STA 630: Bayesian Inference - Chapter 7

Zeya Wang

University of Kentucky

Spring 2025

Mixture Models

- **Clustering**: a very broad set of techniques for finding **subgroups**, or **clusters**, in a data set; partition data into distinct groups so that the observations within each group are quite similar to each other.
- **Mixture models:** we look for soft cluster assignment in a probabilistic way so that the level of uncertainty over the most appropriate assignment can be quantified.
- Characterize behavior of data that arise from a mixture of subpopulations.
- So far, we focussed our attention on simple distributions: Gaussian or exponential family in general
- Multimodel distributions; Semiparametric perspective where mixtures are basis approximations of unknown distributions

Finite Mixture Models

We can define a flexible continuous density as a mixture of simple parametric densities,

$$f(\theta) = \sum_{k=1}^{K} \pi_k f_k(\theta)$$

or more simply using the same functional form for each component but with different parameters

$$f(\theta) = \sum_{k=1}^{K} \pi_k f\left(\cdot \mid \theta_k\right)$$

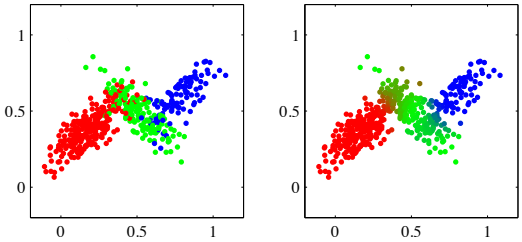
- π_k are called mixture weights, are in $(0, 1)$ and sum up to 1 .
- Most common: Gaussian mixtures
- Infinite (continuous) mixture models, replace sum with integral

Applications

- Identifying subpopulations
- Density estimation
- Clustering
- Classification
- Prediction with missing data
- Normal mixtures widely used in classification, clustering, density estimation
- Poisson mixtures used in spatial statistics

Gaussian mixture models: soft assignment

500 observations drawn from the mixture of three 2-dimensional Gaussians.



- Left: true labels indicated by red, green and blue.
- Right: soft assignment with proportions of red, blue, and green colors.

Gaussian mixture models: definition

Let $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})$ and let $N(\cdot|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ be a multivariate Gaussian density with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. A Gaussian mixture model with K components is a weighted average of K Gaussian densities

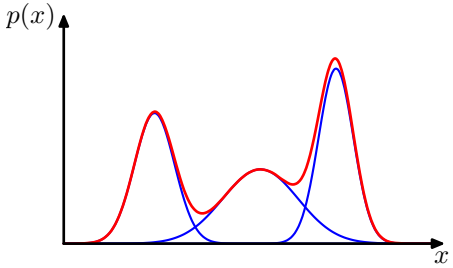
$$p(\boldsymbol{x}_i) = \sum_{k=1}^{K} \pi_k N(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

with

$$\sum_{k=1}^{K} \pi_k = 1$$

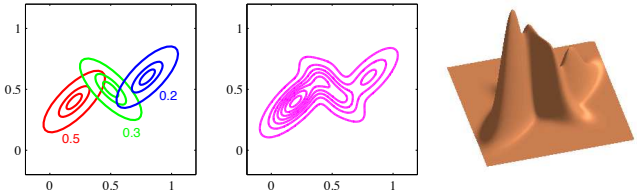
- $N(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is a **component** of the mixture.
- $0 \leq \pi_k \leq 1$ is the **mixture weight** or **mixing coefficients**.

One-dimensional example



Three Gaussians (each scaled by a mixture weight) in blue and their sum in red.

Two-dimensional example



- Left: contours of constant density for each of the 3 Gaussians (red, blue and green), and the mixture weights (numbers below each component).
- Center: contour of the weighted average of 3 Gaussians.
- Right: surface plot of the weighted average of 3 Gaussians.

Latent variable representation

- To make the connection between Gaussian mixtures and clustering concrete, we introduce a latent variable $s_i \in \{1, \ldots, K\}$ for each observation i and assume

$$p(s_i = k) = \pi_k$$
$$p(\boldsymbol{x}_i | s_i = k) = N(\boldsymbol{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

- The marginal distribution of (1) and (2)

$$p(\boldsymbol{x}_i) = \sum_{k=1}^{K} p(\boldsymbol{x}_i | s_i = k) p(s_i = k)$$

  is equivalent to the weighted average representation.

- **Interpretation**: $s_i = k$ indicates observation i belongs to cluster k.

Soft assignment through Bayes theorem

- Since we have a proper probability model, we can calculate conditional probability $p(s_i = k|\boldsymbol{x}_i)$ using Bayes theorem

$$p(s_i = k|\boldsymbol{x}_i) = \frac{\pi_k N(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{l=1}^{K} \pi_l N(\boldsymbol{x}_i|\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)}$$

- **Soft assignment**: $p(s_i = k|\boldsymbol{x}_i)$ is the posterior probability that observation i belongs to cluster k.

Non-Bayesian View: Estimation through EM algorithm

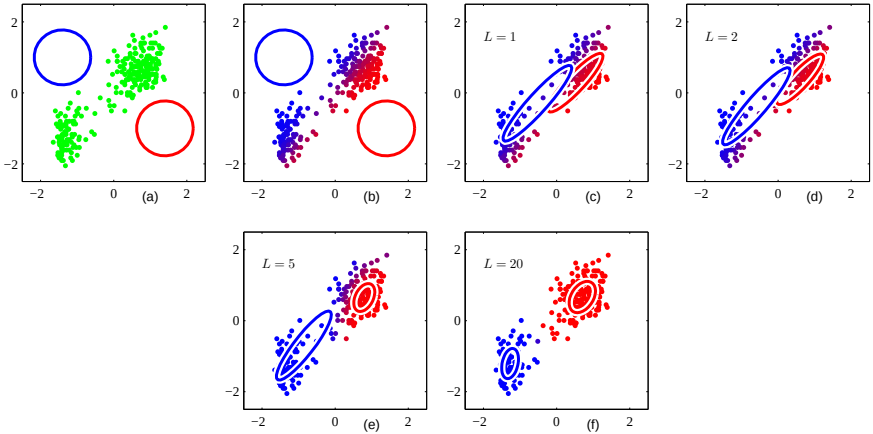
Expectation-maximization algorithm:

