Introduction to Expectation Propagation

Antti Honkela
Helsinki University of Technology
Espoo, Finland
http://www.cis.hut.fi/ahonkela/

Contents

- Approximations and distance measures on distributions
- Limitations of naïve mean field variational Bayes (VB)
- Expectation propagation (EP) and the clutter problem
- Belief networks, loopy belief propagation and EP
- An energy function for EP

Background

- Observations $\mathcal{D}$, model $\mathcal{H}$ with parameters $\boldsymbol{\theta}$
- All information of the parameters is contained in the posterior

$$p(\boldsymbol{\theta}|\mathcal{D},\mathcal{H}) = \frac{p(\mathcal{D}|\boldsymbol{\theta},\mathcal{H})p(\boldsymbol{\theta}|\mathcal{H})}{p(\mathcal{D}|\mathcal{H})},$$

  where $p(\mathcal{D}|\mathcal{H}) = \int_{\boldsymbol{\theta}} p(\mathcal{D}|\boldsymbol{\theta},\mathcal{H})p(\boldsymbol{\theta}|\mathcal{H})d\boldsymbol{\theta}$

- Marginalisation principle:

$$p(\mathbf{x}|\mathcal{D},\mathcal{H}) = \int_{\boldsymbol{\theta}} p(\mathbf{x}|\boldsymbol{\theta},\mathcal{H})p(\boldsymbol{\theta}|\mathcal{D},\mathcal{H})d\boldsymbol{\theta}$$

- How to assess possible approximations $q(\boldsymbol{\theta})$ of the posterior $p(\boldsymbol{\theta}|\mathcal{D},\mathcal{H})$?
- How to approximate $p(\mathcal{D}|\mathcal{H})$?

Bayesian analysis of approximations

- Choosing the best approximation is a decision problem
- Bayesian method: specify utility, maximise expected utility
- For approximations $q(\boldsymbol{\theta}) \in \mathcal{Q}$ and "true parameter values" $\boldsymbol{\theta} \in \Omega$, define a score function $u : \mathcal{Q} \times \Omega \rightarrow \mathbb{R}$
- Expected utility

$$\bar{u}(q) = \int u(q, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta}'$$

Properties of score functions

- The score function is proper, if

$$\sup \bar{u}(q) = \bar{u}(p(\boldsymbol{\theta}|\mathcal{D}))$$

  which is attained only if $q(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\mathcal{D})$

- The score function is local, if

$$u(q, \boldsymbol{\theta}) = u_{\boldsymbol{\theta}}(q(\boldsymbol{\theta}))$$

Score functions

Example. The quadratic score function

$$u(q, \boldsymbol{\theta}) = A\left[2q(\boldsymbol{\theta}) - \int q(\boldsymbol{\theta}')^2 d\boldsymbol{\theta}'\right] + B(\boldsymbol{\theta})$$

corresponding to the expected utility

$$\bar{u}(q) = -\int \left(q(\boldsymbol{\theta}) - p(\boldsymbol{\theta}|\mathcal{D})\right)^2 d\boldsymbol{\theta}$$

is a proper, non-local score function

Bayesian analysis of approximations

Proposition. Smooth, proper, local score functions are of the form

$$u(q, \boldsymbol{\theta}) = A \log q(\boldsymbol{\theta}) + B(\boldsymbol{\theta}),$$

where $A > 0$ and $B(\boldsymbol{\theta})$ are arbitrary.

Proof. We maximise the expected utility

$$\bar{u}(q) = \int u_{\boldsymbol{\theta}}(q(\boldsymbol{\theta}))p(\boldsymbol{\theta}|\mathcal{D})d\boldsymbol{\theta}$$

subject to constraint $\int q(\boldsymbol{\theta})d\boldsymbol{\theta} = 1$. This is done by finding an extremum of

$$F(q(\cdot)) = \bar{u}(q) - A\left[\int q(\boldsymbol{\theta})d\boldsymbol{\theta} - 1\right].$$

Proof contd.

A necessary condition for this follows from the variational principle

$$\frac{\partial}{\partial \alpha} F(q(\cdot) + \alpha \tau(\cdot))\Big|_{\alpha=0} = 0$$

for any function $\tau : \Omega \to \mathbb{R}$. this implies a differential equation

$$u'(q(\boldsymbol{\theta}))p(\boldsymbol{\theta}|\mathcal{D}) - A = 0,$$

which should hold for $q(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\mathcal{D})$. The solutions of this are

$$u(q, \boldsymbol{\theta}) = A \log q(\boldsymbol{\theta}) + B(\boldsymbol{\theta}).$$

Bayesian analysis of approximations

Theorem. Differences of expected utilities under smooth, proper, local score functions are given by the (scaled) Kullback–Leibler (KL) divergence

$$A \cdot D_{KL}(p(\boldsymbol{\theta}|\mathcal{D}) \mid\mid q(\boldsymbol{\theta})) = A \int p(\boldsymbol{\theta}|\mathcal{D}) \log \frac{p(\boldsymbol{\theta}|\mathcal{D})}{q(\boldsymbol{\theta})} d\boldsymbol{\theta}.$$

Proof. Evaluate $\bar{u}(p(\boldsymbol{\theta}|\mathcal{D})) - \bar{u}(q(\boldsymbol{\theta}))$.

