat{z}',\hat{\tau}^*,G))) + \frac{(d-(G-1))}{G} * \log(n) - 2 * K(\hat{z}')$$

$$K(z) = \int p(z|\lambda,G)p(\lambda|G)$$

ILLUSTRATION: VARIATIONAL MIXTURE OF GAUSSIANS

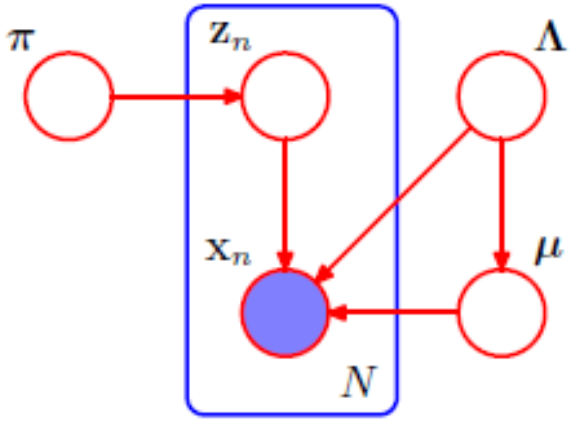
conditional distribution of **Z**, given the mixing coefficients $\boldsymbol{\pi}$

$$p(\mathbf{Z}|\boldsymbol{\pi}) = \prod_{n=1}^{N}\prod_{k=1}^{K}\pi_k^{z_{nk}}.$$

conditional distribution of the observed data vectors, given the latent variables and the component parameters

$$p(\mathbf{X}|\mathbf{Z},\boldsymbol{\mu},\boldsymbol{\Lambda}) = \prod_{n=1}^{N}\prod_{k=1}^{K}\mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k^{-1}\right)^{z_{nk}}$$

where $\boldsymbol{\mu} = \{\boldsymbol{\mu}_k\}$ and $\boldsymbol{\Lambda} = \{\boldsymbol{\Lambda}_k\}$.



Directed acyclic graph representing the Bayesian mixture of Gaussians model, in which the box (plate) denotes a set of *N* i.i.d. observations. where $\boldsymbol{\mu} = \{\boldsymbol{\mu}_k\}$ and $\boldsymbol{\Lambda} = \{\boldsymbol{\Lambda}_k\}$.

INTRODUCE CONJUGATE PRIOR DISTRIBUTION

Dirichlet distribution over the mixing coefficients π

normalization constant for the Dirichlet distribution

$$p(\boldsymbol{\pi}) = \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}_0) = C(\boldsymbol{\alpha}_0)\prod_{k=1}^{K}\pi_k^{\alpha_0-1}$$

Dirichlet distribution

$$f(x_1,\ldots,x_{K-1};\alpha_1,\ldots,\alpha_K) = \frac{1}{\mathrm{B}(\alpha)}\prod_{i=1}^{K}x_i^{\alpha_i-1}$$

$$\mathrm{B}(\alpha) = \frac{\prod_{i=1}^{K}\Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{K}\alpha_i)}, \qquad \alpha = (\alpha_1,\ldots,\alpha_K).$$

α Effective prior number of observations associated with each component of the mixture.

- **α0 is small**, then the posterior distribution will be influenced primarily by the data rather than by the prior

independent **Gaussian-Wishart prior** governing the mean μ and precision Λ of each Gaussian component

$$\begin{aligned} p(\boldsymbol{\mu},\boldsymbol{\Lambda}) &= p(\boldsymbol{\mu}|\boldsymbol{\Lambda})p(\boldsymbol{\Lambda}) \\ &= \prod_{k=1}^{K}\mathcal{N}\left(\boldsymbol{\mu}_k|\mathbf{m}_0,(\beta_0\boldsymbol{\Lambda}_k)^{-1}\right)\mathcal{W}(\boldsymbol{\Lambda}_k|\mathbf{W}_0,\nu_0) \end{aligned}$$

