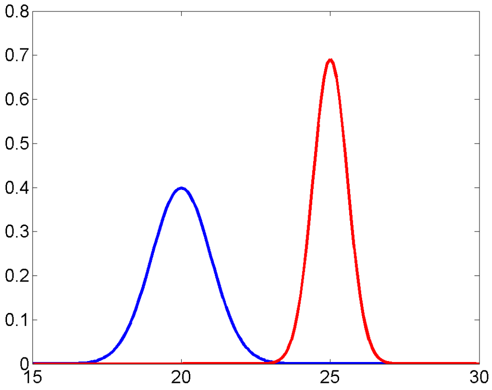
# Inference using Variational Bayes

Will Penny

Workshop on The Free Energy Principle,
UCL, July 5th 2012

# Optimal Data Fusion

For the prior (blue) we have $m_0 = 20$, $\lambda_0 = 1$ and for the likelihood (red) $m_D = 25$ and $\lambda_D = 3$.




Precision, $\lambda$, is inverse variance.

# Bayes rule for Gaussians

For a Gaussian prior with mean $m_0$ and precision $\lambda_0$, and a Gaussian likelihood with mean $m_D$ and precision $\lambda_D$ the posterior is Gaussian with

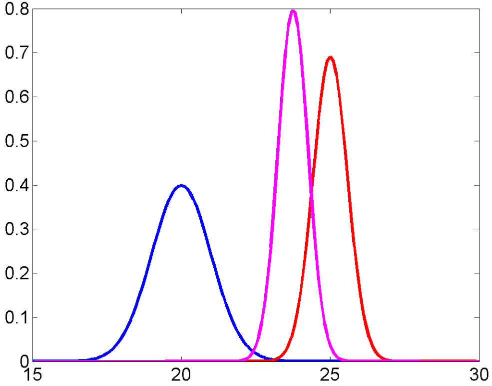
$$\begin{aligned} \lambda &= \lambda_0 + \lambda_D \\ m &= \frac{\lambda_0}{\lambda} m_0 + \frac{\lambda_D}{\lambda} m_D \end{aligned}$$

So,

- Precisions add
- Means are precision-weighted and added

# Bayes rule for Gaussians

For the prior (blue) $m_0 = 20$, $\lambda_0 = 1$ and the likelihood (red) $m_D = 25$ and $\lambda_D = 3$, the posterior (magenta) shows the posterior distribution with $m = 23.75$ and $\lambda = 4$.

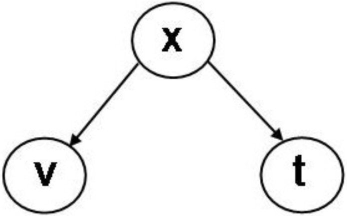



The posterior is closer to the likelihood because the likelihood has higher precision.

# Sensory Integration

Ernst and Banks (2002) asked subjects which of two sequentially presented blocks was the taller. Subjects used either vision alone, touch alone or a combination of the two.

If vision $v$ and touch $t$ information are independent given




an object $x$ then we have

$$p(v,t,x) = p(v|x)p(t|x)p(x)$$

Bayesian fusion of sensory information then produces a posterior density

$$p(x|v,t) = \frac{p(v|x)p(t|x)p(x)}{p(v,t)}$$

# Sensory Integration

In the absence of prior information about block size (ie $p(x)$ is uniform), for Gaussian likelihoods, the posterior will also be a Gaussian with precision $\lambda_{vt}$. From Bayes rule for Gaussians we know that precisions add

$$\lambda_{vt} = \lambda_v + \lambda_t$$

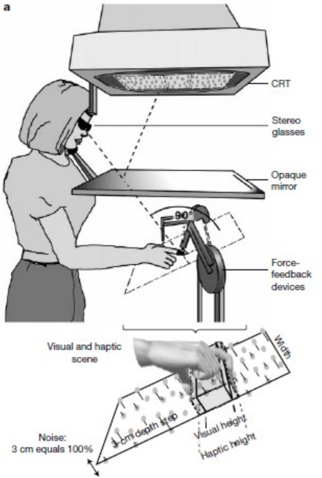
and the posterior mean is a relative-precision weighted combination

$$\begin{aligned} m_{vt} &= \frac{\lambda_v}{\lambda_{vt}} m_v + \frac{\lambda_t}{\lambda_{vt}} m_t \\ m_{vt} &= w_v m_v + w_t m_t \end{aligned}$$

with weights $w_v$ and $w_t$.