1. Initialize the means $\boldsymbol{\mu}_k$, covariances $\boldsymbol{\Sigma}_k$ and mixture weights π_k
2. E step. Evaluate the posterior probabilities using the current parameter values

$$p(s_i = k|\boldsymbol{x}_i) = \frac{\pi_k N(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{l=1}^{K} \pi_l N(\boldsymbol{x}_i|\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)}$$

3. M step. Re-estimate the parameters using the posterior probabilities
   $\boldsymbol{\mu}_k^{new} = \frac{1}{n_k} \sum_{i=1}^{n} p(s_i = k|\boldsymbol{x}_i)\boldsymbol{x}_i$;
   $\boldsymbol{\Sigma}_k^{new} = \frac{1}{n_k} \sum_{i=1}^{n} p(s_i = k|\boldsymbol{x}_i)(\boldsymbol{x}_i - \boldsymbol{\mu}_k^{new})(\boldsymbol{x}_i - \boldsymbol{\mu}_k^{new})^T$; $\pi_k^{new} = \frac{n_k}{n}$ where
   $n_k = \sum_{i=1}^{n} p(s_i = k|\boldsymbol{x}_i)$ is the effective number of observations assigned to cluster k.
4. Check convergence (e.g. through parameters or likelihood). If not convergent, return to step 2.

Illustration of EM algorithm



Comparison with K-means

- K-means outputs hard cluster assignment whereas Gaussian mixture model outputs soft cluster assignment.
- K-means assigns each data point uniquely to one and only one cluster. Some data points may lie roughly midway between cluster centroids. It is not clear that the **hard** assignment to the nearest cluster is the most appropriate.
- **Mixture models** adopt a probabilistic approach and obtain **soft** assignments of data points to clusters in a way that reflects the level of **uncertainty** of cluster assignment.
- If $\boldsymbol{\Sigma}_k = \epsilon \boldsymbol{I}_p$ (variance times an identity matrix) for all $k = 1, \ldots, K$, when ϵ is small, the soft assignment and hard assignment are similar

$$p(s_i = k|\boldsymbol{x}_i) \approx 1 \text{ for } k = \arg\min_k \sum_{j=1}^{p} (x_{ij} - \mu_{kj})^2$$

$$p(s_i = l|\boldsymbol{x}_i) \approx 0 \text{ for } l \neq k$$

Essentially, each observation is assigned to the cluster having the closest mean.

Example

Given the mixture model

$$x_i \sim \pi_1 N(\mu_1, \Sigma_1) + \pi_2 N(\mu_2, \Sigma_2) + \cdots$$

then any sample can come from distributions: $x_i \sim N(\mu_1, \Sigma_1)$ with probability π_1 $x_i \sim N(\mu_2, \Sigma_2)$ with probability π_2, and so on.
In Bayesian mixture modeling, priors are specified for the unknown as

$$\begin{aligned} x_i \mid \pi, \theta \text{ iid } & \sum_{k=1}^{K} \pi_k f(x_i \mid \theta_k) \\ \pi_k & \sim \text{Dir}(\alpha, K) \\ \theta_k & \sim \eta_k(\theta_K) \\ K & \sim \eta(K) \end{aligned}$$

where the mixture weights follow a Dirichlet distribution.

Example

An alternative specification (from a missing data perspective) uses latent variables s_{ik}

$$\begin{aligned} x_i \mid \pi, \theta \text{ iid } & \prod_{k=1}^{K} f\left(x_i \mid \theta_k\right)^{s_{ik}} \\ s_{ik} \sim & \text{ Multinomial}\left(\pi_1, \ldots, \pi_K\right) \\ \pi_k \sim & \text{ Dir}(\alpha, K) \\ \theta_k \sim & \eta_k\left(\theta_K\right) \\ K \sim & \eta(K) \end{aligned}$$

where $s_{ik} = 1$ is x_i is a member of the k-th group and 0 otherwise. This facilitates inference via Gibbs sampler.

Posterior inference for fixed K

Gibbs sampling proceeds at each iteration, given K, with full conditionals:

- $P\left(s_{ik}^{(t)} = 1 \mid \cdot\right) \approx \pi_k^{(t-1)} f\left(x_i \mid \theta_k^{(t-1)}\right)$
- $P\left(\pi_k^{(t)} \mid \cdot\right) = \text{Dir}\left(\alpha_1 + n_1^*, \ldots, \alpha_K + n_K^*\right)$ with $n_k^* = \sum_i s_{ik}^{(t)}$
- $p\left(\theta_k^{(t)} \mid \cdot\right) = \eta_k\left(\theta_k^{(t)}\right) \prod_i f_k\left(x_i \mid \theta_k^{(t)}\right)^{s_{ik}^{(t)}}$

Problems with Gibbs Sampling

The posterior is exchangeable under any permutation of the labels

$$f\left(\pi_1, \ldots, \pi_k, \theta_1, \ldots, \theta_K \mid X\right) = f\left(\pi_{p(1)}, \ldots, \pi_{p(k)}, \theta_{P(1)}, \ldots, \theta_{P(K)} \mid X\right)$$

- A possible solution is to place a constraint on the priors, for example $\theta_1 \leq \ldots \leq \theta_K$, although this may badly constrain the shape of posterior distributions.
- Other solutions: post-MCMC processing.

Number of Components

- What if number of components of mixture unknown
- Complex problem, computationally; model selection problem
- Can average over all possible combinations of K
- Treat K as parameter; model dimension changes

Reversible Jump

- Introduced by Green (1995, Biometrika)
- Allows one to move between parameter spaces with different dimensions; NOTE: most MCMC algorithms assume fixed dimension
- Key application in model comparison: samples both models and parameters nested within models
- Let $M_k, k = 1, \ldots, K$ be the candidate models and θ_k the parameter vector for M_k.
- RJMCMC sets up a Markov Chain on the space of models cross parameters that satisfies detailed balance
- Trick: Additional random variables are introduced to ensure dimension matching
- Needs transition matrix for moving between the discrete models $J(K^* \mid K)$
- Moving from model K to K^*, one introduces an auxiliary random variable, u, with a "jumping" distribution $J(u \mid K, K^*; \theta)$.
- Then $(\theta_{k^*}, u^*) = g_{K,K^*}(\theta_K, u)$ where $d_k + \dim(u) = d_{K^*} + \dim(u^*)$ and g is a deterministic function that relates the parameters of model K to those of K^*.

