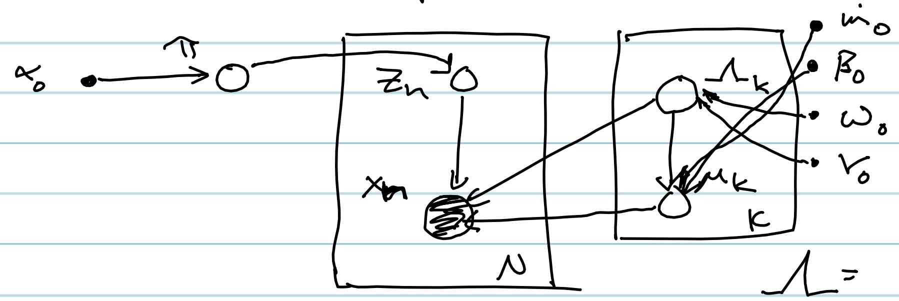
Variational Bayes Updates for a Bayesian Gaussian Mixture Model

B-GMM graphical model



$n = 1 .. N$

$x_n \in \mathbb{R}^D$

$z_n \in \mathbb{Z}$

$\Lambda = \{\Lambda_1, \Lambda_2, \ldots \Lambda_K\}$

K is # of classes

$\Lambda_k \in \mathbb{R}^{D \times D}$, Λ_k pos-semidef.

$\mu = \{\mu_1, \mu_2 \ldots \mu_K\}$

$\mu_k \in \mathbb{R}^D$

$\pi = [\pi_1 \ldots \pi_K]$

class prior probabilities

Latent vars:

π, Z, Λ, μ

Observed vars:

X

Parameters:

$\underbrace{\alpha_0}_{\text{hyperprior for class labels}}, \underbrace{m_0, \beta_0, W_0, \nu_0}_{\text{hyperprior for class means \& covariances}}$

Likelihood of $Z | \pi$, discrete

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

, z_n is vector with one non-zero element equal to 1

Conditional likelihood of $X | Z, \mu, \Lambda$

$$p(X|Z,\mu,\Lambda) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mathcal{N}(x_n | \mu_k, \Lambda_k^{-1})^{z_{nk}}$$

Priors on π, μ, & Λ

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

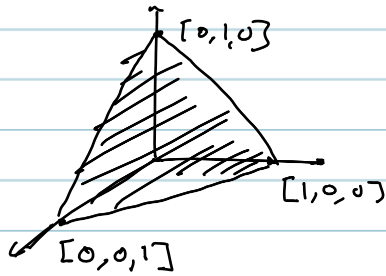
$$\vec{\alpha}_0 = [\underbrace{\alpha_0, \alpha_0, \ldots, \alpha_0}_{K}]$$

$$C(\vec{\alpha}_0) = \frac{\Gamma(K\alpha_0)}{\underbrace{\Gamma(\alpha_0)\Gamma(\alpha_0) \cdots \Gamma(\alpha_0)}_{K}}$$

pg 687 (B.23)

Dirichlet Dist.

Dirichlet distribution is a distribution over distributions



distribution is a vector of numbers btw 0 and 1 whose sum = 1

the set of all such vectors is called the simplex

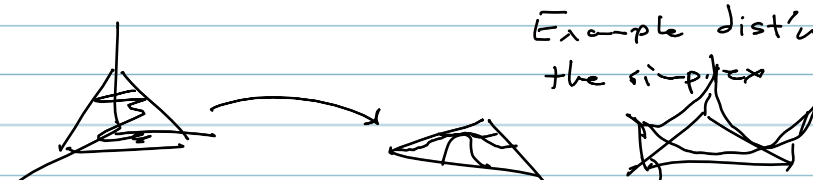
3-d prob. vectors, corners of simplex are

$$[1,0,0], \quad [0,1,0], \quad [0,0,1]$$

Simplex is a 2d plane in 3d.

Dir. dist.

Example dist'ns on the simplex



p([1 0 0]) > p[.33 .33 .33]

this dist encourages sparsity

Prior on class means & covariances

$$p(\mu, \Lambda) = p(\mu | \Lambda) \, p(\Lambda)$$

Choose a Gaussian-Wishart prior governing the mean & precision matrix for each class

$$= \prod_{k=1}^{K} \mathcal{N}(\mu_k | m_0, (\beta_0 \Lambda_k)^{-1}) \, \mathcal{W}(\Lambda_k | W_0, \nu_0)$$

* Note - $\nu_0 >$ D-1 pg 693 (bottom)

B GMM Joint dist'n

$$p(X, Z, \pi, \mu, \Lambda)$$
$$= p(X \mid Z, \mu, \Lambda)\, p(Z \mid \pi)\, p(\pi)\, p(\mu \mid \Lambda)\, p(\Lambda)$$

Note: only X is observed. Everything else is latent

Interpretation of Model: Unsupervised clustering / learning, discovering latent "structure" in the data (clusters?), multi-modal density estimation where the latent density is unknown and complicated.

