# Variational Inference

Will Penny

Bayesian Inference Course,
WTCN, UCL, March 2013

# Information

Shannon (1948) asked how much information is received when we observe a specific value of the variable $x$ ?

If an unlikely event occurs then one would expect the information to be greater. So information must be inversely proportional to $p(x)$, and monotonic.

Shannon also wanted a definition of information such that if $x$ and $y$ are independent then the total information would sum

$$h(x_i, y_j) = h(x_i) + h(y_j)$$

Given that we know that in this case

$$p(x_i, y_j) = p(x_i)p(y_j)$$

then we must have

$$h(x_i) = \log \frac{1}{p(x_i)}$$

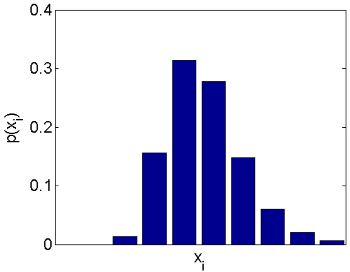
This is the self-information or surprise.

# Entropy

The entropy of a random variable is the average surprise. For discrete variables

$$H(x) = \sum_i p(x_i) \log \frac{1}{p(x_i)}$$

The uniform distribution has maximum entropy.

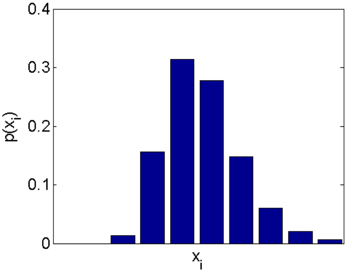



A single peak has minimum entropy. We define

$$0 \log 0 = 0$$

If we take logs to the base 2, entropy is measured in bits.

# Source Coding Theorem




Assigning code-words of length $h(x_i)$ to each symbol $x_i$ results in the maximum rate of information transfer in a noiseless channel. This is the Source Coding Theorem (Shannon, 1948).

$$h(x_i) = \log \frac{1}{p(x_i)}$$

If channel is noisy, see Noisy Channel Coding Theorem (Mackay, 2003)

# Prefix Codes

No code-word is a prefix of another. Use number of bits $b(x_i) = ceil(h(x_i))$. We have

$$
\begin{aligned}
h(x_i) &= \log_2 \frac{1}{p(x_i)} \\
b(x_i) &= \log_2 \frac{1}{q(x_i)}
\end{aligned}
$$

Hence, each code-word has equivalent

$$q(x_i) = 2^{-b(x_i)}$$

| i | $p(x_i)$ | $h(x_i)$ | $b(x_i)$ | $q(x_i)$ | CodeWord |
|---|---|---|---|---|---|
| 1 | 0.016 | 5.97 | 6 | 0.016 | 101110 |
| 2 | 0.189 | 2.43 | 3 | 0.125 | 100 |
| 3 | 0.371 | 1.43 | 2 | 0.250 | 00 |
| 4 | 0.265 | 1.92 | 2 | 0.250 | 01 |
| 5 | 0.115 | 3.12 | 4 | 0.063 | 1010 |
| 6 | 0.035 | 4.83 | 5 | 0.031 | 10110 |
| 7 | 0.010 | 6.67 | 7 | 0.008 | 1011110 |
| 8 | 0.003 | 8.53 | 9 | 0.002 | 101111100 |

# Relative Entropy

Average length of code word

$$
\begin{aligned}
B(x) &= \sum_i p(x_i) b(x_i) \\
&= \sum_i p(x_i) \log \frac{1}{q(x_i)} = 2.65 bits
\end{aligned}
$$

Entropy

$$
\begin{aligned}
H(x) &= \sum_i p(x_i) h(x_i) \\
&= \sum_i p(x_i) \log \frac{1}{p(x_i)} = 2.20 bits
\end{aligned}
$$

