2 - Statistical inference of incomplete data models

S. Robin

INRAE / AgroParisTech / univ. Paris-Saclay
Muséum National d'Histoire Naturelle

Winter School on Mathematical Statistics, Luxembourg, Dec'20

Outline

1 – Models with latent variables in ecology (statistical ecology)

2 – Variational inference for incomplete data models (statistics)

3 – Variational inference for species abundances and network models (statistical ecology)

4 – Beyond variational inference (statistics)

Part 2

Incomplete data models

Variational EM

Variational Bayes EM

Variational inference

Outline

Incomplete data models

Variational EM

Variational Bayes EM

Variational inference

Models with latent variables

Notations.

- Y observed variables (responses)
- x observed covariates (explanatory)
- Z latent (= unobserved, hidden, state) variables
- θ unknown parameters

Models with latent variables

Notations.

Y observed variables (responses)

x observed covariates (explanatory)

Z latent (= unobserved, hidden, state) variables

θ unknown parameters

'Definition' of latent variables.

- Frequentist setting:

$$\text{latent variables} = \text{random}, \qquad \text{parameters} = \text{fixed}$$

Models with latent variables

Notations.

Y observed variables (responses)

x observed covariates (explanatory)

Z latent (= unobserved, hidden, state) variables

θ unknown parameters

'Definition' of latent variables.

- Frequentist setting:

$$\text{latent variables} = \text{random}, \qquad \text{parameters} = \text{fixed}$$

- Bayesian setting:

$$\text{both latent variables and parameters} = \text{random}$$

but

$$\# \text{ latent variables } \simeq \# \text{ data}, \qquad \# \text{ parameters } \ll \# \text{ data}$$

Likelihoods

'Complete' likelihood : both latent and observed variables[1]:

$$p_\theta(Y, Z) = p_\theta(Y, Z; x)$$

$\rightarrow$ often reasonably easy to handle, but involves the unobserved Z

[1] x is dropped for the sake of clarity

[2] We will use $\int \dots \, \mathrm{d}z$ even when Z is discrete (should be $\sum_{z \in \mathcal{Z}}$).

Likelihoods

'Complete' likelihood : both latent and observed variables[1]:

$$p_\theta(Y, Z) = p_\theta(Y, Z; x)$$

$\rightarrow$ often reasonably easy to handle, but involves the unobserved Z

'Observed' likelihood = marginal likelihood of the observed data[2]

$$p_\theta(Y) = \int_{\mathcal{Z}} p_\theta(Y, z) \, \mathrm{d}z$$

$\rightarrow$ involves only the observed Y, but most often intractable

[1] x is dropped for the sake of clarity

[2] We will use $\int \dots \mathrm{d}z$ even when Z is discrete (should be $\sum_{z \in \mathcal{Z}}$).

Maximum likelihood

Maximum likelihood estimate (MLE):

$$\theta_{MLE} = \arg\max_{\theta} \; p_\theta(Y) = \arg\max_{\theta} \; \int p_\theta(Y, z) \, \mathrm{d}z$$

most often intractable

Maximum likelihood

Maximum likelihood estimate (MLE):

$$\theta_{MLE} = \arg\max_\theta \; p_\theta(Y) = \arg\max_\theta \; \int p_\theta(Y, z) \, \mathrm{d}z$$

most often intractable

Decomposition of the log-likelihood [DLR77]:

Maximum likelihood

Maximum likelihood estimate (MLE):

$$\theta_{MLE} = \arg\max_\theta \; p_\theta(Y) = \arg\max_\theta \; \int p_\theta(Y, z) \, \mathrm{d}z$$

most often intractable

Decomposition of the log-likelihood [DLR77]: By definition

$$p_\theta(Z \mid Y) = p_\theta(Y, Z) \,/\, p_\theta(Y)$$

Maximum likelihood

Maximum likelihood estimate (MLE):

$$\theta_{MLE} = \arg\max_\theta \; p_\theta(Y) = \arg\max_\theta \; \int p_\theta(Y, z) \, \mathrm{d}z$$

most often intractable

Decomposition of the log-likelihood [DLR77]: By definition

$$p_\theta(Z \mid Y) = p_\theta(Y, Z) \, / p_\theta(Y)$$

so (reverting the ratio and taking the log)

$$\log p_\theta(Y) = \log p_\theta(Y, Z) - \log p_\theta(Z \mid Y)$$

Maximum likelihood

Maximum likelihood estimate (MLE):

$$\theta_{MLE} = \arg\max_{\theta} \; p_\theta(Y) = \arg\max_{\theta} \; \int p_\theta(Y, z) \, \mathrm{d}z$$

most often intractable

Decomposition of the log-likelihood [DLR77]: By definition

$$p_\theta(Z \mid Y) = p_\theta(Y, Z) \,/\, p_\theta(Y)$$

so (reverting the ratio and taking the log)

$$\log p_\theta(Y) = \log p_\theta(Y, Z) - \log p_\theta(Z \mid Y)$$

and (taking the conditional expectation on both side)

$$\mathbb{E}_\theta[\log p_\theta(Y) \mid Y] = \mathbb{E}_\theta[\log p_\theta(Y, Z) \mid Y] - \mathbb{E}_\theta[\log p_\theta(Z \mid Y) \mid Y]$$

Maximum likelihood

Maximum likelihood estimate (MLE):

$$\theta_{MLE} = \arg\max_\theta \ p_\theta(Y) = \arg\max_\theta \int p_\theta(Y, z) \, \mathrm{d}z$$

most often intractable

Decomposition of the log-likelihood [DLR77]: By definition

$$p_\theta(Z \mid Y) = p_\theta(Y, Z) \,/\, p_\theta(Y)$$

so (reverting the ratio and taking the log)

$$\log p_\theta(Y) = \log p_\theta(Y, Z) - \log p_\theta(Z \mid Y)$$

and (taking the conditional expectation on both side)

$$\mathbb{E}_\theta[\log p_\theta(Y) \mid Y] = \mathbb{E}_\theta[\log p_\theta(Y, Z) \mid Y] - \mathbb{E}_\theta[\log p_\theta(Z \mid Y) \mid Y]$$

that is

$$\log p_\theta(Y) = \mathbb{E}_\theta[\log p_\theta(Y, Z) \mid Y] - \mathbb{E}_\theta[\log p_\theta(Z \mid Y) \mid Y]$$

Decomposition of $\log p_\theta(Y)$

$$\log p_\theta(Y) = \mathbb{E}_\theta[\log p_\theta(Y, Z) \mid Y] - \mathbb{E}_\theta[\log p_\theta(Z \mid Y) \mid Y]$$

$\log p_\theta(Y) =$ (observed) log-likelihood = objective function

$\mathbb{E}_\theta[\log p_\theta(Y, Z) \mid Y] =$ conditional expectation of the 'complete' log-likelihood

$-\mathbb{E}_\theta[\log p_\theta(Z \mid Y) \mid Y] =$ conditional entropy $= \mathcal{H}(p_\theta(Z \mid Y))$