Properties of KL divergence

- In information theory, the KL divergence

$$D_{KL}(p(\boldsymbol{\theta}|\mathcal{D}) \mid\mid q(\boldsymbol{\theta})) = \int p(\boldsymbol{\theta}|\mathcal{D}) \log \frac{p(\boldsymbol{\theta}|\mathcal{D})}{q(\boldsymbol{\theta})} d\boldsymbol{\theta}$$

  measures the overhead when using distribution q to code events following p

- The choice of A reflects the choice of unit of measure, essentially the base of the logarithm

- Natural logarithm $\ln$ yields nats, while $\log_2$ gives bits

Exponential families

Definition A set of distributions with densities

$$p(\boldsymbol{\theta}|\boldsymbol{\xi}) = \frac{1}{Z(\boldsymbol{\xi})} \exp(\boldsymbol{\xi}^T \phi(\boldsymbol{\theta}))$$

is an exponential family with natural parameters $\boldsymbol{\xi}$, sufficient statistics $\phi(\boldsymbol{\theta})$ and partition function $Z(\boldsymbol{\xi})$.

Examples: Gaussian, gamma, multinomial, Dirichlet, . . .

Theorem For exponential families,

$$\nabla_{\boldsymbol{\xi}} \log Z(\boldsymbol{\xi}) = \langle \phi(\boldsymbol{\theta}) \rangle.$$

Properties of the KL divergence

Theorem. Given an approximation in an exponential family

$$q(\boldsymbol{\theta}|\boldsymbol{\xi}) = \frac{1}{Z(\boldsymbol{\xi})} \exp\left(\boldsymbol{\xi}^T \phi(\boldsymbol{\theta})\right),$$

the KL divergence $D_{KL}(p(\boldsymbol{\theta}|\mathcal{D}) \mid\mid q(\boldsymbol{\theta}|\boldsymbol{\xi}))$ is minimized when

$$\langle \phi(\boldsymbol{\theta}) \rangle_{p(\boldsymbol{\theta}|\mathcal{D})} = \langle \phi(\boldsymbol{\theta}) \rangle_{q(\boldsymbol{\theta}|\boldsymbol{\xi})}.$$

Proof. Consider

$$f(\boldsymbol{\xi}) = D_{KL}(p(\boldsymbol{\theta}|\mathcal{D}) \;||\; q(\boldsymbol{\theta}|\boldsymbol{\xi})) = \langle \log p \rangle_p + \langle \log Z(\boldsymbol{\xi}) \rangle_p - \langle \boldsymbol{\xi}^T \phi(\boldsymbol{\theta}) \rangle_p$$
$$= \langle \log p \rangle_p + \log Z(\boldsymbol{\xi}) - \boldsymbol{\xi}^T \langle \phi(\boldsymbol{\theta}) \rangle_p.$$

Zeroing the gradient yields the desired condition, because for exponential families

$$\nabla_{\boldsymbol{\xi}} \log Z(\boldsymbol{\xi}) = \langle \phi(\boldsymbol{\theta}) \rangle.$$

The minimality of the extremum can be checked using the second derivatives.

Properties of the KL divergence

- In VB, the reverse of KL divergence is used:

$$D_{KL}(q(\boldsymbol{\theta}) \mid\mid p(\boldsymbol{\theta}|\mathcal{D})) = \int q(\boldsymbol{\theta}) \log \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\mathcal{D})} d\boldsymbol{\theta}.$$

- Having large $q(\boldsymbol{\theta})$ with very small $p(\boldsymbol{\theta}|\mathcal{D})$ causes large values of the divergence
- Hence the VB approximation will be contained in the true distribution

Limitations of naïve mean field variational Bayes

- The marginal likelihoods and especially rankings evaluated by VB are often quite reliable
- The estimates of the marginals may not be as good, variances can be underestimated
- Sometimes a simpler mode of solution may be preferred because of inadequate approximation

Analysis of variational Bayesian ICA (A. Ilin & H. Valpola)

- Consider the ICA model

$$\mathbf{x} = \mathbf{A}\mathbf{s} + \mathbf{n}$$

- Gaussian noise $\mathbf{n} \sim \mathcal{N}(0, \Sigma_{\mathbf{x}})$
- Non-Gaussian source prior $p(\mathbf{s}) = \prod_i p(s_i)$
- These yield non-diagonal posterior covariance for $\mathbf{s}$:

$$\Sigma_{\mathbf{s}|\mathcal{D}}^{-1} \propto \Sigma_{\mathbf{s}}^{-1} + \mathbf{A}^T \Sigma_{\mathbf{x}}^{-1} \mathbf{A}$$

