The Reverend Thomas Bayes
(1702-1761)

Bayesian Inference

Sudhir Shankar Raman
Translational Neuromodeling Unit,
UZH & ETH

With many thanks for some slides to:
Klaas Enno Stephan & Kay H. Brodersen

Why do I need to learn about Bayesian stats?

Because **SPM** is getting more and more Bayesian:

- Segmentation & spatial normalisation
- Posterior probability maps (PPMs)
- Dynamic Causal Modelling (DCM)
- Bayesian Model Selection (BMS)
- EEG: source reconstruction

Bayesian Inference

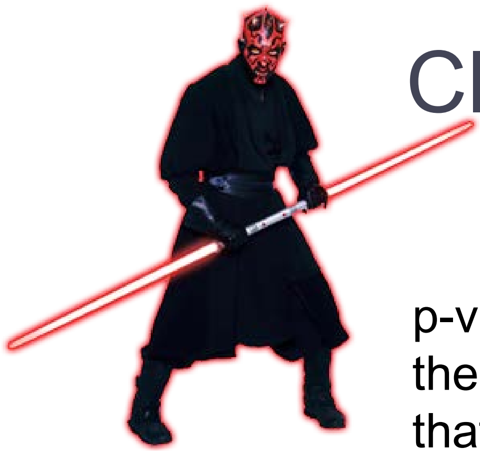
Model Selection

Approximate Inference

Variational Bayes

MCMC Sampling

Bayesian Inference



Classical and Bayesian statistics

p-value: probability of getting the observed data in the effect's absence. If small, reject null hypothesis that there is no effect.

Probability of observing the data *y*, given no effect ($\theta = 0$).

$$H_0 : \theta = 0$$

$$p(y \mid H_0)$$

- ⇨ One can never accept the null hypothesis
- ⇨ Given enough data, one can always demonstrate a significant effect
- ⇨ Correction for multiple comparisons necessary

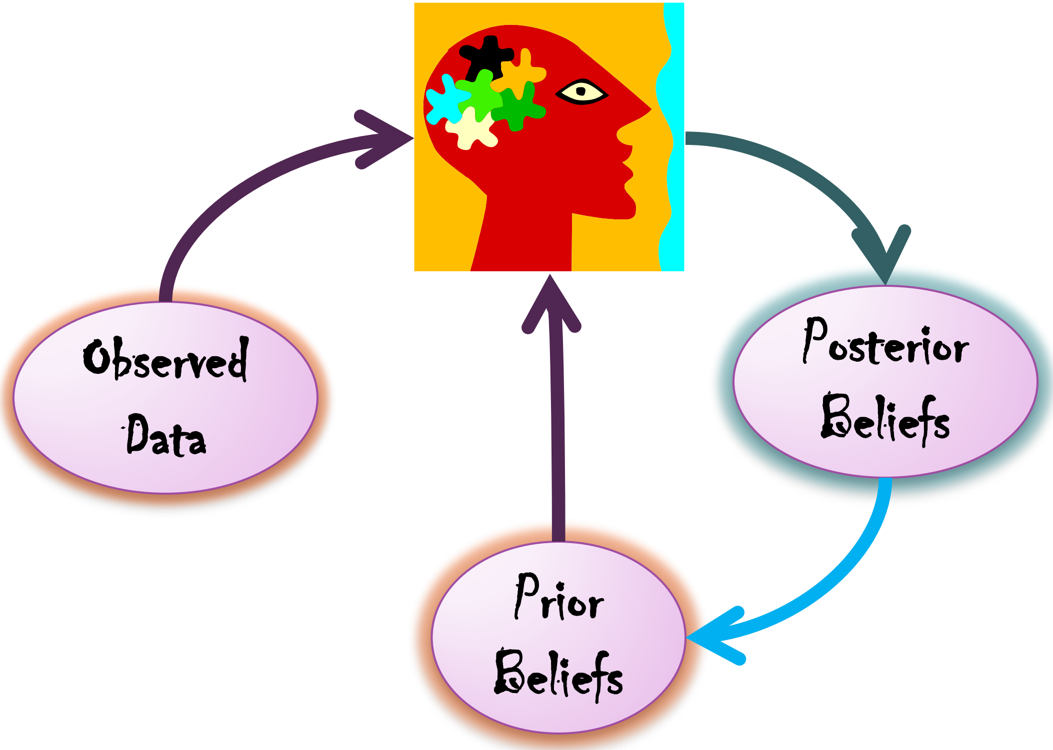
Bayesian Inference

- ⇨ Flexibility in modelling $p(y, \theta)$
- ⇨ Incorporating prior information $p(\theta)$
- ⇨ Posterior probability of effect $p(\theta \mid y)$
- ⇨ Options for model comparison

Statistical analysis and the illusion of objectivity. James O. Berger, Donald A. Berry

Bayes' Theorem

Reverend Thomas Bayes
1702 - 1761



"Bayes' theorem describes, how an ideally rational person processes information."

Bayes' Theorem

Given data *y* and parameters θ, the conditional probabilities are:

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

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

Eliminating $p(y,\theta)$ gives Bayes' rule:

Likelihood

Posterior

$$P(\theta \mid y) = \frac{p(y \mid \theta)\, p(\theta)}{p(y)}$$

Prior

Evidence

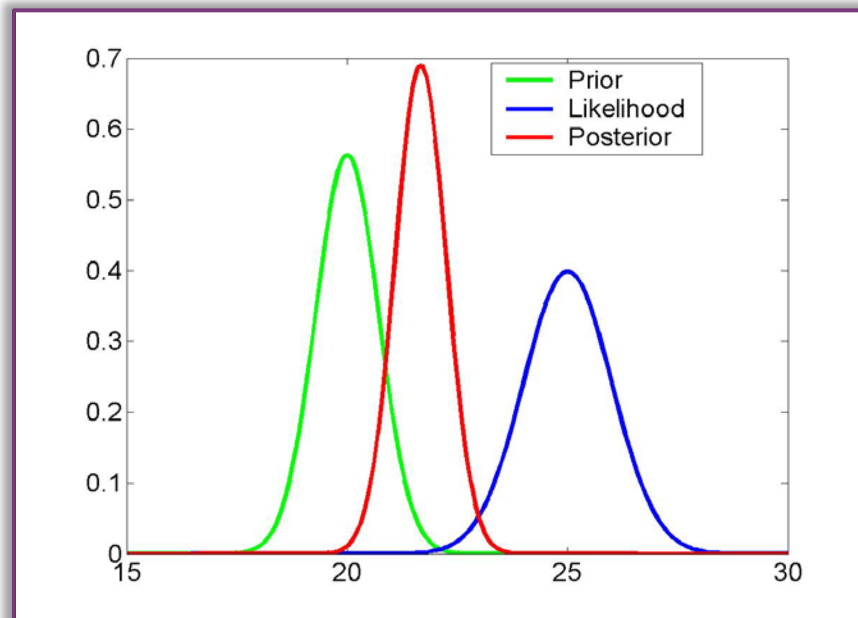
Bayesian statistics

new data: $p(y\,|\,\theta)$

prior knowledge: $p(\theta)$

$$p(\theta\,|\,y) \propto p(y\,|\,\theta)\,p(\theta)$$

posterior $\propto$ likelihood · prior

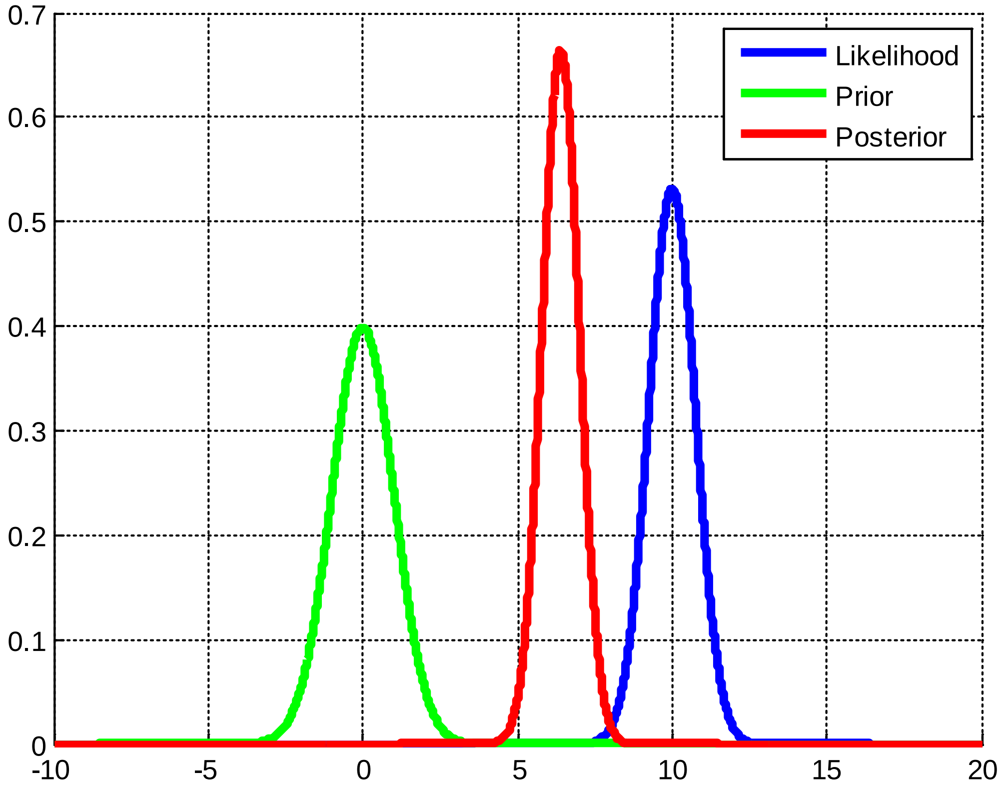


Bayes theorem allows one to formally incorporate prior knowledge into computing statistical probabilities.

Priors can be of different sorts: empirical, principled or shrinkage priors, uninformative.

The “posterior” probability of the parameters given the data is an optimal combination of prior knowledge and new data, weighted by their relative precision.

Bayes in motion - an animation



Principles of Bayesian inference

⇨ Formulation of a generative **model**

Model → Likelihood function $p(y|\theta)$

Model → prior distribution $p(\theta)$

⇨ Observation of **data**

Data → y

⇨ **Model Inversion - Update** of beliefs based upon observations, given a prior state of knowledge

$$p(\theta | y) \propto p(y | \theta) p(\theta)$$

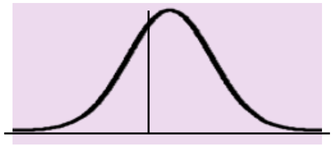
→ Maximum a posteriori (MAP)

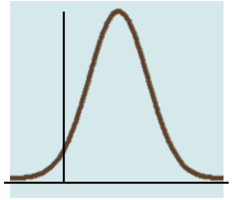
→ Maximum likelihood (ML)

Conjugate Priors

➪ Prior and Posterior have the same form

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





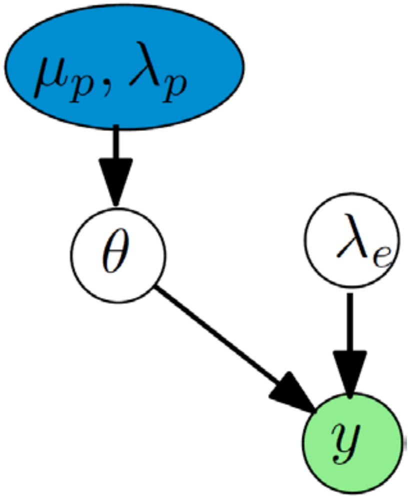
Same form !!

➪ Analytical expression.

➪ Conjugate priors for all exponential family members.

