Bayesian Methods for Latent Variable Models

Olivier Cappé
Télécom ParisTech & CNRS

Apprentissage en traitement du signal et des images
5ème école d'été de Peyresq
26 juin – 2 juillet 2010

2 The Posterior
- Bayesian Inference
- Bayesian Model Selection
- Bayesian Prediction
- Bayesian Estimators
- An Example: Signal in Noise

3 The Prior
- Jeffrey's Priors
- Improper Priors
- The Normalization Penalty
- Conjugate Priors
- Exponential families

4 Latent Variable Models

- Model and Graphical Representation
- Missing data
- Signal Decomposition and *CA Models
- Mixture Models
- Scale Mixture
- Mixture of Mixtures
- State-Space Models

5 MAP Parameter Inference

- Structure of the posterior
- The EM Algorithm
- Variational Inference
- Data Augmentation

6 Posterior Inference

Priors for This Course

Bayesian modelling is a very broad topic; this course takes the following viewpoint

Agnostic Approach to Bayesian Modelling No Bayesian preaching; quantifying the asymptotic statistical performance of Bayesian estimators is a valid approach

Focus on the Information Procesing Context

- Final user is not a statistician but an algorithm that needs to make decisions autonomously
- More interested in the algorithm's output than by the statistical model in itself
- Potentially large amount of data to be processed

Bias Towards Black-Box Modelling In particular, will often trade model expressivity for inference simplicity

Priors for This Course (Contd.)

Bias Towards Exact Inference Conjugate priors, Rao-Blackwellization

or at Least Asymptotically Correct Inference In particular, will cover MCMC rather than variational methods

Bias Towards Sequential Models Where data is indexed by time

Focus on Latent Models Because they are fun and ubiquitous in machine learning, esp. for (partly) unsupervised tasks

Part I

Bayesian Modelling

2 The Posterior

- Bayesian Inference
- Bayesian Model Selection
- Bayesian Prediction
- Bayesian Estimators
- An Example: Signal in Noise

3 The Prior

- Jeffrey's Priors
- Improper Priors
- The Normalization Penalty
- Conjugate Priors
- Exponential families

Bayesian Modelling

Bayesian Model

1. Likelihood The data $\boldsymbol{y}$ is assumed to be generated from a pdf

$$\boldsymbol{y} \sim \ell(y|\theta)$$

called the *likelihood*, when viewed as a function of θ

2. Prior The parameter θ is itself endowed with a *prior pdf*

$$\boldsymbol{\theta} \sim \pi_0(\theta)$$

Considering θ as a random quantity and using π_0 to specify the prior information characterize the Bayesian approach to statistical modelling*

*In the Bayesian literature, approaches that do not follow this principle are referred to as *classical* or *frequentist*

A Note on Notations

1. In most probability and mathematical statistics texts, it is considered to be safer to use different notations for the random variable Y and the values y that it may take*
   - This is usually not the case in Bayesian texts
   - In this course, **boldface** is used to highlight quantities that really need to be interpreted as random variables
2. $\pi(\theta|y)$ should generally be understood as a conditional pdf (dominating measure is Lebesgue or counting*, unless otherwise specified) but will sometimes be used to denote probability measures: $\pi(d\theta|y)$ (continuous) or $\pi(\boldsymbol{\theta}=\theta|y)$ (discrete)
3. Bayesian texts usually make heavy use of *overloading*, denoting all densities by p or π and differentiating them only by their arguments ($\pi(y|\theta)$, $\pi(\theta)$, ...); this is not the default option in this course

*This helps understanding why $\mathrm{P}(Y=y)$ is not necessarily equal to 1!

*Although the integral notation is used by default

Bayesian Inference

The Bayesian Posterior

The Bayesian paradigm provides a principled way to perform inference through the posterior distribution

$$\pi(\theta|y) = \frac{\ell(y|\theta)\pi_0(\theta)}{\int_\Theta \ell(y|\theta')\pi_0(\theta')\,d\theta'} \tag{1}$$

- The normalizing constant $Z(y) = \int_\Theta \ell(y|\theta)\pi_0(\theta)\,d\theta$ is usually called the *(Bayesian) evidence*
- Eq. (1) is often abbreviated to

$$\pi(\theta|y) \propto \ell(y|\theta)\pi_0(\theta)$$

⚠[*]

[*] $\propto$ should not hide factors that depend on θ

The Bayesian approach provides a general framework for choosing between competing models

- Given two models $\mathscr{M}_1$ and $\mathscr{M}_2$ with likelihoods $\ell_1(y|\theta_1)$, $\ell_2(y|\theta_2)$ and priors $\pi_{0,1}(\theta_1)$, $\pi_{0,2}(\theta_2)$ (respectively), define a model indicator $\boldsymbol{m}$ with prior $\mathrm{P}(\boldsymbol{m} = i) = p_{0,i}$
  The posterior π on $\{1\} \times \Theta_1 \cup \{2\} \times \Theta_2$ is

$$\frac{1}{Z(y)} \sum_{i=1}^{2} \mathbb{1}\{m = i\} \ell_i(y|\theta_i) \pi_{0,i}(\theta_i) p_{0,i}$$

  with global evidence

$$Z(y) = \sum_{i=1}^{2} \underbrace{\int_{\Theta_i} \ell_i(y|\theta_i) \pi_{0,i}(\theta_i) \mathrm{d}\theta_i}_{Z_i(y)} p_{0,i}$$

Bayes Factors

The posterior to prior odds ratio

$$\frac{\pi(\boldsymbol{m}=2|y)}{\pi(\boldsymbol{m}=1|y)}\frac{p_{0,1}}{p_{0,2}}=\frac{Z_2(y)}{Z_1(y)}$$

is called the Bayes factor (for model 2 vs model 1)

This is the preferred tool for deciding between the two models but the framework also suggests a different option

Model Averaging

If u is a function of interest that is defined under both $\mathscr{M}_1$ and $\mathscr{M}_\epsilon$, the expected posterior estimate is

$$\mathrm{E}(\boldsymbol{u}|y)=\sum_{i=1}^{2}\frac{Z_i(y)p_{0,i}}{\sum_{j=1}^{2}Z_j(y)p_{0,j}}\int_{\Theta_i}u(i,\theta_i,y)\pi_i(\theta_i|y)\mathrm{d}\theta_i$$

Of particular interest is the case where we seek to predict a new observation $\boldsymbol{y}_\star$ assumed to be an independent replica of $\boldsymbol{y}$ given $\boldsymbol{\theta}$

Bayesian Predictive Distribution

$$\boldsymbol{y}_\star | y \sim \int_\Theta \ell(y_\star|\theta)\pi(\theta|y)\mathrm{d}\theta$$

Similarly, in the model averaging setting

$$\boldsymbol{y}_\star | y \sim \sum_{i=1}^{2} \frac{Z_i(y)p_{0,i}}{\sum_{j=1}^{2} Z_j(y)p_{0,j}} \int_{\Theta_i} \ell(y_\star|\theta_i)\pi_i(\theta_i|y)\mathrm{d}\theta_i$$

Bayesian Sequential Inference

Assume that we are now given a sequence of observations $\boldsymbol{y_1, \ldots, y_n}$ conditionally independent given $\boldsymbol{\theta}$

Sequential Update of the Posterior

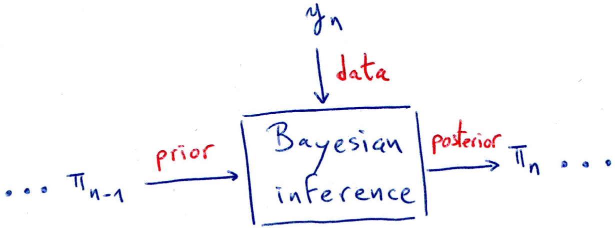
$$\pi_n(\theta|y_{1:n}) = \frac{Z_{n-1}(y_{1:n-1})}{Z_n(y_{1:n})}\ell(y_n|\theta)\pi_{n-1}(\theta|y_{1:n-1})$$

where

$$Z_n(y_{1:n}) = Z_{n-1}(y_{1:n-1})\int_\Theta \ell(y_n|\theta)\pi_{n-1}(\theta|y_{1:n-1})\mathrm{d}\theta$$

with associated predictive distribution

$$\int_\Theta \ell(y_{n+1}|\theta)\pi_n(\theta|y_{1:n})\mathrm{d}\theta$$


y_n
data
Bayesian inference
prior
posterior
$\dots \pi_{n-1}$
$\pi_n \dots$

Open Questions

1. How to summarize information from π?
2. What is influence of the prior π_0, how to set π_0?
3. How to determine π (or characteristics of it)?