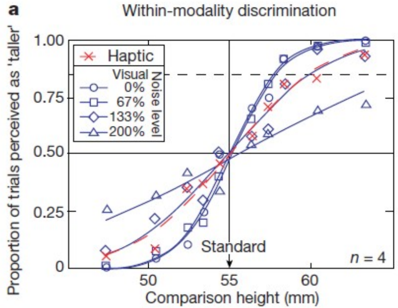
# Vision and Touch

*Ernst and Banks, Nature, 2002* asked subjects which of two sequentially presented blocks was the taller. Subjects used either vision alone, touch alone or a combination of the two.



# Vision and Touch Separately

They recorded the accuracy with which discrimination could be made and plotted this as a function of difference in block height. This was first done for each condition alone. One can then estimate precisions, $\lambda_v$ and $\lambda_t$ by fitting a cumulative Gaussian density function.

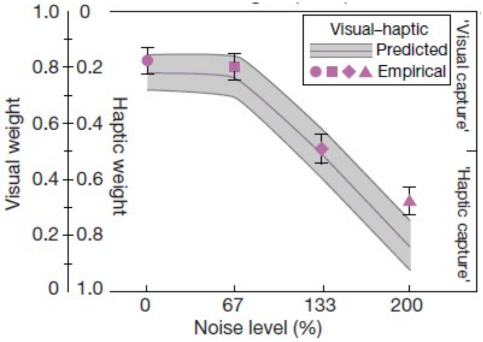



They manipulated the accuracy of the visual discrimination by adding noise onto one of the stereo images.

# Vision and Touch Together

Optimal fusion predicts weights from Bayes rule

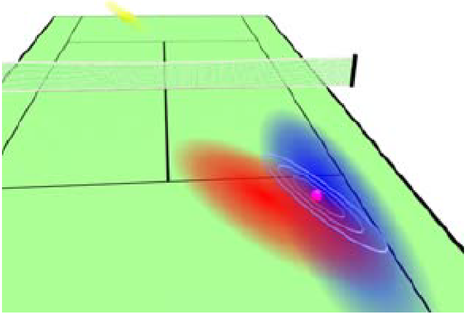
$$
\begin{aligned}
\lambda_{vt} &= \lambda_v + \lambda_t \\
m_{vt} &= \frac{\lambda_v}{\lambda_{vt}} m_v + \frac{\lambda_t}{\lambda_{vt}} m_t \\
m_{vt} &= w_v m_v + w_t m_t
\end{aligned}
$$




They observed visual capture at low levels of visual noise and haptic capture at high levels.

# Higher Dimensions

From *Wolpert and Ghahramani (2006)*



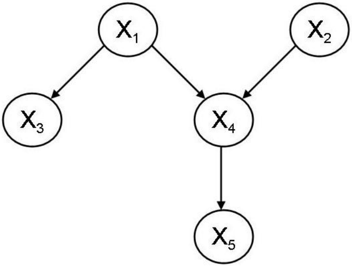
For Gaussian densities we have

$$
\begin{aligned}
\Lambda &= \Lambda_0 + \Lambda_D \\
m &= \Lambda^{-1}(\Lambda_0 m_0 + \Lambda_D m_D)
\end{aligned}
$$

with precision matrices $\Lambda$.

# Generative Models

For a probabilistic generative model




The joint probability of all variables, $x$, can be written down as

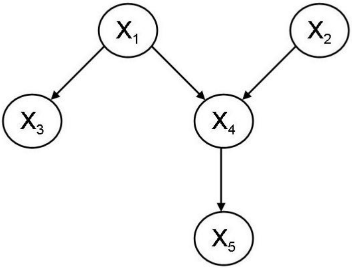
$$p(x) = \prod_{i=1}^{5} p(x_i | pa[x_i])$$

where $pa[x_i]$ are the parents of $x_i$.