VARIATIONAL DISTRIBUTION

- Joint distribution of all of the random variables

$$p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda})p(\mathbf{Z}|\boldsymbol{\pi})p(\boldsymbol{\pi})p(\boldsymbol{\mu}|\boldsymbol{\Lambda})p(\boldsymbol{\Lambda})$$

- Variational distribution which **assumed to factorizes** between the latent variables and the parameters

$$q(\mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = q(\mathbf{Z})q(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$$

- Derivation of the update equation for the factor $q(\mathbf{Z})$

$$\begin{aligned}\ln q^\star(\mathbf{Z}) &= \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda}}[\ln p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})] + \text{const.} \\ &= \mathbb{E}_{\boldsymbol{\pi}}[\ln p(\mathbf{Z}|\boldsymbol{\pi})] + \mathbb{E}_{\boldsymbol{\mu},\boldsymbol{\Lambda}}[\ln p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda})] + \text{const.} \\ &= \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk} \ln \rho_{nk} + \text{const}\end{aligned}$$

$$\begin{aligned}\ln \rho_{nk} &= \mathbb{E}[\ln \pi_k] + \frac{1}{2}\mathbb{E}[\ln |\boldsymbol{\Lambda}_k|] - \frac{D}{2}\ln(2\pi) \\ &\quad -\frac{1}{2}\mathbb{E}_{\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k}\left[(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathrm{T}}\boldsymbol{\Lambda}_k(\mathbf{x}_n - \boldsymbol{\mu}_k)\right]\end{aligned}$$

same functional form as the prior $p(\mathbf{Z}|\boldsymbol{\pi})$.

$$q^\star(\mathbf{Z}) = \prod_{n=1}^{N}\prod_{k=1}^{K} r_{nk}^{z_{nk}}$$

$$r_{nk} = \frac{\rho_{nk}}{\sum_{j=1}^{K}\rho_{nj}}.$$

responsibilities

define three statistics of the observed data set evaluated with respect to the **responsibilities**

$$
\begin{aligned}
N_k &= \sum_{n=1}^{N} r_{nk} \\
\overline{\mathbf{x}}_k &= \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} \mathbf{x}_n \\
\mathbf{S}_k &= \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} (\mathbf{x}_n - \overline{\mathbf{x}}_k)(\mathbf{x}_n - \overline{\mathbf{x}}_k)^{\mathrm{T}}.
\end{aligned}
$$

$$q(\mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = q(\mathbf{Z})q(\pi, \mu, \Lambda)$$

$$\ln q^\star(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = \ln p(\boldsymbol{\pi}) + \sum_{k=1}^{K} \ln p(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) + \mathbb{E}_{\mathbf{Z}} \left[\ln p(\mathbf{Z}|\boldsymbol{\pi})\right]$$
$$+ \sum_{k=1}^{K}\sum_{n=1}^{N} \mathbb{E}[z_{nk}] \ln \mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}\right) + \text{const.}$$

$$q(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = q(\boldsymbol{\pi}) \prod_{k=1}^{K} q(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k).$$

$$q^\star(\boldsymbol{\pi}) = \text{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha})$$

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

UPDATE RULE:

Update equations are analogous to the M-step equations of the EM

$$q^\star(\boldsymbol{\pi}) = \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha})$$

$$\alpha_k = \alpha_0 + N_k.$$

$$\begin{aligned} N_k &= \sum_{n=1}^{N} r_{nk} \\ \overline{\mathbf{x}}_k &= \frac{1}{N_k}\sum_{n=1}^{N} r_{nk}\mathbf{x}_n \\ \mathbf{S}_k &= \frac{1}{N_k}\sum_{n=1}^{N} r_{nk}(\mathbf{x}_n - \overline{\mathbf{x}}_k)(\mathbf{x}_n - \overline{\mathbf{x}}_k)^{\mathrm{T}}. \end{aligned}$$