How to summarize information from π?

Decision Theoretic Framework

- Given a loss function $c(\theta, \theta')$ that represents the cost of confounding θ and θ'
- Estimate θ by minimizing the *Bayesian risk*

$$\hat{\theta} = \arg\min_{t \in \Theta} \underbrace{\int_{\Theta} c(\theta, t)\pi(\theta|y)\mathrm{d}\theta}_{R_B(y,t)}$$

Posterior Mean Estimate

$c(\theta, \theta') = \|\theta - \theta'\|^2$ gives

$$\hat{\theta} = \mathrm{E}\left[\boldsymbol{\theta}|y\right]$$

with associated Bayes risk $\mathrm{trace}(\mathrm{Cov}\left[\boldsymbol{\theta}|y\right])$

The Bayesian estimator associated with the loss c also minimizes the total risk

$$\hat{\theta}(y) = \arg \min_{t:\mathsf{Y}\to\Theta} R_B(t)$$

where

$$\begin{aligned} R_B(t) &= \int_{\mathsf{Y}} \int_{\Theta} c(\theta, t(y))\ell(y|\theta)\pi_0(\theta)\mathrm{d}y\mathrm{d}\theta \\ &= \int_{\mathsf{Y}} \left(\int_{\Theta} \ell(y|\theta')\pi_0(\theta')\mathrm{d}\theta' \right) \int_{\Theta} c(\theta, t(y))\pi(\theta|y)\mathrm{d}\theta \,\mathrm{d}y \\ &= \int_{\Theta} \underbrace{\int_{\mathsf{Y}} c(\theta, t(y))\ell(y|\theta)\mathrm{d}y}_{R_F(\theta,t)} \pi_0(\theta)\mathrm{d}\theta \end{aligned}$$

and $R_F(\theta,\delta)$ is the *frequentist (or classical) risk*

$$R_F(\theta, t) = \mathrm{E}\left[\, c(\theta, t(\boldsymbol{y})) \,\middle|\, \theta \right]$$

Classical approaches focus on

- setups where the minimizer of $R_F(\theta, t)$ does not depend on θ (e.g., Gauss-Markov theorem for the linear model)
- the worst-case *("minimax")* approach $\arg\min_t \max_\theta r_F(\theta, t)$

The Bayesian Estimator

weights the frequentist risk according to the prior π_0

👍 Introduce a total ordering on the set of estimators t

👎 Depends on the choice of π_0

(in)Famous Counter-Example

The MAP (Maximum A Posteriori) Estimator

$$\hat{\theta}_{\mathsf{MAP}} = \arg\max_{\theta \in \Theta} \pi(\theta | y)$$

also called posterior mode estimate

- May be interpreted in the previous framework, when θ is a discrete parameter and c is the 0–1 loss

$$c(\theta, \theta') = \begin{cases} 0 & \text{if } \theta' = \theta \\ 1 & \text{otherwise} \end{cases}$$

- No decision-theoretic justification of the MAP when θ is a continuous parameter

Why the MAP?

- 👍 Computational Convenience and Ease of Interpretation

$$\hat{\theta}_{\mathsf{MAP}} = \arg\max_{\theta \in \Theta} \{\log \ell(y|\theta) + \log \pi_0(\theta)\}$$

  This is just *penalized* Maximum-Likelihood (ML) estimation with $-\log \pi_0(\theta)$ interpreted as a penalty on the parameter estimate

- 👍 Often viewed as an acceptable proxy for the posterior mean (based on asymptotic arguments)
- 👎 Not justified from a Bayesian perspective, often implies that π_0 be tuned from $\boldsymbol{y}$ to achieve desired results (so-called *"empirical Bayes"* approaches)
- 👎 May lead to incorrect decisions in the model selection context

A Different MAP: the Marginal MAP

When the parameter θ consists of a discrete θ_d and a continuous θ_c components, the global posterior mean is most often useless

Usual Two-Step Approach for this Case*

1. Marginal MAP Estimation of θ_d

$$\hat{\theta}_d = \arg\max_{\theta_d \in \Theta_d} \underbrace{\int_{\Theta_c} \pi(\theta_d, \theta_c | y) \mathrm{d}\theta_c}_{\pi(\theta_d | y)}$$

2. Conditional Posterior Mean Estimation of θ_c

$$\hat{\theta}_c = \mathrm{E}[\theta_c | y, \boldsymbol{\theta_d} = \hat{\theta}_d]$$

*Not the only option, check that the loss $c[(\theta_d,\theta_c),(t_d,t_c)] = \mathbb{1}\{t_d \neq \theta_d\}(t_c - \theta_c)^2$ implies a slightly different solution ✎

Please bear with me Although the details may seem a bit involved, the following example is very important:

- Posterior calculations are key to the Bayesian approach
- Many observations that pertain to this simple model are valid in great generality

Signal in Noise Model

Assume that we observe

$$\boldsymbol{y_i} = \boldsymbol{\theta_d}\boldsymbol{\theta_c} s_i + \boldsymbol{u_i} \qquad \text{for } i = 1, \ldots, n$$

where

- $(\boldsymbol{u_i})_{i \geq 1}$ is an iid $\mathcal{N}(\cdot|0, w)$-distributed noise sequence
- $(s_i)_{i \geq 1}$ is a known (deterministic) signal
- $\boldsymbol{\theta_d} \in \{0, 1\}$ is the signal presence indicator parameter
- $\boldsymbol{\theta_c} \in \mathbb{R}$ is the (unknown) signal amplitude parameter

For $(\boldsymbol{\theta_d}, \boldsymbol{\theta_c})$ we assume the independent prior[*]

$$\pi_0(\theta_d, \theta_c) = \underbrace{p^{\theta_d}(1-p)^{1-\theta_d}}_{\text{Bernoulli}} \underbrace{\mathcal{N}(\theta_c|0, v_0)}_{\text{Gaussian}}$$

[*] Equivalent model specification $\pi_0(\mathrm{d}\theta) = p\delta_0(\mathrm{d}\theta) + (1-p)\mathcal{N}(\theta|0, v_0)\mathrm{d}\theta$

Signal in Noise Model

The Posterior

$$\pi(\boldsymbol{\theta}_d = 0|y) = \frac{1-p}{(1-p) + p\sqrt{\frac{v_n}{v_0}}\exp\left(\frac{v_n\langle s,y\rangle^2}{2w^2}\right)}$$

$$\pi(\theta_c|\boldsymbol{\theta}_d = 0, y) = \mathcal{N}(\theta_c|0, v_0) = \pi_0(\theta_c)$$

$$\pi(\theta_c|\boldsymbol{\theta}_d = 1, y) = \mathcal{N}\left(\theta_c\middle|\frac{v_n\langle s,y\rangle}{w^2}, v_n\right)$$

where

$$\begin{cases} v_n = \left(\frac{1}{v_0} + \frac{\|s\|^2}{w}\right)^{-1} \\ \|s\|^2 = \sum_{i=1}^n s_i^2 \\ \langle s,y\rangle = \sum_{i=1}^n s_i y_i \end{cases}$$

Details...