Variational Approximation ~~to GMM Posterior~~

Choose a "magic" factorization (this choose is not always obvious a priori).

$$q(Z, \pi, \mu, \Lambda) = q(Z)\, q(\pi, \mu, \Lambda)$$

Only necessary assumption to achieve practical results.

The functional forms of $q(Z)$ and $q(\pi, \mu, \Lambda)$ will be determined automatically.

Remember coupled "sequential/iterative" updates can be derived by applying update eqn. for each factor

Remember 10.9

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

Recipe

Apply this rule to BGMM with given factors and do lots of algebra.

Start with update rule for factor $q(Z)$.
- Want to discover $q^*(Z)$'s functional / distributional form.
- Want to be able to compute the parameters of this distribution ~~by~~ using expectations computed from other factors

$$\ln q^*(Z) = \mathbb{E}_{\pi,\mu,\Lambda}\left[\ln p(X, Z, \mu, \Lambda)\right] + \text{const}$$

- note as we go along, all factors that are not a function of Z with be absorbed into the const.

- plug in joint dist'n

$$\ln q^*(Z) = \mathbb{E}_{\mu,\pi,\Lambda}\left[\ln p(Z|\pi) + \ln p(X|Z,\mu,\Lambda)\right] + \text{const}$$

other terms in the joint

$$= \mathbb{E}_{\pi}\left[\ln p(Z|\pi)\right] + \mathbb{E}_{\mu,\Lambda}\left[\ln P(X|Z,\mu,\Lambda)\right] + \text{const}$$

plugin def's of each

$$= \mathbb{E}_{\pi}\left[\sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk} \ln \pi_k\right] + \mathbb{E}_{\mu,\Lambda}\left[\sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk} \cdot \ln \mathcal{N}(x_n|\mu_k, \Lambda_k^{-1})\right]$$

evaluate the log ~~under~~ of Normal distribution

Normal dist.

$$\mathcal{N}(x|\mu,\Sigma) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \cdot \exp\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right\}$$

and (from C.13) $|A^{-1}| = \frac{1}{|A|}$

$\ln q^*(Z)$

copy from previous page

$$= \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\, \mathbb{E}_{\pi}[\ln \pi_k] + \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk} \cdot \ln \mathcal{N}(x_n \mid \mu_k, \Lambda_k^{-1})$$

$$= \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\, \mathbb{E}_{\pi}[\ln \pi_k] + \text{const}$$

$$+ \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk} \left(-\frac{D}{2}\ln(2\pi) - \frac{1}{2}\mathbb{E}_{\Lambda}\left[\ln|\Lambda_k^{-1}|\right] - \frac{1}{2}\mathbb{E}_{\mu_k,\Lambda_k}\left[(x_n-\mu_k)^T\Lambda_k(x_n-\mu_k)\right]\right)$$

Let's define

note sign change

$$\ln \rho_{nk} = \mathbb{E}_{\pi}[\ln \pi_k] + \frac{1}{2}\mathbb{E}\left[\ln|\Lambda_k|\right] - \frac{D}{2}\ln(2\pi) - \frac{1}{2}\mathbb{E}_{\mu_k,\Lambda_k}\left[(x_n-\mu_k)^T\Lambda_k(x_n-\mu_k)\right]$$

then we can write

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

and by exponentiating

$$q^*(Z) = C \cdot \prod_{n=1}^{N}\prod_{k=1}^{K} \rho_{nk}^{z_{nk}}$$

- How do we normalize this distribution?
- How do we compute the terms in (expectations?)

1) Normalization: $q^*(Z)$ (unnormalized version). is N iid R.V.'s that are binary indicators over K states
   - this means that for each n $q^*(z_n)$ ← roughly can be normalized by summing over all K states, i.e.

$$r_{nk} = \frac{\rho_{nk}}{\sum_j \rho_{nj}}$$

Then $$\boxed{q^*(Z) = \prod_n \prod_k r_{nk}^{z_{nk}}}$$ ← "distributional form" ✓

2) computation?

Note: ρ_{nk} is an exponent of a real value, so always non-negative

corollary - r_{nk} also non-negative and will sum to 1

Note: $\boxed{\mathbb{E}[z_{nk}] = 0 \cdot (1-r_{nk}) + 1(r_{nk}) = r_{nk}}$

r_{nk} can be interpretted as a "responsibility"

- We have $q^*(z)$ which is a function of z but requires expectations of various quantities taken w.r.t. the other factors.

