Approximate Bayesian Inference for Machine Learning [1,2]

Louis Ellam
l.ellam@warwick.ac.uk
Warwick Centre of Predictive Modelling (WCPM)
The University of Warwick

13 October 2015

http://www2.warwick.ac.uk/wcpm/

[1] Based on Louis Ellam, "Approximate Inference for Machine Learning", MSc Dissertation, Centre for Scientific Computing, University of Warwick

[2] Warwick Centre of Predictive Modelling (WCPM) Seminar Series, University of Warwick

Outline

Outline

Machine Learning

- Machine Learning is the study of algorithms that can learn from data

Regression

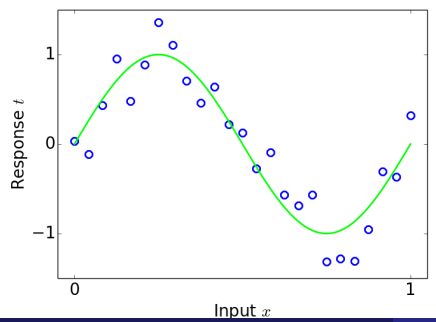
Given $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N, t_1, \ldots, t_N\}$, find a model that allows us to accurately estimate the response of any input.

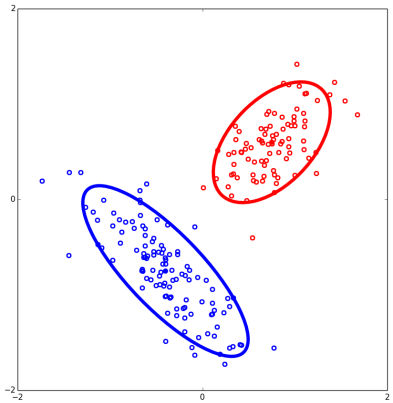
Clustering

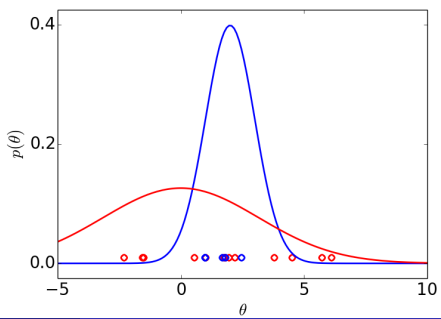
Given data $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ and some fixed number of clusters K, assign each data point to a cluster so that points in each cluster are close together.

The Clutter Problem

Given data $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ comprising of a noisy signal that is immersed in clutter, recover the original signal $\boldsymbol{\theta}$ i.e. determine $\mathbb{E}[\boldsymbol{\theta}]$ and $p(\mathcal{D})$. Observations are drawn from $p(\boldsymbol{x}|\boldsymbol{\theta}) = (1-w)\mathcal{N}(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{I}) + w\mathcal{N}(\boldsymbol{x}|\boldsymbol{0}, a\boldsymbol{I})$.







Outline

Classical Approach to Regression

Linear Model

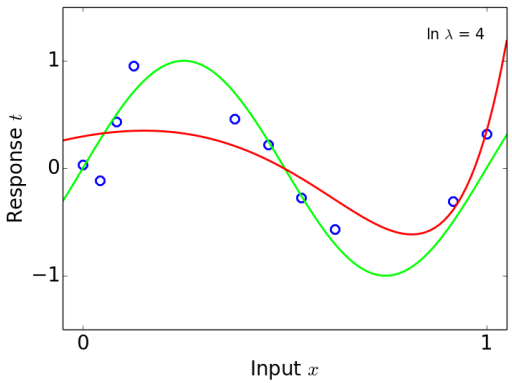
Use a linear model f defined over $M+1$ linearly independent basis functions $\phi_0(\boldsymbol{x}), \ldots, \phi_M(\boldsymbol{x})$ with weights $w_0, \ldots, w_M$

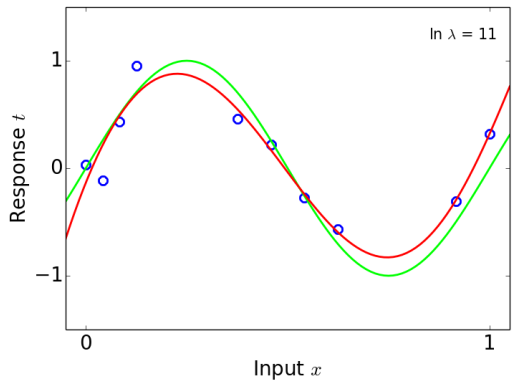
$$f(\boldsymbol{x}, \boldsymbol{w}) = \sum_{m=0}^{M} w_m \phi_m(\boldsymbol{x}).$$

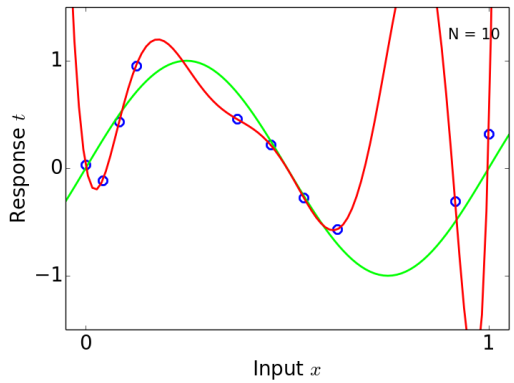
Tikhonov Regularisation

$$E(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{\Phi}\boldsymbol{w} - \boldsymbol{t}\|^2 + \frac{\lambda}{2}\|\boldsymbol{w}\|^2,$$

$$\boldsymbol{w}^* = \underset{\boldsymbol{w}}{\operatorname{argmin}} E(\boldsymbol{w}) = (\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda \boldsymbol{I})^{-1}\boldsymbol{\Phi}^T\boldsymbol{t}.$$







C. Bishop (2006), D. Calvetti et al. (2007)

Outline

Statistical Approach to Regression

Bayes Theorem

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

Likelihood and Prior

We expect that an observation to satisfy

$$t = f(\boldsymbol{x}, \boldsymbol{w}) + \eta,$$

where η denotes mean centered Gaussian noise with precision parameter β. Thus the likelihood function is defined as

$$p(\boldsymbol{t}|\boldsymbol{x}, \boldsymbol{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|\boldsymbol{w}^T\phi(\boldsymbol{x}_n), \beta^{-1}).$$

$$p(\boldsymbol{w}|\alpha) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{0},\alpha^{-1}\boldsymbol{I}).$$

Posterior

$$\ln p(\boldsymbol{w}|, \boldsymbol{x}, \boldsymbol{t}, \alpha, \beta) = -\frac{\beta}{2}\sum_{n=1}^{N}(t_n - \boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_n))^2 - \frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w} + \text{const.}$$