Reversible Jump

The general idea for reversible jump, given unknown K and θ_k, is:

- From a starting state (K, θ_k) propose a new model with probability J_{K,K^*} and generate an augmenting random u from the proposal $J(u \mid K, K^*, \theta_k)$.
- Determine the proposed model parameters $\theta_{K^*} = g_{K,K^*}(\theta_K, u)$
- Accept the new model with probability $\min(r, 1)$ where

$$r = \frac{p(y \mid \theta_{K^*})\, \pi(\theta_{K^*})\, \pi_{K^*} J_{K|K^*} J(u \mid K^*, K, \theta_{K^*})}{p(y \mid \theta_K)\, \pi(\theta_K)\, \pi_K J_{K^*|K} J(u \mid K, K^*, \theta_K)} \times \mid \text{ Jacobian } \mid$$

$$\text{with Jacobian } = \frac{\nabla g_{K,K^*}(\theta_K, u)}{\nabla(\theta_k, u)}$$

Summary

- Reversible jump can be thought of as a generalization of MH sampler
- MH sampler: $r = \{$ likelihood x prior x proposal ratios $\}$
- RJ sampler: $r = \{$ likelihood x prior x proposal ratios x Jacobian $\}$
- Usually implementation has three kind of moves
- BIRTH: Move to dimension $k + 1$
- DEATH: Move to dimension $k - 1$
- MOVE: Move within dimension k
- Not necessarily nested models

Variable Selection for Mixture Models

- Simultaneous variable selection and sample clustering
- Cluster structure of samples confined to a small subset of variables. Noisy variables mask the recovery of the clusters.
- Proposed methodology:
- Use multivariate normal mixture model with an unknown number of components to determine cluster structure of the samples.
- Use stochastic search techniques to examine the space of variable subsets and identify most probable models.
- Also, infinite mixture models via Dirichlet process priors.
- Genomic data: Identify disease subtypes and select the discriminating genes.

Revisit: Finite Mixture Models

- Discriminating variables define a mixture of K distributions

$$f\left(\mathbf{x}_i \mid \pi, \theta\right) = \sum_{k=1}^{K} \pi_k f\left(\mathbf{x}_i \mid \theta_k\right)$$

- We consider $f\left(\mathbf{x}_i \mid \theta_k\right)$ multivariate normal with $\theta_k = (\mu_k, \Sigma_k)$.
- Cluster assignments: $y = (y_1, \ldots, y_n)'$, where $y_i = k$ if the i^{th} observation comes from cluster k

$$p\left(y_i = k\right) = \pi_k.$$

Variable Selection

- Need to select discriminating variables.
- Introduce latent p-vector γ with binary entries

$$\begin{cases} \gamma_j = 1 & \text{if variable } j \text{ defines a mixture distribution} \\ \gamma_j = 0 & \text{otherwise} \end{cases}$$

- The likelihood function is given by

$$L(K, \gamma, \pi, \mu, \Sigma, \eta, \Omega \mid \mathbf{X}, y) = \prod_{k=1}^{K} (2\pi)^{\frac{-pn_k}{2}} |\Sigma_k|^{\frac{-n_k}{2}} \pi_k^{n_k}$$
$$\times \exp\left\{ -\frac{1}{2} \sum_{x_i \in C_k} \left(\mathbf{x}_{(\gamma)i} - \mu_{(\gamma)k}\right)^T \Sigma_{(\gamma)k}^{-1} \left(\mathbf{x}_{(\gamma)i} - \mu_{(\gamma)k}\right) \right\}$$
$$\times \phi\left(X_{(\gamma^c)} \mid \eta_{(\gamma^c)}, \Omega_{(\gamma^c)}\right)$$

where $C_k = \{x_i \mid y_i = k\}$ with cardinality n_k, $\phi(\cdot)$ is multivariate normal density.

Prior Model

- Assume γ_j 's are independent Bernoulli variables
- Number of components, K, can be assumed to follow a truncated Poisson or a discrete Uniform on [2, .., $K_{\max}$].
- $\pi \mid K \sim \text{Dirichlet}(\alpha, \ldots, \alpha)$.
- $$\begin{cases} \mu_{k(\gamma)} \mid \Sigma_{k(\gamma)}, K \sim \mathcal{N}\left(\mu_{0(\gamma)}, h\Sigma_{k(\gamma)}\right) \\ \Sigma_{k(\gamma)} \mid K \sim \mathcal{IW}\left(\delta; Q_\gamma\right) \end{cases}$$
  where (γ) indicates the covariates with $\gamma_j = 1$.

Model Fitting

1. Update γ by Metropolis algorithm (add/delete and swap moves).
2. Update π from its full conditional (Dirichlet draw).
3. Update y from its full conditional (multinomial draw).
4. Split one cluster into two, or merge two into one.
5. Birth or death of an empty component.

Steps (4) and (5) via reversible jump MCMC (Green, 1995).

Posterior Inference for y

- Number of clusters, K, estimated by value most frequently visited by MCMC sampler.
- Estimate marginal posterior probabilities $p(y_i = k \mid X, K)$. Posterior allocation of sample i estimated as

$$\widehat{y}_i = \max_{1 \leq k \leq K} \{p(y_i = k \mid \mathbf{X}, K)\}.$$

Posterior Inference for γ

- Select variables with largest marginal posterior probability

$$p\left(\gamma_j = 1 \mid \mathbf{X}, K\right)$$

- Select variables that are in the "best" models

$$\widehat{\gamma}* = \text{argmax}_t \left\{ p\left(\gamma^{(t)} \mid \mathbf{X}, K, \widehat{\pi}, \widehat{y}\right)\right\},$$

with $\widehat{y}$ the estimated sample allocations and $\widehat{\pi} = \frac{1}{M}\sum_{t=1}^{M} \pi^{(t)}$.