$$\pi(\theta_d,\theta_c|y) \propto \left(\frac{1}{2\pi w}\right)^{n/2} \exp\left(-\frac{1}{2w}\sum_{i=1}^{n}\left(y_i - s_i\theta_d\theta_c\right)^2\right) p^{\theta_d}(1-p)^{1-\theta_d}\left(\frac{1}{2\pi v_0}\right)^{1/2}\exp\left(-\frac{1}{2v_0}\theta_c^2\right)$$

$$\pi(0,\theta_c|y) \propto \cancel{\left(\frac{1}{2\pi w}\right)^{n/2}}\,\cancel{\exp\left(-\frac{1}{2w}\sum_{i=1}^{n} y_i^2\right)}(1-p)\boxed{\left(\frac{1}{2\pi v_0}\right)^{1/2}\exp\left(-\frac{1}{2v_0}\theta_c^2\right)}$$

$$\pi(1,\theta_c|y) \propto \cancel{\left(\frac{1}{2\pi w}\right)^{n/2}}\exp\left\{-\frac{1}{2w}\sum_{i=1}^{n}\left(\cancel{y_i^2} - 2y_i s_i\theta_c + s_i^2\theta_c^2\right)\right\} p\left(\frac{1}{2\pi v_0}\right)^{1/2}\exp\left(-\frac{1}{2v_0}\theta_c^2\right)$$

$$= \frac{p}{\sqrt{2\pi v_0}}\exp\left\{-\frac{1}{2}\left[\theta_c^2\left(\frac{1}{v_0}+\frac{\|s\|^2}{w}\right) - 2\,\theta_c\frac{<s,y>}{w}\right\}$$

$$= \frac{p}{\sqrt{v_0}}\boxed{\sqrt{\frac{v_n}{2\pi\, v_n}}\exp\left\{-\frac{1}{2v_n}\left(\theta_c - \frac{v_n<s,y>}{w}\right)^2\right\}}\exp\left(\frac{1}{2}\,\frac{v_n<s,y>^2}{w^2}\right)$$

The Marginal MAP + Conditional Posterior Mean

$$\hat{\theta}_d = \begin{cases} 1 & \text{if } \sqrt{\frac{v_n}{v_0}} \exp\left(\frac{v_n \langle s,y \rangle^2}{2w^2}\right) > \frac{1-p}{p} \\ 0 & \text{otherwise} \end{cases}$$

$$\hat{\theta}_c = \begin{cases} \frac{v_n}{w^2} \langle s, y \rangle & \text{if } \hat{\theta}_d = 1 \\ 0 & \text{if } \hat{\theta}_d = 0 \end{cases}$$

- The global posterior means yields

$$(\hat{\theta}_d, \hat{\theta}_c) = \left(\pi(\boldsymbol{\theta}_d = 1|y) \, , \, \pi(\boldsymbol{\theta}_d = 1|y) \frac{v_n}{w^2} \langle s, y \rangle \right)$$

- The global MAP chose $\hat{\theta}_d = 1$, $\theta_c = v_n / (\langle s, y \rangle w^2)$ when

$$\exp\left(\frac{v_n \langle s, y \rangle^2}{2w^2} \right) > \frac{1-p}{p}$$

and $(\hat{\theta}_d = 0, \theta_c = 0)$ otherwise

Interpretation of the Results

The Frequentist Perspective

Assuming a persistent signal, $\frac{1}{n}\sum_{i=1}^{n} s_i^2 \underset{n\to\infty}{\longrightarrow} \rho > 0$, and under $\mathrm{P}_{(\theta_d,\theta_c)}$

1. The Maximum Likelihood Estimator (MLE) is consistent:

$$\langle s, \boldsymbol{y}\rangle / \|s\|^2 \xrightarrow{\text{as.}} \theta_d\theta_c$$

2. θ_d may be estimated from the Generalized Likelihood Ratio (GLR) test $\boldsymbol{\hat{\theta}_d} = \mathbb{1}\{\boldsymbol{D} > s\}$ where

$$\boldsymbol{D} = 2\log\frac{\exp\left(-\frac{1}{2w}\sum_{i=1}^{n}(\boldsymbol{y_i} - \langle s, \boldsymbol{y}\rangle / \|s\|^2 s_i)^2\right)}{\exp\left(-\frac{1}{2w}\sum_{i=1}^{n}\boldsymbol{y_i}^2\right)} = \frac{\langle s, \boldsymbol{y}\rangle^2}{w\|s\|^2}$$

If s is kept fixed with n, the probability of wrongly deciding $\theta_d = 0$ tends to zero, while the probability of wrongly deciding $\theta_d = 1$ tends to $1 - F(s)$, where F is the cdf of the chi-square distribution with one degree of freedom

For the Bayesian Marginal MAP Estimator

As $n \to \infty$,

$$v_n = \left(\frac{1}{v_0} + \frac{\|s\|^2}{w} \right)^{-1} \equiv \frac{w}{\|s\|^2}$$

Hence,

- When $\hat{\theta}_d = 1$, $\hat{\theta}_c \equiv \langle s, y \rangle / \|s\|^2$, the Bayesian estimator is equivalent to the MLE
- The decision region for $\hat{\theta}_d = 1$ is approximately given by

$$\frac{\langle s, y \rangle^2}{w\|s\|^2} > 2 \log \frac{(1-p)\sqrt{v_0}}{p\sqrt{w}} + \log \|s\|^2$$

and equivalent to the GLR statistic with an increasing threshold; the Bayesian estimator is a consistent estimator of θ_d[*]

[*] When $\theta_d = 0$, $\langle s, \mathbf{y} \rangle^2 / (w\|s\|^2) \xrightarrow{L} \chi_1^2$ and when $\theta_d = 1$, $\mathrm{E}\langle s, \mathbf{y} \rangle^2 / (w\|s\|^2) = \theta_c^2 \|s\|^2 / w + 1 = O(n)$

The Global Posterior Mean

$$(\hat{\theta}_d, \hat{\theta}_c) = \left(\pi(\boldsymbol{\theta}_d = 1|y)\,, \pi(\boldsymbol{\theta}_d = 1|y)\,\frac{v_n}{w^2}\langle s, y\rangle\right)$$

Also estimates consistently θ_d and θ_c but with a significant over-shrinkage for θ_c

The Global MAP

chose $\hat{\theta}_d = 1$, $\theta_c = v_n/(\langle s, y\rangle w^2)$ when

$$\exp\left(\frac{v_n\langle s, y\rangle^2}{2w^2}\right) > \frac{1-p}{p}$$

Correctly detects that $\theta_d = 1$ but eventually fail with positive probability when $\theta_d = 0$

Role of the Prior

Shrinkage Effect $v_n < w / \|s\|^2$ and hence

$$|\mathrm{E}(\boldsymbol{\theta_c} | \boldsymbol{\theta}_d = 1, y)| < |\hat{\theta}_{\mathsf{ML}}|$$

Decreasing v_0 makes the shrinkage more aggressive, letting $v_0 \to \infty$ makes both estimators equivalent (and not only asymptotically equivalent)

Complexity Penalty

- Increasing values of v_0 renders the most complex alternative $(\boldsymbol{\theta_d} = 1)$ less likely due to the normalization penalty: larger spaces = bigger normalization constants
- If (s_i) was a large-dimensional signal, this effect would even be prevalent over the (more obvious) effect of p (smaller p makes $\boldsymbol{\theta_d} = 1$ less likely)
- Letting $v_0 \to \infty$ causes the alternative $\boldsymbol{\theta_d} = 1$ to be never accepted ⚠

Part I

Bayesian Modelling

2 The Posterior

3 The Prior

- Jeffrey's Priors
- Improper Priors
- The Normalization Penalty
- Conjugate Priors
- Exponential families

The Prior

- Choosing the Prior is a very important issue in some contexts
- But for information processing one generally sticks to conjugate prior families, tuning them to be somewhat noninformative*

Here we will just discuss

1. Jeffrey's prior
2. Improper prior
3. Conjuguate priors

*Without being to careful about what the term exactly means

Jeffreys' Rule

When the Fisher information $I(\theta) = \mathrm{E}\left[\frac{\mathrm{d}\log \ell(\boldsymbol{y}|\theta)}{\mathrm{d}\theta}\right]^2$ is flat (does not depend on θ), a noninformative choice for the prior is to take a flat prior as well