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

$$\begin{aligned} \beta_k &= \beta_0 + N_k \\ \mathbf{m}_k &= \frac{1}{\beta_k}(\beta_0\mathbf{m}_0 + N_k\overline{\mathbf{x}}_k) \\ \mathbf{W}_k^{-1} &= \mathbf{W}_0^{-1} + N_k\mathbf{S}_k + \frac{\beta_0 N_k}{\beta_0 + N_k}(\overline{\mathbf{x}}_k - \mathbf{m}_0)(\overline{\mathbf{x}}_k - \mathbf{m}_0)^{\mathrm{T}} \\ \nu_k &= \nu_0 + N_k. \end{aligned}$$

In order to perform this variational M step,
expectations $E[z_{nk}] = r_{nk}$ representing the **responsibilities** is needed.

obtained by normalizing the ρ_{nk}

$$\ln \rho_{nk} = \mathbb{E}[\ln \pi_k] + \frac{1}{2}\mathbb{E}[\ln |\Lambda_k|] - \frac{D}{2}\ln(2\pi) - \frac{1}{2}\mathbb{E}_{\mu_k,\Lambda_k}\left[(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathrm{T}}\boldsymbol{\Lambda}_k(\mathbf{x}_n - \boldsymbol{\mu}_k)\right]$$

$$\mathbb{E}_{\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k}\left[(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathrm{T}}\boldsymbol{\Lambda}_k(\mathbf{x}_n - \boldsymbol{\mu}_k)\right] = D\beta_k^{-1} + \nu_k(\mathbf{x}_n - \mathbf{m}_k)^{\mathrm{T}}\mathbf{W}_k(\mathbf{x}_n - \mathbf{m}_k)$$

$$\ln \widetilde{\Lambda}_k \equiv \mathbb{E}[\ln |\boldsymbol{\Lambda}_k|] = \sum_{i=1}^{D}\psi\left(\frac{\nu_k + 1 - i}{2}\right) + D\ln 2 + \ln |\mathbf{W}_k|$$

$$\ln \widetilde{\pi}_k \equiv \mathbb{E}[\ln \pi_k] = \psi(\alpha_k) - \psi(\widehat{\alpha})$$

$$r_{nk} \propto \widetilde{\pi}_k \widetilde{\Lambda}_k^{1/2} \exp\left\{-\frac{D}{2\beta_k} - \frac{\nu_k}{2}(\mathbf{x}_n - \mathbf{m}_k)^{\mathrm{T}}\mathbf{W}_k(\mathbf{x}_n - \mathbf{m}_k)\right\}.$$

VARIATIONAL EQUIVALENT EM STEPS

E-like step: use the current distributions over the model parameters to **evaluate the moments** in

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k}\left[(\mathbf{x}_n-\boldsymbol{\mu}_k)^{\mathrm{T}}\boldsymbol{\Lambda}_k(\mathbf{x}_n-\boldsymbol{\mu}_k)\right] &= D\beta_k^{-1}+\nu_k(\mathbf{x}_n-\mathbf{m}_k)^{\mathrm{T}}\mathbf{W}_k(\mathbf{x}_n-\mathbf{m}_k)\\
\ln\widetilde{\Lambda}_k \equiv \mathbb{E}\left[\ln|\boldsymbol{\Lambda}_k|\right] &= \sum_{i=1}^{D}\psi\left(\frac{\nu_k+1-i}{2}\right)+D\ln 2+\ln|\mathbf{W}_k|\\
\ln\widetilde{\pi}_k \equiv \mathbb{E}\left[\ln\pi_k\right] &= \psi(\alpha_k)-\psi(\widehat{\alpha})
\end{aligned}$$