[1] Bayesian Variable Selection in Clustering High-Dimensional Data, Tadesse, Sha and Vannucci (JASA, 2005)

Application to Simulated Data

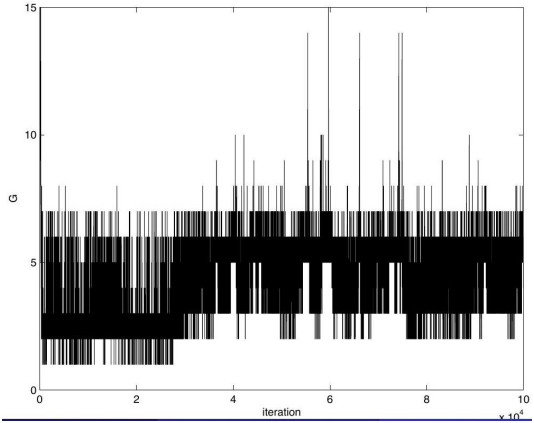
- 15 samples, 4 multivariate normal densities, 20 variables

$$x_{ij} \sim I_{\{1\leq i\leq 4\}}\mathcal{N}\left(\mu_1, \sigma_1^2\right) + I_{\{5\leq i\leq 7\}}\mathcal{N}\left(\mu_2, \sigma_2^2\right) +$$
$$I_{\{8\leq i\leq 13\}}\mathcal{N}\left(\mu_3, \sigma_3^2\right) + I_{\{14\leq i\leq 15\}}\mathcal{N}\left(\mu_4, \sigma_4^2\right),$$
$$i = 1, \ldots, 15, \quad j = 1, \ldots, 20, \mu_k \in [-5, 5], \sigma_k^2 \in [.1, 2]$$

- Cluster sizes: 4-3-6-2
- Additional set of 980 noisy variables drawn from a standard normal density
- Weakly informative priors for model parameters. ($\delta = 3, \alpha = 1, h = 100, Q = kI$)
- Truncated Poisson prior for K with $K_{\max} = 10$.
- MCMC with 100,000 iterations - starting model with 1 randomly selected γ_j set to 1 .

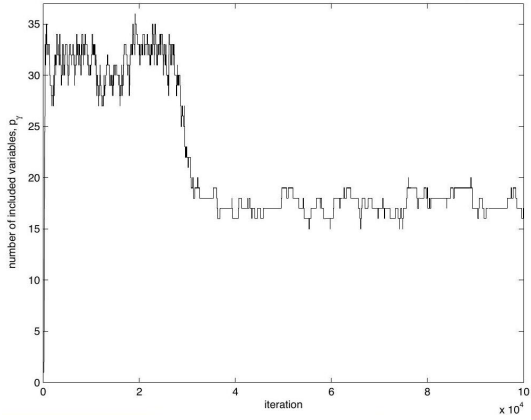
Application to Simulated Data

Trace plot of number of clusters, K



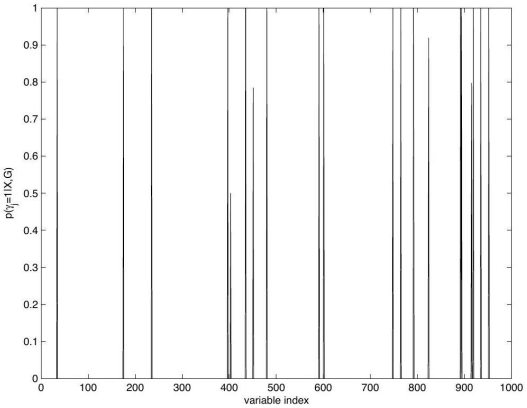
Application to Simulated Data

Trace plot for number of included variables, p_γ



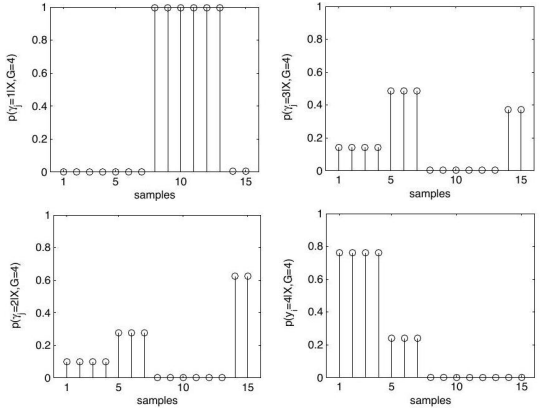
Application to Simulated Data

Marginal posterior probabilities, $p(\gamma_j = 1 \mid \mathbf{X}, \boldsymbol{K} = 4)$



Application to Simulated Data

Marginal posterior probabilities of sample allocations,
$p\left(y_i = k \mid \mathbf{X}, K = 4\right), i = 1, \ldots, 15, k = 1, \ldots, 4$



Results

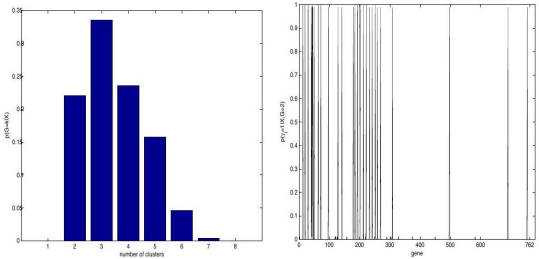
- $K = 4$ had stronger support
- All sample allocations corresponded to the true cluster structure
- There were 16 variables with marginal probability $> .7$ (15 were correct)
- Very little sensitivity to model parameters, with the exception of the covariance hyperparameters

Simultaneous Class Discovery and Gene Selection

- Endometrial cancer: Most common gynecologic malignancy in the US.
- 10 tumor and 4 normal tissues collected from hysterectomy specimens, examined with Affymetrix Hu6800 arrays.
- Probe sets with unreliable readings (< 20 and $> 16,000$) removed $\Rightarrow p = 762$.
- Gene expressions were log-transformed and scaled by their range.
- Specified weakly informative priors for model parameters.
- Used truncated Poisson prior for K with $K_{\max} = n$.
- $p(\gamma_j) \sim \text{Bernoulli}(\varphi = 10/p)$.
- Ran four MCMC chains with widely different starting points: (a) 1; (b) 10; (c) 25; (d) 50 randomly selected γ_j' s set to 1.