# Joint Probability

A DAG specifies the joint probability of all variables.

$$p(x_1, x_2, x_3, x_4, x_5) = p(x_1)p(x_2)p(x_3|x_1)p(x_4|x_1, x_2)p(x_5|x_4)$$

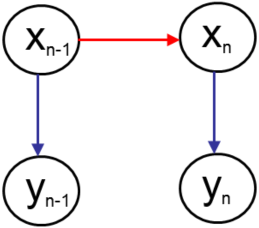



All other variables can be gotten from the joint probability via marginalisation. For later

$$GibbsEnergy = -\log p(x)$$

# Exact Inference

Exact Bayesian Inference is not possible for interesting models.

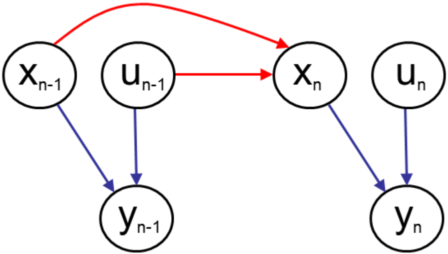



Hidden state $x_n$, Observations $y_n$

For Nonlinear Dynamics or Nonlinear Observation functions.

# Exact Inference

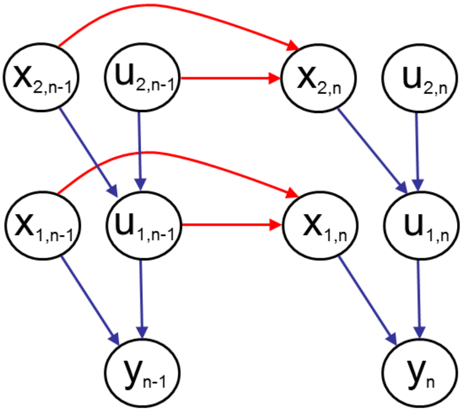
Exact Bayesian Inference is not possible for interesting models.




For Nonlinear Dynamics or Nonlinear Observation functions.

# Exact Inference

Exact Bayesian Inference is not possible for interesting models.




For Nonlinear Dynamics or Nonlinear Observation functions.

# Approximate Inference

There is one way implement exact Bayesian inference, but many methods for approximate inference. How should we quantify approximate ?

True posterior $p(x)$, approximate posterior $q(x)$.

For densities $q(x)$ and $p(x)$ the Kullback-Liebler (KL) divergence from $q$ to $p$ is

$$KL[q||p] = \int q(x) \log \frac{q(x)}{p(x)} dx$$

See *Neal and Hinton, Kluwer, 1993; Dayan et al. Neural Comp, 1995; Mackay, NIPS, 1995*.

# Kullback-Liebler Divergence

For densities $q(x)$ and $p(x)$ the Kullback-Liebler (KL) divergence from $q$ to $p$ is

$$KL[q||p] = \int q(x) \log \frac{q(x)}{p(x)} dx$$

The KL-divergence satisfies Gibbs' inequality

$$KL[q||p] \geq 0$$

with equality only if $q = p$.

In general $KL[q||p] \neq KL[p||q]$, so KL is not a distance measure. See *Mackay, Information Theory, 2003.*

Which should we use ?

# Univariate Gaussians





For Gaussians

$$
\begin{aligned}
p(x) &= \mathrm{N}(x; \mu_p, \sigma_p^2) \\
q(x) &= \mathrm{N}(x; \mu_q, \sigma_q^2)
\end{aligned}
$$

we have

$$
KL(q||p) = \frac{(\mu_q - \mu_p)^2}{2\sigma_p^2} + \frac{1}{2} \log \left( \frac{\sigma_p^2}{\sigma_q^2} \right) + \frac{\sigma_q^2}{2\sigma_p^2} - \frac{1}{2}
$$