Difference is relative entropy

$$
\begin{aligned}
KL(p||q) &= B(x) - H(x) \\
&= \sum_i p(x_i) \log \frac{p(x_i)}{q(x_i)} \\
&= 0.45 bits
\end{aligned}
$$

# Continuous Variables

For continuous variables the (differential) entropy is

$$H(x) = \int p(x) \log \frac{1}{p(x)} dx$$

Out of all distributions with mean $m$ and standard deviation $\sigma$ the Gaussian distribution has the maximum entropy. This is

$$H(x) = \frac{1}{2}(1 + \log 2\pi) + \frac{1}{2} \log \sigma^2$$

# Relative Entropy

We can write the Kullback-Liebler (KL) divergence

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

as a difference in entropies

$$KL(q||p) = \int q(x) \log \frac{1}{p(x)} dx - \int q(x) \log \frac{1}{q(x)} dx$$

This is the average surprise assuming information is encoded under $p(x)$ minus the average surprise under $q(x)$. Its the extra number of bits/nats required to transmit messages.

# 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

$$e = \mu_q - \mu_p$$

# Asymmetry

For densities $q(x)$ and $p(x)$ the Relative Entropy or 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

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

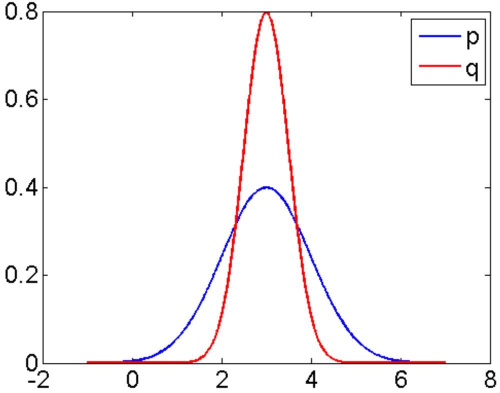
with equality only if $q = p$.

In general $KL[q||p] \neq KL[p||q]$, so KL is not a distance measure.

# Different Variance - Asymmetry

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

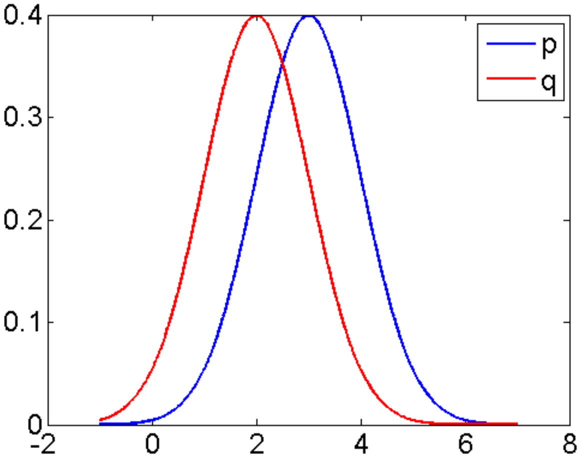
If $\sigma_q \neq \sigma_p$ then $KL(q||p) \neq KL(p||q)$




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

# Same Variance - Symmetry

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




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

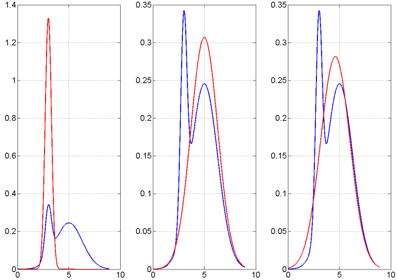
# Approximating multimodal with unimodal





We approximate the density $p$ (blue), which is a Gaussian mixture, with a Gaussian density $q$ (red).

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

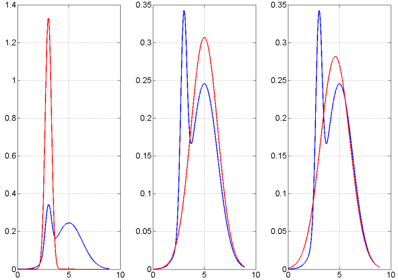


Minimising either KL produces the moment-matched solution.