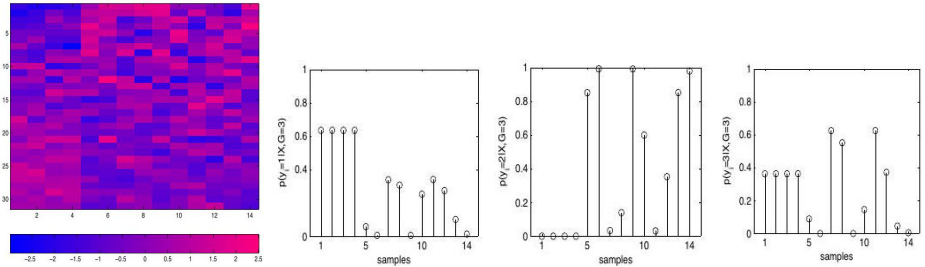
Simultaneous Class Discovery and Gene Selection

- Posterior distribution of K
- Union of 4 chains $-p\left(\gamma_j = 1 \mid \mathbf{X}, K = 3\right)$



Simultaneous Class Discovery and Gene Selection

- We have identified 3 classes and a set of 31 genes that can distinguish subtypes of the disease.



Variational Inference

- Variational inference
- Factorized distributions
- Properties of factorized distributions
- Example: Univariate Gaussian
- Example: Variational Gaussian mixture
- Variational lower bound

Variational inference

- Evaluate posterior distribution $p(\mathbf{Z} \mid \mathbf{X})$ of latent variables $\mathbf{Z}$ given observed data $\mathbf{X}$
- Expectations w.r.t. the posterior distribution:
  - e.g. EM algorithm: expectation of the complete-data log-likelihood
  - Evaluation of the posterior may be infeasible: (1) high dimensionality (2) complex posterior distributions lacking analytical form
- Approximation schemes are required when exact inference is infeasible
  1. **Stochastic approximations:**
     - MCMC: (1) computationally demanding, often limiting use to small-scale problems (2) ensuring independent samples from the target distribution can be challenging
  2. **Deterministic approximations:**
     Cannot produce exact results but offer complementary strengths
     - Scalability
     - Analytical approximations/assume parametric forms

Kullback-Leibler (KL) Divergence

Definition: measures how one probability distribution $q(x)$ diverges from a second, reference distribution $p(x)$.

Definition (Discrete)

$$\mathrm{KL}(q\|p) = \sum_x q(x) \log \frac{q(x)}{p(x)}$$

Definition (Continuous)

$$\mathrm{KL}(q\|p) = \int q(x) \log \frac{q(x)}{p(x)} \, dx$$

- **Non-negative:** $\mathrm{KL}(q\|p) \geq 0$
- **Zero iff:** $q(x) = p(x)$ almost everywhere
- **Not symmetric:** $\mathrm{KL}(q\|p) \neq \mathrm{KL}(p\|q)$

Variational inference

- All latent variables and parameters: $\mathbf{Z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_N\}$
- Goal: find an approximation $q(\mathbf{Z})$ for the posterior distribution $p(\mathbf{Z} \mid \mathbf{X})$
- Decompose the log marginal probability of $\mathbf{X}$ (log evidence) using

$$\ln p(\mathbf{X}) = \mathcal{L}(q) + \mathrm{KL}(q \| p)$$

where

$$\mathcal{L}(q) = \int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \right\} \mathrm{d}\mathbf{Z}$$

$$\mathrm{KL}(q \| p) = -\int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z} \mid \mathbf{X})}{q(\mathbf{Z})} \right\} \mathrm{d}\mathbf{Z}$$

- Focus on continuous variables; however, the analysis goes through unchanged for discrete variables are by replacing the integrations with summations

Variational inference

- $\mathcal{L}(q) \leq \ln p(\mathbf{X})$
- $\mathcal{L}(q)$: **Evidence Lower Bound (ELBO)**
- Maximize the $\mathcal{L}(q)$ w.r.t. $q(\mathbf{Z}) \equiv$ minimize $\mathrm{KL}(q\|p)$
- Ideal case: $q(\mathbf{Z}) = p(\mathbf{Z} \mid \mathbf{X})$
- Solution: consider a restricted family of distributions $q(\mathbf{Z})$ and then seek the member of this family for which $\mathrm{KL}(q\|p)$ is minimized
- Restrict the family sufficiently:
  - Comprise only tractable distributions
  - Sufficiently rich and flexible for a good approximation to the true posterior distribution
- Use a parametric distribution: $q(\mathbf{Z} \mid \boldsymbol{\omega})$
- $\mathcal{L}(q)$ becomes a function of $\boldsymbol{\omega}$
- Nonlinear optimization techniques, e.g., gradient descent

Factorized distributions

- An approximation framework in physics: *mean field theory* (Parisi, 1988)
- Partition the elements of $\mathbf{Z}$ into disjoint groups $\mathbf{Z}_i$ where $i = 1, \ldots, M$

$$q(\mathbf{Z}) = \prod_{i=1}^{M} q_i(\mathbf{Z}_i)$$

- Variational optimization of $\mathcal{L}(q)$ w.r.t all $q_i(\mathbf{Z}_i)$,
- Denote $q_j(\mathbf{Z}_j)$ by simply q_j

$$\mathcal{L}(q) = \int \prod_i q_i \left\{ \ln p(\mathbf{X}, \mathbf{Z}) - \sum_i \ln q_i \right\} \mathrm{d}\mathbf{Z}$$

Factorized distributions

- Dissect out the dependence on one of the factors $q_j(\mathbf{Z}_j)$

$$\mathcal{L}(q) = \int q_j \left\{ \int \ln p(\mathbf{X}, \mathbf{Z}) \prod_{i \neq j} q_i \, \mathrm{d}\mathbf{Z}_i \right\} \mathrm{d}\mathbf{Z}_j - \int q_j \ln q_j \, \mathrm{d}\mathbf{Z}_j + \text{ const}$$

$$= \int q_j \ln \widetilde{p}(\mathbf{X}, \mathbf{Z}_j) \, \mathrm{d}\mathbf{Z}_j - \int q_j \ln q_j \, \mathrm{d}\mathbf{Z}_j + \text{ const}$$