# Multivariate Gaussians

For Gaussians

$$\begin{aligned} p(x) &= \mathrm{N}(x; \mu_p, C_p) \\ q(x) &= \mathrm{N}(x; \mu_q, C_q) \end{aligned}$$

we have

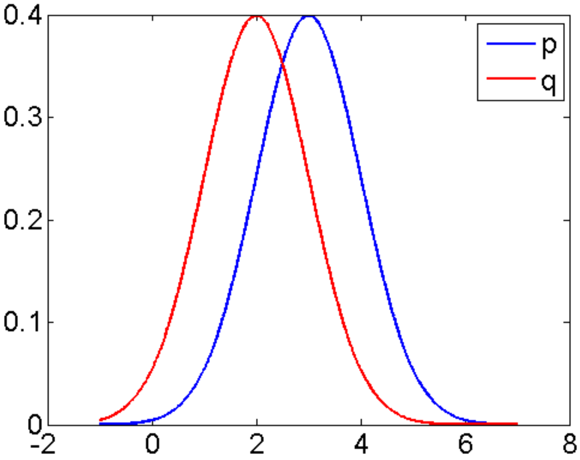
$$KL(q||p) = \frac{1}{2} e^T C_p^{-1} e + \frac{1}{2} \log \frac{|C_p|}{|C_q|} + \frac{1}{2} \mathrm{Tr}\left( C_p^{-1} C_q \right) - \frac{d}{2}$$

where $d = dim(x)$ and

$$e = \mu_q - \mu_p$$

# Same Variance - Symmetry

If $\sigma_q = \sigma_p$ then $KL(q||p) = KL(p||q)$ eg. distributions that just have a different mean

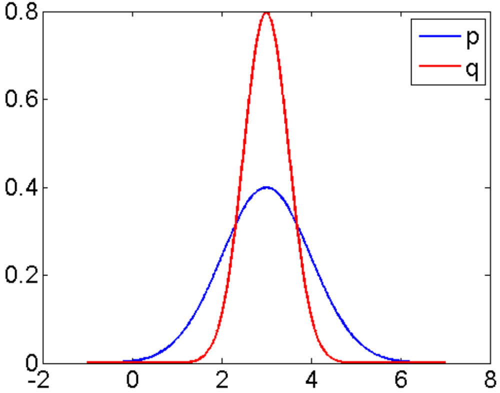



Here $KL(q||p) = KL(p||q) = 0.12$.

# Different Variance - Asymmetry

$$KL[q||p] = \int q(x) \log \frac{q(x)}{p(x)} dx$$

If $\sigma_q \neq \sigma_p$ then $KL(q||p) \neq KL(p||q)$

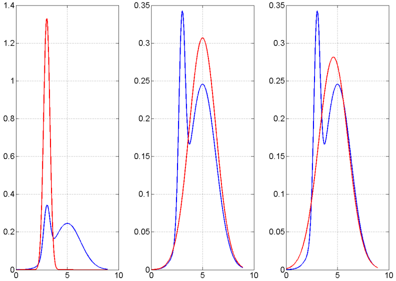



Here $KL(q||p) = 0.32$ but $KL(p||q) = 0.81$.

# Approximating multimodal with unimodal

True posterior $p$ (blue), approximate posterior $q$ (red).
Gaussian approx at mode is a Laplace approximation.

| | Left Mode | Right Mode | Moment Matched |
|---|---|---|---|
| KL(q,p) | 1.17 | 0.09 | 0.07 |
| KL(p,q) | 23.2 | 0.12 | 0.07 |

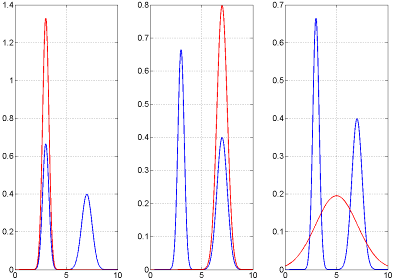



Minimising either KL produces the moment-matched solution.

# Distant Modes

True posterior $p$ (blue), approximate posterior $q$ (red).
Gaussian approx at mode is a Laplace approximation.