Expectation-maximization (EM) algorithm (1/2)

Iterative algorithm [DLR77]: denoting θ^h the estimate at step h, repeat until convergence

$$\theta^{h+1} = \arg\max_{\theta} \mathbb{E}_{\theta^h}[\log p_{\theta}(Y, Z) \mid Y]$$

which requires to (sub-)steps:

Expectation-maximization (EM) algorithm (1/2)

Iterative algorithm [DLR77]: denoting θ^h the estimate at step h, repeat until convergence

$$\theta^{h+1} = \arg\max_{\theta} \mathbb{E}_{\theta^h}[\log p_{\theta}(Y, Z) \mid Y]$$

which requires to (sub-)steps:

Expectation step = computation of all moments needed to evaluate $\mathbb{E}_{\theta^h}[\cdot \mid Y]$

Expectation-maximization (EM) algorithm (1/2)

Iterative algorithm [DLR77]: denoting θ^h the estimate at step h, repeat until convergence

$$\theta^{h+1} = \arg\max_{\theta} \mathbb{E}_{\theta^h}[\log p_{\theta}(Y, Z) \mid Y]$$

which requires to (sub-)steps:

Expectation step = computation of all moments needed to evaluate $\mathbb{E}_{\theta^h}[\cdot \mid Y]$

Maximization step = update the estimate as $\arg\max_{\theta}$

Expectation-maximization (EM) algorithm (1/2)

Iterative algorithm [DLR77]: denoting θ^h the estimate at step h, repeat until convergence

$$\theta^{h+1} = \arg\max_{\theta} \mathbb{E}_{\theta^h}[\log p_{\theta}(Y, Z) \mid Y]$$

which requires to (sub-)steps:

Expectation step = computation of all moments needed to evaluate $\mathbb{E}_{\theta^h}[\cdot \mid Y]$

Maximization step = update the estimate as $\arg\max_{\theta}$

Main property:

$$\log p_{\theta^{h+1}}(Y) \geq \log p_{\theta^h}(Y)$$

$\rightarrow$ Proof in #32.

Expectation-maximization (EM) algorithm (2/2)

$$\theta^{h+1} = \underbrace{\arg\max_\theta}_{\text{M step}} \; \underbrace{\mathbb{E}_{\theta^h}}_{\text{E step}} \left[\log p_\theta(Y, Z) \mid Y\right]$$

Some remarks.

Expectation-maximization (EM) algorithm (2/2)

$$\theta^{h+1} = \underbrace{\arg\max_\theta}_{\text{M step}} \underbrace{\mathbb{E}_{\theta^h}}_{\text{E step}} [\log p_\theta(Y, Z) \mid Y]$$

Some remarks.

1. θ occurs twice in the formula

Expectation-maximization (EM) algorithm (2/2)

$$\theta^{h+1} = \underbrace{\arg\max_\theta}_{\text{M step}} \underbrace{\mathbb{E}_{\theta^h}}_{\text{E step}} \left[\log p_\theta(Y, Z) \mid Y\right]$$

Some remarks.

1. θ occurs twice in the formula

2. Relies on the 'complete' (= joint): easier to handle

Expectation-maximization (EM) algorithm (2/2)

$$\theta^{h+1} = \underbrace{\arg\max_\theta}_{\text{M step}} \underbrace{\mathbb{E}_{\theta^h}}_{\text{E step}} [\log p_\theta(Y, Z) \mid Y]$$

Some remarks.

1. θ occurs twice in the formula
2. Relies on the 'complete' (= joint): easier to handle
3. The objective function $\log p_\theta(Y)$ is never evaluated

Expectation-maximization (EM) algorithm (2/2)

$$\theta^{h+1} = \underbrace{\arg\max_{\theta}}_{\text{M step}} \underbrace{\mathbb{E}_{\theta^h}}_{\text{E step}} [\log p_\theta(Y, Z) \mid Y]$$

Some remarks.

1. θ occurs twice in the formula
2. Relies on the 'complete' (= joint): easier to handle
3. The objective function $\log p_\theta(Y)$ is never evaluated
4. Actually, no need to maximize wrt θ:

$$\mathbb{E}_{\theta^h}[\log p_{\theta^h}(Y, Z) \mid Y] \geq \mathbb{E}_{\theta^h}[\log p_{\theta^{h+1}}(Y, Z) \mid Y]$$

suffices ('generalized' EM = GEM)

M step

Most of the time, same difficulty as maximum likelihood in absence of latent variables

[3] which includes most PLN, SBM and LBM.

M step

Most of the time, same difficulty as maximum likelihood in absence of latent variables

Ex.: Exponential family. If the joint likelihood belongs to the exponential family[3]

$$\log p_\theta(Y, Z) = t(Y, Z)^\intercal \theta - a(Y, Z) - b(\theta)$$

then

$$\mathbb{E}_\theta[\log p_\theta(Y, Z) \mid Y] = \mathbb{E}_\theta[t(Y, Z) \mid Y]^\intercal \theta - \mathbb{E}_\theta[a(Y, Z) \mid Y] - b(\theta)$$

[3] which includes most PLN, SBM and LBM.

M step

Most of the time, same difficulty as maximum likelihood in absence of latent variables

Ex.: Exponential family. If the joint likelihood belongs to the exponential family[3]

$$\log p_\theta(Y, Z) = t(Y, Z)^\intercal \theta - a(Y, Z) - b(\theta)$$

then

$$\mathbb{E}_\theta[\log p_\theta(Y, Z) \mid Y] = \mathbb{E}_\theta[t(Y, Z) \mid Y]^\intercal \theta - \mathbb{E}_\theta[a(Y, Z) \mid Y] - b(\theta)$$

- Usual MLE for θ
- Provided that $\mathbb{E}_\theta[t(Y, Z) \mid Y]$ and $\mathbb{E}_\theta[a(Y, Z) \mid Y]$ can be evaluated

[3] which includes most PLN, SBM and LBM.

E step

Critical step: requires to compute some moments of

$$p_\theta(Z \mid Y) = \frac{p_\theta(Y, Z)}{p_\theta(Y)}$$

Three situations.

E step

Critical step: requires to compute some moments of

$$p_\theta(Z \mid Y) = \frac{p_\theta(Y, Z)}{p_\theta(Y)}$$

Three situations.

- Easy cases: explicit E step
  $\rightarrow$ mixture models (Bayes formula), simple mixed models (close form conditional)

E step

Critical step: requires to compute some moments of

$$p_\theta(Z \mid Y) = \frac{p_\theta(Y, Z)}{p_\theta(Y)}$$

Three situations.

- Easy cases: explicit E step
  $\rightarrow$ mixture models (Bayes formula), simple mixed models (close form conditional)

- Tricky cases: non-explicit, but still exact E step, ...
  $\rightarrow$ hidden Markov models (forward-backward recursions), evolutionary models (upward-downward), belief propagation on trees...

E step

Critical step: requires to compute some moments of

$$p_\theta(Z \mid Y) = \frac{p_\theta(Y, Z)}{p_\theta(Y)}$$

Three situations.

