COMP9418 – Advanced Topics in Statistical Machine Learning

W4 – Variational Inference (Part II)

Instructor: Edwin V. Bonilla

School of Computer Science and Engineering

Last update: 22/8/17 11:10:52 pm

IV. Variational Inference in Bayesian GMMs

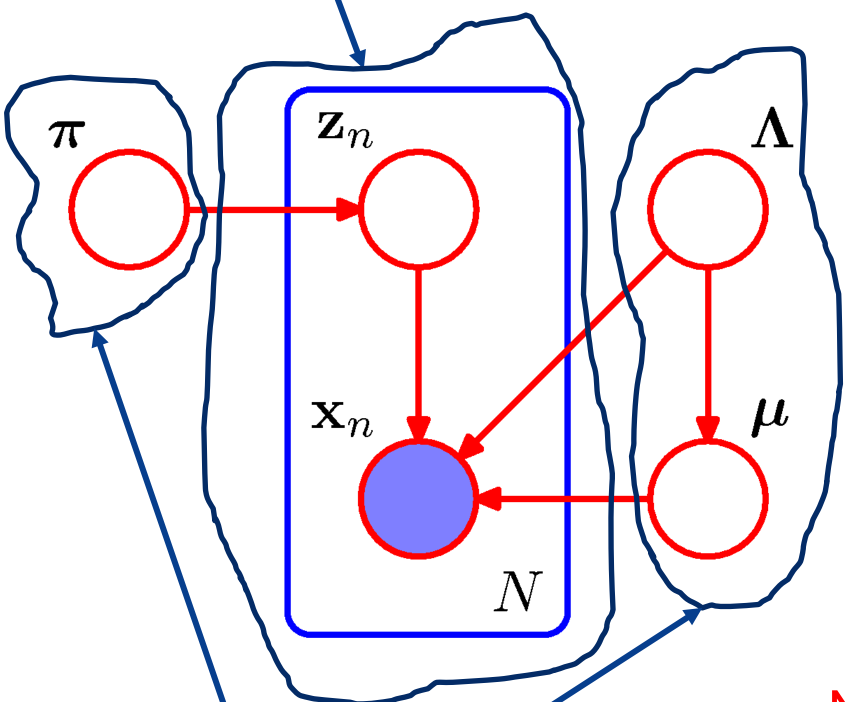
Bayesian GMMs (1)

$\mathbf{x}^{(n)}$: Observed variable for data point n → $\mathbf{X} = \{\mathbf{x}^{(n)}\}$

$\mathbf{z}^{(n)}$: Hidden or missing variable → $\mathbf{Z} = \{\mathbf{z}^{(n)}\}$

K : Number of mixture components

Standard mixture model



Priors over GMM parameters

One-hot encoding

Prior over latent variables

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

Precision matrix

Conditional likelihood

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

Need a prior over $\{\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}\}$

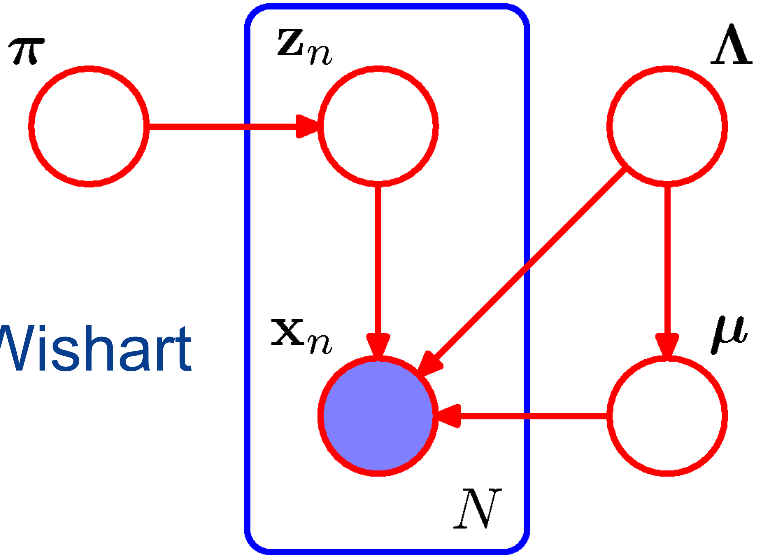
Bayesian GMMs (2)

- Dirichlet prior over mixture weights:

$$p(\boldsymbol{\pi}) = \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}) = \frac{1}{\mathrm{B}(\boldsymbol{\alpha})} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1}$$

- Independent (Conjugate) Gaussian-Wishart prior over mean and precision:

$$p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = p(\boldsymbol{\mu}|\boldsymbol{\Lambda}) p(\boldsymbol{\Lambda})$$