R. Kass (1995)

Predictive Distribution

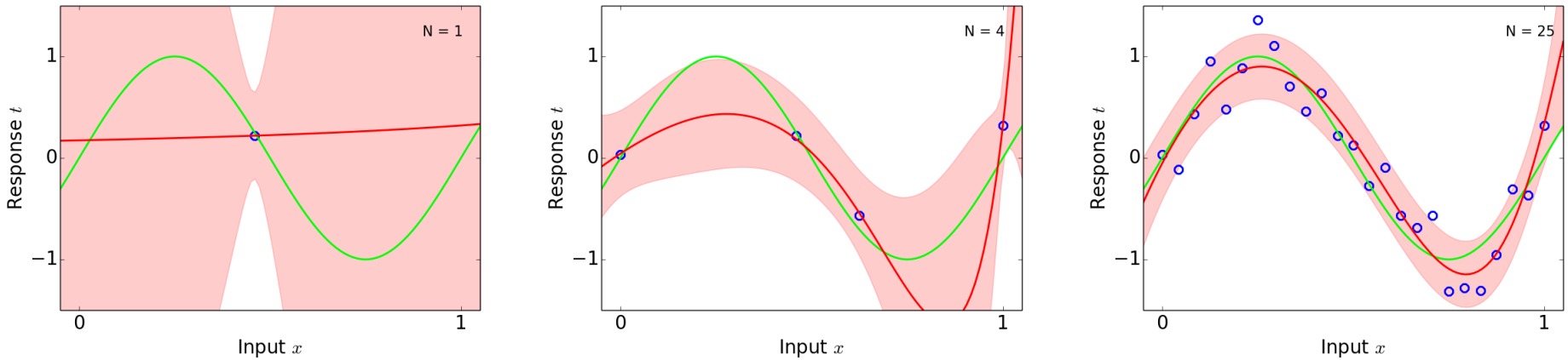
$$p(t|\boldsymbol{x}, \boldsymbol{t}, \alpha, \beta) = \int p(t|\boldsymbol{w}) p(\boldsymbol{w}|\boldsymbol{x}, \boldsymbol{t}, \alpha, \beta) d\boldsymbol{w} = \mathcal{N}(t|\boldsymbol{m}_N^T \phi(\boldsymbol{x}), \sigma_N^2(\boldsymbol{x})),$$

where

$$\sigma_N^2 = \frac{1}{\beta} + \phi(\boldsymbol{x})^T \boldsymbol{S}_N \phi(\boldsymbol{x}),$$

$$\boldsymbol{S}_N^{-1} = \alpha \boldsymbol{I} + \beta \boldsymbol{\Phi}^T \boldsymbol{\Phi},$$

$$\boldsymbol{m}_N = \beta \boldsymbol{S}_N \boldsymbol{\Phi}^T \boldsymbol{y}.$$



- 'More Bayesian' if α is treated as a random variable, but analytically intractable.
- Classical integration methods are not suitable for high dimensional probability density functions.

C. Bishop (2006)

Outline

Monte Carlo Estimate

$$\mathbb{E}_f[q(\boldsymbol{\theta})] = \int q(\boldsymbol{\theta}) f(\boldsymbol{\theta}) d\boldsymbol{\theta} \approx \frac{1}{N} \sum_{n=1}^{N} q(\boldsymbol{X}_n) \text{ where } \boldsymbol{X}_n \overset{iid}{\sim} f(\boldsymbol{\theta}).$$

$$\mathbb{E}_f[q(\boldsymbol{\theta})] = \int q(\boldsymbol{\theta}) \underbrace{\frac{f(\boldsymbol{\theta})}{g(\boldsymbol{\theta})}}_{=:w(\boldsymbol{\theta})} g(\boldsymbol{\theta}) d\boldsymbol{\theta} = \int q(\boldsymbol{\theta}) w(\boldsymbol{\theta}) g(\boldsymbol{\theta}) d\boldsymbol{\theta}$$

Importance Sampling Estimate

$$\mathbb{E}_f[q(\boldsymbol{\theta})] \approx \frac{1}{N} \sum_{n=1}^{N} w(\boldsymbol{Z}_n) q(\boldsymbol{Z}_n) \text{ where } \boldsymbol{Z}_n \overset{iid}{\sim} g(\boldsymbol{\theta}).$$

Self-Normalised Importance Sampling Estimate

$$\mathbb{E}_f[q(\boldsymbol{\theta})] \approx \frac{1}{\sum_{m=1}^{N} \tilde{w}(\boldsymbol{Z}_m)} \sum_{n=1}^{N} \tilde{w}(\boldsymbol{Z}_n) q(\boldsymbol{Z}_n) \text{ where } \boldsymbol{Z}_n \overset{iid}{\sim} g(\boldsymbol{\theta}).$$

Monte Carlo Integration for the Clutter Problem

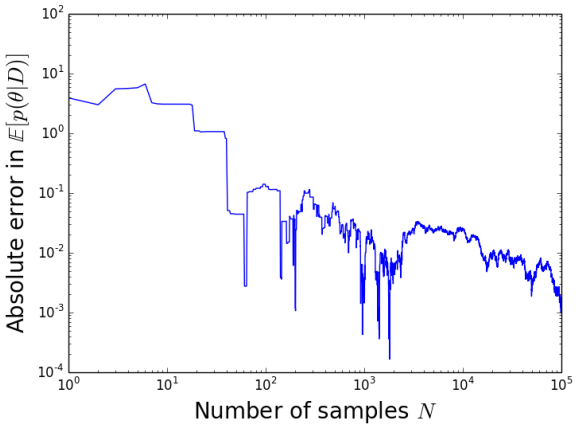
Define importance weights

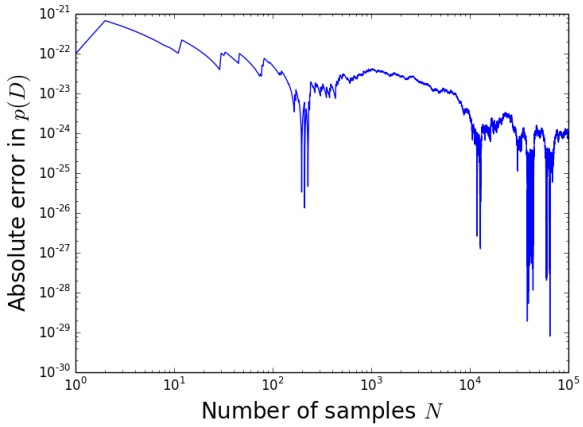
$$w(\boldsymbol{\theta}) = \frac{\tilde{f}(\boldsymbol{\theta})}{\tilde{g}(\boldsymbol{\theta})} = \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{\theta})} = p(\mathcal{D}|\boldsymbol{\theta}).$$

Expectation (SNIS) and Evidence ('Vanilla' MC)

$$\mathbb{E}_{p(\boldsymbol{\theta}|\mathcal{D})}[\boldsymbol{\theta}] \approx \frac{1}{\sum_{m=1}^{N} p(\mathcal{D}|\boldsymbol{X}_m)} \sum_{n=1}^{N} \boldsymbol{X}_n p(\mathcal{D}|\boldsymbol{X}_n) \text{ where } \boldsymbol{X}_n \overset{iid}{\sim} p(\boldsymbol{\theta}),$$

$$p(\mathcal{D}) = \mathbb{E}_{p(\boldsymbol{\theta})}[p(\mathcal{D}|\boldsymbol{\theta})] \approx \frac{1}{N} \sum_{n=1}^{N} p(\mathcal{D}|\boldsymbol{X}_n) \text{ where } \boldsymbol{X}_n \overset{iid}{\sim} p(\boldsymbol{\theta}).$$





T. Minka (2001a), J. Liu (2008)

Outline

Markov Chain Monte Carlo

- A discrete Markov chain is defined as a series of random variables $\boldsymbol{Z}^{(1)}, \ldots, \boldsymbol{Z}^{(M)}$

$$p(\boldsymbol{Z}^{(M)}|\boldsymbol{Z}^{(1)}, \ldots, \boldsymbol{Z}^{(M-1)}) = p(\boldsymbol{Z}^{(M)}|\boldsymbol{Z}^{(M-1)}).$$

- Draw new sample from proposal distribution and accept with probability

$$A(\boldsymbol{Z}|\boldsymbol{Z}^{(M-1)}) = \min\left\{1, \frac{f(\boldsymbol{Z})g(\boldsymbol{Z}^{(M-1)}|\boldsymbol{Z})}{f(\boldsymbol{Z}^{(M-1)})g(\boldsymbol{Z}|\boldsymbol{Z}^{(M-1)})}\right\}.$$