Limitations of variational Bayes

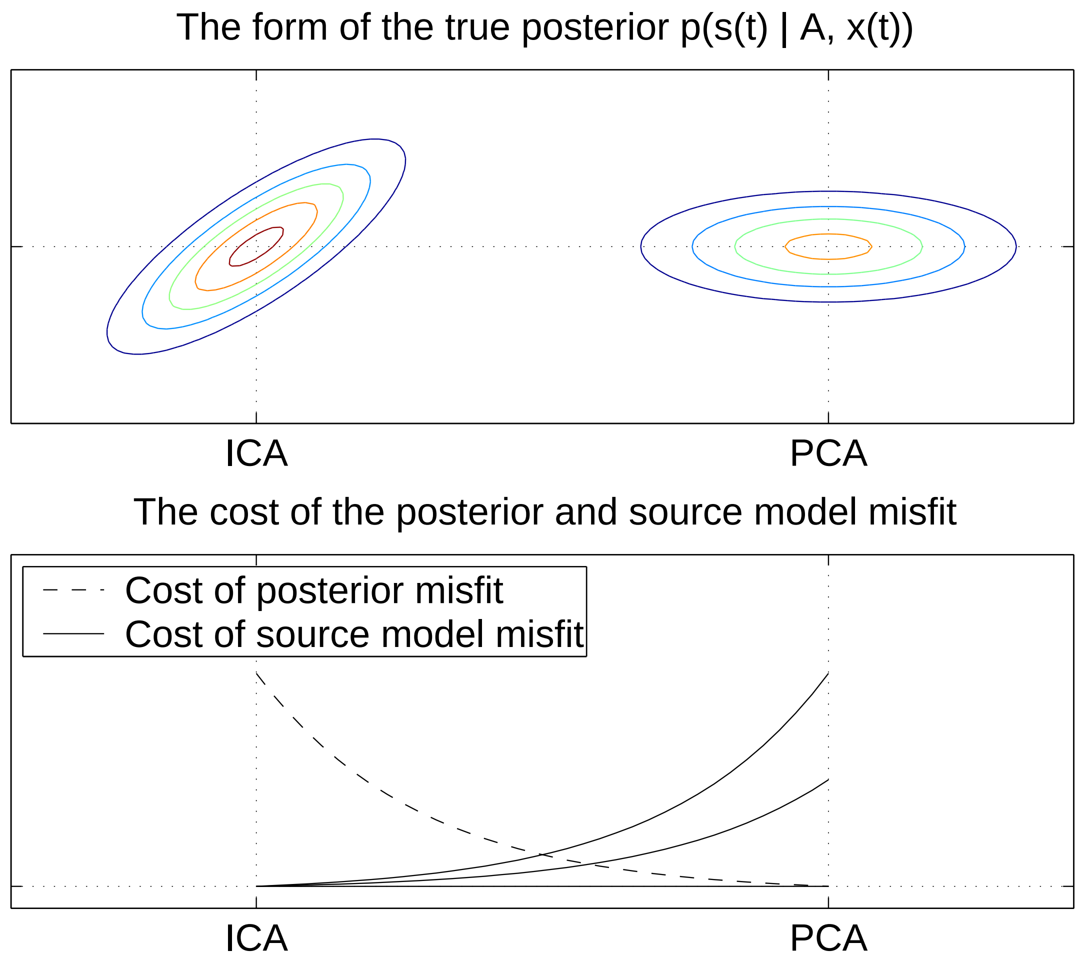
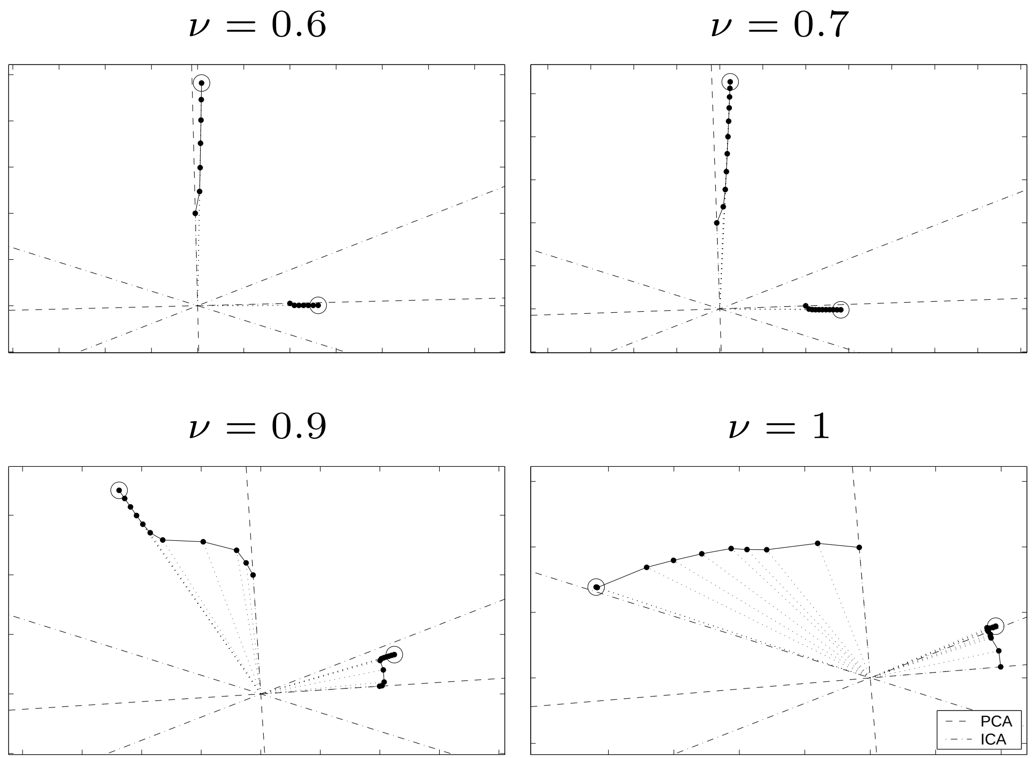


Illustration of the trade-offs between the ICA and PCA solutions.

Limitations of variational Bayes



VB solutions to ICA problem as a function of non-Gaussianity of the sources

Expectation propagation

- An approximate inference method proposed by Thomas Minka in 2001
- Suitable for approximating product forms

$$\prod_{i=0}^{N} t_i(\boldsymbol{\theta}) \approx \prod_{i=0}^{N} \tilde{t}_i(\boldsymbol{\theta})$$

- Iterative refinement of the terms $\tilde{t}_i(\boldsymbol{\theta})$

Expectation propagation

- The parameter posterior is

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{1}{p(\mathcal{D})} p(\boldsymbol{\theta}) \prod_{i=1}^{N} p(\mathbf{x}_i|\boldsymbol{\theta})$$

- As a function of $\boldsymbol{\theta}$, this can be written as

$$p(\boldsymbol{\theta}) \prod_{i=1}^{N} p(\mathbf{x}_i|\boldsymbol{\theta}) = \prod_{i=0}^{N} t_i(\boldsymbol{\theta})$$

where $t_0(\boldsymbol{\theta}) = p(\boldsymbol{\theta})$ and $t_i(\boldsymbol{\theta}) = p(\mathbf{x}_i|\boldsymbol{\theta})$