- Easy cases: explicit E step
  $\rightarrow$ mixture models (Bayes formula), simple mixed models (close form conditional)
- Tricky cases: non-explicit, but still exact E step, ...
  $\rightarrow$ hidden Markov models (forward-backward recursions), evolutionary models (upward-downward), belief propagation on trees...
- Bad cases: no exact evaluation
  $\rightarrow$ either sample from $p_\theta(Z \mid Y)$ (Monte-Carlo)
  $\rightarrow$ or approximate $q(Z) \simeq p_\theta(Z \mid Y)$ (variational approximations)

Poisson log-normal model

Univariate case. ($p = 1$ species)

- $Z \sim \mathcal{N}(0, \sigma^2)$
- $Y \sim \mathcal{P}\left(e^{\mu+Z}\right)$

$\rightarrow$ Z is marginally Gaussian (- -)

Poisson log-normal model

Univariate case. ($p = 1$ species)

- $Z \sim \mathcal{N}(0, \sigma^2)$
- $Y \sim \mathcal{P}\left(e^{\mu+Z}\right)$

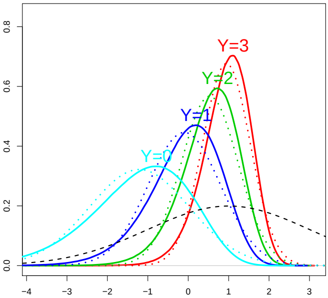
$\rightarrow$ Z is marginally Gaussian (- -)

Conditional distribution.

$$p(z \mid Y = y) \propto \exp\left(-\frac{z^2}{2\sigma^2} - e^{\mu+z} + y(\mu + z)\right)$$

$\rightarrow$ no close form

$\rightarrow$ Z is not conditionaly Gaussian ($-$ vs $\cdots$)



$\mu = 1, \quad \sigma = 2$

Stochastic block-model

Poisson model. (no covariate)

- $\{Z_i\}$ iid $\sim \mathcal{M}(1, \pi)$
- $Y_{ij} \sim \mathcal{P}\left(e^{\alpha_{Z_i Z_j}}\right)$

$\rightarrow$ The Z_i are marginally independent

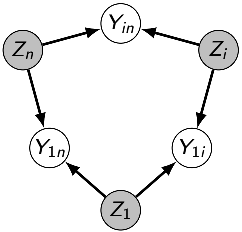
Stochastic block-model

Poisson model. (no covariate)

- $\{Z_i\}$ iid $\sim \mathcal{M}(1, \pi)$
- $Y_{ij} \sim \mathcal{P}\left(e^{\alpha_{Z_i Z_j}}\right)$

$\rightarrow$ The Z_i are marginally independent

Directed graphical model



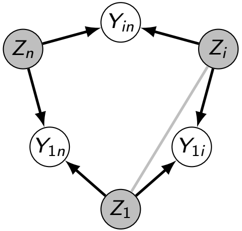
Stochastic block-model

Poisson model. (no covariate)

- $\{Z_i\}$ iid $\sim \mathcal{M}(1, \pi)$
- $Y_{ij} \sim \mathcal{P}\left(e^{\alpha_{Z_i Z_j}}\right)$

$\rightarrow$ The Z_i are marginally independent

Moralization of (Z_1, Z_i)



Moralization. [Lau96]

$$p(Z_i, Z_j \mid Y_{ij}) = \frac{p(Z_i)p(Z_j)p(Y_{ij} \mid Z_i, Z_j)}{p(Y_{ij})}$$

does not factorize in (Z_i, Z_j).

Stochastic block-model

Poisson model. (no covariate)

- $\{Z_i\}$ iid $\sim \mathcal{M}(1, \pi)$
- $Y_{ij} \sim \mathcal{P}\left(e^{\alpha_{Z_i Z_j}}\right)$

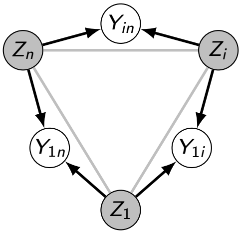
$\rightarrow$ The Z_i are marginally independent

Moralization. [Lau96]

$$p(Z_i, Z_j \mid Y_{ij}) = \frac{p(Z_i)p(Z_j)p(Y_{ij} \mid Z_i, Z_j)}{p(Y_{ij})}$$

does not factorize in (Z_i, Z_j).

Moralization for all pairs



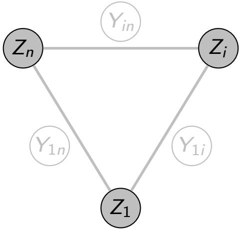
Stochastic block-model

Poisson model. (no covariate)

- $\{Z_i\}$ iid $\sim \mathcal{M}(1, \pi)$
- $Y_{ij} \sim \mathcal{P}\left(e^{\alpha_{Z_i Z_j}}\right)$

$\rightarrow$ The Z_i are marginally independent

Conditional graphical model



Moralization. [Lau96]

$$p(Z_i, Z_j \mid Y_{ij}) = \frac{p(Z_i)p(Z_j)p(Y_{ij} \mid Z_i, Z_j)}{p(Y_{ij})}$$

does not factorize in (Z_i, Z_j).

$\rightarrow$ The Z_i are all conditionally dependent

Outline

Incomplete data models

Variational EM

Variational Bayes EM

Variational inference

General aim

Problem. $p_\theta(Z \mid Y)$ being intractable, we look for a 'good' approximation of it:

$$q(Z) \approx p_\theta(Z \mid Y)$$

More specifically, given

General aim

Problem. $p_\theta(Z \mid Y)$ being intractable, we look for a 'good' approximation of it:

$$q(Z) \approx p_\theta(Z \mid Y)$$

More specifically, given

- a set of approximating distributions $\mathcal{Q}$ and

General aim

Problem. $p_\theta(Z \mid Y)$ being intractable, we look for a 'good' approximation of it:

$$q(Z) \approx p_\theta(Z \mid Y)$$

More specifically, given

- a set of approximating distributions $\mathcal{Q}$ and
- a divergence measure $D[q||p]$,

General aim

Problem. $p_\theta(Z \mid Y)$ being intractable, we look for a 'good' approximation of it:

$$q(Z) \approx p_\theta(Z \mid Y)$$

More specifically, given

- a set of approximating distributions $\mathcal{Q}$ and
- a divergence measure $D[q \| p]$,

we look for

$$q^* = \arg\min_{q \in \mathcal{Q}} D\left[q(Z) \| p_\theta(Z \mid Y)\right]$$

Variational approximations

References. Huge literature; see [WJ08] for a general introduction or [BKM17] for a more recent and concise review

Variational approximations

References. Huge literature; see [WJ08] for a general introduction or [BKM17] for a more recent and concise review

Not all methods enter the framework described above

- loopy belief propagation [MWJ99]
- minimization of Bethe's free energy [YFW01]

Variational approximations

References. Huge literature; see [WJ08] for a general introduction or [BKM17] for a more recent and concise review

Not all methods enter the framework described above

- loopy belief propagation [MWJ99]
- minimization of Bethe's free energy [YFW01]

Choice of the divergence measure.