- Let $\mathbb{E}_{i \neq j}[\cdots]$ denote an expectation w.r.t. the q distributions over all variables $\mathbf{z}_i$ for $i \neq j$

$$\mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})] = \int \ln p(\mathbf{X}, \mathbf{Z}) \prod_{i \neq j} q_i \, \mathrm{d}\mathbf{Z}_i$$

- Define a new distribution $\widetilde{p}(\mathbf{X}, \mathbf{Z}_j)$ by the relation

$$\ln \widetilde{p}(\mathbf{X}, \mathbf{Z}_j) = \mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})] + \text{ const.}$$

Factorized distributions

- Keep $\{q_{i\neq j}\}$ fixed and maximize $\mathcal{L}(q)$ w.r.t. $q_j(\mathbf{Z}_j)$
- $\int q_j \ln \widetilde{p}(\mathbf{X}, \mathbf{Z}_j)\,\mathrm{d}\mathbf{Z}_j - \int q_j \ln q_j\,\mathrm{d}\mathbf{Z}_j$: negative KL divergence between $q_j(\mathbf{Z}_j)$ and $\widetilde{p}(\mathbf{X}, \mathbf{Z}_j)$
- $\max \mathcal{L}(q) \equiv \min \mathrm{KL}(q_j(\mathbf{Z}_j) | \widetilde{p}(\mathbf{X}, \mathbf{Z}_j))$
- The minimum occurs when $q_j(\mathbf{Z}_j) = \widetilde{p}(\mathbf{X}, \mathbf{Z}_j)$, yielding the optimal solution $q_j^\star(\mathbf{Z}_j)$

$$\ln q_j^\star(\mathbf{Z}_j) = \mathbb{E}_{i\neq j}[\ln p(\mathbf{X}, \mathbf{Z})] + \text{ const.}$$

- Take the exponential of both sides and normalize, we have

$$q_j^\star(\mathbf{Z}_j) = \frac{\exp\left(\mathbb{E}_{i\neq j}[\ln p(\mathbf{X}, \mathbf{Z})]\right)}{\int \exp\left(\mathbb{E}_{i\neq j}[\ln p(\mathbf{X}, \mathbf{Z})]\right)\mathrm{d}\mathbf{Z}_j}$$

Factorized distributions

- Use $\ln q_j^\star(\mathbf{Z}_j) = \mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})] + \text{ const.}$
- $\mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})]$ depends on expectations computed w.r.t. $q_i(\mathbf{Z}_i)$ for $i \neq j$
- Iterative approach
  1. Initialize all $q_i(\mathbf{Z}_i)$ appropriately
  2. Iterative update (coordinate descent): $q_j^{(t)}(\mathbf{Z}_j) \propto \exp\left(\mathbb{E}_{q_{i \neq j}^{(t-1)}}[\ln p(\mathbf{X}, \mathbf{Z})]\right)$
- Convergence is guaranteed because bound is convex w.r.t. each of the factors $q_i(\mathbf{Z}_i)$ (Boyd and Vandenberghe, 2004)

Example: Univariate Gaussian

- Infer the posterior distribution for the mean μ and precision τ, given $\mathcal{D} = \{x_1, \ldots, x_N\}$
- Likelihood:

$$p(\mathcal{D} \mid \mu, \tau) = \left(\frac{\tau}{2\pi}\right)^{N/2} \exp\left\{-\frac{\tau}{2}\sum_{n=1}^{N}(x_n - \mu)^2\right\}$$

- Conjugate prior distributions for μ and τ:

$$p(\mu \mid \tau) = \mathcal{N}\left(\mu \mid \mu_0, (\lambda_0\tau)^{-1}\right)$$
$$p(\tau) = \text{Gam}\left(\tau \mid a_0, b_0\right)$$

- Factorized variational approximation to the posterior distribution:

$$q(\mu, \tau) = q_\mu(\mu) q_\tau(\tau)$$

Example: Univariate Gaussian

- Note that the true posterior distribution does not factorize in this way.
- Perform variational approximation

$$\begin{aligned}\ln q_\mu^\star(\mu) &= \mathbb{E}_\tau[\ln p(\mathcal{D} \mid \mu, \tau) + \ln p(\mu \mid \tau)] + \text{ const} \\ &= -\frac{\mathbb{E}[\tau]}{2}\left\{\lambda_0 (\mu - \mu_0)^2 + \sum_{n=1}^{N} (x_n - \mu)^2\right\} + \text{ const.}\end{aligned}$$

- $q_\mu(\mu)$ is a Gaussian $\mathcal{N}(\mu|\mu_N, \lambda_N^{-1})$:

$$\mu_N = \frac{\lambda_0\mu_0 + Nx}{\lambda_0 + N}$$

$$\lambda_N = (\lambda_0 + N)\mathbb{E}[\tau].$$

- $N \to \infty$ gives MLE result in which $\mu_N = x$ and the precision is infinite.

Example: Univariate Gaussian

- The optimal solution for $q_\tau(\tau)$ is given by

$$\begin{aligned}\ln q_\tau(\tau) &= \mathbb{E}_\mu[\ln p(D|\mu,\tau) + \ln p(\mu|\tau)] + \ln p(\tau) + \text{const} \\ &= (a_0 - 1)\ln\tau - b_0\tau + \frac{N}{2}\ln\tau - \frac{\tau}{2}\mathbb{E}_\mu\left[\sum_{n=1}^{N}(x_n - \mu)^2 + \lambda_0(\mu - \mu_0)^2\right] + \text{const}\end{aligned}$$

- $q_\tau(\tau)$ is a gamma distribution $\text{Gam}(\tau|a_N, b_N)$

$$a_N = a_0 + \frac{N}{2}$$

$$b_N = b_0 + \frac{1}{2}\mathbb{E}_\mu\left[\sum_{n=1}^{N}(x_n - \mu)^2 + \lambda_0(\mu - \mu_0)^2\right].$$