Brief Digression

The Gaussian-Wishart distribution

Assume:

$$\boldsymbol{\mu}_k \sim \mathcal{N}(\boldsymbol{\mu}_k | \tilde{\mathbf{m}}_k, (\tilde{\beta}_k \boldsymbol{\Lambda}_k)^{-1})$$

Wishart distribution (over PD matrices)

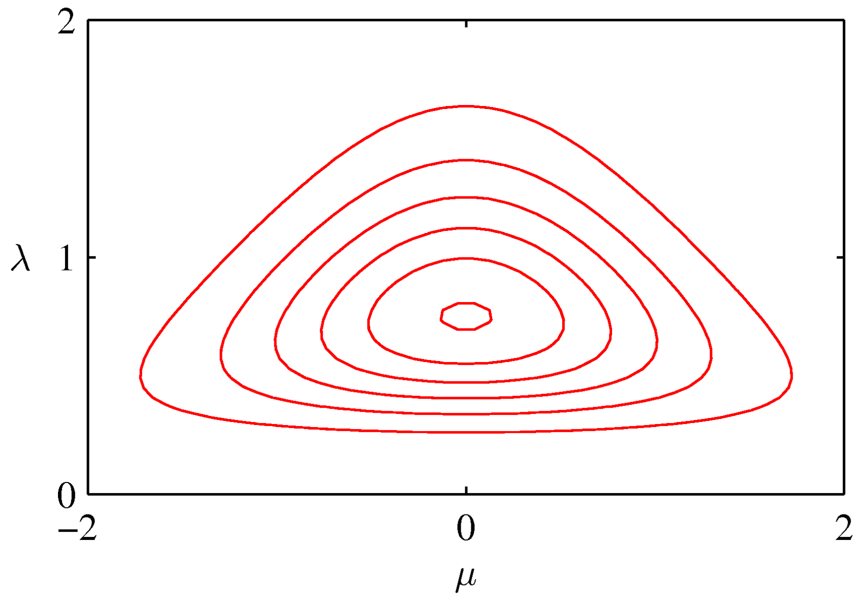
Generalization of the gamma distribution (to D>1)

$$\boldsymbol{\Lambda}_k \sim \mathcal{W}(\boldsymbol{\Lambda}_k | \tilde{\mathbf{W}}_k, \tilde{\nu}_k)$$

Then
$$\begin{aligned} p(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) &= \mathcal{NW}(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k | \tilde{\mathbf{m}}_k, \tilde{\beta}_k, \tilde{\mathbf{W}}_k, \tilde{\nu}_k) \\ &= \mathcal{N}(\boldsymbol{\mu}_k | \tilde{\mathbf{m}}_k, (\tilde{\beta}_k \boldsymbol{\Lambda}_k)^{-1}) \mathcal{W}(\boldsymbol{\Lambda}_k | \tilde{\mathbf{W}}_k, \tilde{\nu}_k) \end{aligned}$$

Gaussian-Wishart

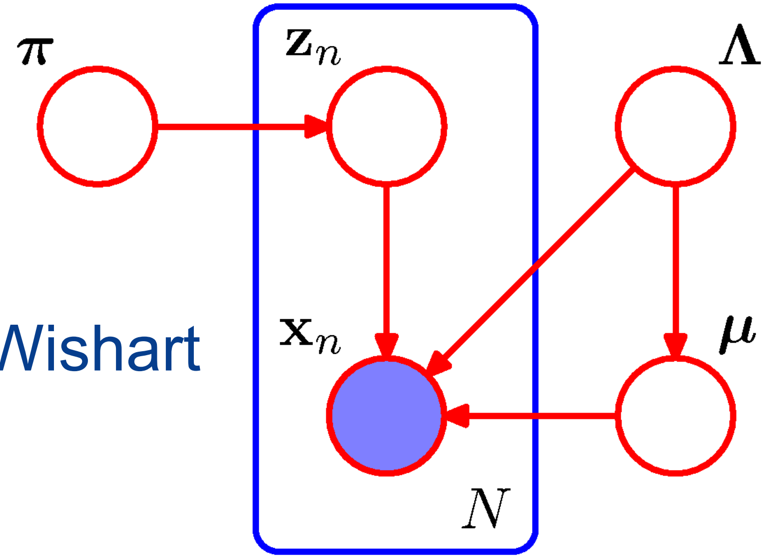
- Joint distribution over mean and precision
- Generalization of the Gaussian-gamma distribution (to D>1)



Bayesian GMMs (2)

- Dirichlet prior over mixture weights:

$$p(\boldsymbol{\pi}) = \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}) = \frac{1}{\mathrm{B}(\boldsymbol{\alpha})}\prod_{k=1}^{K} \pi_k^{\alpha_k - 1}$$



- Independent (Conjugate) Gaussian-Wishart prior over mean and precision:

$$p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = p(\boldsymbol{\mu}|\boldsymbol{\Lambda})p(\boldsymbol{\Lambda})$$

$$= \prod_{k=1}^{K} \mathcal{N}(\boldsymbol{\mu}_k|\mathbf{m}_0, (\beta_0\boldsymbol{\Lambda}_k)^{-1})\mathcal{W}(\boldsymbol{\Lambda}_k|\mathbf{W}_0, \nu_0)$$

- Full joint distribution:

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

  - Note only X are observed
    - » Difference between Z and GMM parameters?