# Approximate Bayesian Inference

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 |

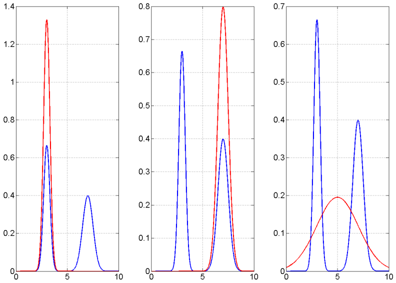


Minimising either KL produces the moment-matched solution.

# Distant Modes

We approximate the density $p$ (blue), which is a Gaussian mixture, with a Gaussian density $q$ (red).

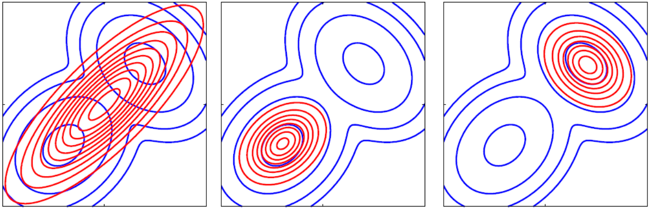
| | Left Mode | Right Mode | Moment Matched |
|---|---|---|---|
| KL(q,p) | 0.69 | 0.69 | 3.45 |
| KL(p,q) | 43.9 | 15.4 | 0.97 |




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).



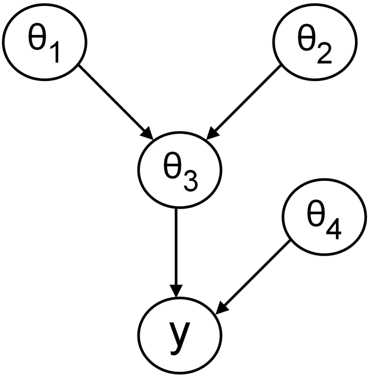
(a) (b) (c)

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*

# Joint Probability

$$p(Y,\theta) = p(y|\theta_3,\theta_4)p(\theta_3|\theta_2,\theta_1)p(\theta_1)p(\theta_2)p(\theta_4)$$




Energy

$$E = -\log p(Y,\theta)$$

# Model Evidence

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

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

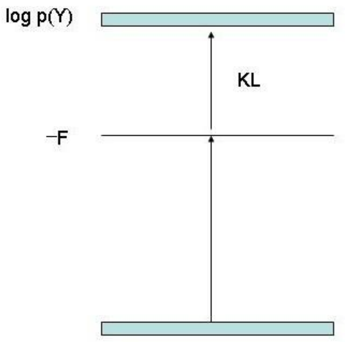
This is an energy term, minus an entropy term, hence 'free energy'.

# 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 Energy 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.

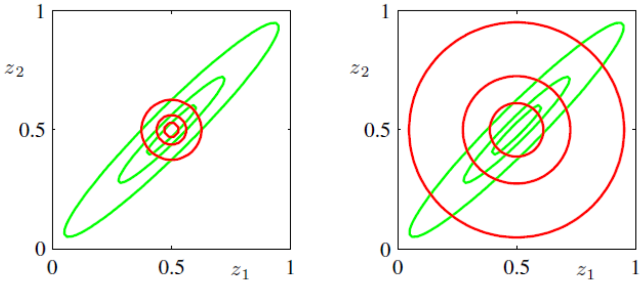
For proof see Bishop (2006) or SPM book. Think of above two equations as an approximate version of Bayes rule.

# 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 produces approximations on left rather than right. That is, uncertainty is underestimated. See Minka (2005) for other divergences.

