Bayesian Inference

Sudhir Shankar Raman
Translational Neuromodeling Unit,
UZH & ETH

With many thanks for materials 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
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.

\[ H_0 : \theta = 0 \]

\[ p(y \mid H_0) \]

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

\[ p(y, \theta) \]

\[ p(\theta) \]

\[ p(\theta \mid y) \]

Bayesian Inference

- Flexibility in modelling
- Incorporating prior information
- Posterior probability of effect
- Options for model comparison

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

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 $\theta$, 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:

$$P(\theta \mid y) = \frac{p(y \mid \theta) \cdot p(\theta)}{p(y)}$$
Bayesian statistics

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 $p(y | \theta)$
- 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

\[ p(\theta | y) = \frac{p(y | \theta) \ p(\theta)}{p(y)} \]

⇒ Analytical expression.

⇒ Conjugate priors for all exponential family members.

⇒ Example – Gaussian Likelihood , Gaussian prior over mean
Gaussian Model

Likelihood & prior

\[ p(y \mid \theta) = N(y \mid \theta, \lambda_e^{-1}) \]
\[ p(\theta) = N(\theta \mid \mu_p, \lambda_p^{-1}) \]

Posterior: \[ p(\theta \mid y) = N(\theta \mid \mu, \lambda^{-1}) \]

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

Relative precision weighting

\[ y = \theta + \varepsilon \]
Bayesian regression: univariate case

Normal densities

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

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

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

Relative precision weighting

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

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

Univariate linear model

\[
y = x\theta + \varepsilon
\]
Bayesian GLM: multivariate case

Normal densities

\[ p(\theta) = N(\theta; \eta_p, C_p) \]

\[ p(y | \theta) = N(y; X\theta, C_e) \]

\[ p(\theta | y) = N(\theta; \eta_{\theta|y}, C_{\theta|y}) \]

\[ C_{\theta|y}^{-1} = X^T C_e^{-1} X + C_p^{-1} \]

\[ \eta_{\theta|y} = C_{\theta|y} \left( X^T C_e y + C_p^{-1} \eta_p \right) \]

- One step if \( C_e \) is known.
- Otherwise define conjugate prior or perform iterative estimation with EM.
Bayesian Inference

Approximate Inference

Variational Bayes

MCMC Sampling

Model Selection

Model Selection

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:

\[
p(\theta | y, m) = \frac{p(y \mid \theta, m) p(\theta \mid m)}{p(y \mid m)}
\]

Model evidence:

\[
p(y \mid m) = \int p(y \mid \theta, m) \cdot p(\theta \mid m) \, d\theta
\]

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

Model comparison via Bayes factor:

\[
\frac{p(m_1 \mid y)}{p(m_2 \mid y)} = \frac{p(y \mid m_1) p(m_1)}{p(y \mid m_2) p(m_2)}
\]

\[
BF = \frac{p(y \mid m_1)}{p(y \mid m_2)}
\]

Model averaging:

\[
p(\theta \mid y) = \sum_{m} p(\theta \mid y, m) p(m \mid y)
\]

<table>
<thead>
<tr>
<th>BF$_{10}$</th>
<th>Evidence against $H_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 to 3</td>
<td>Not worth more than a bare mention</td>
</tr>
<tr>
<td>3 to 20</td>
<td>Positive</td>
</tr>
<tr>
<td>20 to 150</td>
<td>Strong</td>
</tr>
<tr>
<td>&gt; 150</td>
<td>Decisive</td>
</tr>
</tbody>
</table>

Bayesian model selection (BMS)

Various Approximations:

• Akaike Information Criterion (AIC) – Akaike, 1974

\[ \ln p(D | \theta_{ML}) - M \]

• Bayesian Information Criterion (BIC) – Schwarz, 1978

\[ \ln p(D) \cong \ln p(D | \theta_{MAP}) - \frac{1}{2}M \ln(N) \]

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

• Path Sampling (Thermodynamic Integration) - MCMC
Approximate Bayesian inference

Bayesian inference formalizes \textit{model inversion}, the process of passing from a prior to a posterior in light of data.

\[
p(\theta | y) = \frac{p(y | \theta) \cdot p(\theta)}{\int p(y, \theta) d\theta}
\]

\textbf{posterior} \quad \textbf{likelihood} \quad \textbf{prior} \quad \textbf{marginal likelihood} \quad \textbf{p}(y) \quad \textbf{(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

Approximate Bayesian inference

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

**Deterministic approximate inference**
in particular variational Bayes

1. find an analytical proxy $q(\theta)$ that is maximally similar to $p(\theta | y)$
2. 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

1. design an algorithm that draws samples $\theta^{(1)}, ..., \theta^{(m)}$ from $p(\theta | y)$
2. inspect sample statistics (e.g., histogram, sample quantiles, ...)