- Now approximate each term separately to get

$$q(\boldsymbol{\theta}) = \prod_{i=0}^{N} \tilde{t}_i(\boldsymbol{\theta})$$

- Fit the approximation by finding

$$\min_{\tilde{t}_i(\boldsymbol{\theta})} D_{KL}(t_i(\boldsymbol{\theta}) \prod_{j \neq i} \tilde{t}_j(\boldsymbol{\theta}) \mid\mid \tilde{t}_i(\boldsymbol{\theta}) \prod_{j \neq i} \tilde{t}_j(\boldsymbol{\theta}))$$

Expectation propagation algorithm

Input $t_0(\boldsymbol{\theta}), \ldots, t_N(\boldsymbol{\theta})$

Initialise $\tilde{t}_0(\boldsymbol{\theta}) = t_0(\boldsymbol{\theta}), \tilde{t}_i(\boldsymbol{\theta}) = 1$ for $i > 0$, $q(\boldsymbol{\theta}) = \prod_{i=0}^{N} \tilde{t}_i(\boldsymbol{\theta})$

repeat

    for $i = 0, \ldots, N$ **do**

        Deletion: $q_{\setminus i}(\boldsymbol{\theta}) \propto \frac{q(\boldsymbol{\theta})}{\tilde{t}_i(\boldsymbol{\theta})} = \prod_{j \neq i} \tilde{t}_j(\boldsymbol{\theta})$

        Projection: $\tilde{t}_i^{\text{new}}(\boldsymbol{\theta}) \leftarrow \arg\min_{\tilde{t}_i(\boldsymbol{\theta})} D_{KL}(t_i(\boldsymbol{\theta}) q_{\setminus i}(\boldsymbol{\theta}) \,||\, \tilde{t}_i(\boldsymbol{\theta}) q_{\setminus i}(\boldsymbol{\theta}))$

        Inclusion: $q(\boldsymbol{\theta}) \leftarrow \tilde{t}_i^{\text{new}}(\boldsymbol{\theta}) q_{\setminus i}(\boldsymbol{\theta})$

    end for

until convergence

Expectation propagation algorithm (2)

Input $t_0(\boldsymbol{\theta}), \ldots, t_N(\boldsymbol{\theta})$

Initialise $\tilde{t}_0(\boldsymbol{\theta}) = t_0(\boldsymbol{\theta}), \tilde{t}_i(\boldsymbol{\theta}) = 1$ for $i > 0$, $q(\boldsymbol{\theta}) = \prod_{i=0}^{N} \tilde{t}_i(\boldsymbol{\theta})$

repeat

for $i = 0, \ldots, N$ **do**

Deletion: $q_{\setminus i}(\boldsymbol{\theta}) \propto \frac{q(\boldsymbol{\theta})}{\tilde{t}_i(\boldsymbol{\theta})} = \prod_{j \neq i} \tilde{t}_j(\boldsymbol{\theta})$

Inclusion: $q(\boldsymbol{\theta}) \leftarrow \arg\min_{q(\boldsymbol{\theta})} D_{KL}(t_i(\boldsymbol{\theta}) q_{\setminus i}(\boldsymbol{\theta}) \,||\, q(\boldsymbol{\theta}))$

Update: $\tilde{t}_i^{\text{new}}(\boldsymbol{\theta}) \leftarrow \frac{q(\boldsymbol{\theta})}{q_{\setminus i}(\boldsymbol{\theta})}$

end for

until convergence

The clutter problem

Consider a simple Gaussian mixture for $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^N$

$$p(\mathbf{x}|\boldsymbol{\theta}) = w\mathcal{N}(\mathbf{x};\ \boldsymbol{\theta}, \mathbf{I}) + (1-w)\mathcal{N}(\mathbf{x};\ \mathbf{0}, 10\mathbf{I})$$
$$p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta};\ \mathbf{0}, 100\mathbf{I}).$$

A suitable exponential family for this is formed by

$$\mathcal{N}(\mathbf{x};\ \mathbf{m}, v\mathbf{I}) = \mathcal{N}(\mathbf{x};\ \boldsymbol{\xi})$$

with sufficient statistics $\phi(\mathbf{x}) = (\mathbf{x}, \mathbf{x}^T\mathbf{x})$, natural parameters $\boldsymbol{\xi} = (v^{-1}\mathbf{m}, -\frac{1}{2}v^{-1})$ and normalisation $Z(\boldsymbol{\xi}) = (2\pi v)^{d/2}\exp(\frac{1}{2v}\mathbf{m}^T\mathbf{m})$.

Expectation propagation algorithm

Input $t_0(\boldsymbol{\theta}), \ldots, t_N(\boldsymbol{\theta})$

Initialise $\tilde{t}_0(\boldsymbol{\theta}) = t_0(\boldsymbol{\theta}), \tilde{t}_i(\boldsymbol{\theta}) = 1$ for $i > 0$, $q(\boldsymbol{\theta}) = \prod_{i=0}^{N} \tilde{t}_i(\boldsymbol{\theta})$

repeat

  for $i = 0, \ldots, N$ **do**

    Deletion: $q_{\setminus i}(\boldsymbol{\theta}) \propto \frac{q(\boldsymbol{\theta})}{\tilde{t}_i(\boldsymbol{\theta})} = \prod_{j \neq i} \tilde{t}_j(\boldsymbol{\theta})$

    Inclusion: $q(\boldsymbol{\theta}) \leftarrow \arg\min_{q(\boldsymbol{\theta})} D_{KL}(t_i(\boldsymbol{\theta}) q_{\setminus i}(\boldsymbol{\theta}) \,||\, q(\boldsymbol{\theta}))$

    Update: $\tilde{t}_i^{\text{new}}(\boldsymbol{\theta}) \leftarrow \frac{q(\boldsymbol{\theta})}{q_{\setminus i}(\boldsymbol{\theta})}$

  end for

until convergence

EP for the clutter problem (1): Initialisation

For the clutter problem, we have

$$
\begin{aligned}
t_0(\boldsymbol{\theta}) &= p(\boldsymbol{\theta}) \\
t_i(\boldsymbol{\theta}) &= p(\mathbf{x}_i|\boldsymbol{\theta}), \quad i = 1, \ldots, N.
\end{aligned}
$$