Algorithm 1 Metropolis-Hastings

Initialise $\boldsymbol{Z}^{(0)} = \{\boldsymbol{Z}_1^{(0)}, \ldots, \boldsymbol{Z}_K^{(0)}\}$
for M=1,...,T **do**
  Draw $\boldsymbol{Z} \sim g(\cdot|\boldsymbol{Z}^{(M-1)})$
  Compute $A(\boldsymbol{Z}|\boldsymbol{Z}^{(M-1)}) = \min\left\{1, \frac{f(\boldsymbol{Z})g(\boldsymbol{Z}^{(M-1)}|\boldsymbol{Z})}{f(\boldsymbol{Z}^{(M-1)})g(\boldsymbol{Z}|\boldsymbol{Z}^{(M-1)})}\right\}$
  With probability $A(\boldsymbol{Z}|\boldsymbol{Z}^{(M-1)})$ set $\boldsymbol{Z}^{(M)} = \boldsymbol{Z}$, otherwise set $\boldsymbol{Z}^{(M)} = \boldsymbol{Z}^{(M-1)}$
end for

Algorithm 2 Gibbs Sampling

Initialise $\boldsymbol{Z}_1^{(0)}, \ldots, \boldsymbol{Z}_K^{(0)}$
for M=1,...,T **do**
  draw $\boldsymbol{Z}_1^{(M)} \sim \boldsymbol{Z}_1|\boldsymbol{Z}_2^{(M-1)}, \ldots, \boldsymbol{Z}_K^{(M-1)}$
  draw $\boldsymbol{Z}_2^{(M)} \sim \boldsymbol{Z}_2|\boldsymbol{Z}_1^{(M)}, \boldsymbol{Z}_3^{(M-1)}, \ldots, \boldsymbol{Z}_K^{(M-1)}$
  $\vdots$
  draw $\boldsymbol{Z}_K^{(M)} \sim \boldsymbol{Z}_K|\boldsymbol{Z}_1^{(M)}, \ldots, \boldsymbol{Z}_{K-1}^{(M)}$
end for

Gibbs Sampling for Clustering

Likelihood for GMM

$$p(\mathcal{D}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \prod_{n=1}^{N} \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

Introduce Latent Variables

$$p(z_{nk} = 1|\pi_k) = \pi_k,$$

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

Joint Distribution for GMM (Augmented)

$$p(\mathcal{D}, \boldsymbol{z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = p(\mathcal{D}|\boldsymbol{z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) p(\boldsymbol{z}|\boldsymbol{\pi}) p(\boldsymbol{\pi}) p(\boldsymbol{\mu}) p(\boldsymbol{\Sigma}).$$

$$p(\mathcal{D}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{z}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_{nk}}.$$

$$p(\boldsymbol{\mu}) p(\boldsymbol{\Sigma}) = \prod_{k=1}^{K} \mathcal{N}(\boldsymbol{\mu}_k|\boldsymbol{m}_0, \boldsymbol{V}_0) \mathcal{IW}(\boldsymbol{\Sigma}_k|\boldsymbol{W}_0, v_0),$$

$$p(\boldsymbol{\pi}) = \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}_0).$$

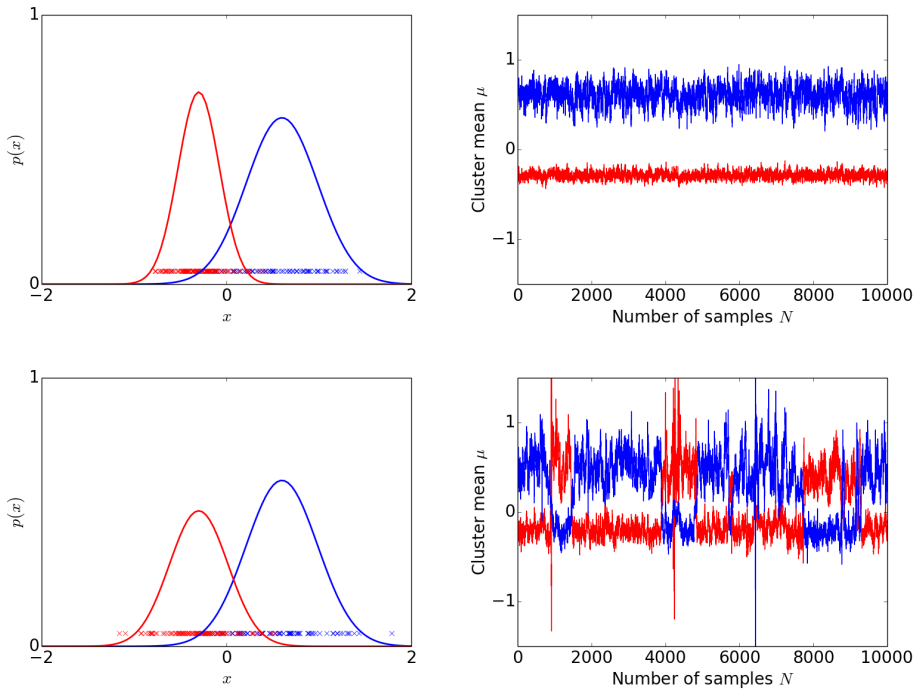
Full Conditionals

$$p(z_{nk} = 1|\boldsymbol{x}_n, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \frac{\pi_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)},$$

$$p(\boldsymbol{\pi}|\boldsymbol{z}) = \text{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}_k),$$

$$p(\boldsymbol{\mu}_k|\boldsymbol{\Sigma}_k, \boldsymbol{z}, \boldsymbol{x}) = \mathcal{N}(\boldsymbol{\mu}_k|\boldsymbol{m}_k, \boldsymbol{V}_k),$$

$$p(\boldsymbol{\Sigma}_k|\boldsymbol{\mu}_k, \boldsymbol{z}, \boldsymbol{x}) = \mathcal{IW}(\boldsymbol{\Sigma}_k|\boldsymbol{W}_k, \nu_k),$$



K. Murphy (2012)

Outline

Laplace Approximation

Seek to approximate probability density function $p(\boldsymbol{z})$ of the form $p(\boldsymbol{z}) = \frac{1}{Z} f(\boldsymbol{z})$.

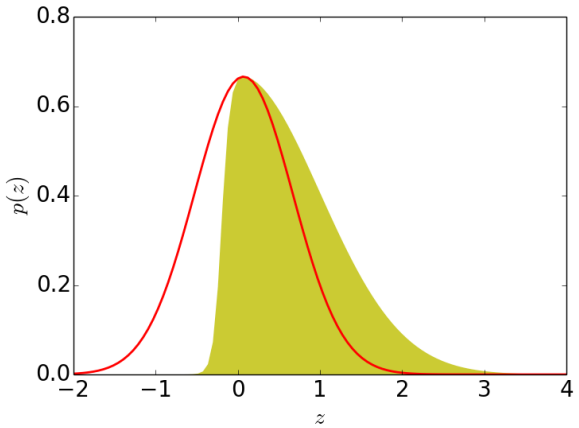
$$\ln f(\boldsymbol{z}) \approx \ln f(\boldsymbol{z}_0) + (\boldsymbol{z} - \boldsymbol{z}_0)^T \nabla \ln f(\boldsymbol{z})|_{z=z_0} + \frac{1}{2}(\boldsymbol{z} - \boldsymbol{z}_0)^T \nabla\nabla \ln f(\boldsymbol{z})|_{z=z_0}(\boldsymbol{z} - \boldsymbol{z}_0),$$
$$\nabla f(\boldsymbol{z})|_{z=z_0} = 0.$$

Laplace Approximation

$$q(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{z}_0, \boldsymbol{A}^{-1}), \text{ where } \boldsymbol{A} = -\nabla\nabla \ln f(\boldsymbol{z})|_{z=z_0},$$