✓ asymptotically exact
✗ computationally expensive
✗ tricky engineering concerns

Bayesian Inference

Approximate Inference

Variational

Bayes

Model Selection

MCMC

Sampling

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

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

normalization constant (unknown) \hspace{2cm} 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) \hspace{2cm} joint density (easy to evaluate)

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:

1. Find the mode of the log-joint:
   $\theta^* = \arg \max_{\theta} \ln p(y, \theta)$

2. Evaluate the curvature of the log-joint at the mode:
   $\nabla^2 \ln p(y, \theta^*)$

3. We obtain a Gaussian approximation:
   $\mathcal{N}(\theta | \mu, \Lambda^{-1})$ with $\mu = \theta^*$
   $\Lambda = -\nabla^2 \ln p(y, \theta^*)$

Limitations of the Laplace approximation

The Laplace approximation is often too strong a simplification.

- ignores global properties of the posterior
- becomes brittle when the posterior is multimodal
- only directly applicable to real-valued parameters

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

\[ KL(q \parallel p) \rightarrow 0 \]

Variational Bayes

Model Evidence
\[ \log p(y \mid m) \]

Free Energy (F)

\[ KL(q \parallel p) \]
Variational Bayesian inference is based on variational calculus.

**Standard calculus**

Newton, Leibniz, and others

- functions $f: x \mapsto f(x)$
- derivatives $\frac{df}{dx}$

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

**Variational calculus**

Euler, Lagrange, and others

- functionals $F: f \mapsto F(f)$
- derivatives $\frac{dF}{df}$

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

Variational calculus and the free energy

Variational calculus lends itself nicely to approximate Bayesian inference.

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

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

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

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

\[
= \int q(\theta) \ln \frac{q(\theta)}{p(\theta | y)} \, d\theta + \int q(\theta) \ln \frac{p(y, \theta)}{q(\theta)} \, d\theta
\]

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

\[
F(q, y) \quad \text{free energy}
\]

Variational calculus and the free energy

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

\[ \ln p(y) = \underbrace{KL[q||p]}_{\text{divergence}} + \underbrace{F(q,y)}_{\text{free energy}} \]

\[ \geq 0 \quad (\text{unknown}) \]

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

- minimizing \( KL[q||p] \)
- 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:

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

$$= \int q(\theta) \ln p(y, \theta) \, d\theta - \int q(\theta) \ln q(\theta) \, d\theta$$

$$= \langle \ln p(y, \theta) \rangle_q + H[q]$$

*expected log-joint*  *Shannon entropy*

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.
Variational algorithm under the mean-field assumption

In summary:
\[ F(q, y) = -KL[q_j \| \exp\left(\langle \ln p(y, \theta) \rangle_{q_j} \right)] + c \]

Suppose the densities \( q_j \equiv q(\theta_j) \) are kept fixed. Then the approximate posterior \( q(\theta_j) \) that maximizes \( F(q, y) \) is given by:

\[
q_j^* = \arg \max_{q_j} F(q, y) = \frac{1}{Z} \exp\left( \langle \ln p(y, \theta) \rangle_{q_j} \right)
\]

Therefore:
\[
\ln q_j^* = \langle \ln p(y, \theta) \rangle_{q_j} - \ln Z =: I(\theta_j)
\]

This implies a straightforward algorithm for variational inference:

1. Initialize all approximate posteriors \( q(\theta_i) \), e.g., by setting them to their priors.
2. Cycle over the parameters, revising each given the current estimates of the others.
3. Loop until convergence.

## Typical strategies in variational inference

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

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:

\[
q(\mu, \tau) = q_\mu(\mu) q_\tau(\tau)
\]

**Example: variational density estimation**

\[
p(\mu | \tau) = \mathcal{N}(\mu | \mu_0, (\lambda_0 \tau)^{-1})
\]

\[
p(\tau) = \text{Ga}(\tau | a_0, b_0)
\]

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


10.1.3; Bishop (2006) PRML
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
\[
\ln \text{Ga}(x|a, b) = a \ln b - \ln \Gamma(a) + (a - 1) \ln x - b x \\
= (a - 1) \ln x - b x + c
\]