The approximation is of the form

$$
\begin{aligned}
\tilde{t}_0(\boldsymbol{\theta}) &= t_0(\boldsymbol{\theta}) = p(\boldsymbol{\theta}) \\
\tilde{t}_i(\boldsymbol{\theta}) &= s_i \exp(\boldsymbol{\xi}_i^T \phi(\boldsymbol{\theta})), \quad i = 1, \ldots, N, \\
q(\boldsymbol{\theta}) &= \prod_{i=0}^{N} \tilde{t}_i(\boldsymbol{\theta}) = s\mathcal{N}(\boldsymbol{\theta};\, \boldsymbol{\xi})
\end{aligned}
$$

Now initialise $\boldsymbol{\xi}_i = \mathbf{0}$ for $i = 1, \ldots, N$.

Expectation propagation algorithm

Input $t_0(\boldsymbol{\theta}), \ldots, t_N(\boldsymbol{\theta})$

Initialise $\tilde{t}_0(\boldsymbol{\theta}) = t_0(\boldsymbol{\theta}), \tilde{t}_i(\boldsymbol{\theta}) = 1$ for $i > 0$, $q(\boldsymbol{\theta}) = \prod_{i=0}^{N} \tilde{t}_i(\boldsymbol{\theta})$

repeat

for $i = 0, \ldots, N$ **do**

Deletion: $q_{\backslash i}(\boldsymbol{\theta}) \propto \frac{q(\boldsymbol{\theta})}{\tilde{t}_i(\boldsymbol{\theta})} = \prod_{j \neq i} \tilde{t}_j(\boldsymbol{\theta})$

Inclusion: $q(\boldsymbol{\theta}) \leftarrow \underset{q(\boldsymbol{\theta})}{\arg\min}\, D_{KL}(t_i(\boldsymbol{\theta}) q_{\backslash i}(\boldsymbol{\theta}) \mid\mid q(\boldsymbol{\theta}))$

Update: $\tilde{t}_i^{\text{new}}(\boldsymbol{\theta}) \leftarrow \frac{q(\boldsymbol{\theta})}{q_{\backslash i}(\boldsymbol{\theta})}$

end for

until convergence

EP for the clutter problem (2): Deletion

When working with natural parameters, the deletion operation

$$q_{\setminus i}(\boldsymbol{\theta}) \propto \frac{q(\boldsymbol{\theta})}{\tilde{t}_i(\boldsymbol{\theta})}$$

is trivial to implement with

$$\boldsymbol{\xi}_{\setminus i} = \boldsymbol{\xi} - \boldsymbol{\xi}_i.$$

Expectation propagation algorithm

Input $t_0(\boldsymbol{\theta}), \ldots, t_N(\boldsymbol{\theta})$

Initialise $\tilde{t}_0(\boldsymbol{\theta}) = t_0(\boldsymbol{\theta}), \tilde{t}_i(\boldsymbol{\theta}) = 1$ for $i > 0$, $q(\boldsymbol{\theta}) = \prod_{i=0}^{N} \tilde{t}_i(\boldsymbol{\theta})$

repeat

    for $i = 0, \ldots, N$ **do**

        Deletion: $q_{\setminus i}(\boldsymbol{\theta}) \propto \frac{q(\boldsymbol{\theta})}{\tilde{t}_i(\boldsymbol{\theta})} = \prod_{j \neq i} \tilde{t}_j(\boldsymbol{\theta})$

        Inclusion: $q(\boldsymbol{\theta}) \leftarrow \arg\min_{q(\boldsymbol{\theta})} D_{KL}(t_i(\boldsymbol{\theta}) q_{\setminus i}(\boldsymbol{\theta}) \,||\, q(\boldsymbol{\theta}))$

        Update: $\tilde{t}_i^{\text{new}}(\boldsymbol{\theta}) \leftarrow \frac{q(\boldsymbol{\theta})}{q_{\setminus i}(\boldsymbol{\theta})}$

    end for

until convergence

EP for the clutter problem (3): Inclusion

The inclusion operation:

$$q(\boldsymbol{\theta}) \leftarrow \arg\min_{q(\boldsymbol{\theta})} D_{KL}(t_i(\boldsymbol{\theta}) q_{\setminus i}(\boldsymbol{\theta}) \;||\; q(\boldsymbol{\theta}))$$

requires matching sufficient statistics of

$$\begin{aligned} t_i(\boldsymbol{\theta}) q_{\setminus i}(\boldsymbol{\theta}) &= (w\mathcal{N}(\mathbf{x}_i;\ \boldsymbol{\theta}, \mathbf{I}) + (1-w)\mathcal{N}(\mathbf{x}_i;\ \mathbf{0}, 10\mathbf{I}))\,\mathcal{N}(\boldsymbol{\theta};\ \boldsymbol{\xi}_{\setminus i}) \\ &= \left(w\mathcal{N}\left(\boldsymbol{\theta};\ \left(\mathbf{x}_i, -\frac{1}{2}\right)\right) + (1-w)\mathcal{N}(\mathbf{x}_i;\ \mathbf{0}, 10\mathbf{I}) \right) \mathcal{N}(\boldsymbol{\theta};\ \boldsymbol{\xi}_{\setminus i}) \\ &= w \frac{Z(\boldsymbol{\xi}^+)}{Z\left(\left(\mathbf{x}_i, -\frac{1}{2}\right)\right) Z(\boldsymbol{\xi}_{\setminus i})} \mathcal{N}(\boldsymbol{\theta};\ \boldsymbol{\xi}^+) + (1-w)\mathcal{N}(\mathbf{x}_i;\ \mathbf{0}, 10\mathbf{I})\mathcal{N}(\boldsymbol{\theta};\ \boldsymbol{\xi}_{\setminus i}) \\ &\propto r\mathcal{N}(\boldsymbol{\theta};\ \boldsymbol{\xi}^+) + (1-r)\mathcal{N}(\boldsymbol{\theta};\ \boldsymbol{\xi}_{\setminus i}), \end{aligned}$$

where $\boldsymbol{\xi}^+ = \boldsymbol{\xi}_{\setminus i} + \left(\mathbf{x}_i, -\frac{1}{2}\right)$

EP for the clutter problem (3): Inclusion (cont.)