- Reasonable
- Suggests a rule that is coherent under *reparameterization*

Jeffreys' Rule

$$\pi_0(\theta) \propto I^{1/2}(\theta)$$

Location Parameter $\ell(y-\theta)$ $\pi_0(\theta) \propto 1$

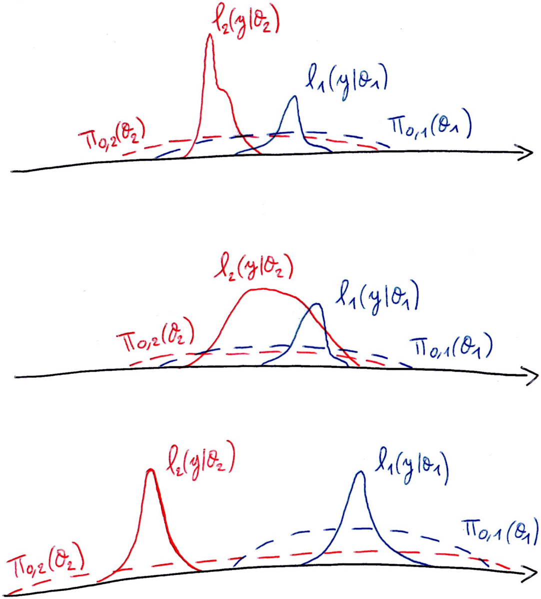
Scale Parameter $\frac{1}{\sigma}\ell(y/\sigma)$ $\pi_0(\sigma) \propto 1/\sigma$

Binomial Probability $p^y(1-p)^{1-y}$ $\pi_0(p) \propto p^{-1/2}(1-p)^{-1/2}$ (Dirichlet or Beta $(1/2,1/2)$ distribution)

Improper Priors

👎 Jeffreys' prior are most often improper, in the sense that they cannot be normalized to a proper pdf such that $\int_\Theta \pi_0(\theta)\mathrm{d}\theta = 1$

- Improper priors can nonetheless be used in cases where $Z(y) = \int_\Theta \ell(y|\theta)\pi_0(\theta)d\theta$ is finite for all y (despite the fact π_0 is not a real pdf)
- They often lead to easier calculations
- As well as to Bayesian estimators that are very close to ML estimators (see our *signal in noise* example)
- However, they usually imply incorrect complexity penalties in the model selection (or testing) context (our *signal in noise* example again)


$l_2(y|\theta_2)$
$l_1(y|\theta_1)$
$\pi_{0,2}(\theta_2)$
$\pi_{0,1}(\theta_1)$
$l_2(y|\theta_2)$
$l_1(y|\theta_1)$
$\pi_{0,2}(\theta_2)$
$\pi_{0,1}(\theta_1)$
$l_2(y|\theta_2)$
$l_1(y|\theta_1)$
$\pi_{0,1}(\theta_1)$
$\pi_{0,2}(\theta_2)$

Conjugate Prior

Conjugacy

Given a likelihood function $\ell(y|\theta)$, the family Π of priors π_0 on Θ is conjugate if the posterior $\pi(\theta|y)$ also belong to Π

In this case, posterior inference is tractable and reduces to updating the hyperparameters* of the prior

*The *hyperparameters* are parameters of the priors; they are most often not treated as a random variables

Discrete/Multinomial & Dirichlet[*]

If the observations consist of positive counts $\boldsymbol{y_1},\ldots,\boldsymbol{y_d}$ modelled by a Multinomial distribution

$$\ell(y|\theta,n)=\frac{n!}{\prod_{i=1}^{d}y_i!}\prod_{i=1}^{d}\theta_i^{y_i}$$

The conjugate family is the $\mathrm{Dirichlet}(\alpha_1,\ldots,\alpha_d)$ distribution

$$\pi_0(\theta|\alpha)=\frac{\Gamma(\sum_{i=1}^{d}\alpha_i)}{\prod_{i=1}^{d}\Gamma(\alpha_i)}\prod_{i}^{d}\theta_i^{\alpha_i-1}$$

defined on the probability simplex $(\theta_i\geq 0,\sum_{i=1}^{d}\theta_i=1)$, where Γ is the gamma function $\Gamma(\alpha)=\int_0^{\infty}t^{\alpha-1}\mathrm{e}^{-t}\mathrm{d}t$ $(\Gamma(k)=(k-1)!$ for integers $k)$

[*] Bernoulli/binomial & Beta, when $d=2$

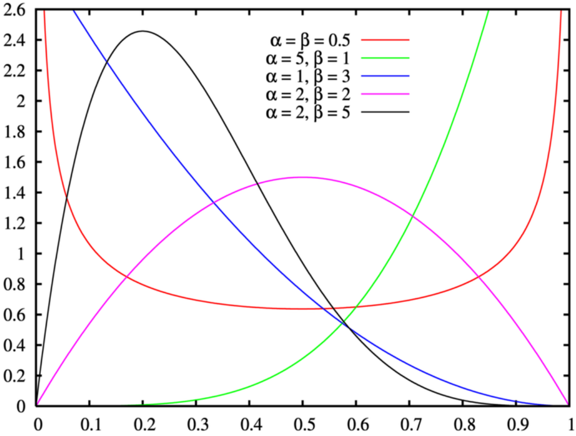


Figure: Dirichlet: 1D marginals

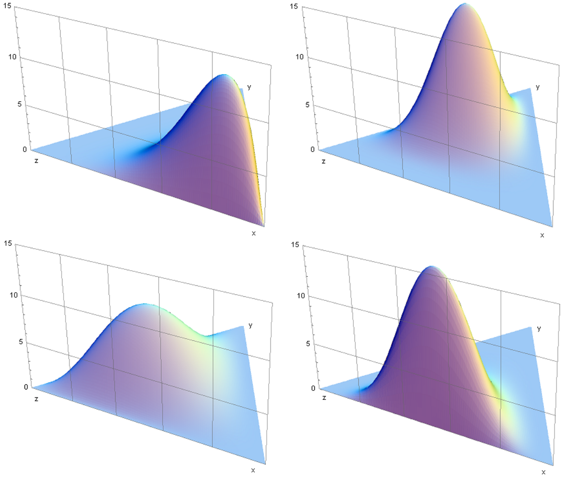


Figure: Dirichlet: 3D examples (projected on two dimensions)

Multinomial Posterior

Posterior

$$\pi_0(\theta|y) = \text{Dirichlet}(y_1 + \alpha_1, \ldots, y_d + \alpha_d)$$

Posterior Mean*

$$\left(\frac{y_i + \alpha_i}{\sum_{j=1}^{d} y_j + \alpha_j} \right)_{1 \le i \le d}$$

MAP

$$\left(\frac{y_i + \alpha_i - 1}{\sum_{j=1}^{d} y_j + \alpha_j - 1} \right)_{1 \le i \le d}$$

if $y_i + \alpha_i > 1$ for $i = 1, \ldots, d$

Evidence

$$Z(y) = \frac{\Gamma(\sum_{i=1}^{d} \alpha_i) \prod_{i=1}^{d} \Gamma(y_i + \alpha_i)}{\prod_{i=1}^{d} \Gamma(\alpha_i) \Gamma(\sum_{i=1}^{d} y_i + \alpha_i)}$$

*Also known as *Laplace smoothing* when $\alpha_i = 1$

Conjugate Priors for the Normal I

Conjugate Prior for the Normal Mean

For the $\mathcal{N}(y|\mu, w)$ distribution with iid observations $\boldsymbol{y_1}, \ldots, \boldsymbol{y_n}$, the conjugate prior for the mean $\boldsymbol{\mu}$ is Gaussian $\mathcal{N}(\mu|m_0, v_0)$:

$$\begin{aligned}\pi(\mu|y_{1:n}) &\propto \exp\left[-(\mu - m_0)^2/2v_0\right] \prod_{k=1}^{n} \exp\left[-(y_k - \mu)^2/2w\right] \\ &\propto \exp\left\{-\frac{1}{2}\left[\mu^2\left(\frac{1}{v_0} + \frac{n}{w}\right) - 2\mu\left(\frac{m_0}{v_0} + \frac{s_n}{w}\right)\right]\right\} \\ &= \mathcal{N}\left(\mu \,\middle|\, \frac{s_n + m_0 w/v_0}{n + w/v_0}, \frac{w}{n + w/v_0}\right)\end{aligned}$$

where $s_n = \sum_{k=1}^{n} y_k$ [*]

[*] And $y_{1:n}$ denotes the collection $y_1, \ldots, y_n$

Conjugate Priors for the Normal II

Conjugate Priors for the Normal Variance

If w is to be estimated and μ is known, the conjugate prior for w is the inverse Gamma distribution $\mathrm{Inv\text{-}Gamma}(w|\alpha_0,\beta_0)$:

$$\pi_0(w|\beta_0,\alpha_0) = \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} w^{-\alpha_0+1} \mathrm{e}^{-\beta_0/w}$$

and

$$\begin{aligned} \pi(w|y_{1:n}) &\propto w^{-(\alpha_0+1)} \mathrm{e}^{-\beta_0/w} \prod_{k=1}^{n} \frac{1}{\sqrt{w}} \exp\left[-(y_k-\mu)^2/2w\right] \\ &= w^{-(n/2+\alpha_0+1)} \exp\left[-(s_n^{(2)}/2+\beta_0)/w\right] \end{aligned}$$

where $s_n^{(2)} = \sum_{k=1}^{n}(Y_k-\mu)^2$.