Variational density estimation: mean $\mu$

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

\[
= \left( \ln \prod_{i} p(y_i | \mu, \tau) \right)_{q(\tau)} + \langle \ln p(\mu | \tau) \rangle_{q(\tau)} + \langle \ln p(\tau) \rangle_{q(\tau)} + c
\]

\[
= \langle \ln \prod \mathcal{N}(y_i | \mu, \tau^{-1}) \rangle_{q(\tau)} + \langle \ln \mathcal{N}(\mu | \mu_0, (\lambda_0 \tau)^{-1}) \rangle_{q(\tau)} + \langle \ln \text{Ga}(\tau | a_0, b_0) \rangle_{q(\tau)} + c
\]

\[
= \sum \left( -\frac{\tau}{2} (y_i - \mu)^2 \right)_{q(\tau)} + \left( -\frac{\lambda_0 \tau}{2} (\mu - \mu_0)^2 \right)_{q(\tau)} + c
\]

\[
= \sum -\frac{\langle \tau \rangle_{q(\tau)}}{2} y_i^2 + \langle \tau \rangle_{q(\tau)} n\bar{y}\mu - n \frac{\langle \tau \rangle_{q(\tau)}}{2} \mu^2 - \frac{\lambda_0 \langle \tau \rangle_{q(\tau)}}{2} \mu^2 + \lambda_0 \mu_0 \langle \tau \rangle_{q(\tau)} - \frac{\lambda_0}{2} \mu_0^2 + c
\]

\[
= -\frac{1}{2} \left\{ n\langle \tau \rangle_{q(\tau)} + \lambda_0 \langle \tau \rangle_{q(\tau)} \right\} \mu^2 + \left\{ n\bar{y}\langle \tau \rangle_{q(\tau)} + \lambda_0 \mu_0 \langle \tau \rangle_{q(\tau)} \right\} \mu + c
\]

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

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

Variational density estimation: precision $\tau$

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

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

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

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

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

\[ \Rightarrow q^*(\tau) = \text{Ga}(\tau | a_n, b_n) \quad \text{with} \]

\[ a_n = a_0 + \frac{n + 1}{2} \]

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

Variational density estimation: illustration

\[ p(\theta | y) \]  
\[ q(\theta) \]

\[ q^*(\theta) \]


Bishop (2006) PRML, p. 472
TAPAS

Variational Bayes Linear Regression

http://www.translationalneuromodeling.org/tapas/
Markov Chain Monte Carlo (MCMC) sampling

- 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

\[ p(\theta^{t+1} | \theta^1, \ldots, \theta^t) = p(\theta^{t+1} | \theta^t) = T_t(\theta^{t+1}, \theta^t) \]

- Invariance

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

- Ergodicity

\[ p^*(\theta) = \lim_{n \to \infty} \left( p(\theta^m) \right) \quad \forall \ p(\theta^0) \]
Metropolis-Hastings Algorithm

- Initialize $\theta$ 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:
  \[
  \begin{align*}
  \theta_1^{t+1} & \sim p(\theta_1 | \theta_2^t, \ldots, \theta_n^t) \\
  \theta_2^{t+1} & \sim p(\theta_2 | \theta_1^{t+1}, \ldots, \theta_n^t) \\
  \vdots & \\
  \theta_n & 
  \end{align*}
  \]
- 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 $m^{th}$ sample to reduce correlation .
- Inspect sample statistics (e.g., histogram, sample quantiles, ...)

$\theta^0 \rightarrow \theta^1 \rightarrow \theta^2 \rightarrow \theta^3 \rightarrow \cdots \rightarrow \theta^N$
Model evidence using MCMC

- Prior arithmetic mean
\[ f(Y) = \frac{1}{M} \sum_{m=1}^{M} p(Y|\theta_m) \]

- Posterior harmonic mean
\[ f(Y) = \frac{1}{\frac{1}{M} \sum_{m=1}^{M} \frac{1}{p(Y|\theta_m)}} \]

- Path Sampling (Thermodynamic Integration)

T=1.0
\( \theta^0 \rightarrow \theta^1 \rightarrow \theta^2 \rightarrow \theta^3 \rightarrow \ldots \rightarrow \theta^N \)

E1.0(logLk)

T=0.8
\( \theta^0 \rightarrow \theta^1 \rightarrow \theta^2 \rightarrow \theta^3 \rightarrow \ldots \rightarrow \theta^N \)

E0.8(logLk)

T=0.5
\( \theta^0 \rightarrow \theta^1 \rightarrow \theta^2 \rightarrow \theta^3 \rightarrow \ldots \rightarrow \theta^N \)

E0.5(logLk)

T=0.0
\( \theta^0 \rightarrow \theta^1 \rightarrow \theta^2 \rightarrow \theta^3 \rightarrow \ldots \rightarrow \theta^N \)

E0.0(logLk)

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
Thank You

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