We wish to match the sufficient statistics of the Gaussian mixture

$$t_i(\boldsymbol{\theta})q_{\setminus i}(\boldsymbol{\theta}) \propto r\mathcal{N}(\boldsymbol{\theta};\ \boldsymbol{\xi}^{+}) + (1-r)\mathcal{N}(\boldsymbol{\theta};\ \boldsymbol{\xi}_{\setminus i}).$$

These are simply

$$\mathbf{m} = r\mathbf{m}^{+} + (1-r)\mathbf{m}_{\setminus i}$$

$$v + \mathbf{m}^T\mathbf{m} = r\left(v^{+} + (\mathbf{m}^{+})^T\mathbf{m}^{+}\right) + (1-r)\left(v_{\setminus i} + \mathbf{m}_{\setminus i}^T\mathbf{m}_{\setminus i}\right)$$

Expectation propagation algorithm

Input $t_0(\boldsymbol{\theta}), \ldots, t_N(\boldsymbol{\theta})$

Initialise $\tilde{t}_0(\boldsymbol{\theta}) = t_0(\boldsymbol{\theta}), \tilde{t}_i(\boldsymbol{\theta}) = 1$ for $i > 0$, $q(\boldsymbol{\theta}) = \prod_{i=0}^{N} \tilde{t}_i(\boldsymbol{\theta})$

repeat

  for $i = 0, \ldots, N$ **do**

    Deletion: $q_{\setminus i}(\boldsymbol{\theta}) \propto \frac{q(\boldsymbol{\theta})}{\tilde{t}_i(\boldsymbol{\theta})} = \prod_{j \neq i} \tilde{t}_j(\boldsymbol{\theta})$

    Inclusion: $q(\boldsymbol{\theta}) \leftarrow \arg\min_{q(\boldsymbol{\theta})} D_{KL}(t_i(\boldsymbol{\theta}) q_{\setminus i}(\boldsymbol{\theta}) \mid\mid q(\boldsymbol{\theta}))$

    Update: $\tilde{t}_i^{\text{new}}(\boldsymbol{\theta}) \leftarrow \frac{q(\boldsymbol{\theta})}{q_{\setminus i}(\boldsymbol{\theta})}$

  end for

until convergence

EP for the clutter problem (4): Update

When working with natural parameters, the update operation

$$\tilde{t}_i^{\text{new}}(\boldsymbol{\theta}) \leftarrow \frac{q(\boldsymbol{\theta})}{q_{\setminus i}(\boldsymbol{\theta})}$$

is again trivial with

$$\boldsymbol{\xi}_i = \boldsymbol{\xi} - \boldsymbol{\xi}_{\setminus i}.$$

Marginal likelihood by EP

- The EP algorithm may be extended to evaluate the marginal likelihood $p(\mathcal{D}|\mathcal{H})$
- To do this, we include a scale on $\tilde{t}_i(\boldsymbol{\theta})$ and through them for $q(\boldsymbol{\theta})$:

$$\tilde{t}_i(\boldsymbol{\theta}) = Z_i \frac{q^*(\boldsymbol{\theta})}{q_{\setminus i}(\boldsymbol{\theta})},$$

  where $q^*(\boldsymbol{\theta})$ is a normalised version of $q(\boldsymbol{\theta})$ and $Z_i = \int q_{\setminus i}(\boldsymbol{\theta}) t_i(\boldsymbol{\theta}) d\boldsymbol{\theta}$

- Finally we approximate

$$p(\mathcal{D}|\mathcal{H}) \approx \int q(\boldsymbol{\theta}) d\boldsymbol{\theta} = \int \prod_i \tilde{t}_i(\boldsymbol{\theta}) d\boldsymbol{\theta}$$

Marginal likelihood for the clutter problem

For the clutter problem

$$\tilde{t}_i(\boldsymbol{\theta}) = Z_i \frac{q^*(\boldsymbol{\theta})}{q_{\setminus i}(\boldsymbol{\theta})}$$

implies

$$s_i = Z_i \frac{Z(\boldsymbol{\xi}_{\setminus i})}{Z(\boldsymbol{\xi})}$$

$$Z_i = w \frac{Z(\boldsymbol{\xi}^+)}{Z\left(\left(\mathbf{x}_i, -\frac{1}{2}\right)\right) Z(\boldsymbol{\xi}_{\setminus i})} + (1-w)\mathcal{N}(\mathbf{x}_i;\ \mathbf{0}, 10\mathbf{I}).$$

And globally

$$p(\mathcal{D}|\mathcal{H}) \approx \int \prod_i \tilde{t}_i(\boldsymbol{\theta}) d\boldsymbol{\theta} = \frac{Z(\boldsymbol{\xi})}{Z(\boldsymbol{\xi}_0)} \prod_{i=1}^{N} s_i$$

EP for belief networks

- A probabilistic model may be represented as a directed graph corresponding to a factorisation of the joint distribution

$$p(\mathbf{x}) = \prod_{x_i \in \mathbf{x}} p(x_i | \text{parents}(x_i))$$

- Derive an EP algorithm using the term factorisation

$$t_i(\mathbf{x}) = p(x_i | \text{parents}(x_i))$$

  and a factorial posterior approximation

$$q(\mathbf{x}) = \prod_k q_k(x_k)$$

- For each term $t_i(\mathbf{x})$ the factorisation implies a factorial approximation

$$\tilde{t}_i(\mathbf{x}) = \prod_{k \in \{i, \mathrm{pa}(i)\}} \tilde{t}_{ik}(x_k)$$