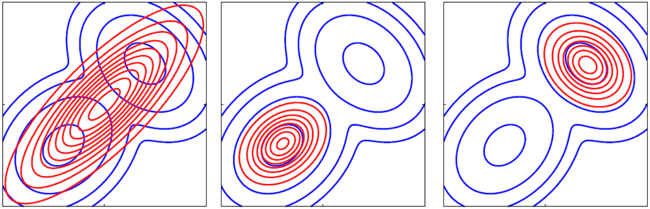
| | Left Mode | Right Mode | Moment Matched |
|---|---|---|---|
| KL(q,p) | 0.69 | 0.69 | 3.45 |
| KL(p,q) | 43.9 | 15.4 | 0.97 |




Minimising $KL(q||p)$ produces mode-seeking. Minimising $KL(p||q)$ produces moment-matching.

# Multiple dimensions

In higher dimensional spaces, unless modes are very close, minimising $KL(p||q)$ produces moment-matching (a) and minimising $KL(q||p)$ produces mode-seeking (b and c).



(a) (b) (c)

Minimising $KL(q||p)$ therefore seems desirable, but how do we do it if we don't know $p$ ?

Figure from *Bishop, Pattern Recognition and Machine Learning, 2006*

# Variational Bayes

Given a probabilistic model of some data, the log of the evidence can be written as

$$\begin{aligned}
\log p(Y) &= \int q(\theta) \log p(Y) d\theta \\
&= \int q(\theta) \log \frac{p(Y,\theta)}{p(\theta|Y)} d\theta \\
&= \int q(\theta) \log \left[ \frac{p(Y,\theta) q(\theta)}{q(\theta) p(\theta|Y)} \right] d\theta \\
&= \int q(\theta) \log \left[ \frac{p(Y,\theta)}{q(\theta)} \right] d\theta \\
&+ \int q(\theta) \log \left[ \frac{q(\theta)}{p(\theta|Y)} \right] d\theta
\end{aligned}$$

where $q(\theta)$ is the approximate posterior. Hence

$$\log p(Y) = -F + KL(q(\theta)||p(\theta|Y))$$

# Free Energy

We have

$$F = -\int q(\theta) \log \frac{p(Y,\theta)}{q(\theta)} d\theta$$

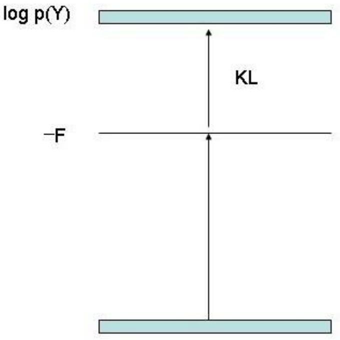
which in statistical physics is known as the variational free energy. We can write

$$F = -\int q(\theta) \log p(Y,\theta) d\theta - \int q(\theta) \log \frac{1}{q(\theta)} d\theta$$

This is an energy term, minus an entropy term, hence 'free energy'.

# Variational Free Energy

Because $KL$ is always positive, due to the Gibbs inequality, $-F$ provides a lower bound on the model evidence. Moreover, because $KL$ is zero when two densities are the same, $-F$ will become equal to the model evidence when $q(\theta)$ is equal to the true posterior. For this reason $q(\theta)$ can be viewed as an *approximate posterior*.




$$\log p(Y) = -F + KL[q(\theta)||p(\theta|Y)]$$

# Factorised Approximations

To obtain a practical learning algorithm we must also ensure that the integrals in $F$ are tractable. One generic procedure for attaining this goal is to assume that the approximating density factorizes over groups of parameters. In physics, this is known as the mean field approximation. Thus, we consider:

$$q(\theta) = \prod_i q(\theta_i)$$

where $\theta_i$ is the $i$th group of parameters. We can also write this as

$$q(\theta) = q(\theta_i) q(\theta_{\setminus i})$$

where $\theta_{\setminus i}$ denotes all parameters *not* in the $i$th group.