Example: Univariate Gaussian

- Explicitly solve $q_\mu(\mu)$ and $q_\tau(\tau)$
- Let's assume noninformative priors in which $\mu_0 = a_0 = b_0 = \lambda_0 = 0$

$$\frac{1}{\mathbb{E}[\tau]} = \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N}(x_n - \mu)^2\right] = \overline{x^2} - 2\bar{x}\mathbb{E}[\mu] + \mathbb{E}\left[\mu^2\right]$$

- Solve for $\mathbb{E}[\tau]$ (leave this with you to do) to give

$$\begin{aligned}\frac{1}{\mathbb{E}[\tau]} &= \frac{1}{N-1}\left(\overline{x^2} - \bar{x}^2\right) \\ &= \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \bar{x})^2\end{aligned}$$

Example: Variational Mixture of Gaussians

- For each observation $\mathbf{x}_n$ we have a corresponding latent variable $\mathbf{z}_n$ comprising a 1 -of- K binary vector with elements z_{nk} for $k = 1, \ldots, K$
- Latent variables: $\mathbf{Z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_N\}$
- Full conditional distribution of $\mathbf{Z}$, given the mixing coefficients π:

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

- Conditional distribution of the observed data vectors

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

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

Example: Variational Mixture of Gaussians

- Priors over the parameters $\boldsymbol{\mu}, \boldsymbol{\Lambda}$ and $\boldsymbol{\pi}$
- Conjugate prior distributions
- Dirichlet distribution over the mixing coefficients π

$$p(\boldsymbol{\pi}) = \text{Dir}\left(\boldsymbol{\pi} \mid \boldsymbol{\alpha}_0\right) = C\left(\boldsymbol{\alpha}_0\right) \prod_{k=1}^{K} \pi_k^{\alpha_0 - 1}$$

where by symmetry we have the same parameter α_0, and $C\left(\boldsymbol{\alpha}_0\right)$ is the normalization constant

- Independent Gaussian-Wishart prior on $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$

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

Example: Variational Mixture of Gaussians

- Joint distribution of all of the random variables:

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

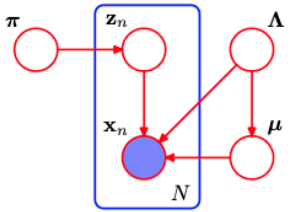


Figure: Directed acyclic graph representing the Bayesian mixture of Gaussians model, in which the box (plate) denotes a set of N i.i.d. observations.

Example: Variational Mixture of Gaussians

- Consider a variational distribution which factorizes between the latent variables and the parameters, we assume

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

- $q(\mathbf{Z})$ and $q(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$ will be determined by optimization of the variational distribution
- For $q(\mathbf{Z})$. The log of the optimized factor is given by

$$\ln q^\star(\mathbf{Z}) = \mathbb{E}_{\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}}[\ln p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})] + \text{ const.}$$

- Discard any terms that do not depend on $\mathbf{Z}$ (only interested in the functional dependence on the variable $\mathbf{Z}$)

$$\ln q^\star(\mathbf{Z}) = \mathbb{E}_{\boldsymbol{\pi}}[\ln p(\mathbf{Z} \mid \boldsymbol{\pi})] + \mathbb{E}_{\boldsymbol{\mu}, \boldsymbol{\Lambda}}[\ln p(\mathbf{X} \mid \mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda})] + \text{ const.}$$

Example: Variational Mixture of Gaussians

- we have

$$\ln q^{\star}(\mathbf{Z}) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \ln \rho_{nk} + \text{const}$$

where we have defined

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

where D is the dimensionality of the data variable $\mathbf{x}$.

Example: Variational Mixture of Gaussians

- Taking the exponential of both sides, we have:

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

- Note that n the quantities z_{nk} are binary and sum to 1 over all values of k:

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

where

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

- $q(\mathbf{Z})$ takes the same functional form as the prior $p(\mathbf{Z} \mid \boldsymbol{\pi})$ and $\mathbb{E}[z_{nk}] = r_{nk}$ (responsibilities)

Example: Variational Mixture of Gaussians

- Simplify the notations for the future use:

$$N_k = \sum_{n=1}^{N} r_{nk}$$

$$\bar{\mathbf{x}}_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} \mathbf{x}_n$$

$$\mathbf{S}_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} \left(\mathbf{x}_n - \bar{\mathbf{x}}_k\right) \left(\mathbf{x}_n - \bar{\mathbf{x}}_k\right)^{\mathrm{T}}$$

Example: Variational Mixture of Gaussians

- Let us consider the factor $q(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$ in the variational posterior distribution

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

- Involving terms with only $\boldsymbol{\pi}$, and terms with $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$ implies that the variational posterior $q(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$ factorizes into $q(\boldsymbol{\pi}) q(\boldsymbol{\mu}, \boldsymbol{\Lambda})$

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

Example: Variational Mixture of Gaussians

- Identifying the terms that depend on π, we have

$$\ln q^\star(\boldsymbol{\pi}) = (\alpha_0 - 1)\sum_{k=1}^{K} \ln \pi_k + \sum_{k=1}^{K}\sum_{n=1}^{N} r_{nk} \ln \pi_k + \text{ const}$$

- Taking the exponential of both sides, we recognize $q^\star(\boldsymbol{\pi})$ as a Dirichlet distribution

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

where $\boldsymbol{\alpha}$ has components α_k given by

$$\alpha_k = \alpha_0 + N_k$$

Example: Variational Mixture of Gaussians

- Write $q^\star(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) = q^\star(\boldsymbol{\mu}_k \mid \boldsymbol{\Lambda}_k)\, q^\star(\boldsymbol{\Lambda}_k)$
- Read off terms involving $\boldsymbol{\mu}_k$ and $\boldsymbol{\Lambda}_k$

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

where we have defined

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

Example: Variational Mixture of Gaussians

- Evaluate this expression involves expectations with respect to the variational distributions of the parameters

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

where we have introduced definitions of $\widetilde{\Lambda}_k$ and $\widetilde{\pi}_k$, and $\psi(\cdot)$ is the digamma function, with $\widehat{\alpha} = \sum_k \alpha_k$

Example: Variational Mixture of Gaussians

- Finally, substitute the items on the left handside for the expression on the right handside, we obtain the following result for the responsibilities