- Equivalently, for each factor $q_k(x_k)$, this corresponds to a regular EP approximation

$$q_k(x_k) = \prod_{i \in \{i, \mathrm{ch}(i)\}} \tilde{t}_{ik}(x_k),$$

EP for belief networks

Input $t_1(\mathbf{x}), \dots, t_N(\mathbf{x})$

Initialise $\tilde{t}_{ik}(x_k) = 1$, $q_k(x_k) = \prod_i \tilde{t}_{ik}(x_k)$

repeat

  for $i = 1, \dots, N$ **do**

    for all k **do**

      Deletion: $q_{\backslash i,k}(x_k) \propto \frac{q_k(x_k)}{\tilde{t}_{ik}(x_k)} = \prod_{j \neq i} \tilde{t}_{jk}(x_k)$

    end for

    for all k **do**

      Projection: $\tilde{t}_{ik}^{\text{new}}(x_k) \leftarrow \sum_{\mathbf{x} \backslash x_k} t_i(\mathbf{x}) \prod_{j \neq k} q_{\backslash i,j}(x_j)$

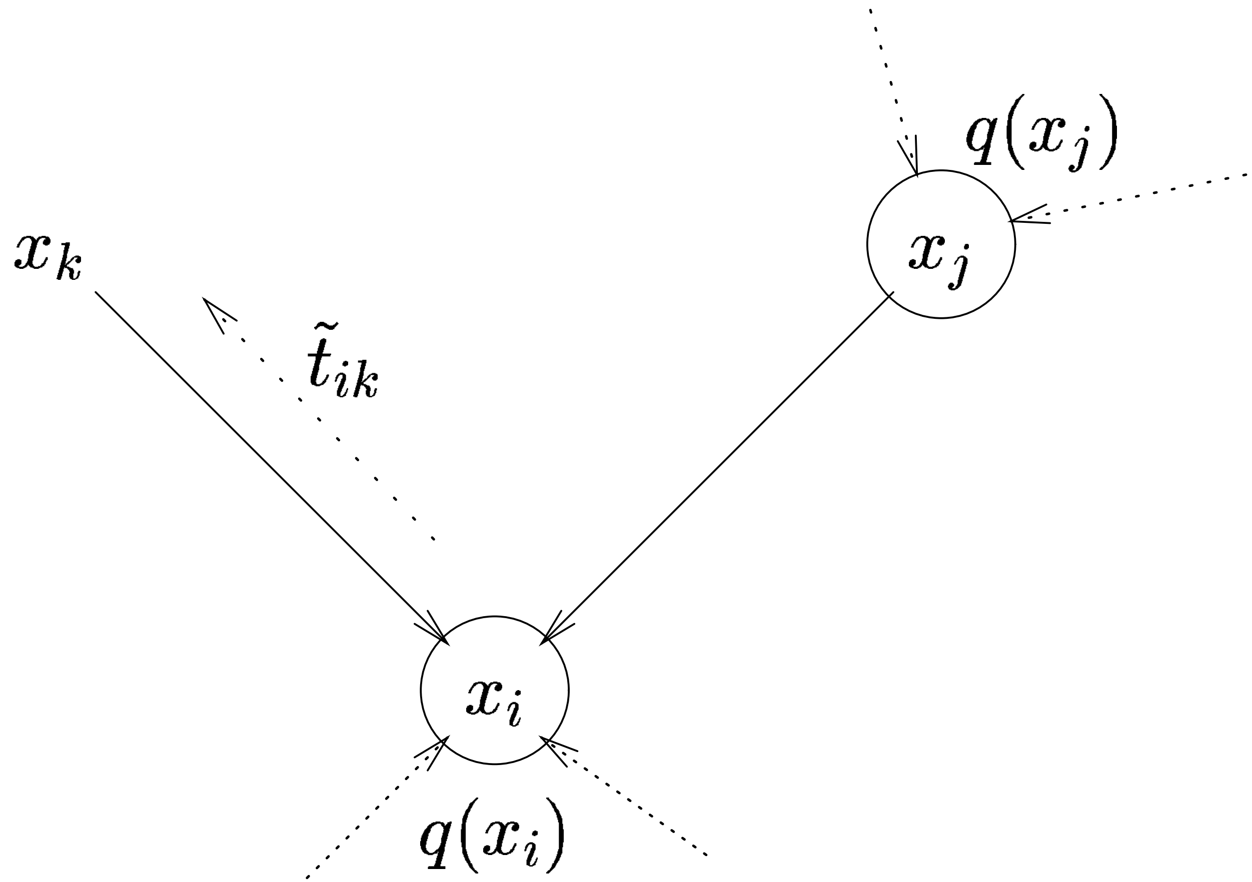
      Inclusion: $q_k(x_k) \leftarrow \tilde{t}_{ik}^{\text{new}}(x_k) q_{\backslash i,k}(x_k)$