# Approximate Posteriors

We define the variational energy for the *i*th partition as

$$I(\theta_i) = -\int q(\theta_{\backslash i}) \log p(Y, \theta) d\theta_{\backslash i}$$

It is the Gibbs Energy (from earlier) averaged over other ensembles. Then the free energy is minimised when

$$q(\theta_i) = \frac{\exp[I(\theta_i)]}{Z}$$

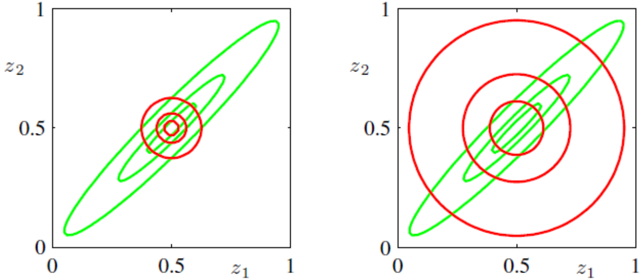
where $Z$ is the normalisation factor needed to make $q(\theta_i)$ a valid probability distribution.

# Factorised Approximations

For

$$q(z) = q(z_1)q(z_2)$$

minimising $KL(q,p)$ where $p$ is green and $q$ is red produces left plot, where minimising $KL(p,q)$ produces right plot.

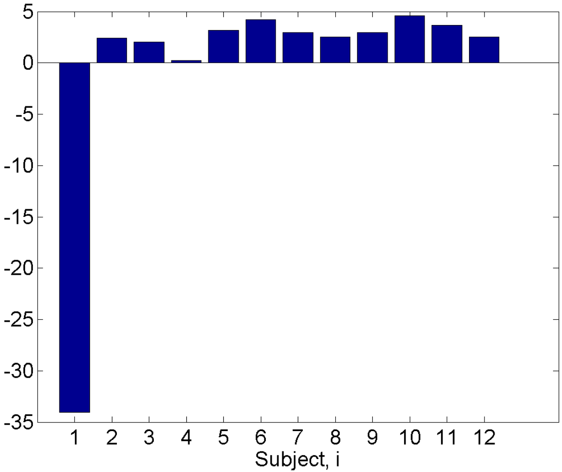



Hence minimising free energy tends to produce approximations on left rather than right. That is, uncertainty can be underestimated in some directions. Implications for FEP ?

# Group Model Inference

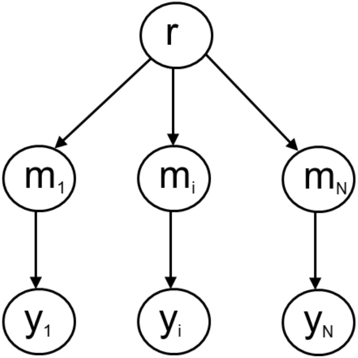
## Log Bayes Factor in favour of model 2

$$\log \frac{p(y_i|m_i = 2)}{p(y_i|m_i = 1)}$$



# Group Model Inference

Model frequencies $r_k$, model assignments $m_i$, subject data $y_i$.