Variational Posterior Distribution (1)

- We need to define a posterior over all unobserved variables:

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

  - Note factorisation (independence) between **Z** and GMM parameters
  - Remarkably, this is the only assumption we will make
    - » No constraints on the functional form
  - Notational simplicity (subscripts on q and dependence on **X** dropped)

- Updates using general result for factorised distributions:

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

  - yielding:

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

  - $q^*(\mathbf{Z})$: same functional form as the prior
  - $\tilde{r}_{nk}$: function of parameters of $q^*(\pi,\mu,\Lambda)$
    - » Automatically satisfies the constraints
    - » Role of responsibilities as $\mathbb{E}_{z_{nk}}[\tilde{r}_{nk}]$

Variational Posterior Distribution (2)

- Expected sufficient statistics (analogous to EM statistics):

$$\tilde{r}_k = \sum_{n=1}^{N} \tilde{r}_{nk} \qquad \tilde{\boldsymbol{\mu}}_k = \frac{1}{\tilde{r}_k} \sum_{n=1}^{N} \tilde{r}_{nk} \mathbf{x}^{(n)} \qquad \tilde{\boldsymbol{\Sigma}}_k = \frac{1}{\tilde{r}_k} \sum_{n=1}^{N} \tilde{r}_{nk} (\mathbf{x}^{(n)} - \tilde{\boldsymbol{\mu}}_k)(\mathbf{x}^{(n)} - \tilde{\boldsymbol{\mu}}_k)^T$$

- For the variational posterior over GMM parameters:

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

  - $q^*(\pi)$: same functional form as the prior

$$\tilde{\alpha}_k = \tilde{r}_k + \alpha_k$$

$$q^\star(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) = \mathcal{N}(\boldsymbol{\mu}_k | \tilde{\mathbf{m}}_k, (\tilde{\beta}_k \boldsymbol{\Lambda}_k)^{-1}) \mathcal{W}(\boldsymbol{\Lambda}_k | \tilde{\mathbf{W}}_k, \tilde{\nu}_k)$$

  - $q^*(\mu_k, \Lambda_k)$: same functional form as the prior
  - $\tilde{m}_k, \tilde{\beta}_k, \tilde{\mathbf{W}}_k, \tilde{\nu}_k$: function of expected sufficient statistics (which include $\tilde{r}_{nk}$)

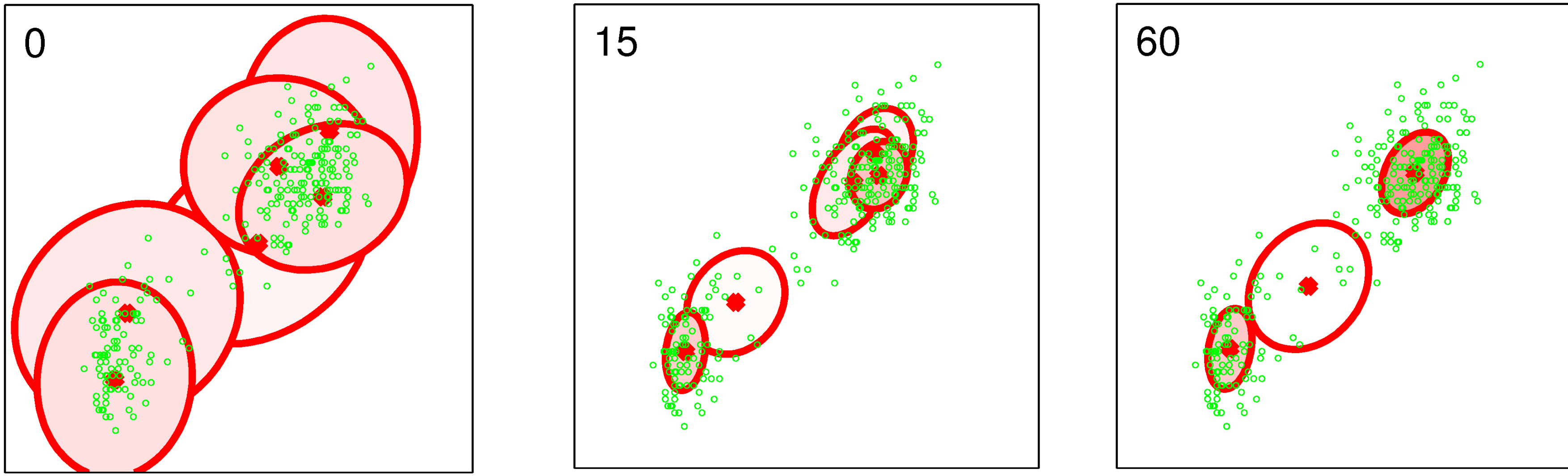
In fact, updates variational responsibilities $\tilde{r}_{nk}$, also analogous to EM's