➪ Example – Gaussian Likelihood , Gaussian prior over mean

Gaussian Model



Likelihood & prior

$$p(y|\theta) = N(y|\theta, \lambda_e^{-1})$$

$$p(\theta) = N(\theta|\mu_p, \lambda_p^{-1})$$

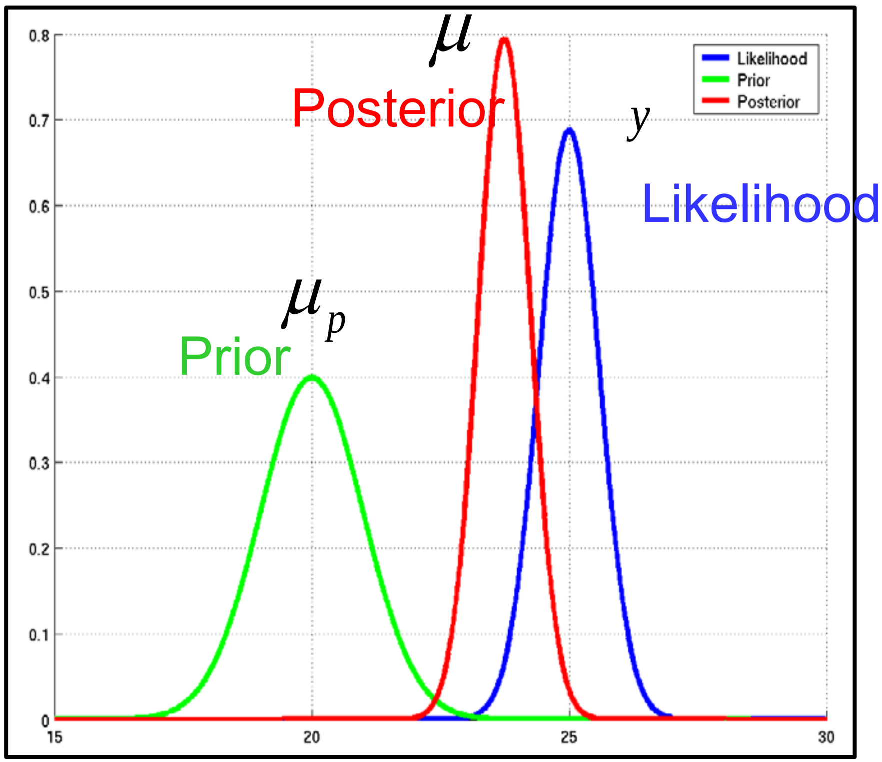
$$y = \theta + \varepsilon$$

Posterior: $p(\theta|y) = N(\theta|\mu, \lambda^{-1})$

$$\lambda = \lambda_e + \lambda_p$$

$$\mu = \frac{\lambda_e}{\lambda} y + \frac{\lambda_p}{\lambda} \mu_p$$

Relative precision weighting



Bayesian regression: univariate case

Normal densities

$$p(\theta) = N(\theta \mid \eta_p, \sigma_p^2)$$

$$p(y \mid \theta) = N(y \mid x\theta, \sigma_e^2)$$

$$p(\theta \mid y) = N(\theta \mid \eta_{\theta|y}, \sigma_{\theta|y}^2)$$

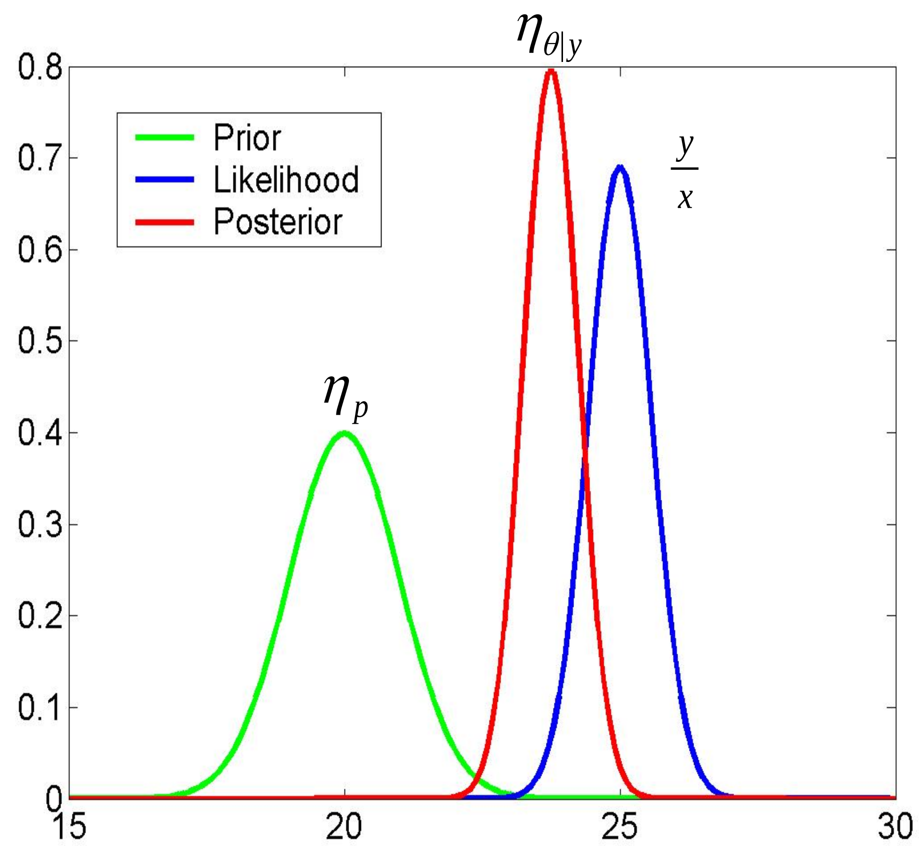
$$\frac{1}{\sigma_{\theta|y}^2} = \frac{x^2}{\sigma_e^2} + \frac{1}{\sigma_p^2}$$

$$\eta_{\theta|y} = \sigma_{\theta|y}^2 \left(\frac{x}{\sigma_e^2} y + \frac{1}{\sigma_p^2} \eta_p \right)$$

Relative precision weighting

Univariate linear model

$$y = x\theta + \varepsilon$$



Bayesian GLM: multivariate case

Normal densities

$$p(\boldsymbol{\theta}) = N(\boldsymbol{\theta}; \boldsymbol{\eta}_p, \mathbf{C}_p)$$

$$p(\mathbf{y} \mid \boldsymbol{\theta}) = N(\mathbf{y}; \mathbf{X}\boldsymbol{\theta}, \mathbf{C}_e)$$

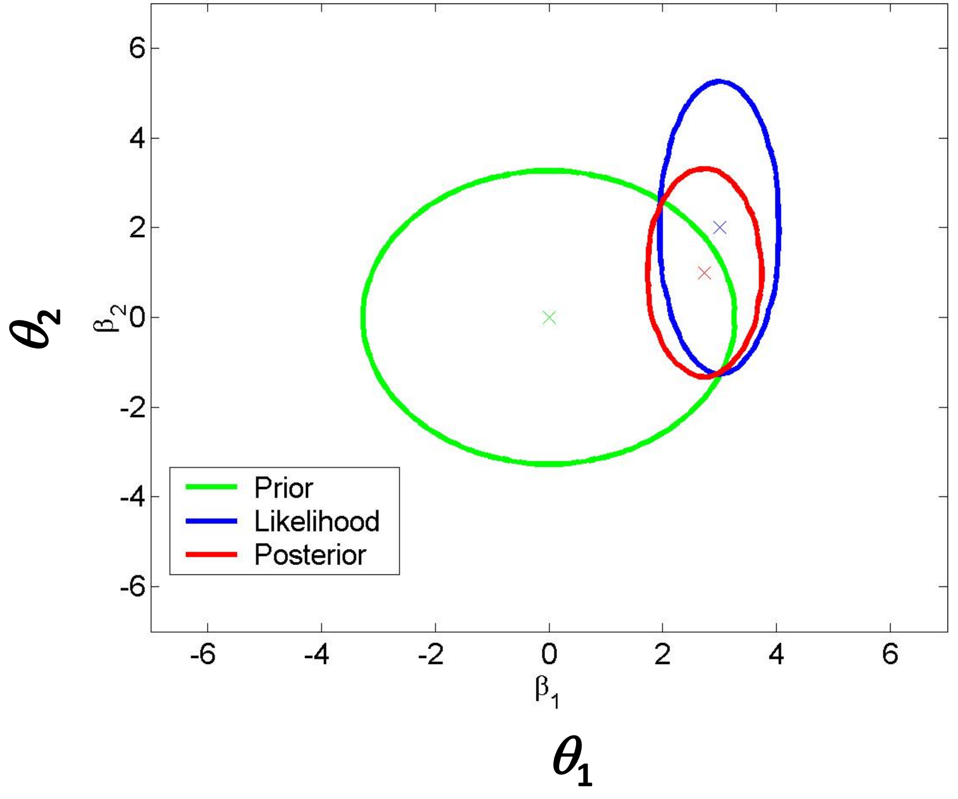
$$p(\boldsymbol{\theta} \mid y) = N(\boldsymbol{\theta}; \boldsymbol{\eta}_{\theta|y}, \mathbf{C}_{\theta|y})$$

$$\mathbf{C}_{\theta|y}{}^{-1} = \mathbf{X}^T \mathbf{C}_e{}^{-1} \mathbf{X} + \mathbf{C}_p{}^{-1}$$

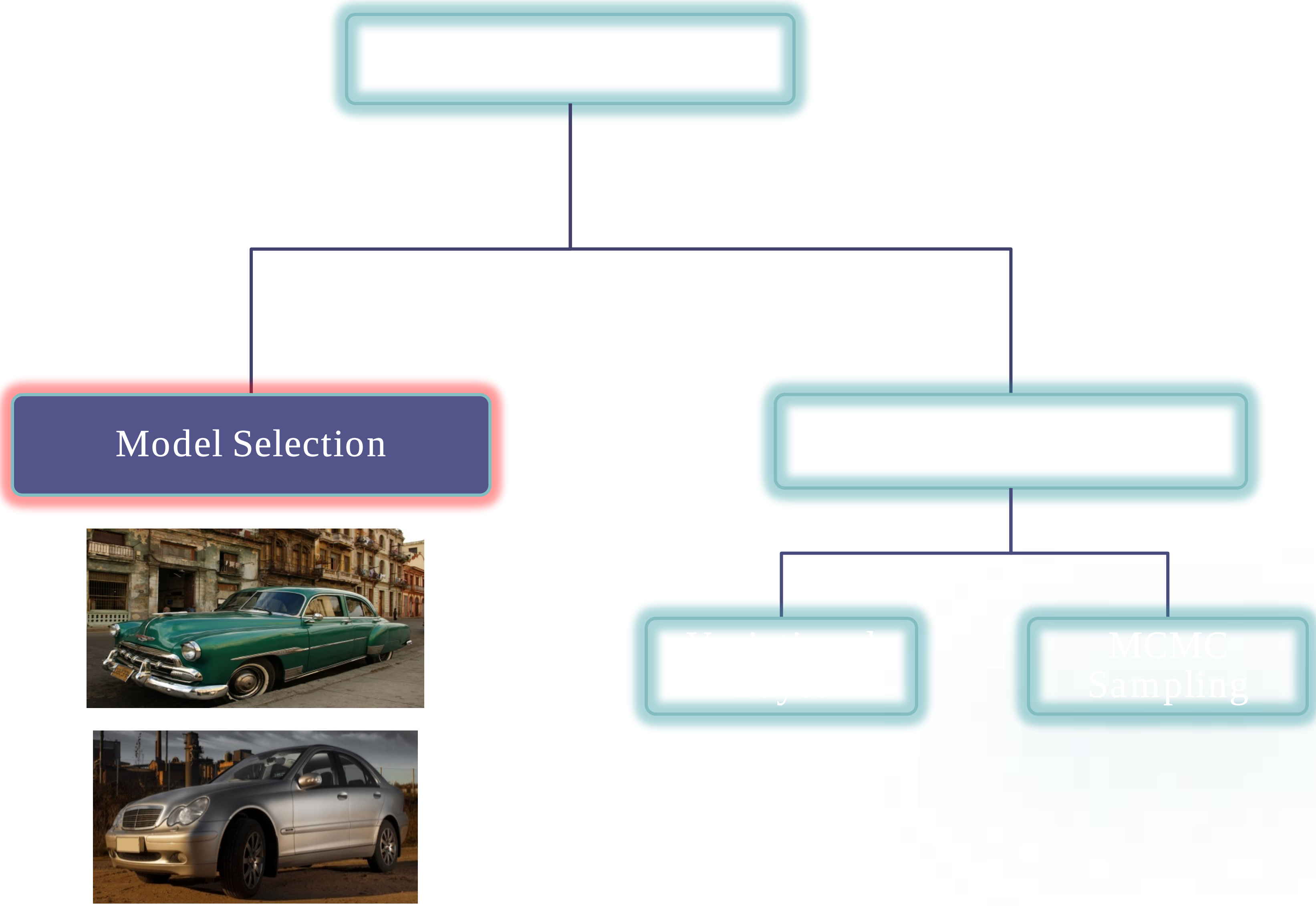
$$\boldsymbol{\eta}_{\theta|y} = \mathbf{C}_{\theta|y} \left(\mathbf{X}^T \mathbf{C}_e{}^{-1} \mathbf{y} + \mathbf{C}_p{}^{-1} \boldsymbol{\eta}_p \right)$$

General Linear Model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \mathbf{e}$$



- One step if C_e is known.
- Otherwise define conjugate prior or perform iterative estimation with EM.

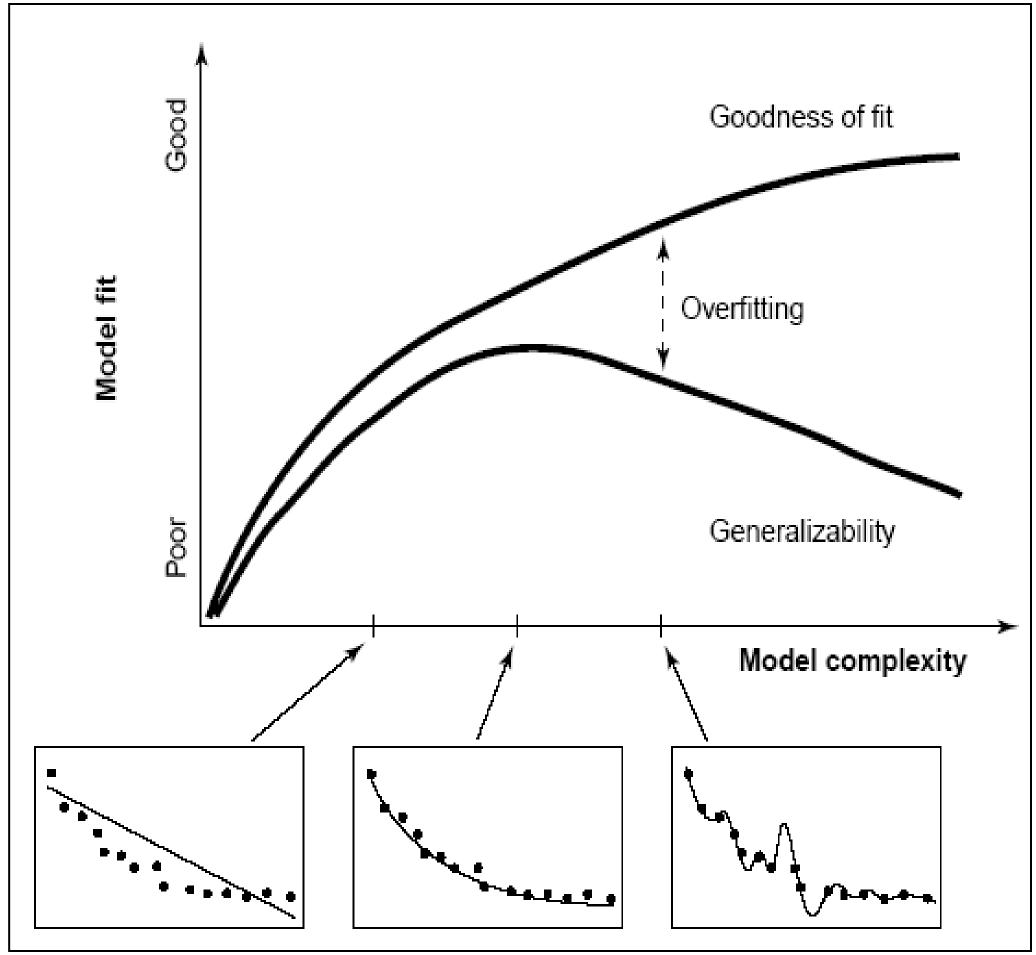

Model Selection
MCMC Sampling

Bayesian model selection (BMS)

Given competing hypotheses on structure & functional mechanisms of a system, which model is the best?