Example

$$p(z) \propto \exp(-z^2/2)\sigma(20z + 4), \text{ where } \sigma(t) = (1 + e^{-t})^{-1}.$$



R. Kass et al. (1995), C. Bishop (2006)

EM Algorithm (GMM)

Want to find mode of GMM

$$\ln p(\mathcal{D}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}.$$

MLE of $\boldsymbol{\mu}_k$ is

$$\boldsymbol{\mu_k} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \boldsymbol{x_n},$$

where

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk}),$$

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}.$$

MLE of $\boldsymbol{\Sigma}_k$ is

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T.$$

A. Dempster et al. (1977), C. Bishop (2006)

EM Algoririthm (GMM)

The maximisation for $\boldsymbol{\pi}$ is constrained to $\sum_{k=1}^{K} \pi_k = 1$ so find MLE of

$$\ln p(\mathcal{D}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) + \lambda\Big(\sum_{k=1}^{K} \pi_k - 1\Big),$$

which yields the following

$$\pi_k = \frac{N_k}{N}.$$

Algorithm 3 EM GMM

Initialise $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k$ for $k = 1, \ldots, K$

repeat

E Step: Evaluate responsibilities

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}.$$

M Step: Define

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk}),$$

then update parameters

$$\boldsymbol{\mu_k} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \boldsymbol{x}_n,$$

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T,$$

$$\pi_k = \frac{N_k}{N}.$$

until convergence

EM Algorithm (GMM)

Link to code / .mp4

Outline

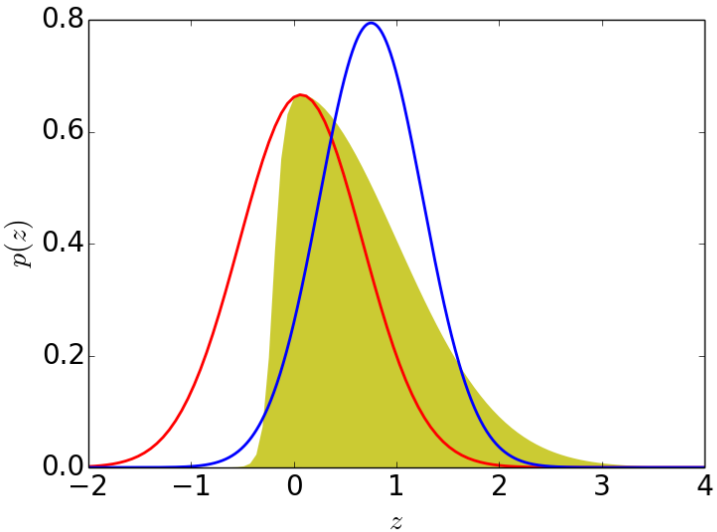
KL Divergence

Choose a proxy function $q(\boldsymbol{\theta})$ and optimise $\boldsymbol{\theta}$ so that q is as close fit as possible to p. One such approach is to minimise

$$\mathrm{KL}(q||p) := -\int q(\boldsymbol{\theta}) \ln \frac{p(\boldsymbol{\theta}|\mathcal{D})}{q(\boldsymbol{\theta})} d\boldsymbol{\theta}.$$

Example

$$p(z) \propto \exp(-z^2/2)\sigma(20z+4), \text{ where } \sigma(t) = (1+e^{-t})^{-1}.$$



H. Attias (2000), Z. Ghahramani et al. (1999), C. Bishop (2006)

$$\begin{aligned}\ln p(\mathcal{D}) &= \ln \frac{p(\mathcal{D},\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\mathcal{D})} = \int q(\boldsymbol{\theta}) \ln \frac{p(\mathcal{D},\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\mathcal{D})} d\boldsymbol{\theta} = \int q(\boldsymbol{\theta}) \ln \frac{p(\mathcal{D},\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\mathcal{D})} \frac{q(\boldsymbol{\theta})}{q(\boldsymbol{\theta})} d\boldsymbol{\theta} \\ &= \int q(\boldsymbol{\theta}) \ln \frac{p(\mathcal{D},\boldsymbol{\theta})}{q(\boldsymbol{\theta})} d\boldsymbol{\theta} - \int q(\boldsymbol{\theta}) \ln \frac{p(\boldsymbol{\theta}|\mathcal{D})}{q(\boldsymbol{\theta})} d\boldsymbol{\theta} \\ &= \mathcal{L}(q) + \mathrm{KL}(q||p).\end{aligned}$$

Lowerbound

Minimising the KL divergence is equivalent to maximising the lower bound

$$\mathcal{L}(q) = \int q(\boldsymbol{\theta}) \ln \frac{p(\mathcal{D},\boldsymbol{\theta})}{q(\boldsymbol{\theta})} d\boldsymbol{\theta},$$

Lowerbound Evidence Approximation

It may be shown that

$$\mathrm{KL}(q||p) \geq 0.$$

hence $p(\mathcal{D})$ may be approximated using

$$\ln p(\mathcal{D}) \geq \mathcal{L}(q),$$

$$p(\mathcal{D}) \to \exp \mathcal{L}(q) \quad \text{as} \quad \mathrm{KL}(q||p) \to 0.$$

Mean Field Approximation

Assume $q(\boldsymbol{\theta})$ can be expressed in factorised form

$$q(\boldsymbol{\theta}) = \prod_{m=1}^{M} q_m(\boldsymbol{\theta}_m).$$

Mean Field Lowerbound

$$\begin{aligned}
\mathcal{L}(q) &= \int \prod_{m=1}^{M} q(\boldsymbol{\theta}_m) \ln \frac{p(\mathcal{D}, \boldsymbol{\theta})}{\prod_{m=1}^{M} q(\boldsymbol{\theta}_m)} d\boldsymbol{\theta} = \int \prod_{m=1}^{M} q(\boldsymbol{\theta}_m) \Big\{ \ln p(\mathcal{D}, \boldsymbol{\theta}) - \sum_{m=1}^{M} \ln q(\boldsymbol{\theta}_m) \Big\} d\boldsymbol{\theta} \\
&= \int q(\boldsymbol{\theta}_j) \Big\{ \int \ln p(\mathcal{D}, \boldsymbol{\theta}) \prod_{\substack{m=1 \\ m \neq j}}^{M} q(\boldsymbol{\theta}_m) d\boldsymbol{\theta}_m \Big\} d\boldsymbol{\theta}_j \quad - \int q(\boldsymbol{\theta}_j) \ln q(\boldsymbol{\theta}_j) d\boldsymbol{\theta}_j + \text{const} \\
&= \int q(\boldsymbol{\theta}_j) \mathbb{E}_{m \neq j}[\ln p(\mathcal{D}, \boldsymbol{\theta})] d\boldsymbol{\theta}_j - \int q(\boldsymbol{\theta}_j) \ln q(\boldsymbol{\theta}_j) d\boldsymbol{\theta}_j + \text{const} \\
&= -\mathrm{KL}(q(\boldsymbol{\theta}_j) || \exp \mathbb{E}_{m \neq j}[\ln p(\mathcal{D}, \boldsymbol{\theta})]) + \text{const}.
\end{aligned}$$

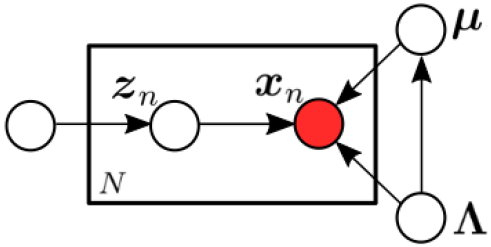
Mean Field Minimised KL Divergence

$$\ln q_j^*(\boldsymbol{\theta}_j) = \mathbb{E}_{m \neq j}[\ln p(\mathcal{D}, \boldsymbol{\theta})] + \text{const}.$$