Thus evaluate $E[z_{nk}] = r_{nk}$

$$\begin{aligned}
N_k &= \sum_{n=1}^{N} r_{nk}\\
\bar{\mathbf{x}}_k &= \frac{1}{N_k}\sum_{n=1}^{N} r_{nk}\mathbf{x}_n\\
\mathbf{S}_k &= \frac{1}{N_k}\sum_{n=1}^{N} r_{nk}(\mathbf{x}_n-\bar{\mathbf{x}}_k)(\mathbf{x}_n-\bar{\mathbf{x}}_k)^{\mathrm{T}}.
\end{aligned}$$

M-like step: keep these **responsibilities fixed** and use them to **re-compute the variational distribution** over the parameters using

$$q(\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda}) = q(\boldsymbol{\pi})\prod_{k=1}^{K} q(\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k).$$

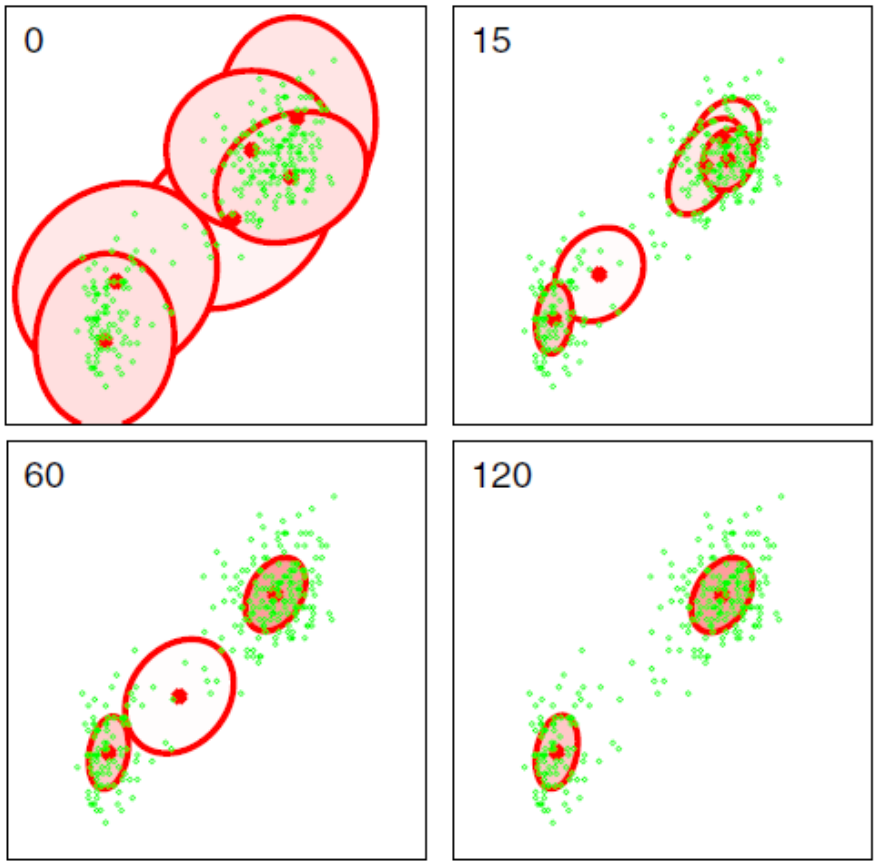
$$q^\star(\boldsymbol{\pi}) = \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha})$$

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

$$\begin{aligned}
\alpha_k &= \alpha_0 + N_k.\\
\beta_k &= \beta_0 + N_k\\
\mathbf{m}_k &= \frac{1}{\beta_k}(\beta_0\mathbf{m}_0+N_k\bar{\mathbf{x}}_k)\\
\mathbf{W}_k^{-1} &= \mathbf{W}_0^{-1}+N_k\mathbf{S}_k+\frac{\beta_0 N_k}{\beta_0+N_k}(\bar{\mathbf{x}}_k-\mathbf{m}_0)(\bar{\mathbf{x}}_k-\mathbf{m}_0)^{\mathrm{T}}\\
\nu_k &= \nu_0+N_k.
\end{aligned}$$

In each case, we see that the variational posterior distribution has the same functional form as the corresponding factor in the joint distribution

$$p(\mathbf{X},\mathbf{Z},\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda}) = p(\mathbf{X}|\mathbf{Z},\boldsymbol{\mu},\boldsymbol{\Lambda})p(\mathbf{Z}|\boldsymbol{\pi})p(\boldsymbol{\pi})p(\boldsymbol{\mu}|\boldsymbol{\Lambda})p(\boldsymbol{\Lambda})$$



Variational Bayesian mixture of $K = 6$ Gaussians applied to the Old Faithful data set.