    end for

  end for

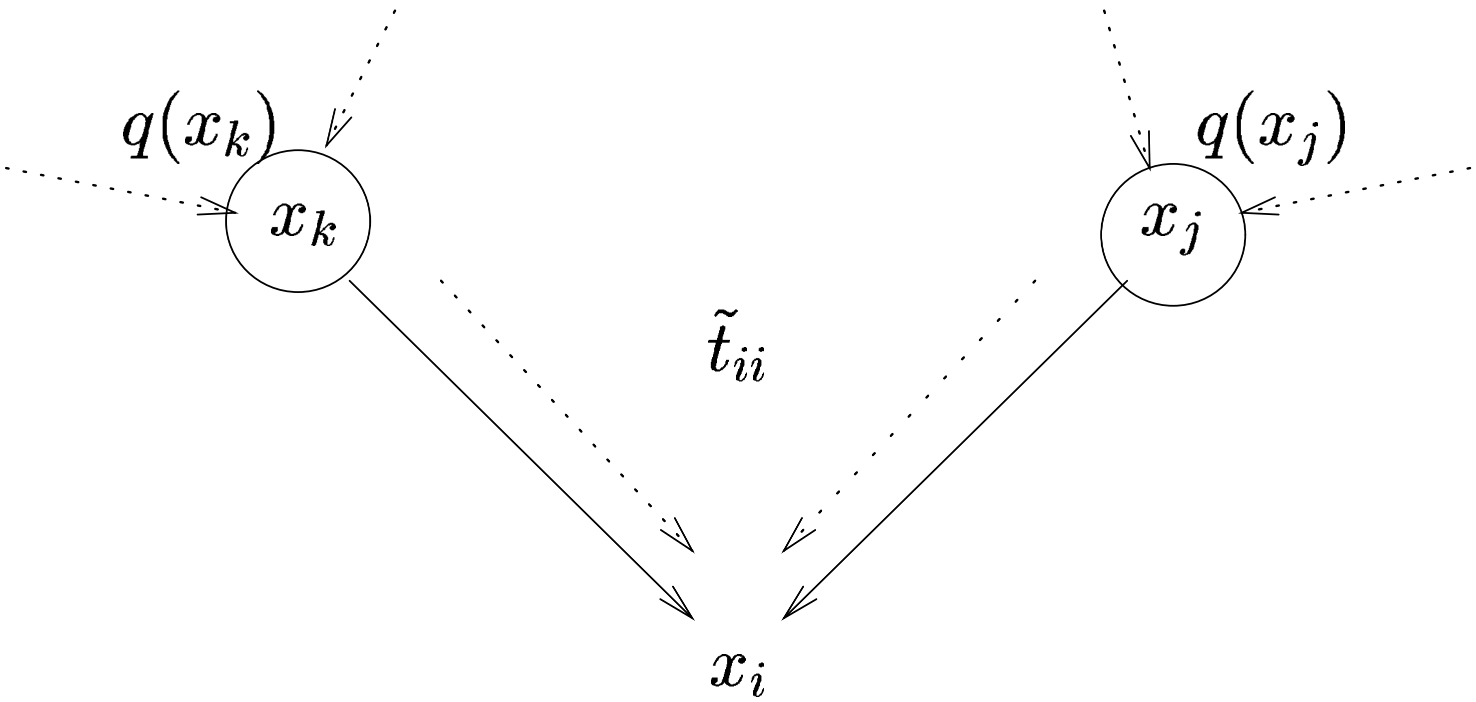
until convergence

EP for belief networks (T. Minka)



$$\tilde{t}_{ik}(x_k) = \sum_{x_i,x_j} p(x_i|x_k, x_j)q_i(x_i)q_j(x_j)$$

EP for belief networks (T. Minka)



$$\tilde{t}_{ii}(x_i) = \sum_{x_k, x_j} p(x_i | x_k, x_j) q_k(x_k) q_j(x_j)$$

EP for belief networks

- The presented EP algorithm is equivalent to a well-known method called (loopy) belief propagation
- For tree structured graphs, it converges in one pass to yield correct marginals
- For general graphs there are no guarantees and it may even diverge

EP for belief networks

- The EP formulation allows simple generalisation to more accurate approximations
- Use fewer more complicated terms $t_i(\mathbf{x})$
- Factorisation $q(\mathbf{x}) = \prod_k q_k(x_k)$ over nodes can still be assumed to only evaluate the marginals

An energy function for EP

- Assume an approximation in an exponential family $\exp(\boldsymbol{\lambda}^T \phi(\boldsymbol{\theta}))$
- With an exact prior,

$$q(\boldsymbol{\theta}) \propto p(\boldsymbol{\theta}) \exp(\boldsymbol{\nu}^T \phi(\boldsymbol{\theta}))$$

  and

$$q_{\backslash i}(\boldsymbol{\theta}) = p(\boldsymbol{\theta}) \exp(\boldsymbol{\lambda}_i^T \phi(\boldsymbol{\theta}))$$

- Let N be the number of terms $t_i(\boldsymbol{\theta})$

- Now, EP fixed points correspond to stationary points of the objective

$$\min_{\nu} \max_{\lambda} \ (N-1) \log \int p(\boldsymbol{\theta}) \exp(\boldsymbol{\nu}^T \phi(\boldsymbol{\theta})) d\boldsymbol{\theta} - \sum_{i=1}^{N} \log \int t_i(\boldsymbol{\theta}) p(\boldsymbol{\theta}) \exp(\boldsymbol{\lambda}_i^T \phi(\boldsymbol{\theta})) d\boldsymbol{\theta}$$

  such that $(N-1)\nu_j = \sum_i \lambda_{ij}$.

- Note: non-convex optimisation problem
- Also other formulations for the energy function

Summary

- Kullback–Leibler divergence $D_{KL}(p(\boldsymbol{\theta}|\mathcal{D}) \,||\, q(\boldsymbol{\theta}))$ is a reasonable measure of goodness of approximation
- EP uses this in a tractable manner to optimise

$$D_{KL}(t_i(\boldsymbol{\theta})q_{\setminus i}(\boldsymbol{\theta}) \,||\, \tilde{t}_i(\boldsymbol{\theta})q_{\setminus i}(\boldsymbol{\theta}))$$

- Provides good approximations of marginals and marginal likelihood
- Alternative interpretation to existing belief net algorithms
- Algorithm may not converge ($\rightarrow$ explicitly minimise the energy?)