Variational Inference GMM

Target Posterior

$$p(\mathcal{D}, \boldsymbol{z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = p(\mathcal{D}|\boldsymbol{z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})p(\boldsymbol{z}|\boldsymbol{\pi})p(\boldsymbol{\pi})p(\boldsymbol{\mu}|\boldsymbol{\Lambda})p(\boldsymbol{\Lambda}).$$



Conjugate Prior

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

Joint Distribution

$$p(\mathcal{D}, \boldsymbol{z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{\pi}) \propto$$
$$\left(\prod_{n=1}^{N}\prod_{k=1}^{K}\left(\pi_k\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)\right)^{z_{nk}}\right) \cdot \mathrm{Dir}(\boldsymbol{\pi}|\alpha_0) \cdot \prod_{k=1}^{K}\mathcal{N}(\boldsymbol{\mu}_k|\boldsymbol{m}_0, \beta_0^{-1}\boldsymbol{\Lambda}_k)\mathcal{W}(\boldsymbol{\Lambda}_k|\boldsymbol{W}_0, \nu_0).$$

Mean Field Approximation

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

Mean Field Minimised KL Divergence

$$\ln q_j^*(\boldsymbol{\theta}_j) = \mathbb{E}_{m \neq j}[\ln p(\mathcal{D}, \boldsymbol{\theta})] + \text{const.}$$

Expression for Responsbilities

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

where

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

which can be used to give an expression of the responsibilities

$$r_{nk} = \mathbb{E}[z_{nk}] = \frac{\rho_{nk}}{\sum_{j=1}^{K} \rho_{nj}}.$$

Variational Inference GMM

Define

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

$$\bar{\boldsymbol{x}}_{\boldsymbol{k}} = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} \boldsymbol{x}_n,$$

$$\boldsymbol{S}_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} (\boldsymbol{x}_n - \bar{\boldsymbol{x}}_k)(\boldsymbol{x}_n - \bar{\boldsymbol{x}}_k)^T.$$

Expression for Model Parameters

$$\ln q^*(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = \mathbb{E}_z[\ln p(\mathcal{D}, \boldsymbol{z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})] + \text{const.}$$

By inspection

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

$$q(\boldsymbol{\pi}|\boldsymbol{\alpha}) = \text{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}),$$

$$q^*(\boldsymbol{\mu}_k|\boldsymbol{\Lambda}_k) = \mathcal{N}(\boldsymbol{\mu}_k|\boldsymbol{m}_k, (\beta_k \boldsymbol{\Lambda}_k)^{-1}),$$

$$q^*(\boldsymbol{\Lambda}_k) = \mathcal{W}(\boldsymbol{\Lambda}_k|\boldsymbol{W}_k, \nu_k).$$

Variational Inference GMM

Expressions for Model Parameters

$$\alpha_k = \alpha_0 + N_k,\ \beta_k = \beta_0 + N_k \text{ and } \nu_k = \nu_0 + N_k,$$

$$\boldsymbol{m}_k = \frac{1}{\beta_k}(\beta_0 \boldsymbol{m}_0 + N_k \bar{\boldsymbol{x}}_k),$$

$$\boldsymbol{W}_k^{-1} = \boldsymbol{W}_0^{-1} + N_k \boldsymbol{S}_k + \frac{\beta_0 N_k}{\beta_0 + N_k}(\bar{\boldsymbol{x}}_k - \boldsymbol{m}_0)(\bar{\boldsymbol{x}}_k - \boldsymbol{m}_0)^T.$$

Expressions for Responsbilities

$$\mathbb{E}_{\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k}\left[(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Lambda}_k (\boldsymbol{x}_n - \boldsymbol{\mu}_k)\right] = \nu_k (\boldsymbol{x}_n - \boldsymbol{m}_k)^T \boldsymbol{W}_k (\boldsymbol{x}_n - \boldsymbol{m}_k) + D\beta_k^{-1},$$

$$\ln \tilde{\pi}_k := \mathbb{E}[\ln |\pi_k|] = \psi(\alpha_k) - \psi(\sum_{k=1}^{K} \alpha_k),$$

$$\ln \tilde{\Lambda}_k := \mathbb{E}[\ln |\boldsymbol{\Lambda}_k|] = \sum_{i=1}^{D} \psi\left(\frac{\nu_k + 1 - i}{2}\right) + D \ln 2 + \ln |\boldsymbol{W}_k|,$$

Mixing Coefficients

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

Variational Inference GMM

Algorithm 4 Variational Bayes GMM

Initialise $\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{m}, \boldsymbol{W}, \beta_0, \nu_0, \alpha_0$

repeat

E Step: Compute the responsibilities as follows

$$r_{nk} \propto \tilde{\pi}_k \tilde{\boldsymbol{\Lambda}}_k^{1/2} \exp\Big\{ -\frac{D}{2\beta_k} - \frac{\nu_k}{2}(\boldsymbol{x}_n - \boldsymbol{m}_k)^T \boldsymbol{W}_k (\boldsymbol{x}_n - \boldsymbol{m}_k)\Big\},$$

$$\ln \tilde{\pi}_k = \psi(\alpha_k) - \psi\Big(\sum_{k=1}^{K} \alpha_k\Big),$$

$$\ln \tilde{\boldsymbol{\Lambda}}_k = \sum_{i=1}^{D} \Big(\psi\Big(\frac{\nu_k + 1 - i}{2}\Big)\Big) + D \ln 2 + \ln |\boldsymbol{W}_k|,$$

M Step: Update model parameters as follows

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

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

$$\boldsymbol{S}_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} (\boldsymbol{x}_n - \bar{\boldsymbol{x}}_k)(\boldsymbol{x}_n - \bar{\boldsymbol{x}}_k)^T,$$

$$\alpha_k = \alpha_0 + N_k, \beta_k = \beta_0 + N_k \text{ and } \nu_k = \nu_0 + N_k,$$

$$\boldsymbol{m}_k = \beta_k^{-1}(\beta_0 \boldsymbol{m}_0 + N_k \bar{\boldsymbol{x}}_k),$$

$$\boldsymbol{W}_k^{-1} = \boldsymbol{W}_0^{-1} + N_k \boldsymbol{S}_k + \frac{\beta_0 N_k}{\beta_0 + N_k} (\bar{\boldsymbol{x}}_k - \boldsymbol{m}_0)(\bar{\boldsymbol{x}}_k - \boldsymbol{m}_0)^T.$$