Components whose expected mixing coefficient are numerically indistinguishable from zero are not plotted.

Components that take essentially no responsibility for explaining the data points have $r_{nk} \cong 0$ and hence $N_k \cong 0$.

So $\alpha_k \cong \alpha_0$ and other parameters revert to their prior values.

Variational Gaussian mixture model :
Expected values of the mixing coefficients in the posterior distribution are

$$\mathbb{E}[\pi_k] = \frac{\alpha_k + N_k}{K\alpha_0 + N}$$

$r_{nk} \cong 0$ and $N_k \cong 0$.

If $\alpha_0 \rightarrow 0$, then $E[\pi_k] \rightarrow 0$ and the component plays no role in the model,

A BAYESIAN METHOD OF MODEL SELECTION CRITERIA FOR GAUSSIAN MIXTURE MODELS

- Finding the number of components
  - Modeling the parameters

1. Use the **lower bound** as the model selection score.

$$P(D,\mu,T,s|\pi) = P(D|\mu,T,s)P(s|\pi)P(\mu)P(T).$$

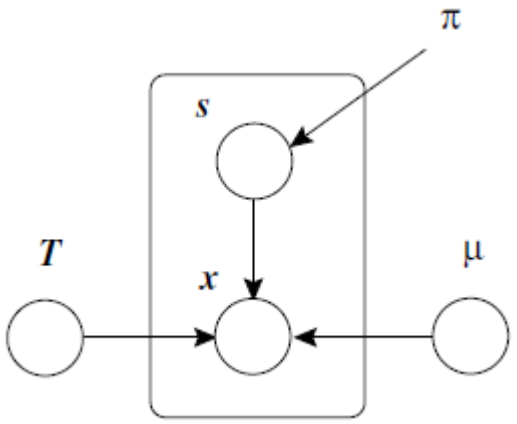
$$P(D|\mu,T,s) = \prod_{n=1}^{N}\prod_{i=1}^{M}\mathcal{N}(x_n|\mu_i,T_i)^{s_{in}}.$$

$$P(s|\pi) = \prod_{i=1}^{M}\prod_{n=1}^{N}\pi_i^{s_{in}}.$$

$$P(\mu) = \prod_{i=1}^{M}\mathcal{N}(\mu_i|0,\beta I)$$

$$P(T) = \prod_{i=1}^{M}\mathcal{W}(T_i|\nu,V)$$

W : Wishart distribution



$$\pi_i = \frac{1}{N}\sum_{n=1}^{N} p_{in}.$$

$$p_{in} = \frac{\widetilde{p}_{in}}{\sum_{j=1}^{M}\widetilde{p}_{jn}}$$

$$\begin{aligned}\widetilde{p}_{in} = \exp(&\langle \ln|T_i|\rangle/2 + \ln\pi_i \\ &-\frac{1}{2}\mathrm{Tr}\{\langle T_i\rangle(x_n x_n^{\mathrm{T}} - \langle\mu_i\rangle x_n^{\mathrm{T}} \\ &-x_n\langle\mu_i\rangle^{\mathrm{T}} + \langle\mu_i\mu_i^{\mathrm{T}}\rangle)\})\end{aligned}$$

2. remove components with very small **mixing coefficients**

A. Corduneanu and C. Bishop. Variational Bayesian Model Selection for Mixture Distributions 2001