Other factors

Because we already know the answer, let's write down the following statistics of the observed data

$$\boxed{\begin{aligned} N_k &= \sum_{n=1}^{N} r_{nk} \\ \bar{x}_k &= \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} x_n \\ S_k &= \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} (x_n - \bar{x}_k)(x_n - \bar{x}_k)^T \end{aligned}}$$

- Note these look like the per-class updates for the EM GMM algorithm.

Next $q(\pi, \mu, \Lambda)$ and remember - use general result

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

- 1 induced factorization will turn out to factor note
- 2 choice of conjugate priors will help tremendously

$$\ln q_j^*(Z_j) = \mathbb{E}_{i\neq j}\left[\ln p(X,Z)\right] + \text{const.}$$

remember

$$p(X,Z,\pi,\mu,\Lambda) = p(X|Z,\mu,\Lambda)\,p(Z|\pi)\,p(\pi)\,p(\mu|\Lambda)\,p(\Lambda)$$

Applying this for $q^*(\pi,\mu,\Lambda)$

$$\ln q^*(\pi,\mu,\Lambda) = \mathbb{E}_{q(z)}\left[\ln\left(p(X|Z,\mu,\Lambda)\,p(Z|\pi)\,p(\pi)\,p(\mu,\Lambda)\right)\right] + \text{con}$$

$$= \mathbb{E}_z\left[\ln(p(Z|X,\mu,\Lambda)\right] + \underbrace{\mathbb{E}_z\left[\ln p(Z|\pi)\right] + \mathbb{E}_z\left[\ln p(\pi)\right]}_{\text{involve only } \pi \,\&\, Z} + \mathbb{E}_z\left[\ln p(\mu,\Lambda)\right] + \text{const}$$

involve only Z,μ,Λ

$$\Rightarrow q(\pi,\mu,\Lambda) = q(\pi)\,q(\mu,\Lambda) \quad \text{(further factorization)}$$

and noting that $p(\mu,\Lambda) = \prod_k p(\mu_k,\Lambda_k)$

$\quad$ & $P(Z|X,\mu,\Lambda) = \prod_k \prod_n \pi_k^{z_{nk}}$

$$\Rightarrow q(\pi,\mu,\Lambda) = q(\pi)\,q(\mu,\Lambda)$$
$$= q(\pi)\prod_{k=1}^{K} q(\mu_k,\Lambda_k)$$

- net effect: can optimize each of these factors <u>independently</u>

To start:

$$q^*(\pi) = \mathbb{E}_z\left[\ln p(Z|\pi)\right] + \ln p(\pi) + \text{const}$$

Remember Dir distribution

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

and

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

Plug these into $\ln q^*(\pi) = \mathbb{E}_Z[\ln p(Z|\pi)] + \ln p(\pi) + \text{const}$

$$\ln q^*(\pi) = (\alpha_0-1)\cdot\sum_{k=1}^{K}\ln\pi_k + \sum_{n=1}^{N}\sum_{k=1}^{K}\mathbb{E}[z_{nk}]\cdot\ln\pi_k + \text{const}$$

rearrange

$$\ln q^*(\pi) = \sum_{k=1}^{K}\left(\alpha_0-1+\sum_{n=1}^{N}\mathbb{E}[z_{nk}]\right)\ln\pi_k + \text{const}$$

now remember that $\mathbb{E}[z_{nk}] = r_{nk}$

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

so

$$\ln q^*(\pi) = \sum_{k=1}^{K}(\alpha_0-1+N_k)\ln\pi_k + \text{const}$$

$$\Rightarrow \quad q^*(\pi) = C\cdot\prod_{k=1}^{K}\pi_k^{(\alpha_0-1+N_k)}$$

examine carefully, this is a Dir dist
- hint: this is true because of conj. choice

$$\boxed{\Rightarrow q^*(\pi) = \mathrm{Dir}(\pi|\vec{\alpha})}$$

$$\vec{\alpha} = [\alpha_0+N_1, \alpha_0+N_2, \dots, \alpha_0+N_K]$$

To compute the responsibilities we need to compute

$$\mathbb{E}_{q^*(\pi)}[\ln\pi_k] = ?$$

This is given in PRML in B Page 687

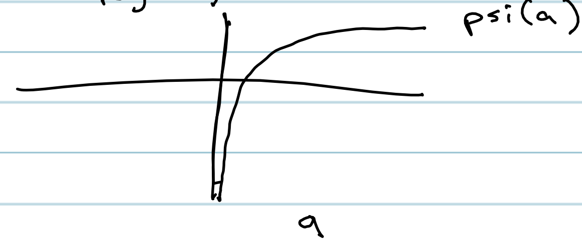
$$\mathbb{E}[\ln\pi_k] = \psi(\alpha_0+N_k) - \psi\left(K\alpha_0 + \sum_{k=1}^{K}N_k\right)$$