# Example

## Approximate posteriors

$$
\begin{aligned}
q(r|Y) &= \mathrm{Dir}(r;\alpha) \\
q(m|Y) &= \prod_{i=1}^{N}\prod_{k=1}^{K} g_{ik}^{m_{ik}}
\end{aligned}
$$

## Update $q(m|Y)$:

$$
\begin{aligned}
u_{ik} &= \exp\left[\log p(y_i|k) + \psi(\alpha_k) - \sum_{k'}\psi(\alpha_{k'})\right] \\
g_{ik} &= \frac{u_{ik}}{\sum_{k'} u_{ik'}}
\end{aligned}
$$

## Update $q(r|Y)$:

$$
\alpha_k = \alpha_k^0 + \sum_i g_{ik}
$$

Sufficient statistics of approximate posteriors are coupled. Update and iterate - see later.

# 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*)

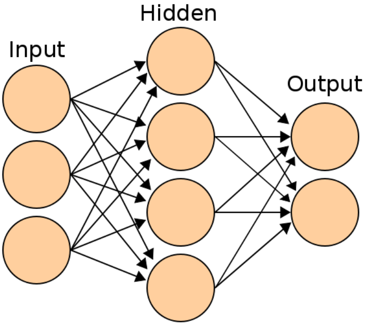
# Applications

- 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*

# Penalised Model Fitting

We can write

$$
\begin{aligned}
F &= -\int q(\theta) \log p(Y|\theta) d\theta + \int q(\theta) \log \frac{q(\theta)}{p(\theta)} d\theta \\
-F &= \int q(\theta) \log p(Y|\theta) d\theta - \int q(\theta) \log \frac{q(\theta)}{p(\theta)} d\theta \\
-F &= Accuracy - Complexity
\end{aligned}
$$

# Model comparison

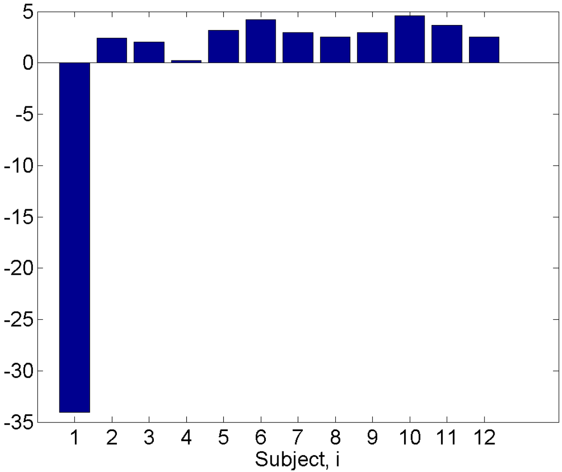
The negative free energy, being an approximation to the model evidence, can also be used for model comparison. See for example

- 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*)

# Group Model Inference

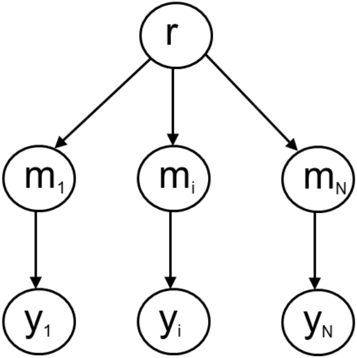
Log Bayes Factor in favour of model 2

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



# 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

## Approximate posteriors

$$
\begin{aligned}
q(r|Y) &= \mathrm{Dir}(r;\alpha) \\
q(m|Y) &= \prod_{i=1}^{N}\prod_{k=1}^{K} g_{ik}^{m_{ik}}
\end{aligned}
$$