Which model represents the best **balance** between model fit and model complexity?

For which model m does p(y|m) become **maximal**?



Pitt & Miyung (2002), *TICS*

Bayesian model selection (BMS)

Bayes' rule: $$p(\theta | y, m) = \frac{p(y|\theta, m)\, p(\theta | m)}{p(y|m)}$$

Model evidence: $$p(y|m) = \int p(y|\theta, m) \cdot p(\theta|m)\, d\theta$$

- accounts for both accuracy and complexity of the model
- allows for inference about structure (generalizability) of the model

Model comparison via Bayes factor:

$$\frac{p(m_1|y)}{p(m_2|y)} = \frac{p(y|m_1)}{p(y|m_2)} \frac{p(m_1)}{p(m_2)}$$

$$BF = \frac{p(y|m_1)}{p(y|m_2)}$$

BF_{10}	Evidence against H_0
1 to 3	Not worth more than a bare mention
3 to 20	Positive
20 to 150	Strong
> 150	Decisive

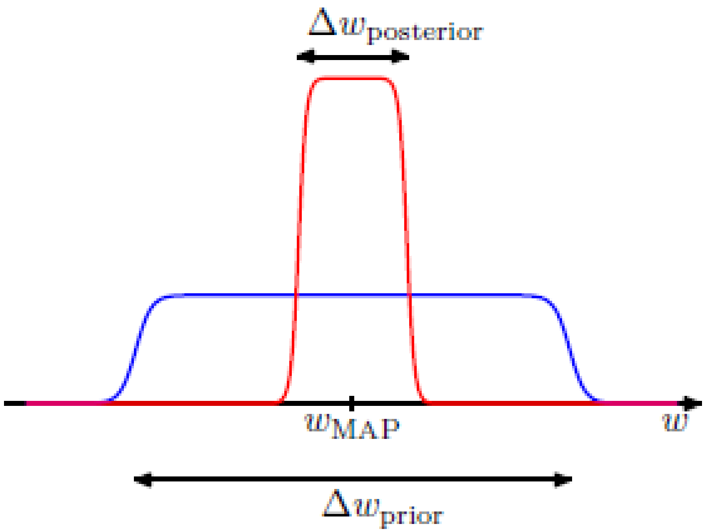
Model averaging

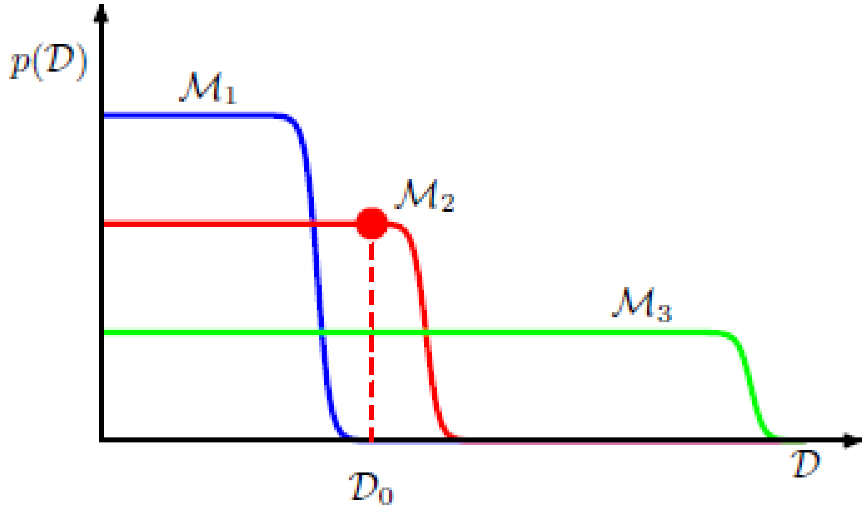
$$p(\theta|y) = \sum_m p(\theta|y, m)\, p(m|y)$$

Kass and Raftery (1995), Penny et al. (2004) *NeuroImage*

Model Evidence

$$p(y|m)=\int p(y|\theta,m)p(\theta|m)\,d\theta$$





$$p(\mathcal{D}) = \int p(\mathcal{D}|w)p(w)\,\mathrm{d}w \simeq p(\mathcal{D}|w_{\text{MAP}})\frac{\Delta w_{\text{posterior}}}{\Delta w_{\text{prior}}}$$

McKay 1992, *Neural Computations*, Bishop *PRML* 2007

Bayesian model selection (BMS)

Various Approximations:

- Akaike Information Criterion (AIC) – Akaike, 1974

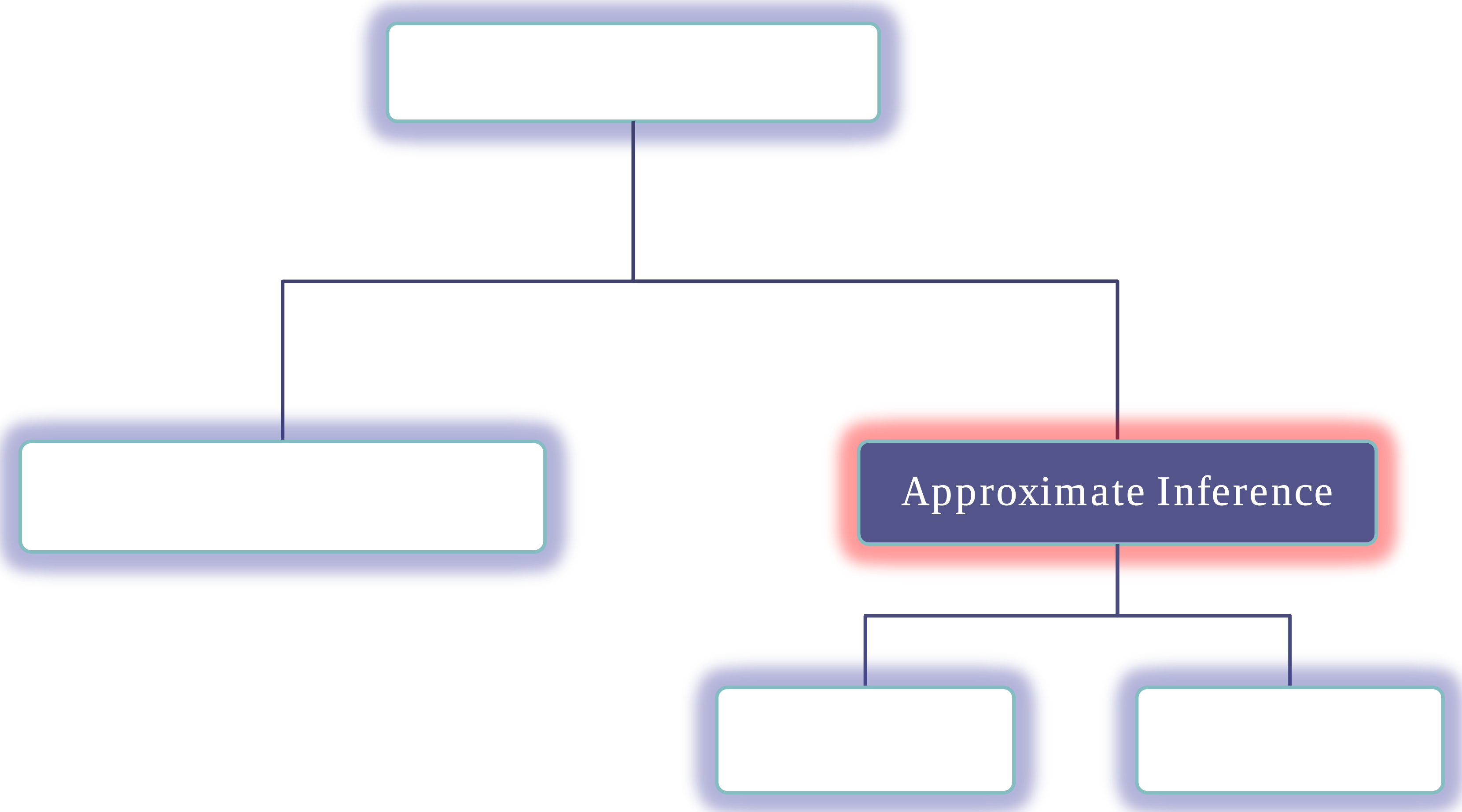
$$\ln p(D) \cong \ln p(D|\theta_{ML}) - M$$

- Bayesian Information Criterion (BIC) – Schwarz, 1978

$$\ln p(D) \cong \ln p(D|\theta_{ML}) - \frac{1}{2} M \ln(N)$$

- Negative free energy (F)
  - A by-product of Variational Bayes

- Path Sampling (Thermodynamic Integration) - MCMC


Approximate Inference

Approximate Bayesian inference

Bayesian inference formalizes *model inversion*, the process of passing from a prior to a posterior in light of data.

$$\underbrace{p(\theta|y)}_{\text{posterior}} = \frac{\overbrace{p(y|\theta)}^{\text{likelihood}}\ \overbrace{p(\theta)}^{\text{prior}}}{\int p(y,\theta)\mathrm{d}\theta}$$

marginal likelihood $p(y)$
(model evidence)

In practice, evaluating the posterior is usually difficult because we cannot easily evaluate $p(y)$, especially when:

- High dimensionality, complex form
- analytical solutions are not available
- numerical integration is too expensive

Source: Kay H. Brodersen, 2013, http://people.inf.ethz.ch/bkay/talks/Brodersen_2013_03_22.pdf

Approximate Bayesian inference

There are two approaches to approximate inference. They have complementary strengths and weaknesses.

Deterministic approximate inference

in particular variational Bayes

❶ find an analytical proxy $q(\theta)$ that is maximally similar to $p(\theta|y)$

❷ inspect distribution statistics of $q(\theta)$ (e.g., mean, quantiles, intervals, …)

☑ often insightful and fast
☒ often hard work to derive
☒ converges to local minima

Stochastic approximate inference

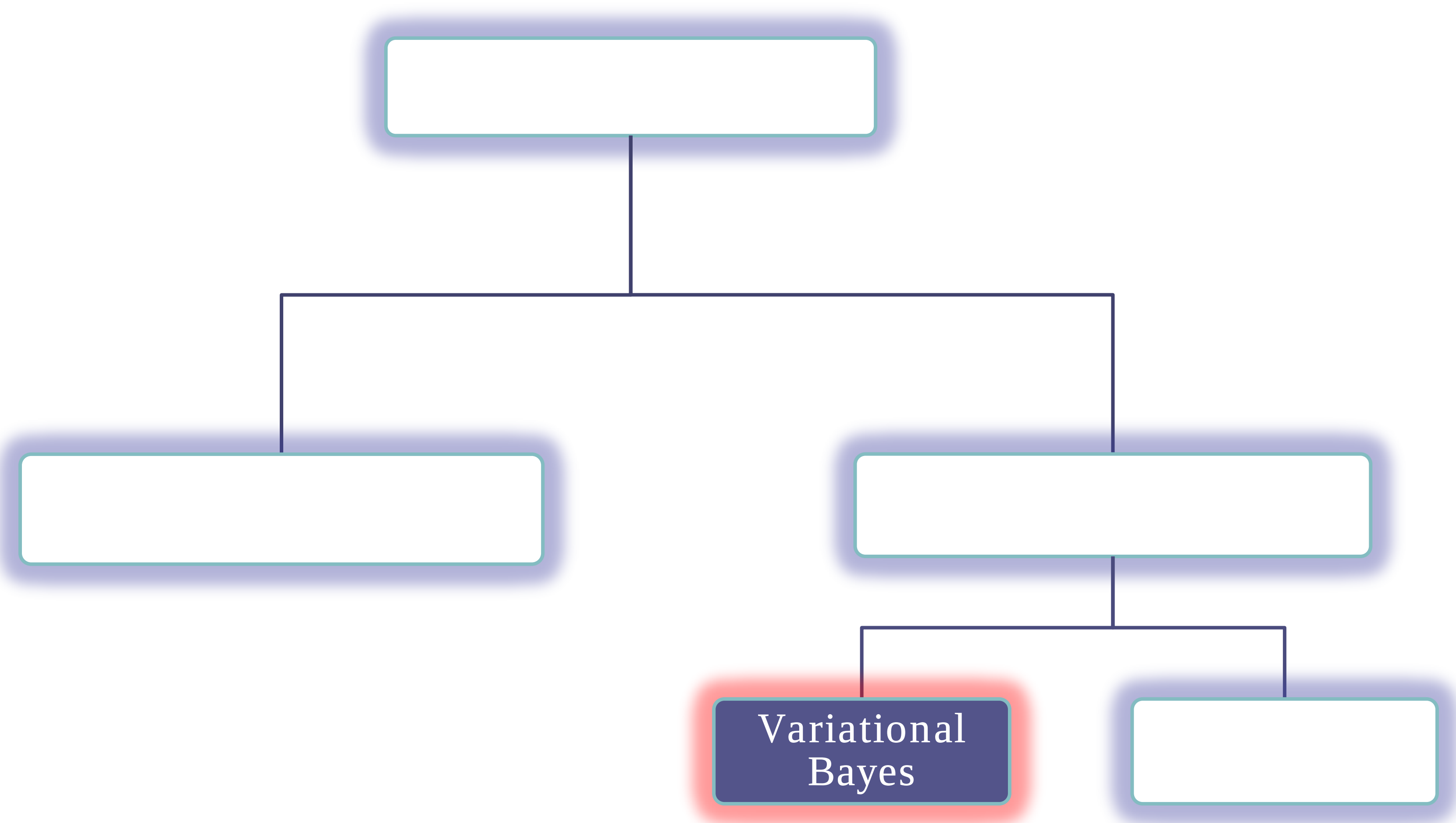
in particular sampling

❶ design an algorithm that draws samples $\theta^{(1)}, \dots, \theta^{(m)}$ from $p(\theta|y)$

❷ inspect sample statistics (e.g., histogram, sample quantiles, …)

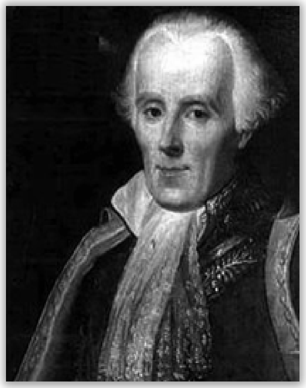
☑ asymptotically exact
☒ computationally expensive
☒ tricky engineering concerns

Source: Kay H. Brodersen, 2013, http://people.inf.ethz.ch/bkay/talks/Brodersen_2013_03_22.pdf


Variational
Bayes

The Laplace approximation

The Laplace approximation provides a way of approximating a density whose normalization constant we cannot evaluate, by fitting a Normal distribution to its mode.