- Most popular choice = Küllback–Leibler:

$$D[q||p] = KL[q||p] = \mathbb{E}_q \log (q/p)$$

$\rightarrow$ the error $\log(q/p)$ is averaged wrt the approximation q itself

Variational approximations

References. Huge literature; see [WJ08] for a general introduction or [BKM17] for a more recent and concise review

Not all methods enter the framework described above

- loopy belief propagation [MWJ99]
- minimization of Bethe's free energy [YFW01]

Choice of the divergence measure.

- Most popular choice = Küllback–Leibler:

$$D[q\|p] = KL[q\|p] = \mathbb{E}_q \log\left(q/p\right)$$

  $\rightarrow$ the error $\log(q/p)$ is averaged wrt the approximation q itself

- Expectation propagation (EP, [Min01]): $D[q\|p] = KL[p\|q]$
  $\rightarrow$ more sensible, but requires integration wrt p

Variational approximations

References. Huge literature; see [WJ08] for a general introduction or [BKM17] for a more recent and concise review

Not all methods enter the framework described above

- loopy belief propagation [MWJ99]
- minimization of Bethe's free energy [YFW01]

Choice of the divergence measure.

- Most popular choice = Küllback–Leibler:

$$D[q||p] = KL[q||p] = \mathbb{E}_q \log\left(q/p\right)$$

$\rightarrow$ the error $\log(q/p)$ is averaged wrt the approximation q itself

- Expectation propagation (EP, [Min01]): $D[q||p] = KL[p||q]$
  $\rightarrow$ more sensible, but requires integration wrt p

- Many others (see e.g. [Min05])

Variational EM algorithm

In a nutshell: replace the E step with an approximation ('VE') step

[1] Actually log-evidence, as the evidence is $p(Y)$

Variational EM algorithm

In a nutshell: replace the E step with an approximation ('VE') step

'Evidence lower bound' (ELBO) = lower bound of the log-likelihood:

$$J_{\theta,q}(Y) = \log p_\theta(Y) - KL\left[q(Z) \| p_\theta(Z \mid Y)\right]$$

[1] Actually log-evidence, as the evidence is $p(Y)$

Variational EM algorithm

In a nutshell: replace the E step with an approximation ('VE') step

'Evidence lower bound' (ELBO) = lower bound of the log-likelihood:

$$J_{\theta,q}(Y) = \log p_\theta(Y) - KL\left[q(Z) \| p_\theta(Z \mid Y)\right]$$

VEM algorithm.

VE step: maximize $J_{\theta,q}(Y)$ wrt q

M step: maximize $J_{\theta,q}(Y)$ wrt θ

[1]Actually log-evidence, as the evidence is $p(Y)$

Variational EM algorithm

In a nutshell: replace the E step with an approximation ('VE') step

'Evidence lower bound' (ELBO) = lower bound of the log-likelihood:

$$J_{\theta,q}(Y) = \log p_\theta(Y) - KL\left[q(Z) \| p_\theta(Z \mid Y)\right]$$

VEM algorithm.

VE step: maximize $J_{\theta,q}(Y)$ wrt q

M step: maximize $J_{\theta,q}(Y)$ wrt θ

Property: $J_{\theta,q}(Y)$ increases at each step.

[1] Actually log-evidence, as the evidence is $p(Y)$

Variational EM algorithm

The ELBO can written in two ways:

$$J_{\theta,q}(Y) = \log p_\theta(Y) - KL\left[q(Z) \| p_\theta(Z \mid Y)\right]$$

$$= \mathbb{E}_q \log p_\theta(Y, Z) - \mathbb{E}_q \log q(Z)$$

$\rightarrow$ See #33

Variational EM algorithm

The ELBO can written in two ways:

$$J_{\theta,q}(Y) = \log p_\theta(Y) - KL\left[q(Z) \| p_\theta(Z \mid Y)\right]$$

$$= \mathbb{E}_q \log p_\theta(Y, Z) - \mathbb{E}_q \log q(Z)$$

$\rightarrow$ See #33

VEM algorithm.

- VE step (approximation):

$$q^{h+1} = \arg\min_{q \in \mathcal{Q}} \; KL\left[q(Z) \| p_{\theta^h}(Z \mid Y)\right]$$

- M step (parameter update):

$$\theta^{h+1} = \arg\max_\theta \; \mathbb{E}_{q^{h+1}} \log p_\theta(Y, Z)$$

EM as a VEM algorithm

We have that

$$\log p_\theta(Y) = \mathbb{E}[\log p_\theta(Y, Z) \mid Y] - \mathbb{E}[\log p_\theta(Z \mid Y) \mid Y] \qquad \text{(EM)}$$

$$J_{\theta,q}(Y) = \mathbb{E}_q[\log p_\theta(Y, Z)] - \mathbb{E}_q[\log q(Z)] \qquad \text{(VEM)}$$

EM as a VEM algorithm

We have that

$$\log p_\theta(Y) = \mathbb{E}[\log p_\theta(Y, Z) \mid Y] - \mathbb{E}[\log p_\theta(Z \mid Y) \mid Y] \qquad \text{(EM)}$$

$$J_{\theta,q}(Y) = \mathbb{E}_q[\log p_\theta(Y, Z)] - \mathbb{E}_q[\log q(Z)] \qquad \text{(VEM)}$$

- Both are the same iff $q(Z) = p_\theta(Z \mid Y)$ (as $KL\left[q^{h+1}(Z) \| p_{\theta^h}(Z \mid Y)\right] = 0$)

EM as a VEM algorithm

We have that

$$\log p_\theta(Y) = \mathbb{E}[\log p_\theta(Y, Z) \mid Y] - \mathbb{E}[\log p_\theta(Z \mid Y) \mid Y] \qquad \text{(EM)}$$

$$J_{\theta,q}(Y) = \mathbb{E}_q[\log p_\theta(Y, Z)] - \mathbb{E}_q[\log q(Z)] \qquad \text{(VEM)}$$

- Both are the same iff $q(Z) = p_\theta(Z \mid Y)$ (as $KL\left[q^{h+1}(Z) \| p_{\theta^h}(Z \mid Y)\right] = 0$)
- This happens when $\mathcal{Q}$ is unrestricted, that is

$$q^{h+1}(Z) = \arg\min_q \; KL\left[q(Z) \| p_{\theta^h}(Z \mid Y)\right] = p_{\theta^h}(Z \mid Y)$$

EM as a VEM algorithm

We have that

$$\log p_\theta(Y) = \mathbb{E}[\log p_\theta(Y, Z) \mid Y] - \mathbb{E}[\log p_\theta(Z \mid Y) \mid Y] \qquad \text{(EM)}$$

$$J_{\theta,q}(Y) = \mathbb{E}_q[\log p_\theta(Y, Z)] - \mathbb{E}_q[\log q(Z)] \qquad \text{(VEM)}$$