## Update $q(m|Y)$:

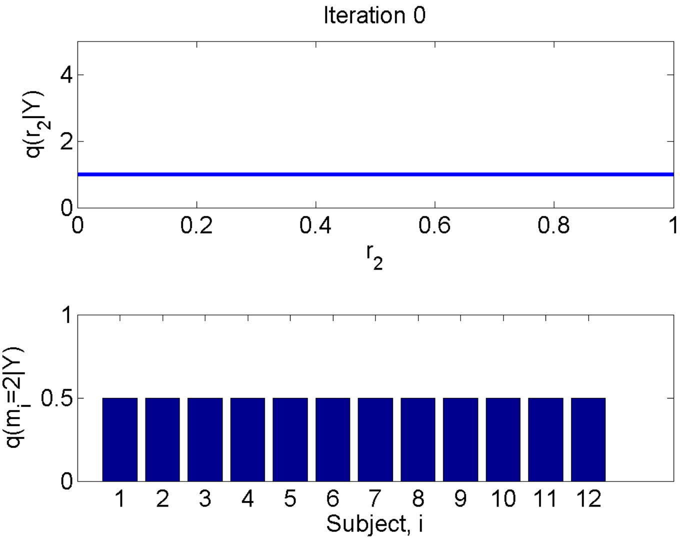
$$
\begin{aligned}
u_{ik} &= \exp\left[\log p(y_i|k) + \psi(\alpha_k) - \sum_{k'} \psi(\alpha_{k'})\right] \\
g_{ik} &= \frac{u_{ik}}{\sum_{k'} u_{ik'}}
\end{aligned}
$$

## Update $q(r|Y)$:

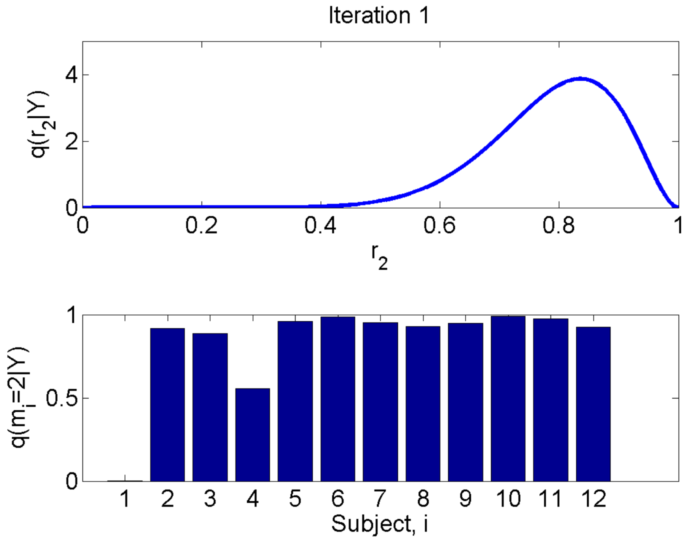
$$
\alpha_k = \alpha_k^0 + \sum_i g_{ik}
$$

Here $\log p(y_i|k)$ is the entry in the log evidence table from the $i$th subject (row) and $k$th model (column).
The quantity $g_{ik}$ is the posterior probability that subject $i$ used the $k$th model.