The Gamma, Chi-Square and Inverses

The Gamma Distribution[*]

$$\text{Gamma}(\theta|\alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)}\theta^{\alpha-1}\mathrm{e}^{-\beta\theta}$$

where α is the shape and β the inverse scale parameter $\left(\mathrm{E}(\boldsymbol{\theta}) = \alpha/\beta,\ \mathrm{Var}(\boldsymbol{\theta}) = \alpha/\beta^2\right)$

- $\boldsymbol{\theta} \sim \text{Inv-Gamma}(\theta|\alpha,\beta)$: $1/\boldsymbol{\theta} \sim \text{Gamma}(\theta|\alpha,\beta)$
- $\boldsymbol{\theta} \sim \text{Chi-square}(\theta|\nu)$: $\boldsymbol{\theta} \sim \text{Gamma}(\theta|\nu/2, 1/2)$
- $\boldsymbol{\theta} \sim \text{Inv-Chi-square}(\theta|\nu)$:

  $1/\boldsymbol{\theta} \sim \text{Chi-Square}(\theta|\nu)$ or $\boldsymbol{\theta} \sim \text{Inv-Gamma}(\theta|\nu/2, 1/2)$

[*] MATLAB's convention is to use `gam*(a,b)`, where $b = 1/\beta$ is the scale parameter

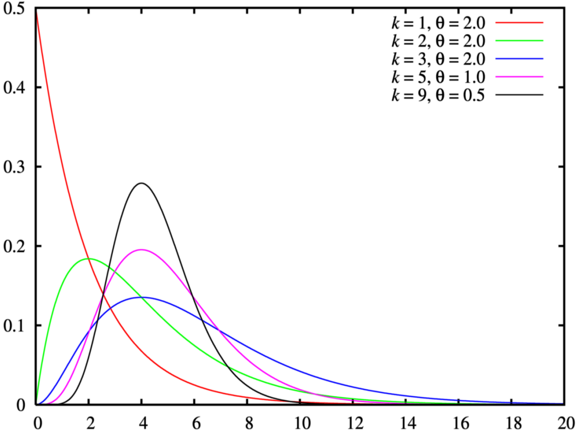


Figure: Gamma pdf $(k = \alpha, \theta = 1/\beta)$

Conjugate Priors for the Normal III

Conjugate Priors are However Available Only in Simple Cases

In the previous example there is no (useful) conjugate prior when both μ and w are unknown.

- Hence, it is very common to resort to independent marginally conjugate priors: eg., in the Gaussian case, take $\mathcal{N}(\mu|m_0, v_0)\,\text{Inv-Gamma}(w|\alpha_0, \beta_0)$ as prior, then $\pi(\mu|w, y)$ is Gaussian, $\pi(w|\mu, y)$ is inverse-gamma but $\pi(\mu, w|y)$ does not belong to a known family[*]
- There nonetheless exists some important multivariate extensions : Bayesian normal linear model, inverse-Wishart distribution for covariance matrices

[*] Although closed-form expressions for $\pi(\mu|y)$ and $\pi(w|y)$ are available

The previous examples are instances of a general framework

Exponential Family Distributions

$$\ell(y|\theta) = h(y) \exp\left[\langle s(y), \psi(\theta)\rangle - B(\theta)\right]$$

where $s(y)$ are the sufficient statistics

If ψ is an invertible mapping it is possible through the reparameterization $\eta = \psi(\theta)$ to rewrite the above in canonical (or natural) form

$$\ell(y|\eta) = h(y) \exp\left[\langle s(y), \eta\rangle - A(\eta)\right]$$

and A is called the log-partition function

Exponential family distributions play a very important role in statistics

1. Any likelihood-based estimator of θ (incl. Bayesian estimators) can only depend on y through the statistic $s(y)$
2. $\nabla^2 A(\eta) = \mathrm{Cov}(s(\boldsymbol{y})|\eta)$ and hence $\ell(y|\eta)$ is a log-concave function of η[*] ✎
3. $\nabla A(\eta) = \mathrm{E}(s(\boldsymbol{y})|\eta)$ and hence the maximum likelihood estimator $\hat{\eta}_{\mathrm{ML}}$ corresponding to independent observation $y_1, \ldots, y_n$ is the unique solution of the equation

$$\mathrm{E}(s(\boldsymbol{y})|\eta) = \frac{1}{n}\sum_{i=1}^{n} s(y_i)$$

[*] $\nabla^2 A(\eta)$ is also equal to the fisher information matrix for η

Conjugacy in Exponential Families

The conjugate distribution for

$$\ell(y|\theta) = h(y)\exp\left[\langle s(y), \psi(\theta)\rangle - B(\theta)\right]$$

is

$$\pi_0(\theta|\mu_0, \lambda_0) = Z_0^{-1}(\mu_0, \lambda_0)\exp\left[\langle \mu_0, \psi(\theta)\rangle - \lambda_0 B(\theta)\right]$$

where μ_0 has the same dimension as $s(y)$ and $\lambda_0 \in \mathbb{R}_+$[*]

After seeing n independent observations $y_1, \dots, y_n$, the posterior update consists in

$$\mu \longleftarrow \mu_0 + \sum_{i=1}^{n} s(y_i)$$

$$\lambda \longleftarrow \lambda_0 + n$$

[*] May be improper for some value of μ_0, λ_0

Part II

Latent Variable Models

4 Latent Variable Models
- Model and Graphical Representation
- Missing data
- Signal Decomposition and *CA Models
- Mixture Models
- Scale Mixture
- Mixture of Mixtures
- State-Space Models

5 MAP Parameter Inference
- Structure of the posterior
- The EM Algorithm
- Variational Inference
- Data Augmentation

6 Posterior Inference

Latent Variable Model

The observation $\boldsymbol{y}$ is viewed as the marginal outcome of a larger scale random experiment which involves an unobservable component $\boldsymbol{x}$

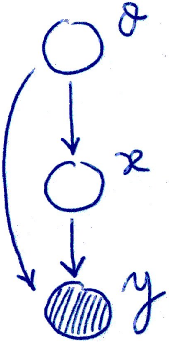
$$\ell(y|\theta) = \int_{\mathsf{X}} f(x, y|\theta)\mathrm{d}x$$

$\boldsymbol{x}$ is referred to as latent, missing or hidden data and $(\boldsymbol{x}, \boldsymbol{y})$ is complete the data

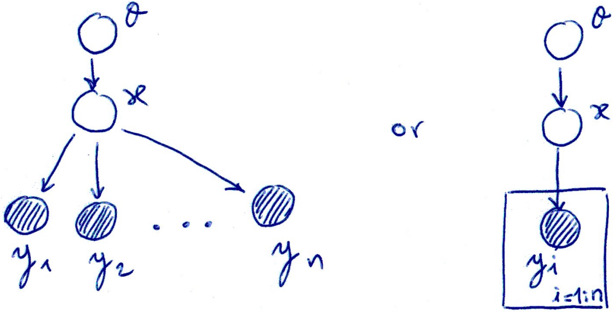
Usually, the model is naturally specified in a hierarchic fashion also called generative model

$$\boldsymbol{x} \sim q(x|\theta)$$
$$\boldsymbol{y}|x \sim \ell(y|x, \theta)$$

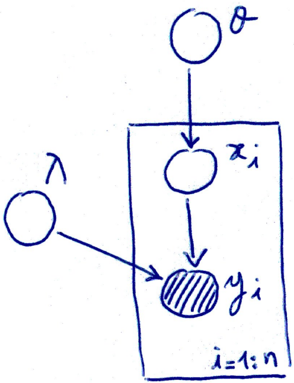
Graphical Representation: Bayesian Networks (Directed Graphical Models)



Plates for Conditionally Independent Replications



A Fairly Common Graph



Interpolation of Corrupted Audio Samples

[Ó Ruanaidh and Fitzgerald, 1996; Godsill and Rayner, 1998]

$$\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \sim \text{Gaussian AR Model}(a_1, \ldots, a_r, v)$$

$$\boldsymbol{y}_k = \boldsymbol{x}_k \quad \text{unless } k \text{ is the index of a corrupted sample}$$

Given some priors on $a_1, \ldots, a_r$ and v, how do we reconstruct the signal $\boldsymbol{x}_j$ at indices j where the corresponding observation is missing?