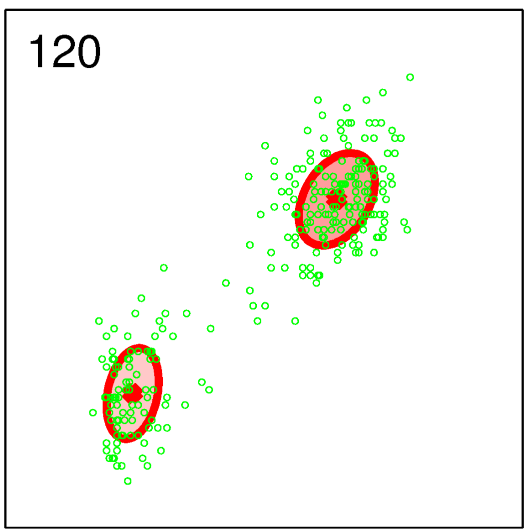
Variational EM-like Algorithm

1. VE-step: Update variational responsibilities $\tilde{r}_{nk}$ Using current distributions over model parameters
2. VM-step: Update posterior over model parameters using current responsibilities

- *Variational posteriors with same form as prior a consequence of conjugate priors*
  - *This is not the case for general priors or non-factorised distributions*
- *Deterministic approximations using bounds*
- *Assumptions on posterior functional form may be needed*
  - » *stochastic approximations of the variational objective*
  - » *Gradient-based optimization*

Variational Inference in Action



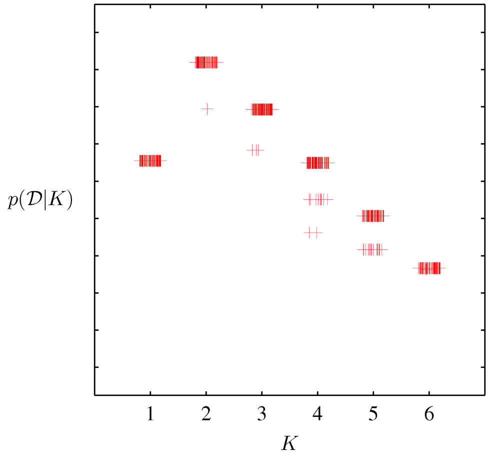


- Bayesian mixture with K=6
  - Ellipses: 1 std. dev. Contours
  - Density of red shade: mean of posterior mixing weights
  - Dirichlet prior to encourage sparsity: $\alpha=10^{-3}$
    - » Only two components with non-zero posterior mean weights

V. Final Remarks

Variational Inference: Remarks

- Close resemblance variational solution and EM's
  - As $N \to \infty$ → ML solution by EM
  - Similar computational cost
  - Singularity problems do not appear in the Bayesian treatment
  - No overfitting
  - Can evaluate lower bound for converge assessment and debugging
- Alternative derivations
  - Write down lower bound and optimise wrt posterior parameters
    - » Exploit knowledge about conjugacy
- Determining number of components
  - Model selection needs to consider K! settings
    - » Adding penalty term log K! to the lower bound
    - » Alternative to cross-validation



Advantages and Disadvantages of VI

(Slide based on Shakir Mohamed's tutorial on VI)

Disadvantages

- Never exact, only approximate posterior
- Typically underestimate the variance
- Can get stuck in local optima
- Limited theory

Advantages

- Applicable to a large class of models
- Can assess convergence
- Can do model selection
- Usually scalable and fast
- Compact representation of posterior

Conclusions

- Variational inference (VI) as a deterministic approximate algorithm for posterior estimation
  - Integration problem transformed into an optimisation problem
  - Optimisation of the variational objective (evidence lower bound) equivalent to minimizing KL(q||p)
  - usually fast and scalable but inexact
- Possible solutions
  - Free-form posterior
    - » Factorised distributions and mean-field solution
    - » Closed-form updates through conjugacy
  - Parameterized posterior
- Reading
  - Bishop (PRML, 2006): Ch. 10 (except Sec. 10.4, 10.5, 10.7), Sec 2.3.6
  - Murphy (MLaPP, 2012): Ch 21 (except Sec. 21.4, 21.7, 21.8)