Pierre-Simon Laplace
(1749 – 1827)
French mathematician and astronomer

$$p(z) = \frac{1}{Z} \times f(z)$$

normalization constant (unknown) — main part of the density (easy to evaluate)

This is exactly the situation we face in Bayesian inference:

$$p(\theta|y) = \frac{1}{p(y)} \times p(y,\theta)$$

model evidence (unknown) — joint density (easy to evaluate)

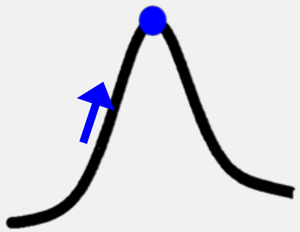
Source: Kay H. Brodersen, 2013, http://people.inf.ethz.ch/bkay/talks/Brodersen_2013_03_22.pdf

Applying the Laplace approximation

Given a model with parameters $\theta = (\theta_1, \ldots, \theta_p)$, the Laplace approximation reduces to a simple three-step procedure:

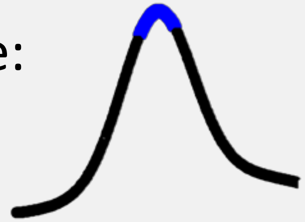
❶ Find the mode of the log-joint:

$$\theta^* = \arg\max_\theta \ \ln p(y, \theta)$$



❷ Evaluate the curvature of the log-joint at the mode:

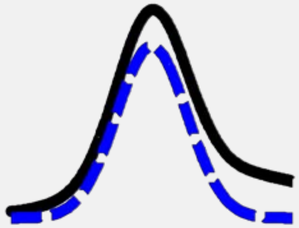
$$\nabla\nabla \ln p(y, \theta^*)$$



❸ We obtain a Gaussian approximation:

$\mathcal{N}(\theta | \mu, \Lambda^{-1})$ with $\mu = \theta^*$

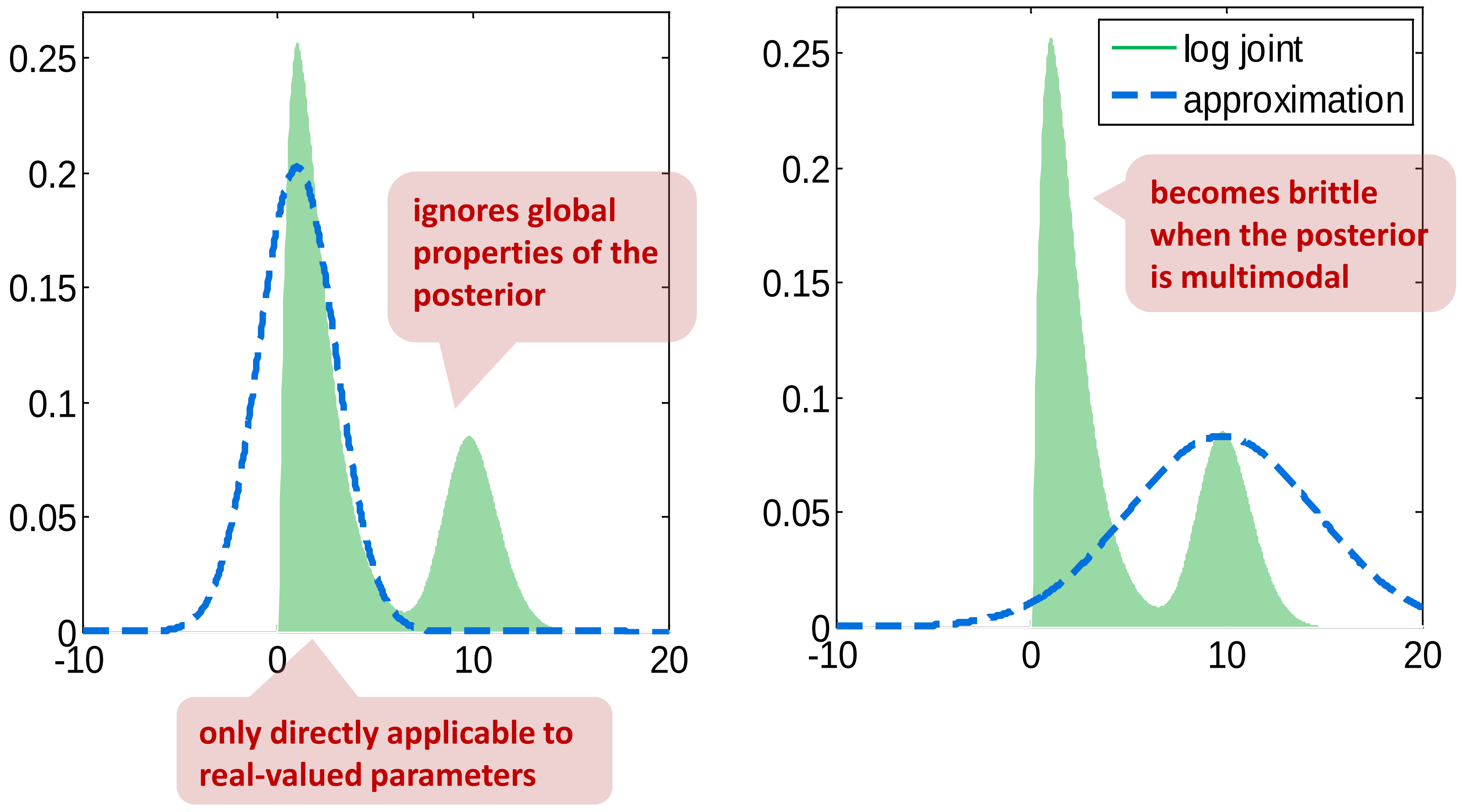
$$\Lambda = -\nabla\nabla \ln p(y, \theta^*)$$



Source: Kay H. Brodersen, 2013, http://people.inf.ethz.ch/bkay/talks/Brodersen_2013_03_22.pdf

Limitations of the Laplace approximation

The Laplace approximation is often too strong a simplification.

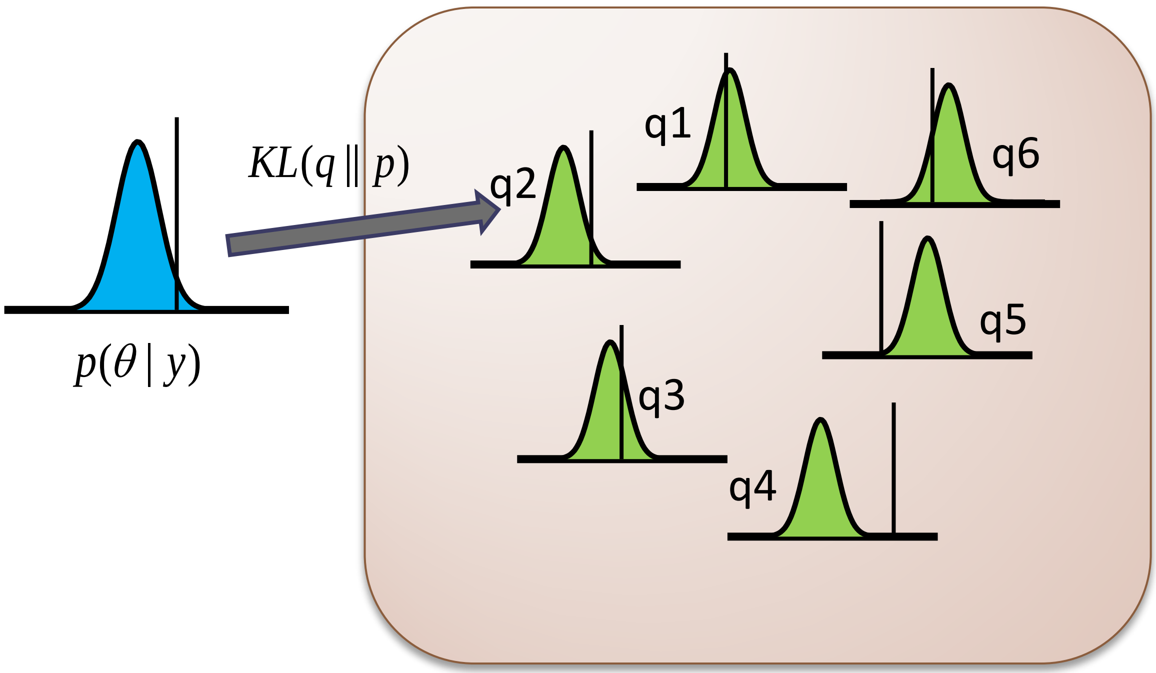


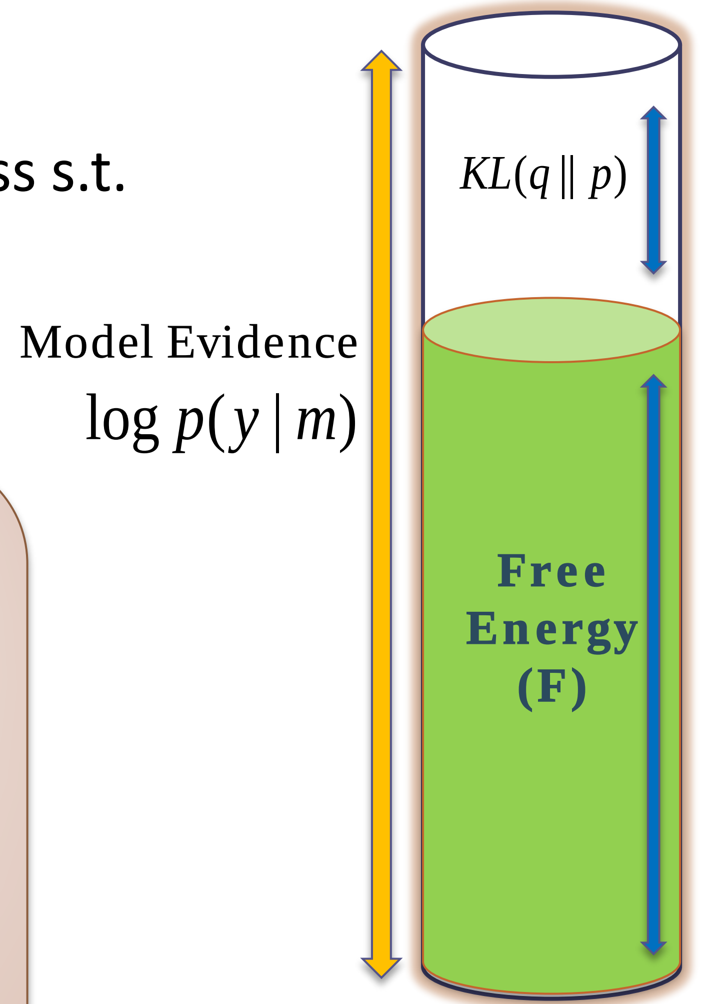
Source: Kay H. Brodersen, 2013, http://people.inf.ethz.ch/bkay/talks/Brodersen_2013_03_22.pdf

Variational Bayes

Goal: Choose q from a hypothesis class s.t.

$$KL(q \,||\, p) \longrightarrow 0$$





Variational calculus

Variational Bayesian inference is based on variational calculus.

Standard calculus

Newton, Leibniz, and others

- functions $f: x \mapsto f(x)$
- derivatives $\frac{\mathrm{d}f}{\mathrm{d}x}$

Example: maximize the likelihood expression $p(y|\theta)$ w.r.t. θ

Variational calculus

Euler, Lagrange, and others

- functionals $F: f \mapsto F(f)$
- derivatives $\frac{\mathrm{d}F}{\mathrm{d}f}$

Example: maximize the entropy $H[p]$ w.r.t. a probability distribution $p(x)$

Leonhard Euler
(1707 – 1783)

Swiss mathematician, 'Elementa Calculi Variationum'

Source: Kay H. Brodersen, 2013, http://people.inf.ethz.ch/bkay/talks/Brodersen_2013_03_22.pdf

Variational calculus and the free energy

Variational calculus lends itself nicely to approximate Bayesian inference.

$$\ln p(y) = \ln \frac{p(y,\theta)}{p(\theta|y)}$$

$$= \int q(\theta) \ \ln \frac{p(y,\theta)}{p(\theta|y)} \ \mathrm{d}\theta$$

$$= \int q(\theta) \ln \frac{p(y,\theta)}{p(\theta|y)} \frac{q(\theta)}{q(\theta)} \ \mathrm{d}\theta$$

$$= \int q(\theta) \left(\ln \frac{q(\theta)}{p(\theta|y)} + \ln \frac{p(y,\theta)}{q(\theta)} \right) \mathrm{d}\theta$$

$$= \underbrace{\int q(\theta) \ln \frac{q(\theta)}{p(\theta|y)} \mathrm{d}\theta}_{\mathrm{KL}[q||p]} + \underbrace{\int q(\theta) \ln \frac{p(y,\theta)}{q(\theta)} \mathrm{d}\theta}_{F(q,y)}$$

$\mathrm{KL}[q||p]$
divergence between $q(\theta)$ and $p(\theta|y)$

$F(q, y)$
free energy

Source: Kay H. Brodersen, 2013, http://people.inf.ethz.ch/bkay/talks/Brodersen_2013_03_22.pdf