Here the goal is to recover $(\boldsymbol{x}_j)$ rather than to the estimate the autoregressive parameters

Estimation/Detection of Multicomponents Signals in Noise

[Andrieu and Doucet, 1999]

$$\boldsymbol{y_k} = \overbrace{\sum_{i=1}^{r} a_i \cos(\omega_i k + \varphi_i)}^{\boldsymbol{x_k}} + \boldsymbol{u_k}$$

where $(\boldsymbol{u_k})$ is a Gaussian white noise of variance v

- Given some priors on v, (a_i), (ω_i), (φ_i) and r, how do we estimate the number of components and their frequencies?
- How to recover the noiseless signal $(\boldsymbol{x_k})$? In this case, it is possible to bypass parameter estimation and use model averaging computing $\mathrm{E}\left[\boldsymbol{x_k} | y_{1:n}\right]$

Probabilistic PCA

[Tipping and Bishop, 1999] $\mathrm{Cov}(\boldsymbol{y}_k) \approx FF'$ where F is a rank r matrix is interpreted as

$$\boldsymbol{x}_\mathbf{k} \sim \mathcal{N}(0, I_r)$$
$$\boldsymbol{y}_\mathbf{k} = F\boldsymbol{x}_\mathbf{k} + \boldsymbol{u}_\mathbf{k}$$

where $\boldsymbol{u}_\mathbf{k} \sim \mathcal{N}(0, vI_d)$ (and $d \gg r$).

The above is fully equivalent to assuming that $\boldsymbol{y}_k$ is $\mathcal{N}(0, FF' + vI_d)$–distributed but the latent variable view suggests an algorithm for estimating FF' and v* as well as extensions

- Methods for estimating r
- Methods to deal with missing data
- Models with different marginal distribution

*Note that another important direct observation in this context is the convexity of $K \mapsto -\log|K| + \mathrm{trace}(KC)$

Finite Mixture Model

Mixture PDF

$$f(y) = \sum_{i=1}^{r} \alpha_i f_i(y)$$

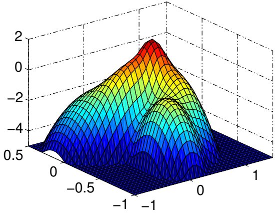
Missing Date Interpretation

$$\mathrm{P}(\boldsymbol{x_k} = i) = \alpha_i$$

$$\boldsymbol{y_k} | \boldsymbol{x_k} = i \sim f_i(y)$$

Mixture modelling is used in a variety of applications

- As a flexible tool for modelling densities
- As a clustering method



Mixture of 8 Gaussians (2D projection) trained from speech data

Scale Mixture

Eg., model for “sparse” regression [Tipping, 2001]

$$\boldsymbol{y} = v'\boldsymbol{x} + \boldsymbol{u}$$

where v contains observed covariates and the vector $\boldsymbol{x}$ of regression coefficients $\boldsymbol{x_i}$ is given an independent heavy-tailed prior specified in hierarchical form*:

1. $\boldsymbol{s_i} \sim \mathrm{Gamma}(s|\alpha_0, \beta_0)$
2. $\boldsymbol{x_i}|s_i \sim \mathcal{N}(x|0, 1/s_i)$

Also often used to model heavy-tailed observation noise, etc.

*Equivalent to $\pi_0(x) \propto \left(1 + \frac{x^2}{2\beta_0}\right)^{-(\alpha_0+1/2)}$, which is a scaled Student-t distribution when $\beta_0 = 1/2$

Admixtures, Simplicial Mixtures, Partial Membership Models

$$\boldsymbol{x_k} \sim \mathrm{Dirichlet}(\alpha_1, \ldots, \alpha_r)$$
$$\boldsymbol{y_k} | x_k \sim f(y | B x_k)$$

If the columns of B are interpreted as parameters of clusters, $\boldsymbol{y_k}$ is allowed to be explained by a convex combination of these clusters defined by the latent variable $\boldsymbol{x_k}$

- If f is such that $\mathrm{E}[\boldsymbol{y_k} | x_k] = B x_k$, this may be interpreted as a probabilistic variant of *CA decomposition
- If, in addition, $B \geq 0$, this is a form of (probabilistic) Non-Negative Matrix Factorization
- In some settings, normalization of $\boldsymbol{x}$ may be restrictive and the $\boldsymbol{x_i}$ are gamma-distributed[*] [Buntine and Jakulin, 2006]
- Most natural for exponential family f [Heller et al., 2008]

[*] A normalized vector of independent gamma-distributed variables is Dirichlet-distributed

Latent Dirichlet Association

[Blei et al., 2002; Griffiths and Steyvers, 2002] The "document" $\boldsymbol{y_k}$ consists of a vector of "word" counts and the columns of B are word frequency patterns (ie., $B_{ij} \geq 0$ and $\sum_{i=1}^{d} B_{ij} = 1$)

$$\boldsymbol{x_k} \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_d)$$

$$\boldsymbol{y_k} | x_k, n_k \sim \text{Multinomial}(y | n_k, Bx_k)$$

This is equivalent to the more usual generative representation

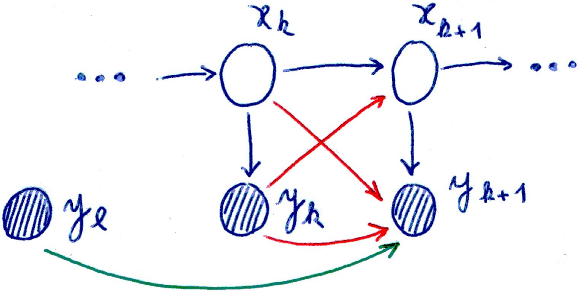
For $i = 1, \ldots, n_k$

1. Draw one "theme" $\boldsymbol{t_{k,i}}$ in $\{1, \ldots, r\}$ with probabilities $\boldsymbol{x_1}, \ldots, \boldsymbol{x_r}$
2. Draw one word $\boldsymbol{w_{k,i}}$ with probabilities given by the $\boldsymbol{t_{k,i}}$–th column of B

Collect these in the word count vector $\boldsymbol{y_k}$

The document is a *bag-of-words* drawn from different theme-specific word distributions, where the latent variable $\boldsymbol{x_k}$ represents the document-level repartition of the different themes

State-Space Models, Hidden Markov Models (HMMs), Switching Autoregresions, . . .



Part II

Latent Variable Models

We start by examining the simpler case where we want to compute the MAP estimate of θ

The simplest possible case

Assume that $(\boldsymbol{x},\boldsymbol{y})^*$ has an exponential family distribution in natural parameterization

$$f(x,y|\theta) = h(x,y)\exp\left[\langle s(x,y),\theta\rangle - nA(\theta)\right]$$

and that we use a conjugate prior

$$\pi_0(\theta|\mu_0,\lambda_0) \propto \exp\left[\langle \mu_0,\theta\rangle - \lambda_0 A(\theta)\right]$$

The posterior is given by

$$\pi(\theta|y) \propto \int h(x,y)\exp\left[\langle s(x,y)+\mu_0,\theta\rangle - (n+\lambda_0)A(\theta)\right]\mathrm{d}x$$

[*] To simplify the notations, we assume that we have n independent observations but use $\boldsymbol{x}$ and $\boldsymbol{y}$ to denote the collections of latent and observed variables respectively; thus $s(x,y) = \sum_{i=1}^{n} s(x_i,y_i)$

Problem $\pi(\theta|y)$ is not log-concave any more
Optimizing $\log\pi(\theta|y)$ wrt θ is a complex numerical optimization task and the presence of local maxima is, to some extent, unavoidable