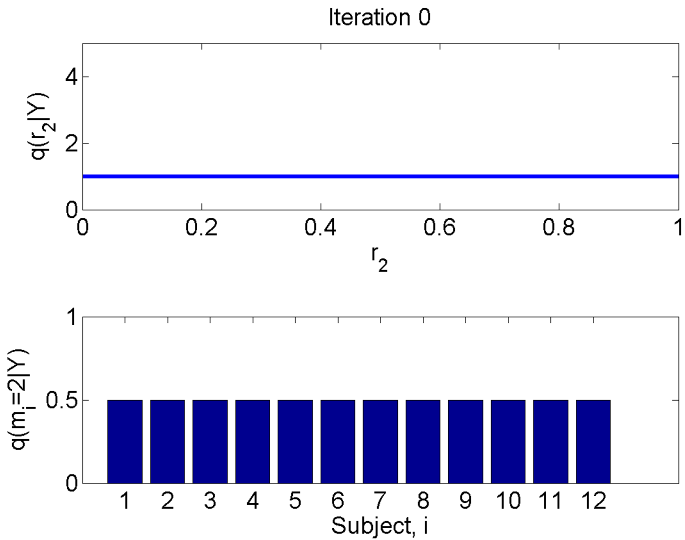



Approximate posterior

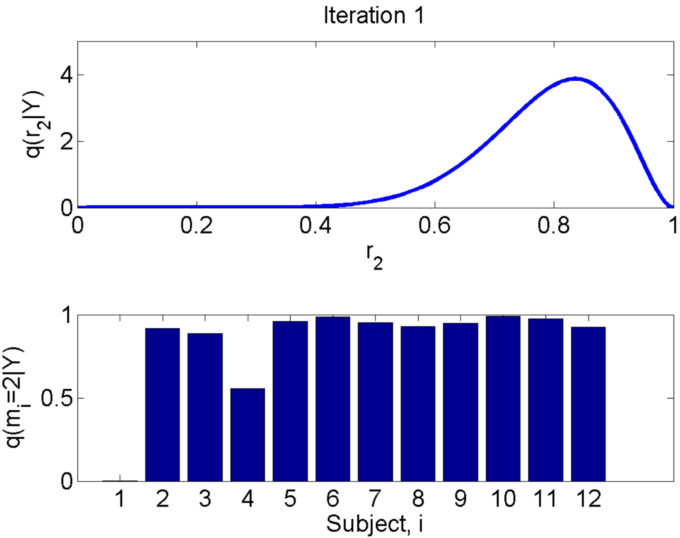
$$q(r, m|Y) = q(r|Y)q(m|Y)$$

*Stephan, Neuroimage, 2009.*

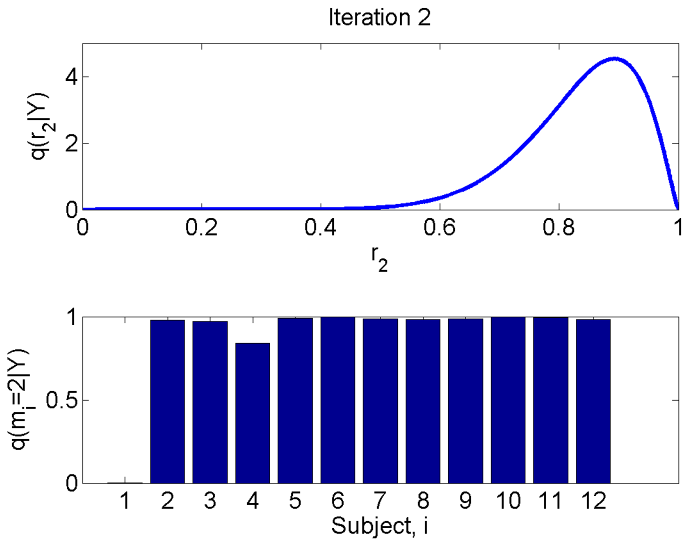
# Group Model Inference



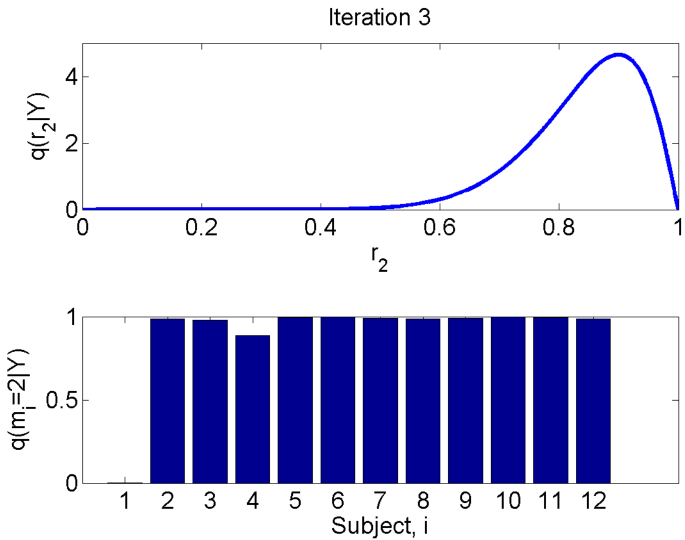
# Group Model Inference



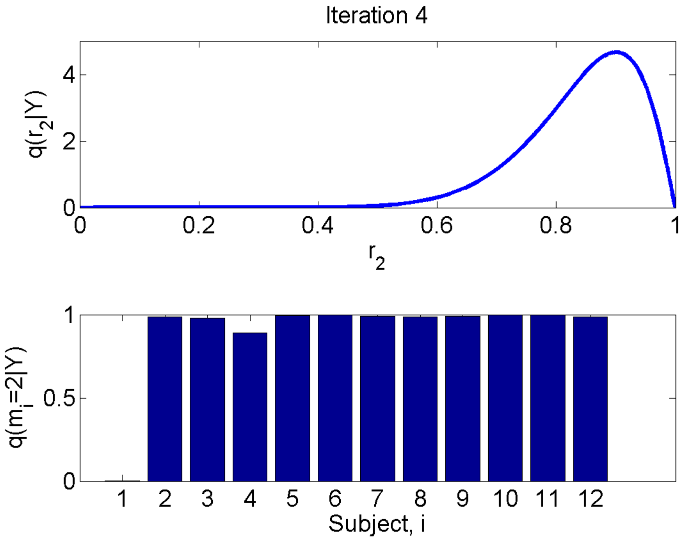
# Group Model Inference



# Group Model Inference



# Group Model Inference



# Applications

Variational Inference has been applied to

- Hidden Markov Models (*Mackay, Cambridge, 1997*)
- Graphical Models (*Jordan, Machine Learning, 1999*)
- Logistic Regression (*Jaakola and Jordan, Stats and Computing, 2000*)
- Gaussian Mixture Models, (*Attias, UAI, 1999*)
- Independent Component Analysis, (*Attias, UAI, 1999*)
- Dynamic Trees, (*Storkey, UAI, 2000*)

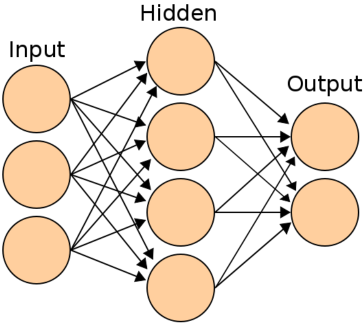
# Applications

- Relevance Vector Machines, (*Bishop and Tipping, 2000*)
- Linear Dynamical Systems (*Ghahramani and Beal, NIPS, 2001*)
- Nonlinear Autoregressive Models (*Roberts and Penny, IEEE SP, 2002*)
- Canonical Correlation Analysis (*Wang, IEEE TNN, 2007*)
- Dynamic Causal Models (*Friston, Neuroimage, 2007*)
- Nonlinear Dynamic Systems (*Daunizeau, PRL, 2009*)