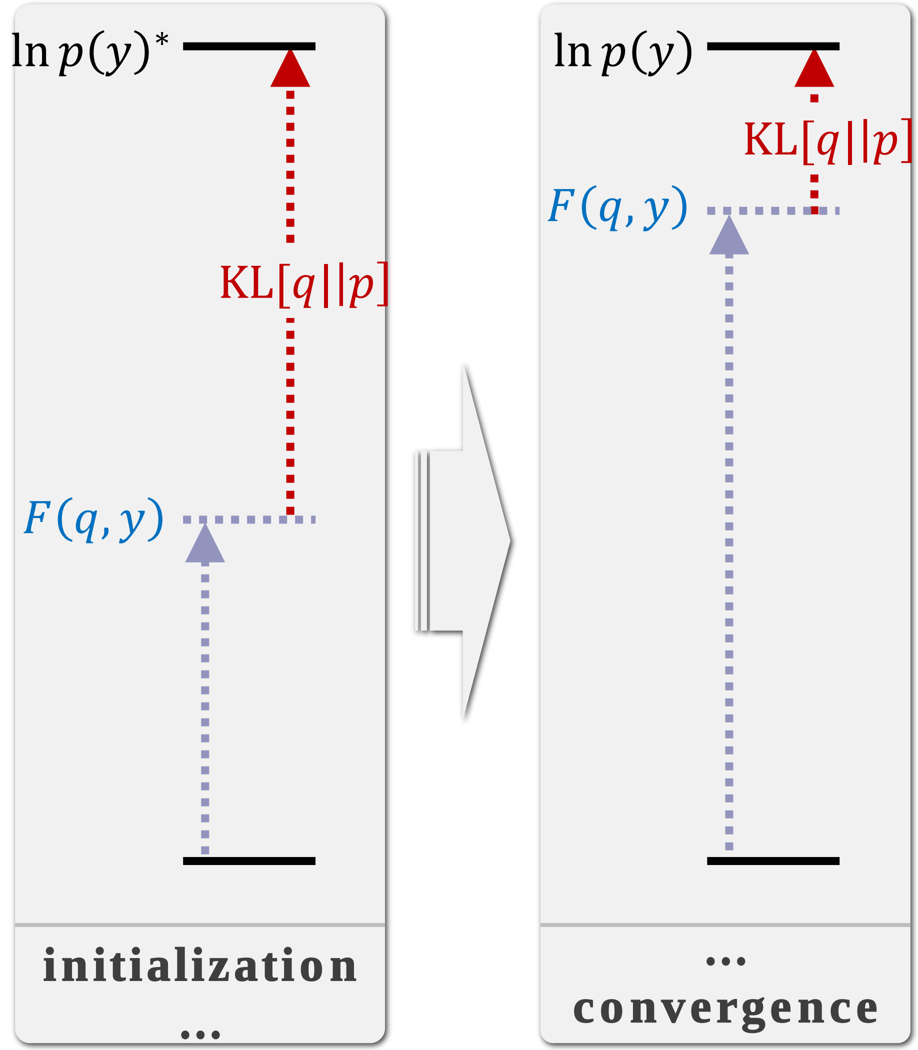
Variational calculus and the free energy

In summary, the log model evidence can be expressed as:

$$\ln p(y) = \underbrace{\mathrm{KL}[q||p]}_{\substack{\textbf{divergence} \\ \geq\ 0 \\ \text{(unknown)}}} + \underbrace{F(q,y)}_{\substack{\textbf{free energy} \\ \text{(easy to evaluate} \\ \text{for a given } q)}}$$

Maximizing $F(q,y)$ is equivalent to:

- minimizing $\mathrm{KL}[q||p]$
- tightening $F(q,y)$ as a lower bound to the log model evidence



Computing the free energy

We can decompose the free energy $F(q,y)$ as follows:

$$
\begin{aligned}
F(q,y) &= \int q(\theta) \ln \frac{p(y,\theta)}{q(\theta)} \mathrm{d}\theta \\
&= \int q(\theta) \ln p(y,\theta)\ d\theta - \int q(\theta) \ln q(\theta)\ d\theta \\
&= \underbrace{\langle \ln p(y,\theta) \rangle_q}_{\text{expected log-joint}} + \underbrace{H[q]}_{\text{Shannon entropy}}
\end{aligned}
$$

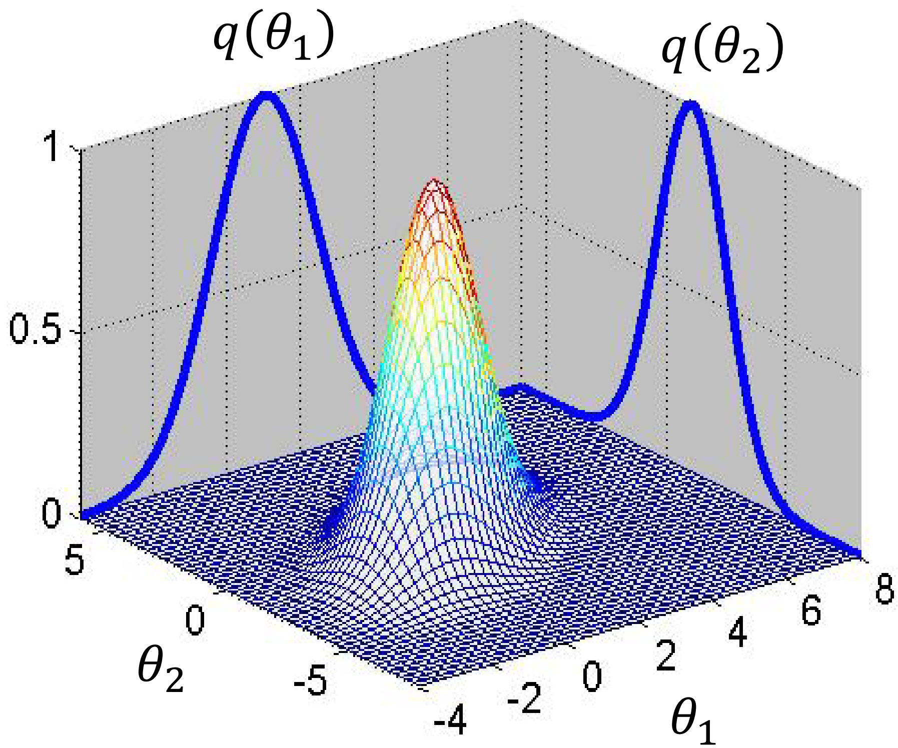
Source: Kay H. Brodersen, 2013, http://people.inf.ethz.ch/bkay/talks/Brodersen_2013_03_22.pdf

The mean-field assumption

When inverting models with several parameters, a common way of restricting the class of approximate posteriors $q(\theta)$ is to consider those posteriors that factorize into independent partitions,

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

where $q_i(\theta_i)$ is the approximate posterior for the i^{th} subset of parameters.



Jean Daunizeau, www.fil.ion.ucl.ac.uk/~jdaunize/presentations/Bayes2.pdf

Source: Kay H. Brodersen, 2013, http://people.inf.ethz.ch/bkay/talks/Brodersen_2013_03_22.pdf

Variational inference under the mean-field assumption

$$
\begin{aligned}
F(q,y) &= \int q(\theta) \ln \frac{p(y,\theta)}{q(\theta)} \, \mathrm{d}\theta \\
&= \int \prod_i q_i \times \left(\ln p(y,\theta) - \sum_i \ln q_i \right) \mathrm{d}\theta \\
&= \int q_j \prod_{\setminus j} q_i \left(\ln p(y,\theta) - \ln q_j \right) \mathrm{d}\theta - \int q_j \prod_{\setminus j} q_i \sum_{\setminus j} \ln q_i \, \mathrm{d}\theta \\
&= \int q_j \left(\underbrace{\int \prod_{\setminus j} q_i \ln p(y,\theta) \, \mathrm{d}\theta_{\setminus j}}_{\langle \ln p(y,\theta) \rangle_{q_{\setminus j}}} - \ln q_j \right) \mathrm{d}\theta_j - \int q_j \int \prod_{\setminus j} q_i \ln \prod_{\setminus j} q_i \, \mathrm{d}\theta_{\setminus j} \, \mathrm{d}\theta_j \\
&= \int q_j \ln \frac{\exp\left(\langle \ln p(y,\theta) \rangle_{q_{\setminus j}} \right)}{q_j} \mathrm{d}\theta_j \quad + \quad c \\
&= -\mathrm{KL}\left[q_j \,||\, \exp\left(\langle \ln p(y,\theta) \rangle_{q_{\setminus j}} \right) \right] \quad + \quad c
\end{aligned}
$$

mean-field assumption: $q(\theta) = \prod_i q_i(\theta_i)$

Variational algorithm under the mean-field assumption

In summary:

$$F(q,y) = -\text{KL}\left[q_j || \exp\left(\langle \ln p(y,\theta)\rangle_{q_{\setminus j}}\right)\right] + c$$

Suppose the densities $q_{\setminus j} \equiv q(\theta_{\setminus j})$ are kept fixed. Then the approximate posterior $q(\theta_j)$ that maximizes $F(q,y)$ is given by:

$$\begin{aligned} q_j^* &= \arg\max_{q_j} F(q,y) \\ &= \frac{1}{Z}\exp\left(\langle \ln p(y,\theta)\rangle_{q_{\setminus j}}\right) \end{aligned}$$

Therefore:

$$\ln q_j^* = \underbrace{\langle \ln p(y,\theta)\rangle_{q_{\setminus j}}}_{=:I(\theta_j)} - \ln Z$$

This implies a straightforward algorithm for variational inference:

❶ Initialize all approximate posteriors $q(\theta_i)$, e.g., by setting them to their priors.

❷ Cycle over the parameters, revising each given the current estimates of the others.

❸ Loop until convergence.

Source: Kay H. Brodersen, 2013, http://people.inf.ethz.ch/bkay/talks/Brodersen_2013_03_22.pdf

Typical strategies in variational inference

	no parametric assumptions	**parametric assumptions** $q(\theta) = F(\theta \mid \delta)$
no mean-field assumption	(variational inference = exact inference)	fixed-form optimization of moments
mean-field assumption $q(\theta) = \prod q(\theta_i)$	iterative free-form variational optimization	iterative fixed-form variational optimization

Source: Kay H. Brodersen, 2013, http://people.inf.ethz.ch/bkay/talks/Brodersen_2013_03_22.pdf

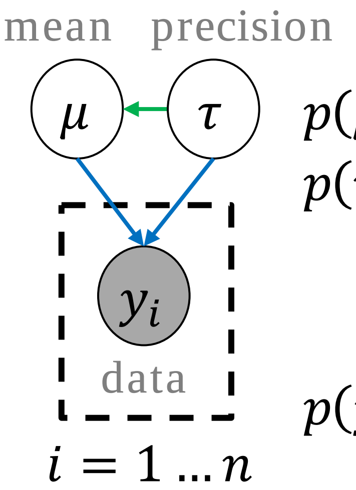
Example: variational density estimation

We are given a univariate dataset $\{y_1, \ldots, y_n\}$ which we model by a simple univariate Gaussian distribution. We wish to infer on its mean and precision:

$$p(\mu, \tau | y)$$

Although in this case a closed-form solution exists*, we shall pretend it does not. Instead, we consider approximations that satisfy the mean-field assumption:

$$q(\mu, \tau) = q_\mu(\mu)\, q_\tau(\tau)$$



$$p(\mu|\tau) = \mathcal{N}(\mu | \mu_0, (\lambda_0 \tau)^{-1})$$
$$p(\tau) = \mathrm{Ga}(\tau | a_0, b_0)$$

$$p(y_i | \mu, \tau) = \mathcal{N}(y_i | \mu, \tau^{-1})$$

10.1.3; Bishop (2006) PRML

Source: Kay H. Brodersen, 2013, http://people.inf.ethz.ch/bkay/talks/Brodersen_2013_03_22.pdf

Recurring expressions in Bayesian inference

Univariate normal distribution

$$\ln \mathcal{N}(x|\mu,\lambda^{-1}) = \frac{1}{2}\ln\lambda - \frac{1}{2}\ln\pi - \frac{\lambda}{2}(x-\mu)^2$$
$$= -\frac{1}{2}\lambda x^2 + \lambda\mu x + c$$

Multivariate normal distribution

$$\ln \mathcal{N}_d(x|\mu,\Lambda^{-1}) = -\frac{1}{2}\ln|\Lambda^{-1}| - \frac{d}{2}\ln 2\pi - \frac{1}{2}(x-\mu)^T\Lambda(x-\mu)$$
$$= -\frac{1}{2}x^T\Lambda x + x^T\Lambda\mu + c$$

Gamma distribution

$$\ln \mathrm{Ga}(x|a,b) = a\ln b - \ln\Gamma(a) + (a-1)\ln x - b\,x$$
$$= (a-1)\ln x - b\,x + c$$

Source: Kay H. Brodersen, 2013, http://people.inf.ethz.ch/bkay/talks/Brodersen_2013_03_22.pdf

Variational density estimation: mean μ

$$\begin{aligned}
\ln q^*(\mu) &= \langle \ln p(y,\mu,\tau)\rangle_{q(\tau)} + c \\
&= \left\langle \ln \prod_i^n p(y_i|\mu,\tau) \right\rangle_{q(\tau)} + \langle \ln p(\mu|\tau)\rangle_{q(\tau)} + \langle \ln p(\tau)\rangle_{q(\tau)} + c \\
&= \langle \ln \prod \mathcal{N}(y_i|\mu,\tau^{-1})\rangle_{q(\tau)} + \langle \ln \mathcal{N}(\mu|\mu_0,(\lambda_0\tau)^{-1})\rangle_{q(\tau)} + \langle \ln \mathrm{Ga}(\tau|a_0,b_0)\rangle_{q(\tau)} + c \\
&= \sum \left\langle -\frac{\tau}{2}(y_i-\mu)^2 \right\rangle_{q(\tau)} + \left\langle -\frac{\lambda_0\tau}{2}(\mu-\mu_0)^2 \right\rangle_{q(\tau)} + c \\
&= \sum -\frac{\langle\tau\rangle_{q(\tau)}}{2} y_i^2 + \langle\tau\rangle_{q(\tau)} n\bar{y}\mu - n\frac{\langle\tau\rangle_{q(\tau)}}{2}\mu^2 - \frac{\lambda_0\langle\tau\rangle_{q(\tau)}}{2}\mu^2 + \lambda_0\mu\mu_0\langle\tau\rangle_{q(\tau)} - \frac{\lambda_0}{2}\mu_0^2 + c \\
&= -\frac{1}{2}\left\{ n\langle\tau\rangle_{q(\tau)} + \lambda_0\langle\tau\rangle_{q(\tau)} \right\}\mu^2 + \left\{ n\bar{y}\langle\tau\rangle_{q(\tau)} + \lambda_0\mu_0\langle\tau\rangle_{q(\tau)} \right\}\mu + c
\end{aligned}$$

reinstation by inspection

$$\Longrightarrow q^*(\mu) = \mathcal{N}(\mu|\mu_n, \lambda_n^{-1}) \quad \text{with} \quad \lambda_n = (\lambda_0 + n)\langle\tau\rangle_{q(\tau)}$$

$$\mu_n = \frac{n\bar{y}\langle\tau\rangle_{q(\tau)} + \lambda_0\mu_0\langle\tau\rangle_{q(\tau)}}{\lambda_n} = \frac{\lambda_0\mu_0 + n\bar{y}}{\lambda_0 + n}$$