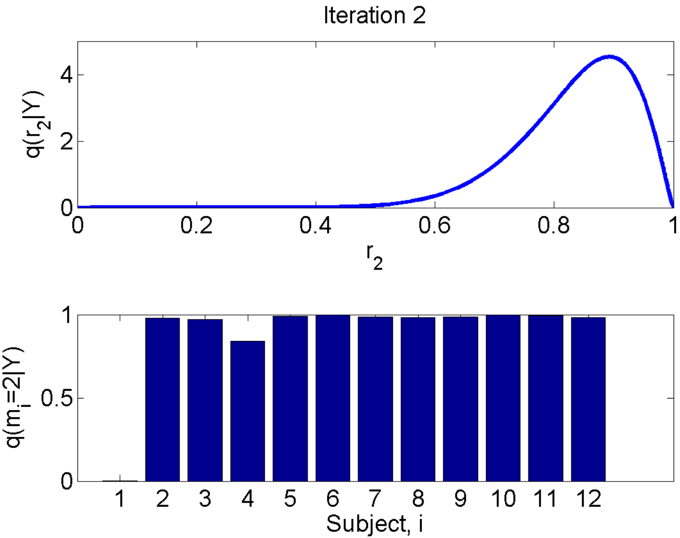
# Group Model Inference



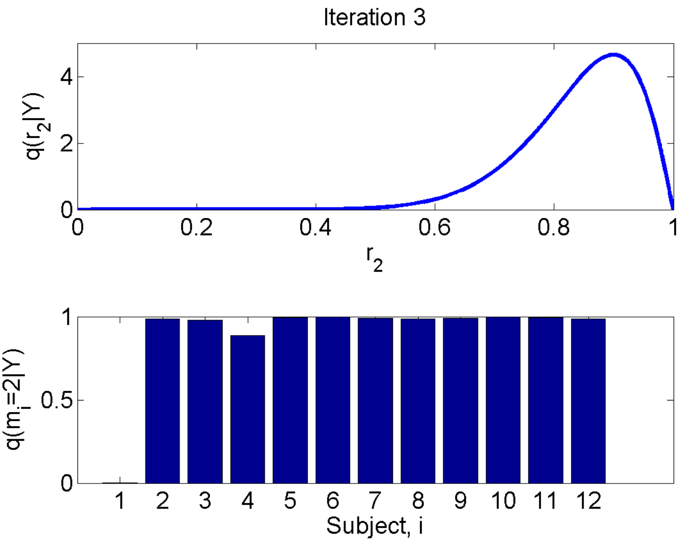
# Group Model Inference



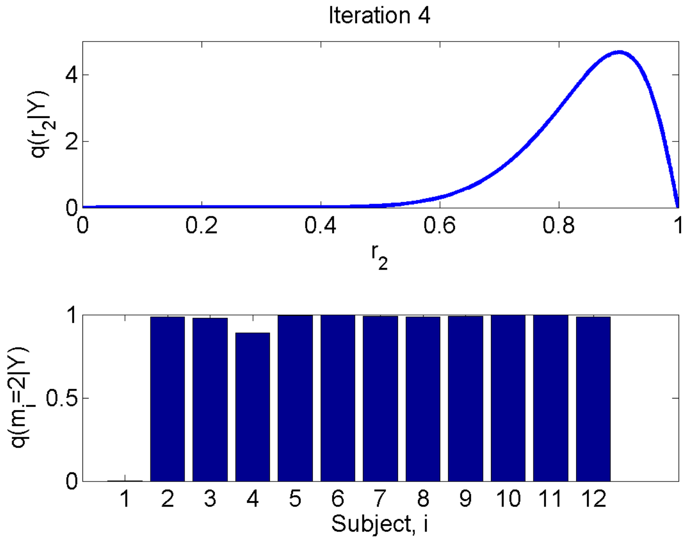
# Group Model Inference



# Group Model Inference



# Group Model Inference



# 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

# Summary

Entropy:

$$H(\theta) = \int q(\theta) \log \frac{1}{q(\theta)} d\theta$$

KL-Divergence:

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

Energy:

$$E = -\log p(Y, \theta)$$

Free Energy is Energy minus Entropy:

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

Model Evidence is Negative Free Energy + KL:

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

Negative Free Energy is Accuracy minus Complexity:

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

# References

M. Beal (2003) PhD Thesis. Gatsby Computational Neuroscience Unit, UCL.

C. Bishop (2006) Pattern Recognition and Machine Learning, Springer.

G. Deco et al. (2008) The Dynamic Brain: From Spiking Neurons to Neural Masses and Cortical Fields. PLoS CB 4(8), e1000092.

D. Mackay (2003) Information Theory, Inference and Learning Algorithms, Cambridge.

T. Minka et al. (2005) Divergence Measures and Message Passing. Microsoft Research Cambridge.

S. Roberts and W. Penny (2002). Variational Bayes for generalised autoregressive models. IEEE transactions on signal processing. 50(9), 2245-2257.

W. Penny (2006) Variational Bayes. In SPM Book, Elsevier.

D. Valente and C. Wellekens (2004) Scoring unknown speaker clustering: VB versus BIC. ICSLP 2004, Korea.