- Both are the same iff $q(Z) = p_\theta(Z \mid Y)$ (as $KL\left[q^{h+1}(Z) \| p_{\theta^h}(Z \mid Y)\right] = 0$)
- This happens when $\mathcal{Q}$ is unrestricted, that is

$$q^{h+1}(Z) = \arg\min_q \; KL\left[q(Z) \| p_{\theta^h}(Z \mid Y)\right] = p_{\theta^h}(Z \mid Y)$$

- This provides us with a second proof of EM's main property

'Mean-field' approximations

'Mean-field' approximations

Choice of the approximation class. A popular choice is

$$\mathcal{Q}_{\text{fact}} = \{\text{factorable distributions}\} = \{q : q(Z) = \prod_i q_i(Z_i)\}$$

'Mean-field' approximations

Choice of the approximation class. A popular choice is

$$\mathcal{Q}_{\text{fact}} = \{\text{factorable distributions}\} = \{q : q(Z) = \prod_i q_i(Z_i)\}$$

Property. For a given distribution $p(Z)$,

$$q^* = \arg\min_{q \in \mathcal{Q}_{\text{fact}}} KL[q \| p]$$

satisfies

$$q_i^*(Z_i) \propto \exp\left(\mathbb{E}_{\bigotimes_{j \neq i} q_j^*} \log p(Z)\right)$$

$\rightarrow$ Proof in [Bea03] (sketch in #34)

'Mean-field' approximations

Choice of the approximation class. A popular choice is

$$\mathcal{Q}_{\text{fact}} = \{\text{factorable distributions}\} = \{q : q(Z) = \prod_i q_i(Z_i)\}$$

Property. For a given distribution $p(Z)$,

$$q^* = \arg\min_{q \in \mathcal{Q}_{\text{fact}}} KL[q \| p]$$

satisfies

$$q_i^*(Z_i) \propto \exp\left(\mathbb{E}_{\bigotimes_{j \neq i} q_j^*} \log p(Z)\right)$$

→ Proof in [Bea03] (sketch in #34)

- $\log q_i^*(Z_i)$ is obtained by setting the $\{Z_j\}_{j \neq i}$ 'to their respective mean' (each wrt to q_j^*).

Outline

Incomplete data models

Variational EM

Variational Bayes EM

Variational inference

Bayesian inference

Bayesian setting: The parameters in θ are random (no latent variable yet)

Bayesian inference

Bayesian setting: The parameters in θ are random (no latent variable yet)

- 'Prior' = marginal distribution of the parameter

$$p(\theta)$$

Bayesian inference

Bayesian setting: The parameters in θ are random (no latent variable yet)

- 'Prior' = marginal distribution of the parameter

$$p(\theta)$$

- 'Likelihood' = conditional distribution of the observations

$$p(Y \mid \theta)$$

Bayesian inference

Bayesian setting: The parameters in θ are random (no latent variable yet)

- 'Prior' = marginal distribution of the parameter

$$p(\theta)$$

- 'Likelihood' = conditional distribution of the observations

$$p(Y \mid \theta)$$

- 'Posterior' = conditional distribution of the parameters given the data

$$p(\theta \mid Y) = \frac{p(\theta)p(Y \mid \theta)}{\int p(\theta)p(Y \mid \theta) \, \mathrm{d}\theta}$$

Variational Bayes

Ideal case: Explicit posterior $\rightarrow$ Conjugate priors

Variational Bayes

Ideal case: Explicit posterior $\rightarrow$ Conjugate priors

Most of the time: No explicit form for $p(\theta \mid Y)$

Variational Bayes

Ideal case: Explicit posterior $\rightarrow$ Conjugate priors

Most of the time: No explicit form for $p(\theta \mid Y)$

- Sample from it, i.e. try to get

$$\{\theta^b\}_{1 \leq b \leq B} \overset{\text{iid}}{\approx} p(\theta \mid Y)$$

$\rightarrow$ Monte-Carlo (MC), MCMC, SMC, HMC, ...

Variational Bayes

Ideal case: Explicit posterior $\rightarrow$ Conjugate priors

Most of the time: No explicit form for $p(\theta \mid Y)$

- Sample from it, i.e. try to get

$$\{\theta^b\}_{1 \leq b \leq B} \overset{\text{iid}}{\approx} p(\theta \mid Y)$$

  $\rightarrow$ Monte-Carlo (MC), MCMC, SMC, HMC, ...

- Approximate it, i.e. look for

$$q(\theta) \simeq p(\theta \mid Y)$$

  $\rightarrow$ Variational Bayes (VB) [Att00]

Variational Bayes

Ideal case: Explicit posterior $\rightarrow$ Conjugate priors

Most of the time: No explicit form for $p(\theta \mid Y)$

- Sample from it, i.e. try to get
$$\{\theta^b\}_{1 \leq b \leq B} \overset{\text{iid}}{\approx} p(\theta \mid Y)$$
$\rightarrow$ Monte-Carlo (MC), MCMC, SMC, HMC, ...

- Approximate it, i.e. look for
$$q(\theta) \simeq p(\theta \mid Y)$$
$\rightarrow$ Variational Bayes (VB) [Att00]

Example. Consider $\mathcal{N} = \{\text{Gaussian distributions}\}$

$$q^*(\theta) = \arg\min_{q \in \mathcal{N}} KL[q(\theta) \mid p(\theta \mid Y)]$$

(or $KL[p(\theta \mid Y) \mid q(\theta)]$)

Including latent variables

Bayesian model with latent variables.

$\theta \sim p(\theta)$	prior distribution
$Z \sim p(Z \mid \theta)$	latent variables
$Y \sim p(Y \mid \theta, Z)$	observed variables

Including latent variables

Bayesian model with latent variables.

$\theta \sim p(\theta)$	prior distribution
$Z \sim p(Z \mid \theta)$	latent variables
$Y \sim p(Y \mid \theta, Z)$	observed variables

Aim of Bayesian inference. Determine the joint conditional distribution

$$p(\theta, Z \mid Y) = \frac{p(\theta)\ p(Z \mid \theta)\ p(Y \mid \theta, Z)}{p(Y)}$$

where

$$p(Y) = \int \int p(\theta)\ p(Z \mid \theta)\ p(Y \mid \theta, Z)\ \mathrm{d}\theta\ \mathrm{d}Z$$

is most often intractable

Variational Bayes EM

Variational approximation of the joint conditional $p(\theta, Z \mid Y)$

$$q(\theta, Z) = \arg\min_{q \in \mathcal{Q}} \; KL[q(\theta, Z) \| p(\theta, Z \mid Y)]$$

taking $\mathcal{Q} = \mathcal{Q}_{\text{fact}} = \{q : q(\theta, Z) = q_\theta(\theta) q_Z(Z)\}$ [Bea03,BG03]