Source: Kay H. Brodersen, 2013, http://people.inf.ethz.ch/bkay/talks/Brodersen_2013_03_22.pdf

Variational density estimation: precision τ

$$\ln q^*(\tau) = \langle \ln p(y,\mu,\tau)\rangle_{q(\mu)} + c$$

$$= \left\langle \ln \prod_{i=1}^{n} \mathcal{N}(y_i|\mu,\tau^{-1}) \right\rangle_{q(\mu)} + \langle \ln \mathcal{N}(\mu|\mu_0,(\lambda_0\tau)^{-1})\rangle_{q(\mu)} + \langle \ln \mathrm{Ga}(\tau|a_0,b_0)\rangle_{q(\mu)} + c$$

$$= \sum_{i=1}^{n} \left\langle \frac{1}{2}\ln\tau - \frac{\tau}{2}(y_i-\mu)^2 \right\rangle_{q(\mu)} + \left\langle \frac{1}{2}\ln(\lambda_0\tau) - \frac{\lambda_0\tau}{2}(\mu-\mu_0)^2 \right\rangle_{q(\mu)}$$
$$+ \langle (a_0-1)\ln\tau - b_0\tau\rangle_{q(\mu)} + c$$

$$= \frac{n}{2}\ln\tau - \frac{\tau}{2}\langle \sum (y_i-\mu)^2\rangle_{q(\mu)} + \frac{1}{2}\ln\lambda_0 + \frac{1}{2}\ln\tau - \frac{\lambda_0\tau}{2}\langle(\mu-\mu_0)^2\rangle_{q(\mu)} + (a_0-1)\ln\tau - b_0\tau + c$$

$$= \left\{\frac{n}{2} + \frac{1}{2} + (a_0-1)\right\}\ln\tau - \left\{\frac{1}{2}\langle \sum (y_i-\mu)^2\rangle_{q(\mu)} + \frac{\lambda_0}{2}\langle(\mu-\mu_0)^2\rangle_{q(\mu)} + b_0\right\}\tau + c$$

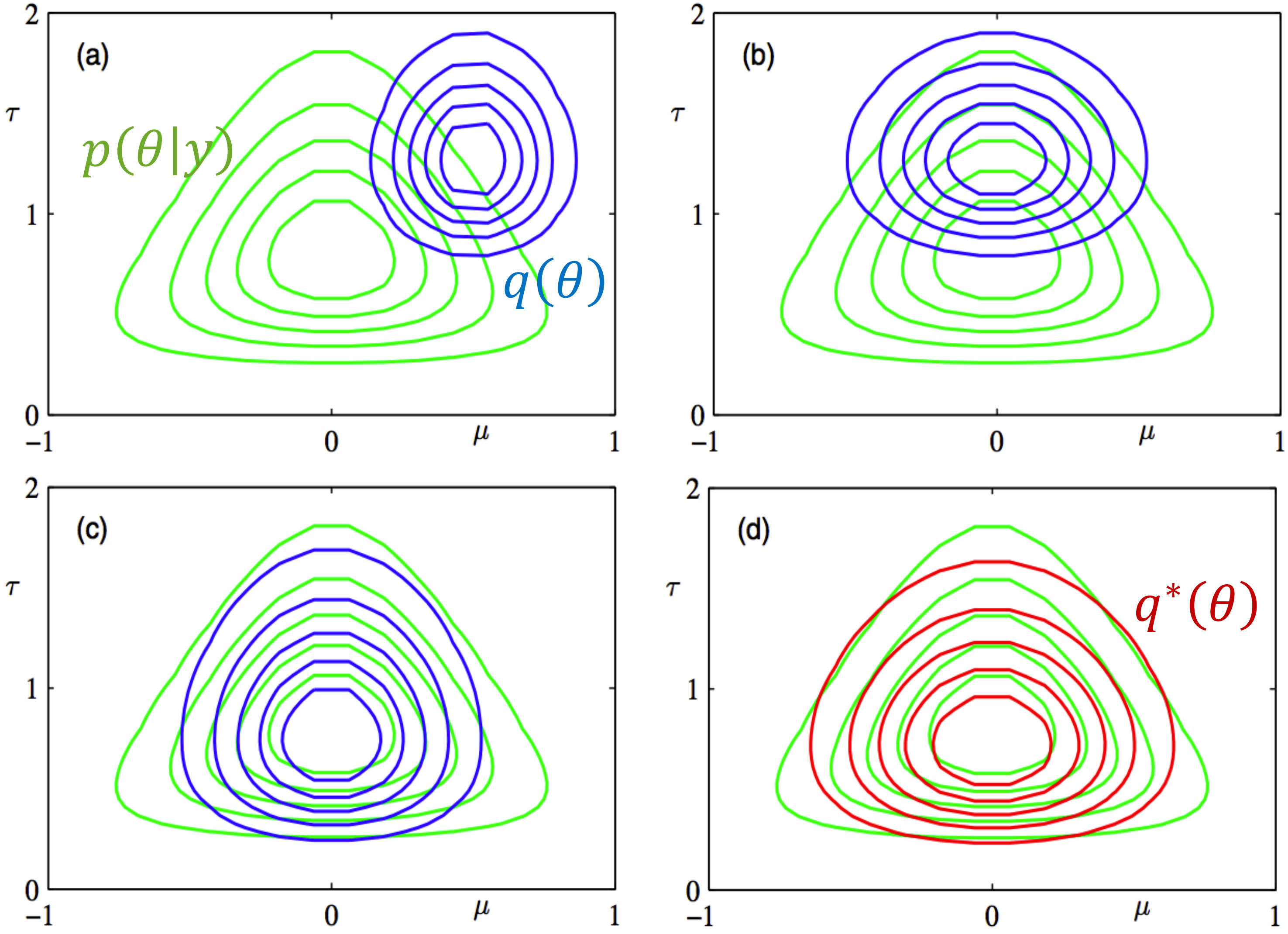
$\Longrightarrow q^*(\tau) = \mathrm{Ga}(\tau|a_n,b_n)$ with

$$a_n = a_0 + \frac{n+1}{2}$$

$$b_n = b_0 + \frac{\lambda_0}{2}\langle(\mu-\mu_0)^2\rangle_{q(\mu)} + \frac{1}{2}\langle \sum (y_i-\mu)^2\rangle_{q(\mu)}$$

Source: Kay H. Brodersen, 2013, http://people.inf.ethz.ch/bkay/talks/Brodersen_2013_03_22.pdf

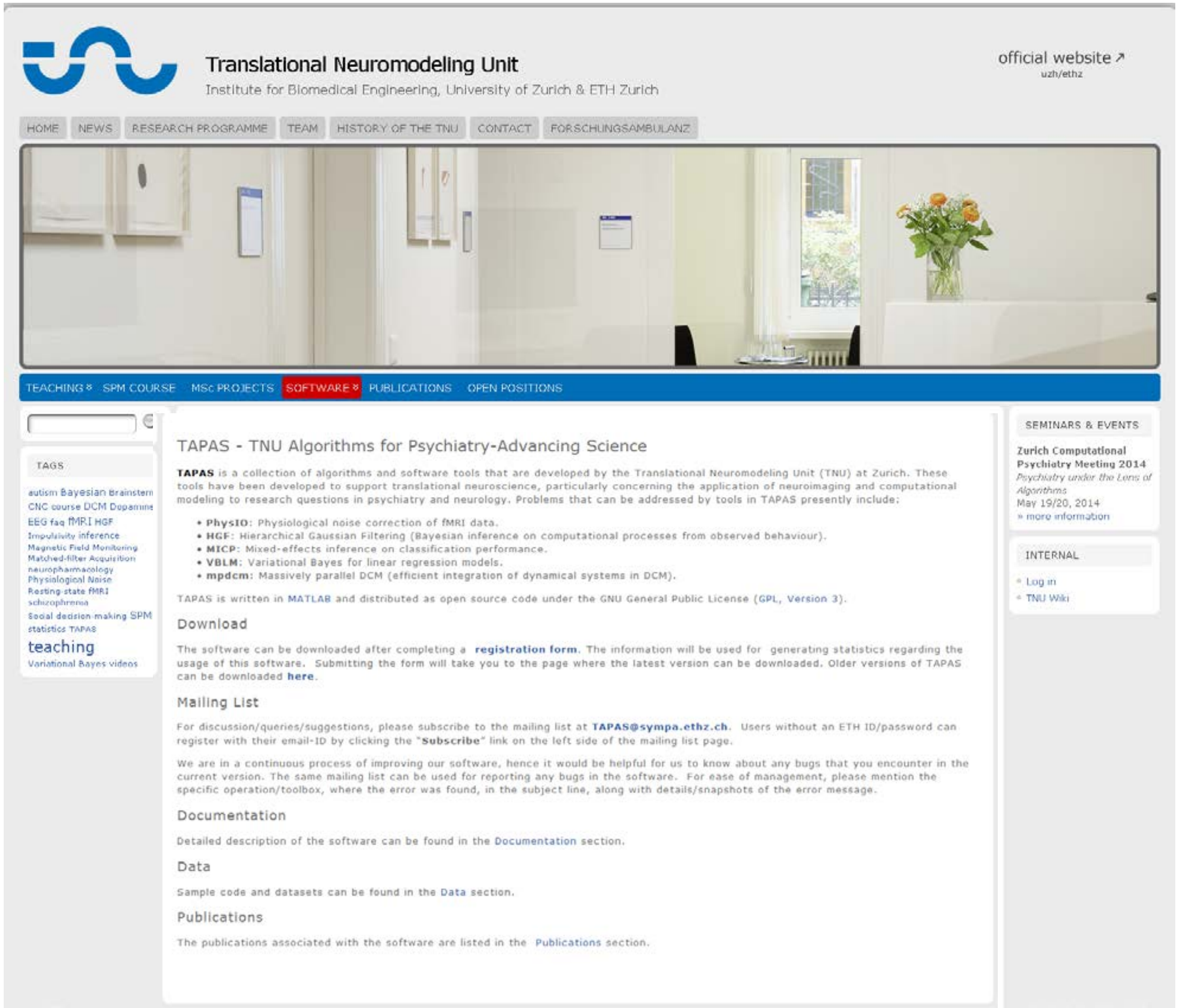
Variational density estimation: illustration



Source: Kay H. Brodersen, 2013, http://people.inf.ethz.ch/bkay/talks/Brodersen_2013_03_22.pdf

Bishop (2006) PRML, p. 472

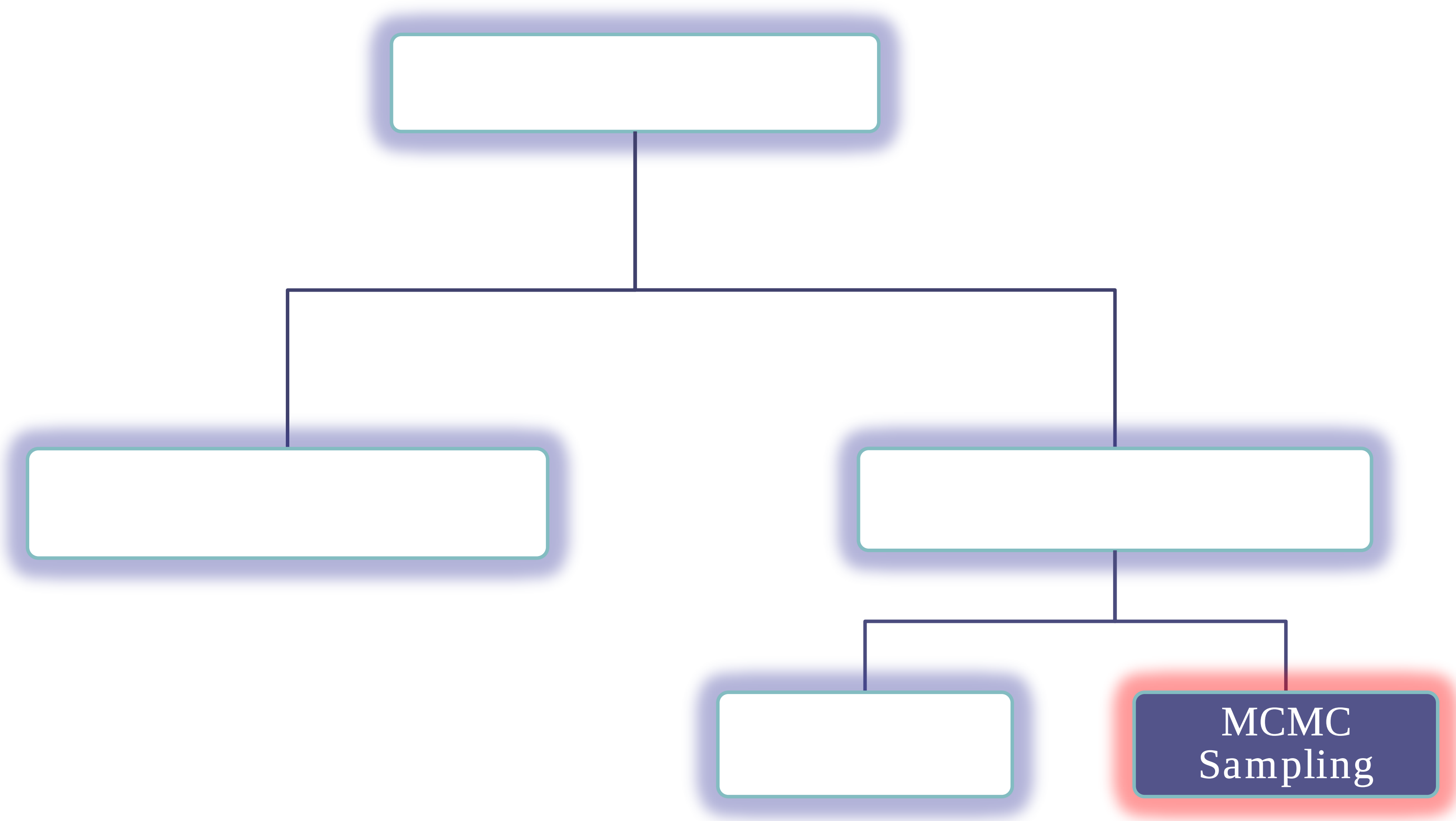
TAPAS



- Variational Bayes Linear Regression
  - http://www.translationalneuromodeling.org/tapas/