until convergence

Variational Inference GMM

Link to code / .mp4

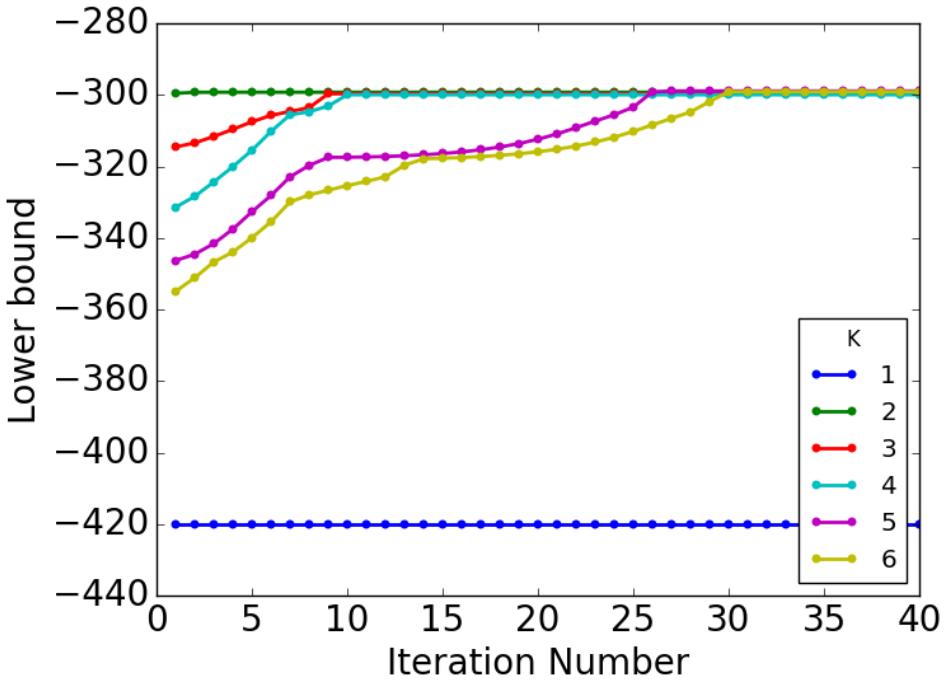
Variational Inference GMM

Link to code / .mp4

Variational Inference GMM Lowerbound

Lowerbound

$$\begin{aligned}\mathcal{L}(q) &= \sum_{z} \iiint q(\boldsymbol{z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \ln \left\{ \frac{p(\mathcal{D}, \boldsymbol{z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})}{q(\boldsymbol{z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})} \right\} d\boldsymbol{\pi} d\boldsymbol{\mu} d\boldsymbol{\Lambda} \\ &= \mathbb{E}[\ln p(\mathcal{D}, \boldsymbol{z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})] - \mathbb{E}[\ln q(\boldsymbol{z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})] \\ &= \mathbb{E}[\ln p(\mathcal{D}|\boldsymbol{z}, \boldsymbol{\mu}, \boldsymbol{\Lambda})] + \mathbb{E}[\ln p(\boldsymbol{z}|\boldsymbol{\pi})] + \mathbb{E}[\ln p(\boldsymbol{\pi})] + \mathbb{E}[\ln p(\boldsymbol{\mu}, \boldsymbol{\Lambda})] \\ &\quad - \mathbb{E}[\ln q(\boldsymbol{z})] - \mathbb{E}[\ln q(\boldsymbol{\pi})] - \mathbb{E}[\ln q(\boldsymbol{\mu}, \boldsymbol{\Lambda})].\end{aligned}$$



Variational Inference GMM Model Selection

Model Selection with Lowerbound

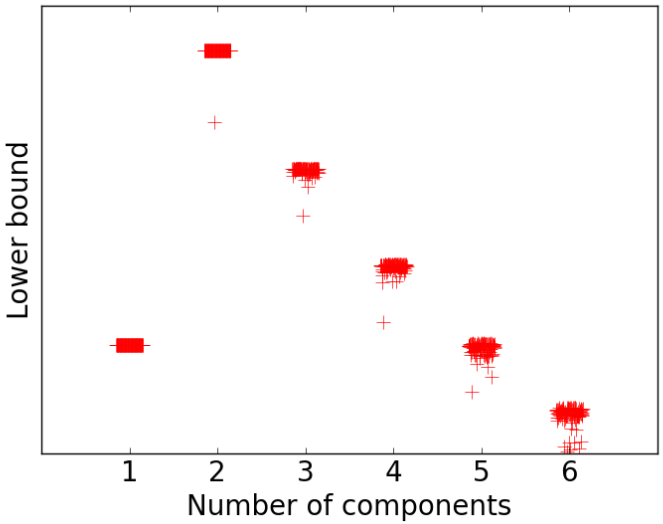
$$p(K|\mathcal{D}) = \frac{K! p(\mathcal{D}|K)}{\sum_{K'} K! p(\mathcal{D}|K')},$$

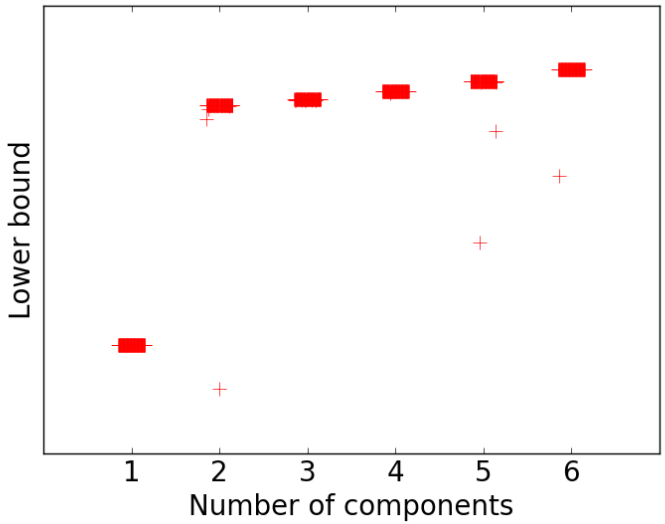
where

$$p(\mathcal{D}|K) \approx \exp \mathcal{L}(K).$$

Alternative View

$$\mathcal{L}_M(K) \approx \mathcal{L}(K) + \ln K!.$$





A. Corduneanu (2001), C. Bishop (2006), K. Murphy (2012)

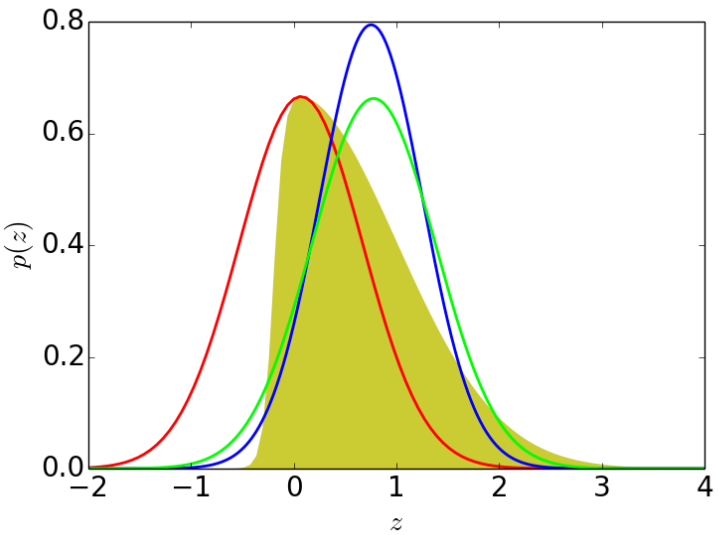
Outline

Minimise Reverse KL Divergence

$$\mathrm{KL}(p||q) = -\int p(\boldsymbol{\theta}|\mathcal{D}) \ln \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\mathcal{D})} d\boldsymbol{\theta}.$$

Example

$$p(z) \propto \exp(-z^2/2)\sigma(20z + 4), \text{ where } \sigma(t) = (1 + e^{-t})^{-1}.$$



T. Minka (2001a), T. Minka (2001b)

Exponential Family

$$q(\boldsymbol{\theta}) = h(\boldsymbol{\theta})g(\boldsymbol{\eta})\exp\{\boldsymbol{\eta}^T\boldsymbol{u}(\boldsymbol{\theta})\},$$

Reverse KL Divergence for Exponential Family

$$\begin{aligned}\mathrm{KL}(p||q) &= \int p(\boldsymbol{\theta})\ln\left\{\frac{p(\boldsymbol{\theta})}{q(\boldsymbol{\theta})}\right\}d\boldsymbol{\theta} \\ &= \int p(\boldsymbol{\theta})\{\ln p(\boldsymbol{\theta}) - \ln\left(h(\boldsymbol{\theta})g(\boldsymbol{\eta})\right) - \boldsymbol{\eta}^T\boldsymbol{u}(\boldsymbol{\theta})\}d\boldsymbol{\theta} \\ &= -\ln g(\boldsymbol{\eta}) - \boldsymbol{\eta}^T\mathbb{E}_{p(\boldsymbol{\theta})}[\boldsymbol{u}(\boldsymbol{\theta})] + \text{const.}\end{aligned}$$

Minimised Reverse KL Divergence for Exponential Family

$$-\nabla \ln g(\boldsymbol{\eta}) = \mathbb{E}_{p(\boldsymbol{\theta})}[\boldsymbol{u}(\boldsymbol{\theta})],$$

i.e. when

$$\mathbb{E}_{q(\boldsymbol{\theta})}[\boldsymbol{u}(\boldsymbol{\theta})] = \mathbb{E}_{p(\boldsymbol{\theta})}[\boldsymbol{u}(\boldsymbol{\theta})].$$

Difficult, as trying to taking expectation w.r.t. unknown posterior.