Variational Bayes EM

Variational approximation of the joint conditional $p(\theta, Z \mid Y)$

$$q(\theta, Z) = \arg\min_{q \in \mathcal{Q}} \; KL[q(\theta, Z) \| p(\theta, Z \mid Y)]$$

taking $\mathcal{Q} = \mathcal{Q}_{\text{fact}} = \{q : q(\theta, Z) = q_\theta(\theta) q_Z(Z)\}$ [Bea03,BG03]

Variational Bayes EM (VBEM) algorithm. Makes use of the mean-field approximation

Variational Bayes EM

Variational approximation of the joint conditional $p(\theta, Z \mid Y)$

$$q(\theta, Z) = \arg\min_{q \in \mathcal{Q}} \; KL[q(\theta, Z) \| p(\theta, Z \mid Y)]$$

taking $\mathcal{Q} = \mathcal{Q}_{\text{fact}} = \{q : q(\theta, Z) = q_\theta(\theta) q_Z(Z)\}$ [Bea03,BG03]

Variational Bayes EM (VBEM) algorithm. Makes use of the mean-field approximation

- VBE step = update of the latent variable distribution

$$q_Z^{h+1}(Z) \propto \exp\left(\mathbb{E}_{q_\theta^h} \log p(Y, Z, \theta)\right)$$

Variational Bayes EM

Variational approximation of the joint conditional $p(\theta, Z \mid Y)$

$$q(\theta, Z) = \arg\min_{q \in \mathcal{Q}} KL[q(\theta, Z) \| p(\theta, Z \mid Y)]$$

taking $\mathcal{Q} = \mathcal{Q}_{\text{fact}} = \{q : q(\theta, Z) = q_\theta(\theta) q_Z(Z)\}$ [Bea03,BG03]

Variational Bayes EM (VBEM) algorithm. Makes use of the mean-field approximation

- VBE step = update of the latent variable distribution

$$q_Z^{h+1}(Z) \propto \exp\left(\mathbb{E}_{q_\theta^h} \log p(Y, Z, \theta)\right)$$

- VBM step = update of the parameter distribution

$$q_\theta^{h+1}(\theta) \propto \exp\left(\mathbb{E}_{q_Z^{h+1}} \log p(Y, Z, \theta)\right)$$

VBEM in practice

Exponential family / conjugate prior. If

$p(Y, Z \mid \theta)$ belongs to the exponential family

and $p(\theta)$ is the corresponding conjugate prior

then both the VBE and VBM steps are completely explicit [BG03]

VBEM in practice

Exponential family / conjugate prior. If

$p(Y, Z \mid \theta)$ belongs to the exponential family

and $p(\theta)$ is the corresponding conjugate prior

then both the VBE and VBM steps are completely explicit [BG03]

Many VBEM's.

- Force further factorization among the Z (see e.g. [LBA12,GDR12,KBCG15] for block-models)
- Use further approximations when conjugacy does not hold [JJ00]

Outline

Incomplete data models

Variational EM

Variational Bayes EM

Variational inference

Variational inference

Variational approximations for conditional distributions $p_\theta(Z \mid Y)$ or $p(\theta, Z \mid Y)$

$\rightarrow$ computationally efficient alternative to Monte-Carlo methods

Variational inference

Variational approximations for conditional distributions $p_\theta(Z \mid Y)$ or $p(\theta, Z \mid Y)$

$\rightarrow$ computationally efficient alternative to Monte-Carlo methods

VEM algorithms are similar to EM algorithms

$\rightarrow$ reasonably easy to implement

Variational inference

Variational approximations for conditional distributions $p_\theta(Z \mid Y)$ or $p(\theta, Z \mid Y)$

$\rightarrow$ computationally efficient alternative to Monte-Carlo methods

VEM algorithms are similar to EM algorithms

$\rightarrow$ reasonably easy to implement

Variational inference is a versatile framework for the inference of incomplete data models

$\rightarrow$ see Part 3 for applications in statistical ecology

Variational inference

Variational approximations for conditional distributions $p_\theta(Z \mid Y)$ or $p(\theta, Z \mid Y)$

$\rightarrow$ computationally efficient alternative to Monte-Carlo methods

VEM algorithms are similar to EM algorithms

$\rightarrow$ reasonably easy to implement

Variational inference is a versatile framework for the inference of incomplete data models

$\rightarrow$ see Part 3 for applications in statistical ecology

Statistical guarantees still need to be established for the resulting estimates

$\rightarrow$ see Part 4

References I

H. Attias. A variational Bayesian framework for graphical models. In *Advances in neural information processing systems*, pages 209–215, 2000.

M. Beal. *Variational algorithms for approximate Bayesian inference*. PhD thesis, university of London, 2003.

J. Beal, M. and Z. Ghahramani. The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures. *Bayesian Statistics*, 7:543–52, 2003.

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 39:1–38, 1977.

S. Gazal, J.-J. Daudin, and S. Robin. Accuracy of variational estimates for random graph mixture models. *Journal of Statistical Computation and Simulation*, 82(6):849–862, 2012.

T. S. Jaakkola and M. I. Jordan. Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10(1):25–37, 2000.

C. Keribin, V. Brault, G. Celeux, and G. Govaert. Estimation and selection for the latent block model on categorical data. *Statistics and Computing*, 25(6):1201–1216, 2015.

S. L. Lauritzen. *Graphical Models*. Oxford Statistical Science Series. Clarendon Press, 1996.

P. Latouche, E. Birmelé, and C. Ambroise. Variational Bayesian inference and complexity control for stochastic block models. *Statis. Model.*, 12(1):93–115, 2012.

T. P Minka. Expectation propagation for approximate Bayesian inference. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pages 362–369. Morgan Kaufmann Publishers Inc., 2001.

T. Minka. Divergence measures and message passing. Technical Report MSR-TR-2005-173, Microsoft Research Ltd, 2005.

References II

K. P. Murphy, Y. Weiss, and M. I Jordan. Loopy belief propagation for approximate inference: An empirical study. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 467–475. Morgan Kaufmann Publishers Inc., 1999.

M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.*, 1(1–2):1–305, 2008.

J. S. Yedidia, W. T Freeman, and Y. Weiss. Bethe free energy, kikuchi approximations, and belief propagation algorithms. *Advances in neural information processing systems*, 13, 2001.