Demo VB

- Linear Regression

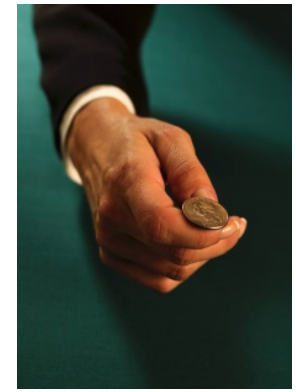

MCMC
Sampling

Markov Chain Monte Carlo(MCMC) sampling

Andrei Markov
(1856 – 1922)
Russian mathematician

Proposal distribution

$q(\theta)$



$\theta^0 \rightarrow \theta^1 \rightarrow \theta^2 \rightarrow \theta^3 \rightarrow \ldots\ldots\ldots\ldots \rightarrow \theta^N$

$\downarrow$

Posterior distribution

$p(\theta|y)$

- A general framework for sampling from a large class of distributions
- Scales well with dimensionality of sample space
- Asymptotically convergent

Markov chain properties

- Transition probabilities – homogeneous

$$p(\theta^{t+1} \mid \theta^1, \dots, \theta^t) = p(\theta^{t+1} \mid \theta^t) = T_t(\theta^{t+1}, \theta^t)$$

- Invariance

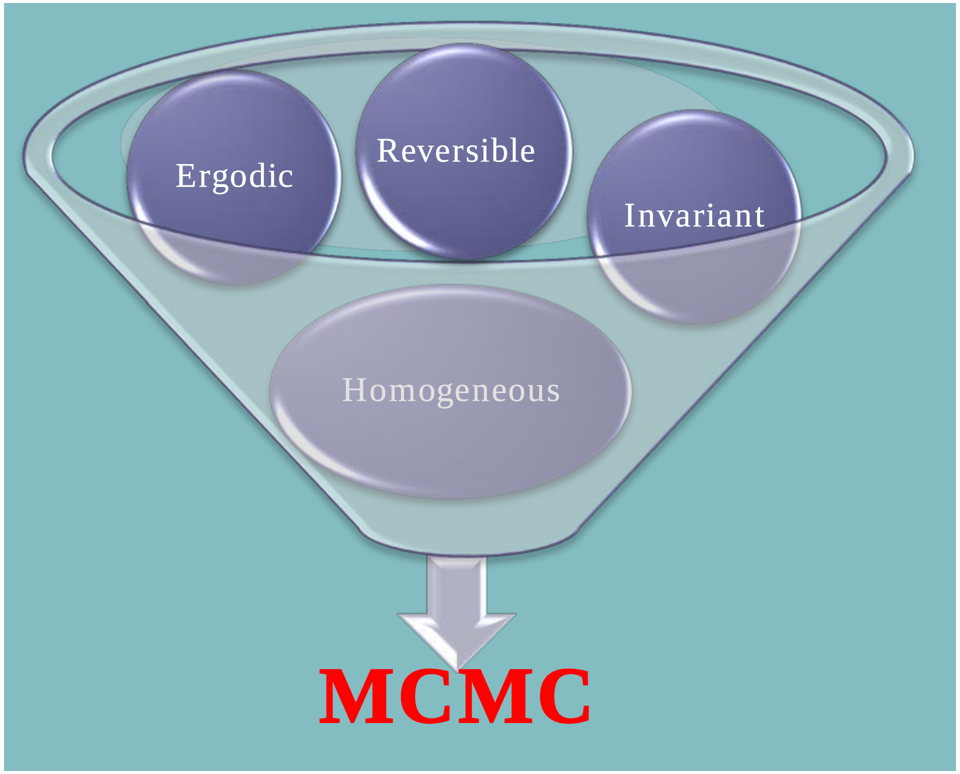
$$p^*(\theta) = \sum_{\theta'} T(\theta', \theta) p^*(\theta')$$

- Detailed Balance

$$T(\theta, \theta') p^*(\theta) = T(\theta', \theta) p^*(\theta')$$

- Ergodicity

$$p^*(\theta) = \lim_{n \to \infty} \left(p(\theta^n) \right) \quad \forall\, p(\theta^0)$$



Metropolis-Hastings Algorithm

- Initialize θ at step 1 - for example, sample from prior
- At step t, sample from the proposal distribution:

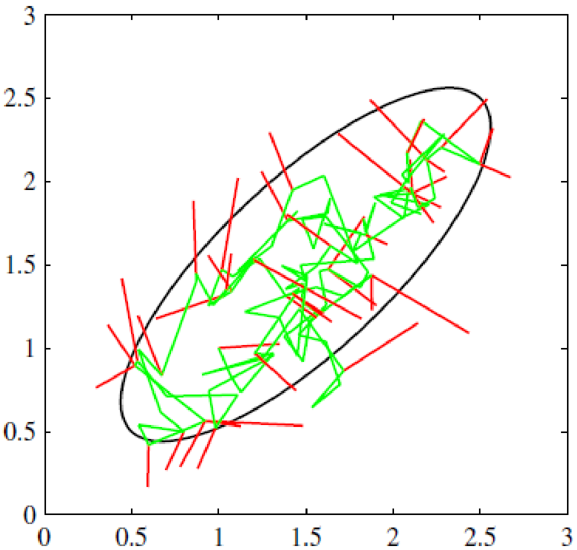
$$\theta^* \sim q(\theta^*|\theta^t)$$

- Accept with probability:

$$A(\theta^*, \theta^t) \sim min\left(1, \frac{p(\theta^*|y)\, q(\theta^t|\theta^*)}{p(\theta^t|y) q(\theta^*|\theta^t)}\right)$$

- Metropolis – Symmetric proposal distribution

$$A(\theta^*, \theta^t) \sim min\left(1, \frac{p(\theta^*|y)}{p(\theta^t|y)}\right)$$



Bishop (2006) PRML, p. 539

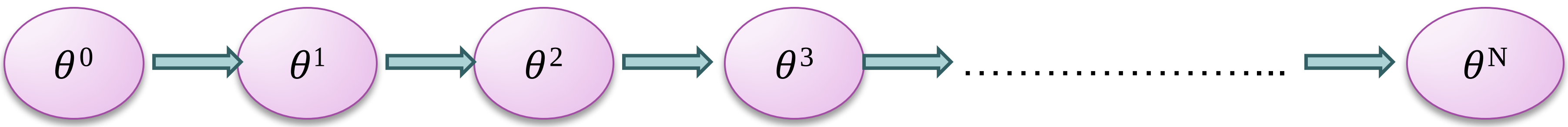
Gibbs Sampling Algorithm

- Special case of Metropolis Hastings
- At step t, sample from the conditional distribution:

$$
\begin{aligned}
\theta_1{}^{t+1} &\sim p(\theta_1 | \theta_2{}^t, \ldots \ldots, \theta_n{}^t) \\
\theta_2{}^{t+1} &\sim p(\theta_2 | \theta_1{}^{t+1}, \ldots \ldots, \theta_n{}^t) \\
&\vdots \\
&\vdots \\
&\vdots \\
\theta_n
\end{aligned}
$$

- Acceptance probability = 1
- Blocked Sampling

Posterior analysis from MCMC

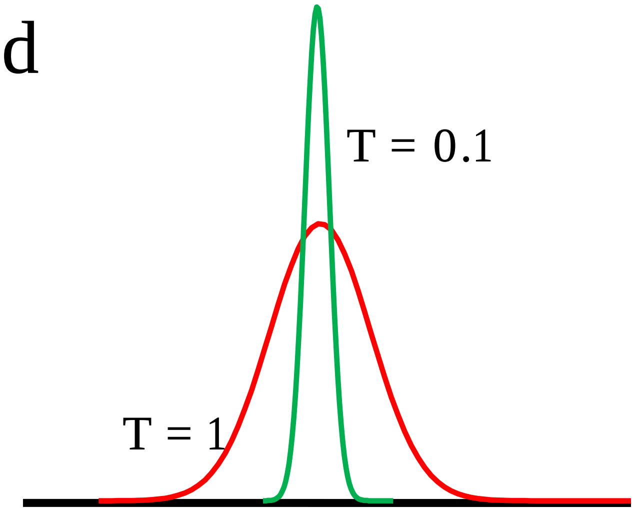


Obtain independent samples:

- Generate samples based on MCMC sampling.
- Discard initial “burn-in” period samples to remove dependence on initialization.
- Thinning- select every mth sample to reduce correlation .
- Inspect sample statistics (e.g., histogram, sample quantiles, …)

MAP estimate via Simulated Annealing

- Add a temperature parameter and schedule to update it



- Algorithm
  - Set T = 1
  - Until convergence
    - For every K iterations sample from: $p^{1/T}(\theta|y)$
    - Reduce T

Convergence Analysis

- Single chain methods
  - Geweke (1992)
  - Raftery-Lewis (1992)

- Multi-chain methods
  - Gelman-Rubin – (1992)
    - Potential Scale Reduction factor

Model evidence using MCMC

- Importance Sampling

$$p(D \mid M) = \frac{E_g\left[\frac{p(D|\theta,M)p(\theta|M)}{g(\theta)}\right]}{E_g\left[\frac{p(\theta|M)}{g(\theta)}\right]},$$

- Prior arithmetic mean

$$\widehat{f(Y)} = \frac{1}{M}\sum_{m=1}^{M} p(Y|\theta_m)$$

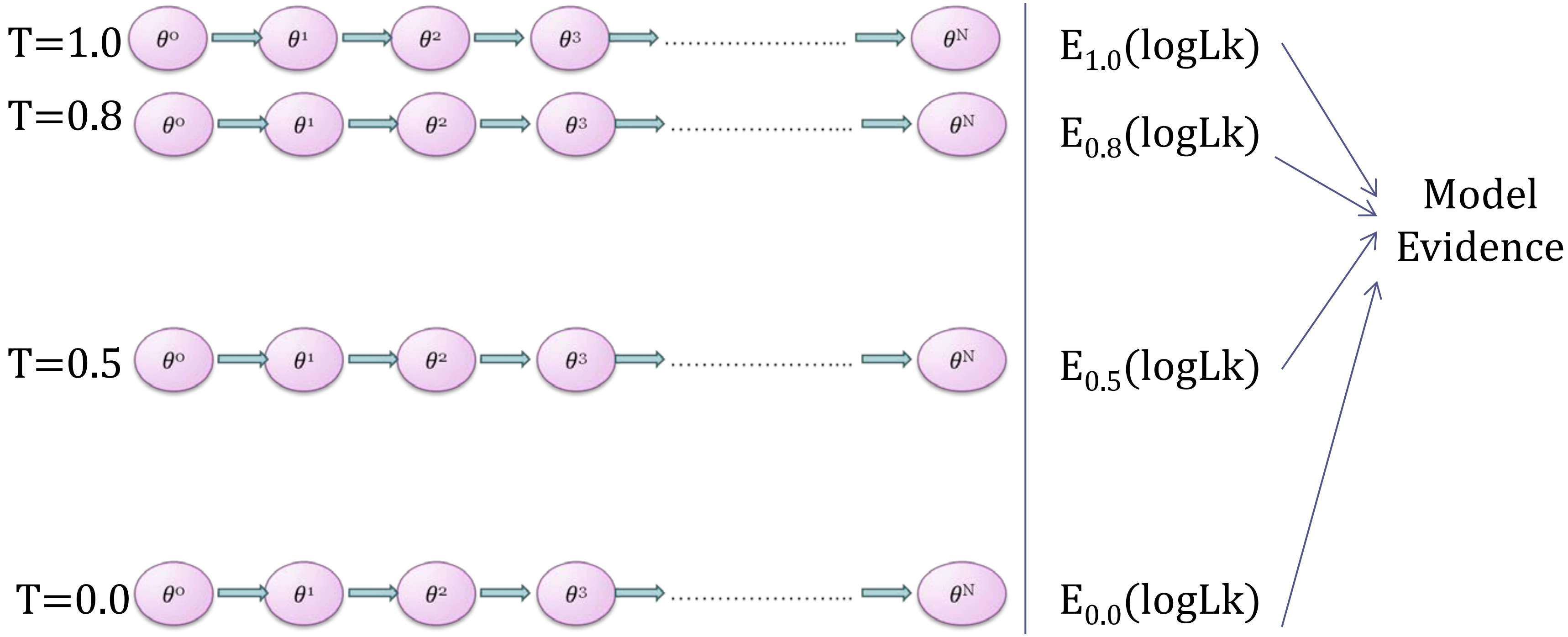
- Posterior harmonic mean

$$\widehat{f(Y)} = \frac{1}{\frac{1}{M}\sum_{m=1}^{M}\frac{1}{p(\mathbf{Y}|\bullet)}},$$

Thermodynamic Integration

- Path Sampling (Thermodynamic Integration)

$$q_\beta(\theta) = p(D \mid \theta, M)^\beta p(\theta \mid M).$$



Ogata ,Y. 1989. Num.Math.55:137-157,
Gelman,A.1998. . Stat . Sci . 13:163 — 185

Derivation - Extra

$$q_\beta(\theta) = p(D \mid \theta, M)^\beta p(\theta \mid M).$$

$$p_\beta(\theta) = \frac{1}{Z_\beta} q_\beta(\theta), \tag{15}$$

$$Z_\beta = \int_\Theta q_\beta(\theta) d\theta. \tag{16}$$

When β tends to 0 (resp. 1), p_β converges pointwise to p_0 (resp. p_1), and Z_β to Z_0 (resp. Z_1).