We do however have a simple closed-form expression of the gradient wrt θ

$$\nabla_\theta \log\pi(\theta|y) = \underbrace{\frac{\int s(x,y)\,h(x,y)\exp\left[\langle s(x,y),\theta\rangle - nA(\theta)\right]\mathrm{d}x}{\int h(x,y)\exp\left[\langle s(x,y),\theta\rangle - nA(\theta)\right]\mathrm{d}x}}_{\mathrm{E}[s(\boldsymbol{x},y)|y,\theta]}$$

$$- n\nabla_\theta A(\theta) + \left\{\mu_0 - \lambda_0\nabla_\theta A(\theta)\right\}$$

$$= \mathrm{E}\left[\nabla_\theta \log f(\boldsymbol{x},y|\theta)\,\middle|\,y,\theta\right] + \nabla_\theta \log\pi_0(\theta)$$

attributed to Fisher (see disc. of [Dempster et al., 1977])

The Expectation-Maximization Algorithm [Dempster et al., 1977]

Given a parameter estimate $\hat{\theta}_k$

1. Compute

$$q_{\hat{\theta}_k}(\theta) = \mathrm{E}\left[\log f(\boldsymbol{x}, y|\theta) \,\middle|\, y, \hat{\theta}_k\right] + \log \pi_0(\theta)$$

2. Update the parameter estimate to

$$\hat{\theta}_{k+1} = \arg\max_{\theta \in \Theta} q_{\hat{\theta}_k}(\theta)$$

Rationale

1. Because of Fisher relation, the algorithm can only stop in a stationary point of the log-posterior $\log \pi(\theta|y)$*
2. It is an ascent algorithm:

$$
\begin{aligned}
& q_{\hat{\theta}_k}(\hat{\theta}_{k+1}) - q_{\hat{\theta}_k}(\hat{\theta}_k) \\
&\qquad = \mathrm{E}\left[\log \frac{f(\boldsymbol{x}, y|\hat{\theta}_{k+1})}{f(\boldsymbol{x}, y|\hat{\theta}_k)} \middle| y, \hat{\theta}_k \right] + \log \frac{\pi_0(\hat{\theta}_{k+1})}{\pi_0(\hat{\theta}_k)} \\
&\qquad\qquad = \underbrace{\mathrm{E}\left[\log \frac{f(\boldsymbol{x}|y, \hat{\theta}_{k+1})}{f(\boldsymbol{x}|y, \hat{\theta}_k)} \middle| y, \hat{\theta}_k \right]}_{\leq 0} + \log \frac{\pi(\hat{\theta}_{k+1}|y)}{\pi(\hat{\theta}_k|y)}
\end{aligned}
$$

*See [Wu, 1983] for necessary topological and regularity assumptions

The EM Intermediate Quantity as a Minorizing Surrogate

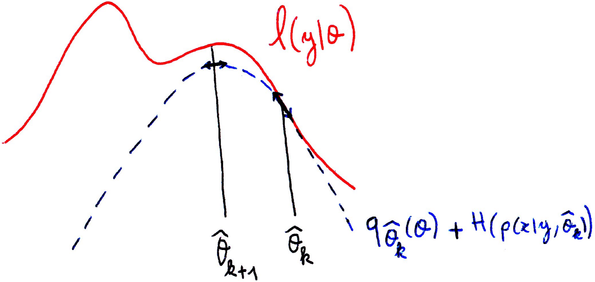


Figure: One EM iteration for ML estimation

Rationale (Contd.)

3. In exponential family models with conjugate priors, the EM principle does define a practical algorithm when
   1. $\mathrm{E}\left[\, s(\boldsymbol{x}, y) \,\middle|\, y, \hat{\theta}_k \right]$ may be evaluated
   2. The complete-data ML problem $\max_{\theta\in\Theta}\{\langle S, \psi(\theta)\rangle - nB(\theta)\}$ may be solved for all feasible S

   Then,

$$\hat{\theta}_{k+1} = \arg\max_{\theta\in\Theta} \left\{ \langle \mathrm{E}\left[\, s(\boldsymbol{x}, y) \,\middle|\, y, \hat{\theta}_k \right] + \mu_0, \psi(\theta)\rangle - (n + \lambda_0) B(\theta) \right\}$$

For the natural parameterization, $\hat{\theta}_{k+1}$ is the unique solution of*

$$\frac{\mathrm{E}\left[\, s(\boldsymbol{x}, \boldsymbol{y}) | \theta \right]}{n} = \frac{\mathrm{E}\left[\, s(\boldsymbol{x}, y) \,\middle|\, y, \hat{\theta}_k \right] + \mu_0}{n + \lambda_0}$$

* $\mathrm{E}\left[\, s(\boldsymbol{x_1}, \boldsymbol{y_1}) \,\middle|\, \theta \right] = (n + \lambda_0)^{-1} \left(\sum_{i=1}^{n} \mathrm{E}\left[\, s(\boldsymbol{x_i}, y_i) \,\middle|\, y_i, \hat{\theta}_k \right] + \mu_0 \right)$ for iid observations

There are many variants

- Partial update of θ
- Limited increase of q
- "Accelerated" methods*
- Monte Carlo EM, ie. approximating $\mathrm{E}\left[\, s(\boldsymbol{x},y) \middle| y,\hat{\theta}_k \right]$ by Monte Carlo averages*
- Iterated Conditional Mode (image MRF), Viterbi training (speech HMM), Classification EM (mixtures) and the likes:

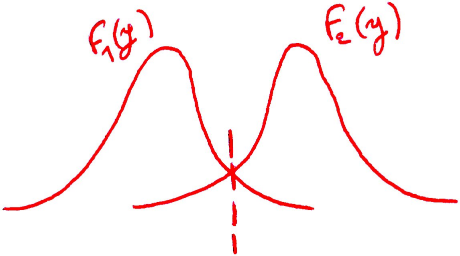
  $$\text{replace } E\left[\, s(\boldsymbol{x},y) \middle| y,\hat{\theta}_k \right] \text{ by } s(x_k^\star, y)$$

  where $x_k^\star$ is the most likely sequence given y and $\hat{\theta}_k$
  May be successful but greedy and biased, possibly unstable in cases where y is poorly informative about x and θ

* Simple one: compute the gradient and use a quasi Newton optimizer

* Requires MCMC simulations

The Imputation Bias



Problematic if parameter estimation is the main objective but not necessarily so in other contexts (eg. clustering)

In addition The term *EM* is often used loosely for algorithms that do not follow the previous principle but use the surrogate minorization function trick (most often using convex inequalities)

[Hunter and Lange, 2004] propose to call these MM algorithms

Sometimes the term EM is also associated to coordinate ascent algorithms (that are not MM algs.)

Variational Approximation [Neal and Hinton, 1999]

$$
\begin{aligned}
\log \pi(\theta|y) = C^{st} + \log \int_{\mathsf{X}} f(x,y|\theta)\pi_0(\theta)\mathrm{d}x \\
\geq \int_{\mathsf{X}} \log \frac{f(x,y|\theta)\pi_0(\theta)}{q(x)} q(x)\mathrm{d}x
\end{aligned}
$$

The variational algorithm proceeds by alternate maximimizations of the rhs wrt $\theta \in \Theta$ and $q \in \mathscr{Q}$

Variational Algorithm in Exponential Family [Jordan et al., 1999]

1. For fixed $\hat{q}_k$

$$\hat{\theta}_{k+1} = \arg\max_{\theta\in\Theta} \left\{ \langle \mathrm{E}_q[s(\boldsymbol{x}, y)] + \mu_0, \psi(\theta)\rangle - (n+\lambda_0)B(\theta) \right\}$$

2. For fixed $\hat{\theta}_{k+1}$

$$\hat{q}_{k+1} = \arg\max_{q\in\mathcal{Q}} \int_{\mathsf{X}} \log \frac{f(x, y|\hat{\theta}_{k+1})\pi_0(\hat{\theta}_{k+1})}{q(x)} q(x)\mathrm{d}x$$

which is a convex optimization problem whenever $\mathcal{Q}$ is a convex set

If $\mathcal{Q} \supset \{p(x|y,\theta); \theta \in \Theta\}$, then $\hat{q}_{k+1} = p(x|y, \hat{\theta}_{k+1})$ and one recovers the EM algorithm (there is then no variational approximation)