EM property

We have to show that

$$\log p_{\theta^{h+1}}(Y) - \log p_{\theta^h}(Y) \geq 0.$$

EM property

We have to show that

$$\log p_{\theta^{h+1}}(Y) - \log p_{\theta^h}(Y) \geq 0.$$

Because $\theta^{h+1} = \arg\max_\theta \mathbb{E}_{\theta^h}[\log p_\theta(Y, Z) \mid Y]$, we have that

Back to #9

EM property

We have to show that

$$\log p_{\theta^{h+1}}(Y) - \log p_{\theta^h}(Y) \geq 0.$$

Because $\theta^{h+1} = \arg\max_\theta \mathbb{E}_{\theta^h}[\log p_\theta(Y, Z) \mid Y]$, we have that

$$0 \leq \mathbb{E}_{\theta^h}[\log p_{\theta^{h+1}}(Y, Z) \mid Y] - \mathbb{E}_{\theta^h}[\log p_{\theta^h}(Y, Z) \mid Y]$$

Back to #9

EM property

We have to show that

$$\log p_{\theta^{h+1}}(Y) - \log p_{\theta^h}(Y) \geq 0.$$

Because $\theta^{h+1} = \arg\max_\theta \mathbb{E}_{\theta^h}[\log p_\theta(Y, Z) \mid Y]$, we have that

$$0 \leq \mathbb{E}_{\theta^h}[\log p_{\theta^{h+1}}(Y, Z) \mid Y] - \mathbb{E}_{\theta^h}[\log p_{\theta^h}(Y, Z) \mid Y]$$

$$= \mathbb{E}_{\theta^h}\left[\log \frac{p_{\theta^{h+1}}(Y, Z)}{p_{\theta^h}(Y, Z)} \mid Y\right]$$

Back to #9

EM property

We have to show that

$$\log p_{\theta^{h+1}}(Y) - \log p_{\theta^h}(Y) \geq 0.$$

Because $\theta^{h+1} = \arg\max_\theta \mathbb{E}_{\theta^h}[\log p_\theta(Y, Z) \mid Y]$, we have that

$$0 \leq \mathbb{E}_{\theta^h}[\log p_{\theta^{h+1}}(Y, Z) \mid Y] - \mathbb{E}_{\theta^h}[\log p_{\theta^h}(Y, Z) \mid Y]$$

$$= \mathbb{E}_{\theta^h}\left[\log \frac{p_{\theta^{h+1}}(Y, Z)}{p_{\theta^h}(Y, Z)} \mid Y\right] \qquad \leq \log\left(\mathbb{E}_{\theta^h}\left[\frac{p_{\theta^{h+1}}(Y, Z)}{p_{\theta^h}(Y, Z)} \mid Y\right]\right) \qquad (\textit{Jensen})$$

Back to #9

EM property

We have to show that

$$\log p_{\theta^{h+1}}(Y) - \log p_{\theta^h}(Y) \geq 0.$$

Because $\theta^{h+1} = \arg\max_\theta \mathbb{E}_{\theta^h}[\log p_\theta(Y, Z) \mid Y]$, we have that

$$0 \leq \mathbb{E}_{\theta^h}[\log p_{\theta^{h+1}}(Y, Z) \mid Y] - \mathbb{E}_{\theta^h}[\log p_{\theta^h}(Y, Z) \mid Y]$$

$$= \mathbb{E}_{\theta^h}\left[\log \frac{p_{\theta^{h+1}}(Y, Z)}{p_{\theta^h}(Y, Z)} \mid Y\right] \qquad \leq \log\left(\mathbb{E}_{\theta^h}\left[\frac{p_{\theta^{h+1}}(Y, Z)}{p_{\theta^h}(Y, Z)} \mid Y\right]\right) \qquad (\textit{Jensen})$$

$$= \log \int \frac{p_{\theta^h}(Y, Z)}{p_{\theta^h}(Y)} \frac{p_{\theta^{h+1}}(Y, Z)}{p_{\theta^h}(Y, Z)} \, \mathrm{d}Z$$

Back to #9

EM property

We have to show that

$$\log p_{\theta^{h+1}}(Y) - \log p_{\theta^h}(Y) \geq 0.$$

Because $\theta^{h+1} = \arg\max_\theta \mathbb{E}_{\theta^h}[\log p_\theta(Y, Z) \mid Y]$, we have that

$$0 \leq \mathbb{E}_{\theta^h}[\log p_{\theta^{h+1}}(Y, Z) \mid Y] - \mathbb{E}_{\theta^h}[\log p_{\theta^h}(Y, Z) \mid Y]$$

$$= \mathbb{E}_{\theta^h}\left[\log \frac{p_{\theta^{h+1}}(Y, Z)}{p_{\theta^h}(Y, Z)} \mid Y\right] \qquad \leq \log\left(\mathbb{E}_{\theta^h}\left[\frac{p_{\theta^{h+1}}(Y, Z)}{p_{\theta^h}(Y, Z)} \mid Y\right]\right) \qquad (\textit{Jensen})$$

$$= \log \int \frac{p_{\theta^h}(Y, Z)}{p_{\theta^h}(Y)} \frac{p_{\theta^{h+1}}(Y, Z)}{p_{\theta^h}(Y, Z)} \, \mathrm{d}Z \qquad = \log\left(\frac{1}{p_{\theta^h}(Y)} \int p_{\theta^{h+1}}(Y, Z) \, \mathrm{d}Z\right)$$

Back to #9

EM property

We have to show that

$$\log p_{\theta^{h+1}}(Y) - \log p_{\theta^h}(Y) \geq 0.$$

Because $\theta^{h+1} = \arg\max_\theta \mathbb{E}_{\theta^h}[\log p_\theta(Y, Z) \mid Y]$, we have that

$$0 \leq \mathbb{E}_{\theta^h}[\log p_{\theta^{h+1}}(Y, Z) \mid Y] - \mathbb{E}_{\theta^h}[\log p_{\theta^h}(Y, Z) \mid Y]$$

$$= \mathbb{E}_{\theta^h}\left[\log \frac{p_{\theta^{h+1}}(Y, Z)}{p_{\theta^h}(Y, Z)} \mid Y\right] \qquad \leq \log\left(\mathbb{E}_{\theta^h}\left[\frac{p_{\theta^{h+1}}(Y, Z)}{p_{\theta^h}(Y, Z)} \mid Y\right]\right) \qquad (\textit{Jensen})$$

$$= \log \int \frac{p_{\theta^h}(Y, Z)}{p_{\theta^h}(Y)} \frac{p_{\theta^{h+1}}(Y, Z)}{p_{\theta^h}(Y, Z)} \, \mathrm{d}Z \qquad = \log\left(\frac{1}{p_{\theta^h}(Y)} \int p_{\theta^{h+1}}(Y, Z) \, \mathrm{d}Z\right)$$

$$= \log \frac{p_{\theta^{h+1}}(Y)}{p_{\theta^h}(Y)}$$

Back to #9

EM property

We have to show that

$$\log p_{\theta^{h+1}}(Y) - \log p_{\theta^h}(Y) \geq 0.$$

Because $\theta^{h+1} = \arg\max_\theta \mathbb{E}_{\theta^h}[\log p_\theta(Y, Z) \mid Y]$, we have that

$$0 \leq \mathbb{E}_{\theta^h}[\log p_{\theta^{h+1}}(Y, Z) \mid Y] - \mathbb{E}_{\theta^h}[\log p_{\theta^h}(Y, Z) \mid Y]$$

$$= \mathbb{E}_{\theta^h}\left[\log \frac{p_{\theta^{h+1}}(Y, Z)}{p_{\theta^h}(Y, Z)} \mid Y\right] \qquad \leq \log\left(\mathbb{E}_{\theta^h}\left[\frac{p_{\theta^{h+1}}(Y, Z)}{p_{\theta^h}(Y, Z)} \mid Y\right]\right) \qquad (\textit{Jensen})$$

$$= \log \int \frac{p_{\theta^h}(Y, Z)}{p_{\theta^h}(Y)} \frac{p_{\theta^{h+1}}(Y, Z)}{p_{\theta^h}(Y, Z)} \, \mathrm{d}Z \qquad = \log\left(\frac{1}{p_{\theta^h}(Y)} \int p_{\theta^{h+1}}(Y, Z) \, \mathrm{d}Z\right)$$

$$= \log \frac{p_{\theta^{h+1}}(Y)}{p_{\theta^h}(Y)} \qquad = \log p_{\theta^{h+1}}(Y) - \log p_{\theta^h}(Y)$$