Assumed Form of Joint Distribution

$$p(\boldsymbol{\theta}, \mathcal{D}) = \prod_n f_n(\boldsymbol{\theta}),$$

Approximating Distribution

An approximating posterior belonging to the **exponential family** is sought of the following form where Z is a normalisation constant

$$q(\boldsymbol{\theta}) = \frac{1}{Z} \prod_n \tilde{f}_n(\boldsymbol{\theta}).$$

- The idea is to refine each factor factor $\tilde{f}_n$ by making the approximating distribution

$$q^{\text{new}}(\boldsymbol{\theta}) \propto \tilde{f}_n(\boldsymbol{\theta}) \prod_{j \neq n} \tilde{f}_j(\boldsymbol{\theta}),$$

a closer approximation to

$$f_n(\boldsymbol{\theta}) \prod_{j \neq n} \tilde{f}_j(\boldsymbol{\theta}).$$

Expectation Propagation

- To proceed, factor n is first removed from the approximating distribution by

$$q^{\backslash n}(\boldsymbol{\theta}) = \frac{q(\boldsymbol{\theta})}{\tilde{f}_n(\boldsymbol{\theta})}.$$

- A new distribution is defined by

$$p^n(\boldsymbol{\theta}) = \frac{1}{Z_n} f_n(\boldsymbol{\theta}) q^{\backslash n}(\boldsymbol{\theta}),$$

  with normalisation constant

$$Z_n = \int f_n(\boldsymbol{\theta}) q^{\backslash n}(\boldsymbol{\theta}) d\boldsymbol{\theta}.$$

- The factor $\tilde{f}_n$ is updated such that the sufficient statistics of q are matched to those of p^n.

Algorithm 5 Expectation Propagation

Initialise approximating factors $\tilde{f}_n(\boldsymbol{\theta})$
Initialise approximating function $q(\boldsymbol{\theta}) \propto \prod_n \tilde{f}_n(\boldsymbol{\theta})$
repeat
  Choose a factor $\tilde{f}_n(\boldsymbol{\theta})$ for refinement
  Set $q^{\backslash n}(\boldsymbol{\theta}) = q(\boldsymbol{\theta})/\tilde{f}_n(\boldsymbol{\theta})$
  Set sufficient statistics of $q^{new}(\boldsymbol{\theta})$ to those of $q^{\backslash n}(\boldsymbol{\theta}) f_n(\boldsymbol{\theta})$ including $Z = \int q^{\backslash n}(\boldsymbol{\theta}) f_n(\boldsymbol{\theta}) d\boldsymbol{\theta}$
  Set $\tilde{f}_n(\boldsymbol{\theta}) = Z_n q^{new}(\boldsymbol{\theta})/q^{\backslash n}(\boldsymbol{\theta})$
until convergence

Expectation Propagation for the Clutter Problem

Factorised Posterior Distribution

$$f_n(\boldsymbol{\theta}) = (1-w)\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\theta}, \boldsymbol{I}) + w\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{0}, a\boldsymbol{I}) \text{ for } n = 1, \ldots, N$$
$$f_0(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}|\boldsymbol{0}, a\boldsymbol{I})$$

Approximating Distribution

Taken to be an isotropic Gaussian of the form

$$q(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}|\boldsymbol{m}, v\boldsymbol{I}),$$

- q is treated as a product of $N+1$ approximating factors

$$q(\boldsymbol{\theta}) \propto \prod_{n=0}^{N} \tilde{f}_n(\boldsymbol{\theta}).$$

- The factors have been defined as follows where factor $\tilde{f}_n$ has scaling constant s_n, mean $\boldsymbol{m}_n$ and variance $v_n\boldsymbol{I}$

$$\tilde{f}_n(\boldsymbol{\theta}) = s_n\mathcal{N}(\boldsymbol{\theta}|\boldsymbol{m}_n, v_n\boldsymbol{I}),$$

  in terms of a scaled Gaussian distribution.

- The approximating distribution is therefore Gaussian

$$q(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}|\boldsymbol{m}, v\boldsymbol{I}).$$

- $p(\mathcal{D}) \approx \int q(\boldsymbol{\theta})d\boldsymbol{\theta}$ is obtained using results for Gaussian products.

Expectation Propagation for Clutter Problem

Algorithm 6 Expectation Propagation for the Clutter Problem

Initialise the prior factor $\tilde{f}_0(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}|\mathbf{0}, b\boldsymbol{I})$
Initialise the remaining factors $\tilde{f}_n = 1, v_n = \infty, \boldsymbol{m}_n = \mathbf{0}$
Initialise the approximating function $\boldsymbol{m} = \mathbf{0}, v = b$
repeat
  for n=1,...,N **do**
    Remove the current factor from $q(\boldsymbol{\theta})$

$$(v^{\backslash n})^{-1} = v^{-1} - v_n^{-1},$$

$$\boldsymbol{m}^{\backslash n} = \boldsymbol{m} + v_n^{-1} v^{\backslash n}(\boldsymbol{m} - \boldsymbol{m}_n).$$

    Evaluate the new posterior $q(\boldsymbol{\theta})$ by matching sufficient statistics

$$Z_n = (1-w)\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{m}^{\backslash n}, (v^{\backslash n}+1)\boldsymbol{I}) + w\mathcal{N}(\boldsymbol{x}_n|\mathbf{0}, a\boldsymbol{I}),$$

$$\rho_n = 1 - \frac{w}{Z_n}\mathcal{N}(\boldsymbol{x}_n|\mathbf{0}, a\boldsymbol{I}),$$

$$\boldsymbol{m} = \boldsymbol{m}^{\backslash n} + \rho_n \frac{v^{\backslash n}}{v^{\backslash n}+1}(\boldsymbol{x}_n - \boldsymbol{m}^{\backslash n}),$$

$$v = v^{\backslash n} - \rho_n \frac{(v^{\backslash n})^2}{v^{\backslash n}+1} + \rho_n(1+\rho_n)\frac{(v^{\backslash n})^2\|\boldsymbol{x}_n - \boldsymbol{m}^{\backslash n}\|^2}{D(v^{\backslash n}+1)^2}.$$