Remember $\sum_{k=1}^{K}N_k = N$

aside

$\psi(a) = \frac{d}{da}\Gamma(a)$ is called the "digamma" function (known as psi in matlab)

a lot like $\log(a)$



In order to complete the VB approximate inference algorithm for GMM's we need the distributional form of $q^*(\mu,\Lambda) = \prod_k q^*(\mu_k,\Lambda_k)$.

This involves a huge amount of algebra, but is similar in spirit to the last two examples.

To make further factorization obvious

$$\ln(q^*(\mu,\Lambda)) = \sum_k \mathbb{E}_z\left[\mathbb{E}_\pi\left[\ln p(\mu_k,\Lambda_k)\right]\right] + \sum_k\sum_n \mathbb{E}_z\left[\mathbb{E}_\pi\left[\ln N(x_n|\mu_k,\Lambda_k^{-1})^{z_{nk}}\right]\right] + \text{const}$$

$$= \underbrace{\sum_k \ln p(\mu_k,\Lambda_k) + \sum_k\sum_n \mathbb{E}_z[z_{nk}] \ln N(x_n|\mu_k,\Lambda_k^{-1})}_{\text{sums over } k \text{ imply } k \text{ products}}$$

and k independent factors $q(\mu_k,\Lambda_k)$

$$\ln(q^*(\mu_k,\Lambda_k)) = \ln p(\mu_k,\Lambda_k) + \sum_n \mathbb{E}_{q^*(z)}[z_{nk}] \ln N(x_n|\mu_k,\Lambda_k^{-1})$$

is nasty algebraically, but with experiance in conjugate families, one can almost guess the answer

$$\boxed{q^*(\mu_k, \Lambda_k) = \mathcal{N}(\mu_k \mid m_k, (\beta_k \Lambda_k)^{-1}) \mathcal{W}(\Lambda_k \mid W_k, \nu_k)}$$

where

$$\beta_k = \beta_0 + N_k$$

$$m_k = \frac{1}{\beta_k}(\beta_0 m_0 + N_k \bar{x}_k)$$

$$W_k^{-1} = W_0^{-1} + N_k S_k + \frac{\beta_0 N_k}{\beta_0 + N_k}(\bar{x}_k - m_0)(\bar{x}_k - m_0)^T$$

$$\nu_k = \nu_0 + N_k + 1$$

Now we have (in boxes) all distributional forms for VB ~~updates~~ factors.

VB updates to factor parameters require deriving or computing

$$\mathbb{E}_\Lambda[\ln|\Lambda_k|] \quad \& \quad \mathbb{E}_{\mu_k, \Lambda_k}[(x_n - \mu_k)^T \Lambda_k (x_n - \mu_k)]$$

$$10.65 \quad = \sum_{i=1}^{D} \psi\left(\frac{\nu_k + 1 - i}{2}\right) + D \cdot \ln 2 + \ln|W_k|$$

$$10.64 \quad = D\beta_k^{-1} + \nu_k (x_n - m_k)^T W_k (x_n - m_k)$$

Now we have everything to compute ~~the~~ $q^*(z)$, and given $q^*(z)$ we can update the parameters for $q^*(\pi)$ & $q^*(\mu_k, \Lambda_k)$ $\forall k$.

Running this until convergence will produce an approximate posterior

$$q(z, \mu, \Lambda, \pi) = q(\pi)\, q(z) \prod_{k=1}^{K} q(\mu_k, \Lambda_k)$$

with parameters given by formulae in boxes.

VB for GMM's

Note: - Coupled update equations roughly correspond to the E & M steps of EM for GMMs
- Functional form of variational factors is a consequence of the choice of conjugate priors

Interesting aside: Model selection : α_0 controls "sparsity" of resulting model, $\alpha_0 < 1$ prefers "sparse" (low component count) models (i.e. posterior distribution over models places high score on models with few components)

Under the VB GMM the expected value of the class probabilities is

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

if $N_k \approx 0$ then as $N \to \infty$ $\mathbb{E}[\pi_k] \to 0$ if α_0 is small. If $\alpha_0 \to \infty$ then $\alpha_k \approx \alpha_0$ and as N gets big $\mathbb{E}[\pi_k] \to 1/K$

Bayesian treatment	ML, EM
No singularities	Singularities
A little computational overhead	Computationally minimal

VB
model selection w/o cross validation