Back to #9

Two version of the ELBO

$$J_{\theta,q}(Y) = \log p_\theta(Y) - KL\left[q(Z) \| p_\theta(Z \mid Y)\right] \qquad \text{(lower bound)}$$

Two version of the ELBO

$$J_{\theta,q}(Y) = \log p_\theta(Y) - KL\left[q(Z) \| p_\theta(Z \mid Y)\right] \qquad \text{(lower bound)}$$

$$= \log p_\theta(Y) - \mathbb{E}_q \log\left(q(Z)/p_\theta(Z \mid Y)\right)$$

Two version of the ELBO

$$
\begin{aligned}
J_{\theta,q}(Y) &= \log p_\theta(Y) - KL\left[q(Z) \| p_\theta(Z \mid Y)\right] && \text{(lower bound)} \\
&= \log p_\theta(Y) - \mathbb{E}_q \log\left(q(Z)/p_\theta(Z \mid Y)\right) \\
&= \log p_\theta(Y) - \mathbb{E}_q \log\left(\frac{q(Z)p_\theta(Y)}{p_\theta(Y, Z)}\right)
\end{aligned}
$$

Two version of the ELBO

$$
\begin{aligned}
J_{\theta,q}(Y) &= \log p_\theta(Y) - KL\left[q(Z) \| p_\theta(Z \mid Y)\right] \qquad \text{(lower bound)} \\
&= \log p_\theta(Y) - \mathbb{E}_q \log\left(q(Z)/p_\theta(Z \mid Y)\right) \\
&= \log p_\theta(Y) - \mathbb{E}_q \log\left(\frac{q(Z)p_\theta(Y)}{p_\theta(Y,Z)}\right) \\
&= \log p_\theta(Y) - \mathbb{E}_q \log q(Z) - \mathbb{E}_q \log p_\theta(Y) + \mathbb{E}_q \log p_\theta(Y,Z)
\end{aligned}
$$

Two version of the ELBO

$$
\begin{aligned}
J_{\theta,q}(Y) &= \log p_\theta(Y) - KL\left[q(Z) \| p_\theta(Z \mid Y)\right] \qquad \text{(lower bound)} \\
&= \log p_\theta(Y) - \mathbb{E}_q \log\left(q(Z)/p_\theta(Z \mid Y)\right) \\
&= \log p_\theta(Y) - \mathbb{E}_q \log\left(\frac{q(Z)p_\theta(Y)}{p_\theta(Y,Z)}\right) \\
&= \log p_\theta(Y) - \mathbb{E}_q \log q(Z) - \mathbb{E}_q \log p_\theta(Y) + \mathbb{E}_q \log p_\theta(Y,Z) \\
&= \mathbb{E}_q \log p_\theta(Y,Z) \underbrace{- \mathbb{E}_q \log q(Z)}_{\text{entropy } \mathcal{H}(q)}
\end{aligned}
$$

Back to #19

Mean-field approximation

- We know that the function q_1 that minimizes

$$F(q_1) = \int L(z_1, q_1(z_1)) \, \mathrm{d}z_1$$

satisfies (see #35 or [Bea03])

$$\partial q_1(z_1) \; L(z_1, q_1(z_1)) = 0$$

Mean-field approximation

- We know that the function q_1 that minimizes

$$F(q_1) = \int L(z_1, q_1(z_1)) \, \mathrm{d}z_1$$

satisfies (see #35 or [Bea03])

$$\partial q_1(z_1) \; L(z_1, q_1(z_1)) = 0$$

- Let us consider $z = (z_1, z_2)$, $q(z) = q_1(z_1) q_2(z_2)$

Mean-field approximation

- We know that the function q_1 that minimizes

$$F(q_1) = \int L(z_1, q_1(z_1)) \, \mathrm{d}z_1$$

satisfies (see #35 or [Bea03])

$$\partial q_1(z_1) \; L(z_1, q_1(z_1)) = 0$$

- Let us consider $z = (z_1, z_2)$, $q(z) = q_1(z_1) q_2(z_2)$ and define

$$L(z_1, q_1(z_1)) = q_1(z_1) \int q_2(z_2) \log \frac{q_1(z_1) q_2(z_2)}{p(z)} \, \mathrm{d}z_2 \qquad \Rightarrow \qquad F(q_1) = KL[q \| p].$$

Mean-field approximation

- We know that the function q_1 that minimizes

$$F(q_1) = \int L(z_1, q_1(z_1)) \,\mathrm{d}z_1$$

satisfies (see #35 or [Bea03])

$$\partial q_1(z_1)\; L(z_1, q_1(z_1)) = 0$$

- Let us consider $z = (z_1, z_2)$, $q(z) = q_1(z_1)q_2(z_2)$ and define

$$L(z_1, q_1(z_1)) = q_1(z_1) \int q_2(z_2) \log \frac{q_1(z_1)q_2(z_2)}{p(z)} \,\mathrm{d}z_2 \qquad \Rightarrow \qquad F(q_1) = KL[q\|p].$$

- Observe that

$$\partial q_1(z_1)\; L(z_1, q_1(z_1)) = \log q_1(z_1) - \int q_2(z_2) \log p(z) \,\mathrm{d}z_2 + \mathrm{cst}$$

Back to #21

Variational lemma

- Consider

$$F(q) = \int L(z, q(z)) \, \mathrm{d}z$$

Variational lemma

- Consider

$$F(q) = \int L(z, q(z)) \, \mathrm{d}z$$

- q is optimal if, for any function h,

$$\partial_t F(q + th)|_{t=0} = 0$$

Variational lemma

- Consider
$$F(q) = \int L(z, q(z)) \, \mathrm{d}z$$

- q is optimal if, for any function h,
$$\partial_t F(q + th)|_{t=0} = 0$$

- Observe that
$$\partial_t F(q + th) = \int h(z) \, \partial_{q(z)} L(z, q(z)) \, \mathrm{d}z$$

Variational lemma

- Consider
$$F(q) = \int L(z, q(z)) \; \mathrm{d}z$$

- q is optimal if, for any function h,
$$\partial_t F(q + th)|_{t=0} = 0$$

- Observe that
$$\partial_t F(q + th) = \int h(z) \; \partial_{q(z)} L(z, q(z)) \; \mathrm{d}z$$

- This must be zero for any function h, meaning that
$$\partial_{q(z)} L(z, q(z)) \equiv 0.$$

Back to #34