Taking the derivative of $\ln Z_\beta$ with respect to β:

$$\frac{\partial \ln Z_\beta}{\partial \beta} = \frac{1}{Z_\beta} \frac{\partial Z_\beta}{\partial \beta} \tag{17}$$

$$= \frac{1}{Z_\beta} \frac{\partial}{\partial \beta} \int_\Theta q_\beta(\theta) d\theta \tag{18}$$

$$= \frac{1}{Z_\beta} \int_\Theta \frac{\partial q_\beta(\theta)}{\partial \beta} d\theta \tag{19}$$

$$= \int_\Theta \frac{1}{q_\beta(\theta)} \frac{\partial q_\beta(\theta)}{\partial \beta} \frac{q_\beta(\theta)}{Z_\beta} d\theta \tag{20}$$

$$= \int_\Theta \frac{\partial \ln q_\beta(\theta)}{\partial \beta} p_\beta(\theta) d\theta \tag{21}$$

$$= E_\beta \left[\frac{\partial \ln q_\beta(\theta)}{\partial \beta} \right], \tag{22}$$

p_β. Defining the *potential*

$$U(\theta) = \frac{\partial \ln q_\beta(\theta)}{\partial \beta}, \tag{23}$$

one has thus the first moment identity:

$$\frac{\partial \ln Z_\beta}{\partial \beta} = E_\beta[U]. \tag{24}$$

Integrating over [0, 1] yields the log-ratio one is looking for:

$$\mu = \ln Z_1 - \ln Z_0 \tag{25}$$

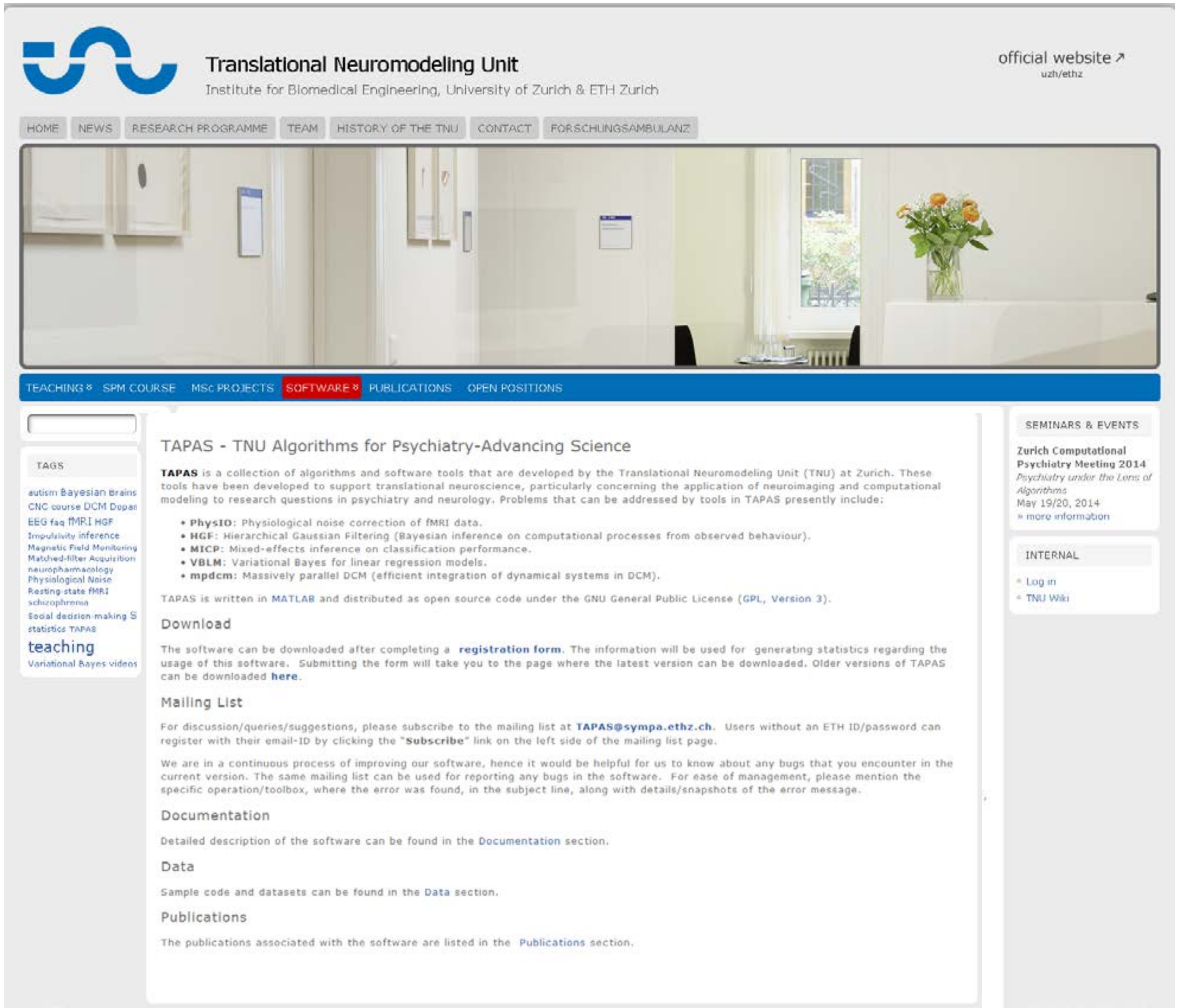
$$= \int_0^1 \frac{\partial \ln Z_\beta}{\partial \beta} d\beta \tag{26}$$

$$= \int_0^1 E_\beta[U] d\beta. \tag{27}$$

Other MCMC variants

- Slice Sampling
- Adaptive MH
- Hamiltonian Monte Carlo
- Population MCMC

TAPAS


Translational Neuromodeling Unit
Institute for Biomedical Engineering, University of Zurich & ETH Zurich
official website ↗
uzh/ethz
HOME NEWS RESEARCH PROGRAMME TEAM HISTORY OF THE TNU CONTACT FORSCHUNGSAMBULANZ
TEACHING SPM COURSE MSc PROJECTS SOFTWARE PUBLICATIONS OPEN POSITIONS

TAGS

SEMINARS & EVENTS

Zurich Computational Psychiatry Meeting 2014
Psychiatry under the Lens of Algorithms
May 19/20, 2014
» more information

INTERNAL

Log in
TNU Wiki

TAPAS - TNU Algorithms for Psychiatry-Advancing Science

TAPAS is a collection of algorithms and software tools that are developed by the Translational Neuromodeling Unit (TNU) at Zurich. These tools have been developed to support translational neuroscience, particularly concerning the application of neuroimaging and computational modeling to research questions in psychiatry and neurology. Problems that can be addressed by tools in TAPAS presently include:

- **PhysIO**: Physiological noise correction of fMRI data.
- **HGF**: Hierarchical Gaussian Filtering (Bayesian inference on computational processes from observed behaviour).
- **MICP**: Mixed-effects inference on classification performance.
- **VBLM**: Variational Bayes for linear regression models.
- **mpdcm**: Massively parallel DCM (efficient integration of dynamical systems in DCM).

TAPAS is written in MATLAB and distributed as open source code under the GNU General Public License (GPL, Version 3).

Download

The software can be downloaded after completing a **registration form**. The information will be used for generating statistics regarding the usage of this software. Submitting the form will take you to the page where the latest version can be downloaded. Older versions of TAPAS can be downloaded **here**.

Mailing List

For discussion/queries/suggestions, please subscribe to the mailing list at **TAPAS@sympa.ethz.ch**. Users without an ETH ID/password can register with their email-ID by clicking the "**Subscribe**" link on the left side of the mailing list page.

We are in a continuous process of improving our software, hence it would be helpful for us to know about any bugs that you encounter in the current version. The same mailing list can be used for reporting any bugs in the software. For ease of management, please mention the specific operation/toolbox, where the error was found, in the subject line, along with details/snapshots of the error message.

Documentation

Detailed description of the software can be found in the Documentation section.

Data

Sample code and datasets can be found in the Data section.

Publications

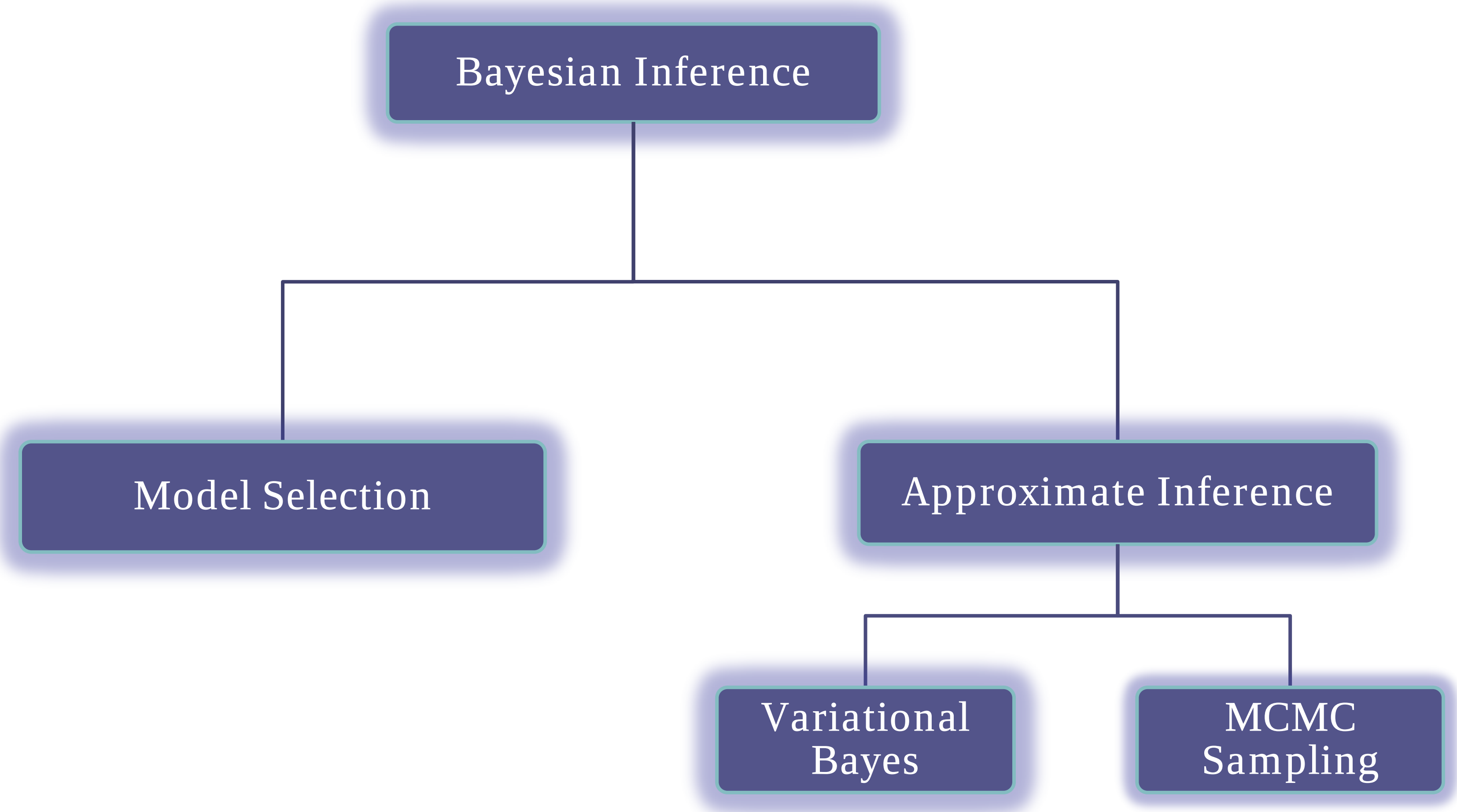
The publications associated with the software are listed in the Publications section.

- mpdcm – GPU based MCMC
  - http://www.translationalneuromodeling.org/tapas/

Demo MCMC

- Linear Regression

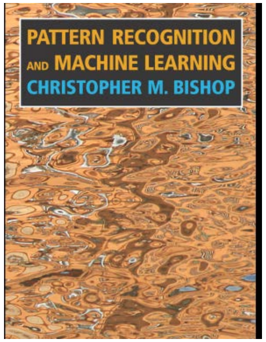
Summary

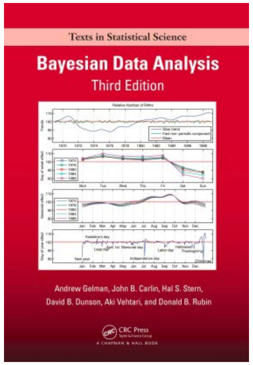


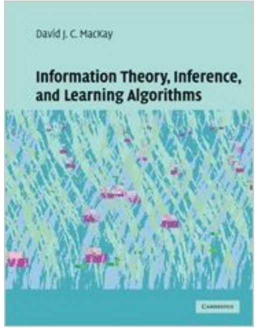
Bayesian Inference
Model Selection
Approximate Inference
Variational Bayes
MCMC Sampling

References

- Pattern Recognition and Machine Learning (2006). Christopher M. Bishop
- Bayesian reasoning and machine learning. David Barber
- Information theory, pattern recognition and learning algorithms. David McKay
- Conjugate Bayesian analysis of the Gaussian distribution. Kevin P. Murphy.
- Videolectures.net – Bayesian inference, MCMC , Variational Bayes
- Akaike, H (1974). A new look at statistical model identification. IEEE transactions on Automatic Control 19, 716-723.
- Schwarz, G. (1978). Estimating the dimension of a model. Annals of Statistics 6, 461-464.
- Kass, R.E. and Raftery, A.E. (1995).Bayes factors Journal of the American Statistical Association, 90, 773-795.
- Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference, Second Edition (Chapman & Hall/CRC Texts in Statistical Science) Dani Gamerman, Hedibert F. Lopes.
- Statistical Parametric Mapping: The Analysis of Functional Brain Images.
- N. Lartillot and H. Philippe, "Computing bayes factors using thermodynamic integration," Systematic Biology, vol. 55, no. 2, pp. 195–207, 2006.
- Andrew Gelman, Donald B. Rubin. Inference from iterative simulation using multiple sequences Stat. Sci., 7 (4) (1992), pp. 457–472
- J. Geweke. Evaluating the accuracy of sampling-based approaches to calculating posterior moments Bayesian Statistics 4, (1992), pp. 169–193
- Adrian E. Raftery, Steven LewisHow many iterations in the Gibbs sampler? Bayesian Statistics 4, Oxford University Press (1992), pp. 763–773







Thank You

http://www.translationalneuromodeling.org/tapas/