    Evaluate and store the new factor $\tilde{f}_n(\boldsymbol{\theta})$

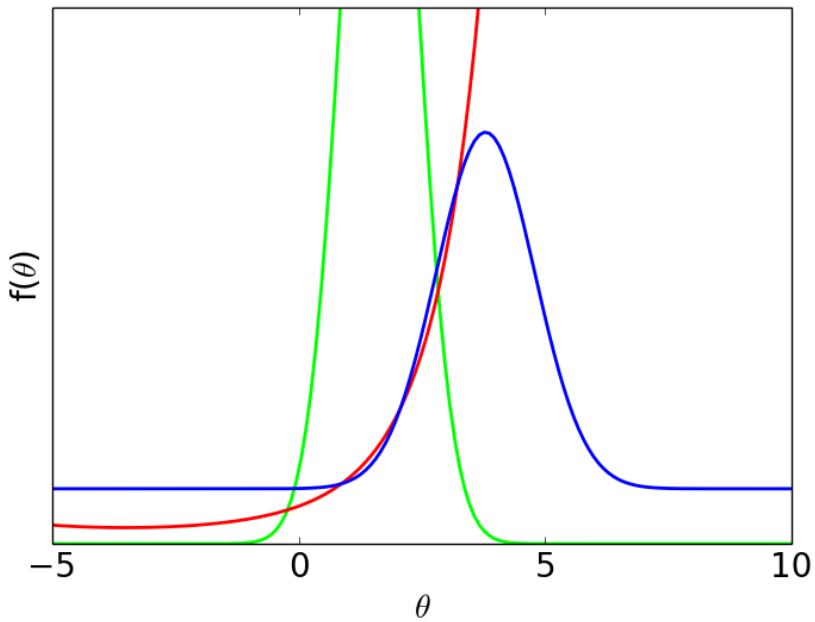
$$v_n^{-1} = v^{-1} - (v^{\backslash n})^{-1},$$

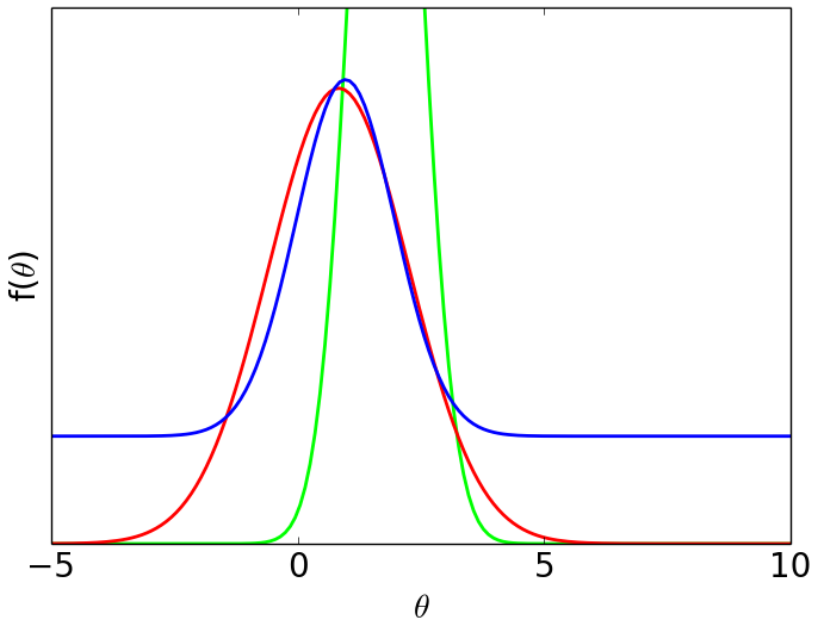
$$\boldsymbol{m}_n = \boldsymbol{m}^{\backslash n} + (v_n + v^{\backslash n})(v^{\backslash n})^{-1}(\boldsymbol{m} - \boldsymbol{m}^{\backslash n}),$$

$$s_n = \frac{Z_n}{\mathcal{N}(\boldsymbol{m}_n|\boldsymbol{m}^{\backslash n}, (v_n + v^{\backslash n})\boldsymbol{I})}.$$

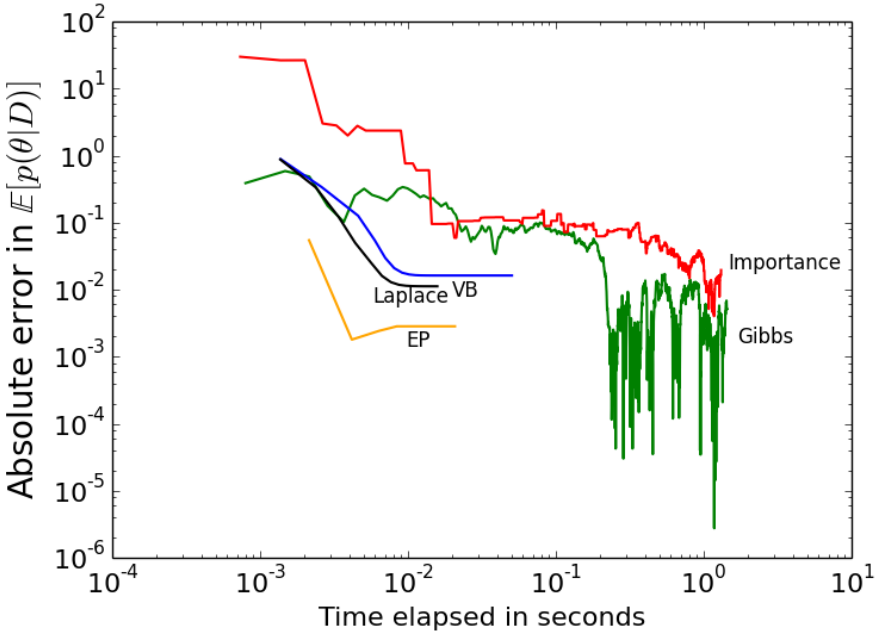
  end for
until convergence

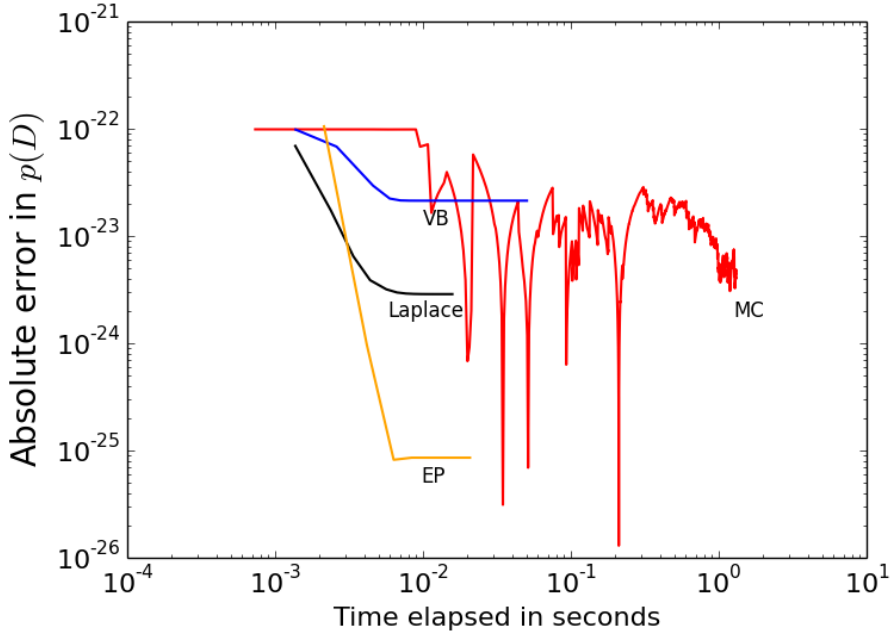
Expectation Propagation for the Clutter Problem





Summary of Methods for the Clutter Problem





T. Minka (2001a), T. Minka (2001b)

MSc Dissertation

L. Ellam.

Approximate Bayesian Inference for Machine Learning.

The MSc dissertation may be downloaded from here.

Software

The Python code for implementing the various algorithms reported in this presentation may be obtained by clicking on the respective figures. All software and datasets may be obtained in zip form from here.

Acknowledgments

- Professor N. Zabaras (Thesis Advisor) and Dr P. Brommer (Second Marker)
- EPSRC Strategic Package Project EP/L027682/1 for research at WCPM