# Penalised Model Fitting

We can write

$$F = -\int q(\theta) \log p(Y|\theta) d\theta + \int q(\theta) \log \frac{q(\theta)}{p(\theta)} d\theta$$




Replace point estimate $\theta$ with an ensemble $q(\theta)$. Keep parameters $\theta$ imprecise by penalizing distance from a prior $p(\theta)$, as measured by KL-divergence.

See *Hinton and van Camp, COLT, 1993*

# Model comparison





The (negative) free energy, being an approximation to the model evidence, can also be used for model comparison. See for example

- Graphical models (*Beal, PhD Gatsby, 2003*)
- Linear dynamical systems (*Ghahramani and Beal, NIPS, 2001*)
- Nonlinear autoregressive models (*Roberts and Penny, IEEE SP, 2002*)
- Hidden Markov Models (*Valente and Wellekens, ICSLP 2004*)
- Dynamic Causal Models (*Penny, Neuroimage, 2011*)

# Generic Approaches

VB for generic models

- *Winn and Bishop, Variational Message Passing, JLMR, 2005*
- *Wainwright and Jordan, A Variational Principle for Graphical Models, 2005*
- *Friston et al. Dynamic Expectation Maximisation, Neuroimage, 2008*

For more see

- http://en.wikipedia.org/wiki/Variational-Bayesian-methods
- http://www.variational-bayes.org/
- http://www.cs.berkeley.edu/jordan/variational.html