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

- Cycling between two stages
  1. Update $q^{\star}(\mathbf{Z})$ (E-step)
  2. Update $q^{\star}(\boldsymbol{\pi})$, $q^{\star}\left(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k\right)$ (M-step)

Variational lower bound

- We can straightforwardly evaluate **ELBO** $\mathcal{L}(q)$
- ELBO should not decrease: test for convergence; check on both the mathematical expressions for the solutions
- For the variational mixture of Gaussians, the lower bound is:

$$\begin{aligned}\mathcal{L} =& \sum_{\mathbf{Z}} \iiint q(\mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})}{q(\mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})} \right\} \mathrm{d}\boldsymbol{\pi}\ \mathrm{d}\boldsymbol{\mu}\ \mathrm{d}\boldsymbol{\Lambda} \\ =& \mathbb{E}[\ln p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})] - \mathbb{E}[\ln q(\mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})] \\ =& \mathbb{E}[\ln p(\mathbf{X} \mid \mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda})] + \mathbb{E}[\ln p(\mathbf{Z} \mid \boldsymbol{\pi})] + \mathbb{E}[\ln p(\boldsymbol{\pi})] + \mathbb{E}[\ln p(\boldsymbol{\mu}, \boldsymbol{\Lambda})] \\ & - \mathbb{E}[\ln q(\mathbf{Z})] - \mathbb{E}[\ln q(\boldsymbol{\pi})] - \mathbb{E}[\ln q(\boldsymbol{\mu}, \boldsymbol{\Lambda})]\end{aligned}$$

- Omitted the $\star$ superscript on the q distributions, along with the subscripts on the expectation operators (each expectation is taken w.r.t. all of the random variables in its argument)

Variational lower bound

- The various terms in the bound

$$\begin{aligned}\mathbb{E}[\ln p(\mathbf{X} \mid \mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda})] =& \frac{1}{2} \sum_{k=1}^{K} N_k \left\{ \ln \widetilde{\Lambda}_k - D\beta_k^{-1} - \nu_k \operatorname{Tr}(\mathbf{S}_k \mathbf{W}_k) \right. \\ & \left. - \nu_k (\overline{\mathbf{x}}_k - \mathbf{m}_k)^{\mathrm{T}} \mathbf{W}_k (\overline{\mathbf{x}}_k - \mathbf{m}_k) - D \ln(2\pi) \right\} \\ \mathbb{E}[\ln p(\mathbf{Z} \mid \boldsymbol{\pi})] =& \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \ln \widetilde{\pi}_k \\ \mathbb{E}[\ln p(\boldsymbol{\pi})] =& \ln C(\boldsymbol{\alpha}_0) + (\alpha_0 - 1) \sum_{k=1}^{K} \ln \widetilde{\pi}_k\end{aligned}$$

Variational lower bound

$$\mathbb{E}[\ln p(\boldsymbol{\mu}, \boldsymbol{\Lambda})] = \frac{1}{2}\sum_{k=1}^{K}\left\{D\ln\left(\beta_0/2\pi\right) + \ln\widetilde{\Lambda}_k - \frac{D\beta_0}{\beta_k}\right.$$

$$\left. -\beta_0\nu_k\left(\mathbf{m}_k - \mathbf{m}_0\right)^{\mathrm{T}}\mathbf{W}_k\left(\mathbf{m}_k - \mathbf{m}_0\right)\right\} + K\ln B\left(\mathbf{W}_0, \nu_0\right)$$

$$+ \frac{\left(\nu_0 - D - 1\right)}{2}\sum_{k=1}^{K}\ln\widetilde{\Lambda}_k - \frac{1}{2}\sum_{k=1}^{K}\nu_k\,\mathrm{Tr}\left(\mathbf{W}_0^{-1}\mathbf{W}_k\right)$$

$$\mathbb{E}[\ln q(\mathbf{Z})] = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\ln r_{nk}$$

$$\mathbb{E}[\ln q(\boldsymbol{\pi})] = \sum_{k=1}^{K}\left(\alpha_k - 1\right)\ln\widetilde{\pi}_k + \ln C(\boldsymbol{\alpha})$$

$$\mathbb{E}[\ln q(\boldsymbol{\mu}, \boldsymbol{\Lambda})] = \sum_{k=1}^{K}\left\{\frac{1}{2}\ln\widetilde{\Lambda}_k + \frac{D}{2}\ln\left(\frac{\beta_k}{2\pi}\right) - \frac{D}{2} - \mathrm{H}\left[q\left(\boldsymbol{\Lambda}_k\right)\right]\right\}$$

Variational lower bound

- D is the dimensionality of $\mathbf{x}$, $\mathrm{H}\left[q\left(\boldsymbol{\Lambda}_k\right)\right]$ is the entropy of the Wishart distribution, and the coefficients $C(\boldsymbol{\alpha})$ and $B(\mathbf{W}, \nu)$ are predefined.
- Some simplifications and combination of terms can be performed when these expressions are summed to give the lower bound.

Infinite Mixture Models via Dirichlet Process Priors

- Integrating over π and taking $K \to \infty$ we get

$$p\left(y_i = k \text{ and } y_l = k \text{ for some } l \neq i \mid \mathbf{y}_{-i}\right) = \frac{n_{-i,k}}{n-1+\alpha}$$
$$p\left(y_i \neq y_l \text{ for all } l \neq i \mid \mathbf{y}_{-i}\right) = \frac{\alpha}{n-1+\alpha}.$$

- MCMC updates γ via Metropolis and y_i from full conditionals

$$p\left(y_i = k \text{ and } y_l = k \text{ for some } l \neq i \mid \mathbf{y}_{-i}, \mathbf{X}, \gamma\right)$$
$$p\left(y_i \neq y_l \text{ for all } l \neq i \mid \mathbf{y}_{-i}, \mathbf{X}, \gamma\right).$$

- Inference on $\mathbf{y}$ by MAP or by estimating $p\left(y_i = y_j \mid \mathbf{X}\right)$. Same as before for γ

References

The content of these slides is based on the references listed below:

1. A first course in Bayesian statistical methods (2009) Springer.
2. Pattern recognition and machine learning (2006) Springer.
3. Rice University STAT 622: Bayesian Analysis.