Latent Variable Models are Used in Two Very Different Contexts

Black-Box or Behavioral Modelling

Mostly $\ell(y|\theta)$ matters and $\boldsymbol{x}$ is essentially fictitious

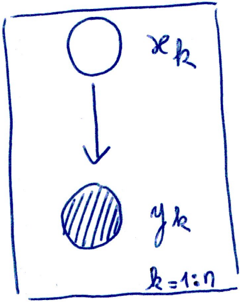
Physical Modelling

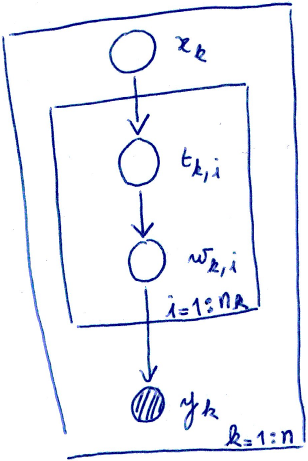
The latent model is used to represent a system with data corruption or loss and the definition of $\boldsymbol{x}$ is motivated by a physical interpretation

Especially in the first case, it is important to remember that there is an infinity of ways in which $\boldsymbol{x}$ could be defined for a given $\ell(y|\theta)$

Different Levels of Data Augmentation

LDA (Latent Dirichlet Association)





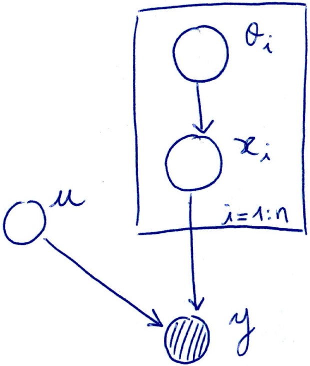
Mixture of Student-t Distributions [Peel and McLachlan, 2000]

For robust mixture estimation, one can replace the Gaussian by (multivariate) Student-t distributions:
$\ell(y|\theta) = \sum_{i=1}^{r} \alpha_i \mathrm{Student}(y|\nu, \mu_i, \Sigma_i)$

To use the EM algorithm in this context, one may use the scale mixture representation of the Student-t distribution

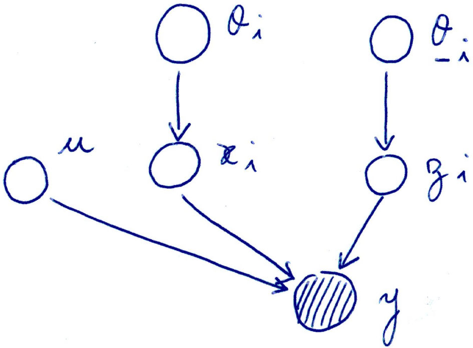
1. $\boldsymbol{z} \sim \mathrm{Student}(z|\nu)$
2. $\boldsymbol{y}|\boldsymbol{z} \sim \mathcal{N}(y|\mu, \nu/z\Sigma)$

Conditional-Dependent Data Augmentation



In the above model assume that we want to update θ_i only, given the other parameters (performing alternate maximizations)

The idea used in [Fessler and Hero, 1995] (for EM) and [Doucet et al., 2005] (for MCMC) is to use non-consistent conditional-dependent completions that preserve $\pi(\theta_i|y,\theta_{-i})$



For instance, if $y = \sum_{i=1}^n x_i + u$, $z_i = \sum_{j\neq i} x_j$ (assuming that the law of z_i given $\theta_{-i} = (\theta_j)_{j\neq i}$ is available)

Part II

Latent Variable Models

4 Latent Variable Models

- Model and Graphical Representation
- Missing data
- Signal Decomposition and *CA Models
- Mixture Models
- Scale Mixture
- Mixture of Mixtures
- State-Space Models

5 MAP Parameter Inference

- Structure of the posterior
- The EM Algorithm
- Variational Inference
- Data Augmentation

6 Posterior Inference

Posterior Inference

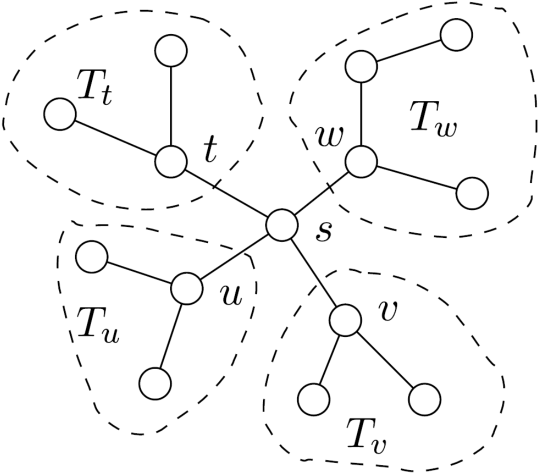
Both latent variable inference and parameter inference through the EM algorithm rely on evaluation of expectations under $p(x|y,\theta)$

Obviously, this usually only requires the use of Bayes' rule

$$p(x|y,\theta) = \frac{\ell(y|x,\theta)\,q(x|\theta)}{\int_{\mathsf{X}} \ell(y|x',\theta)\,q(x'|\theta)\mathrm{d}x'}$$

In models with more complex dependencies, this can be more challenging

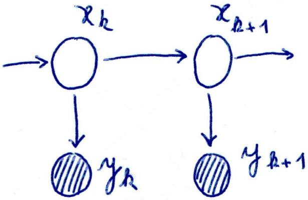
There exists a general algorithm (called sum-product or belief propagation) for doing so in models whose Bayesian network representation forms a tree [Wainwright and Jordan, 2008]



From [Wainwright and Jordan, 2008]

Posterior Inference in State-Space Models

- The filtering pdfs $(p(x_k|y_{1:k},\theta))_{k\geq 1}$ may be determined recursively
- The smoothing pdfs $(p(x_k, x_{k+1}|y_{1:n},\theta))_{1\leq k\leq n-1}$ can be determined from the corresponding filtering pdfs $(p(x_k|y_{1:k},\theta))_{1\leq k\leq n}$ by backward induction[a]
- Proceeding similarly, one can simulate state sequences $\boldsymbol{x_{1:n}}$ from $p(x_{1:n}|y_{1:n},\theta)$
- For a fixed function s, $\mathrm{E}\left[\sum_{j=1}^{k} s(x_j)|y_{1:k},\theta\right]$ may be updated recursively using an auxiliary recursion



[a] Rauch-Tung-Striebel smoothing in linear SSMs

Part III

Markov Chain Monte Carlo Methods

7 MCMC Basics

- Why Do We Need MCMC?
- (Minimal) Markov Chain Theory
- MCMC Essentials

8 The Gibbs Sampler

9 Rao-Blackwellization

10 Auxiliary Targets

Why Do We Need MCMC for Bayesian Inference?

Unfortunately, the methods discussed so far are usually not applicable in more complex models:

- When computation of $\mathrm{E}(s(\boldsymbol{x}, y)|y, \theta)$ is no more feasible
- When the prior $\pi_0(\theta)$ cannot be chosen in a conjugate family
- When the inference involves competing models*

Even in simpler models, estimators other than the MAP (so-called *fully Bayesian inference*) can generally not be computed using the algorithms described so far

*Particularly so when the inference involves a potentially unlimited number of models as in *Bayesian nonparametric models*

Usual solutions include

1. Variational Methods
   - 👍 Computations scale even for larger models and datasets
   - 👎 (Almost) no control over the approximation error
2. Monte Carlo Methods
   - 👍 The random approximation error is controlled by the computation time
   - 👎 Theoretical and practical performance not always guaranteed for complex models and large datasets

- MC clearly wins when dealing with moderate-dimensional problems and in cases where inference bias is not tolerable (statistics, physics, . . .)
- In machine learning the issue is less clear-cut*

*And the preferred answer somewhat subject to hype

In the following, we give a quick introduction to MCMC (Markov Chain Monte Carlo) techniques, focussing on the Gibbs sampler*

We denote by $\pi(z)$ the *target* density, typically this a full posterior $\pi(x,\theta|y)$ or conditional $\pi(x|\theta,y)$ and is known only up to an unknown normalizing constant

*Typically preferred when dealing with behavioral models with conjugate priors. In physics and, to some extent, statistics the situation is almost reversed and Metropolis-Hastings MCMC is the basic tool

Basic Monte Carlo Doesn't Solve the Problem

Self-Normalized Importance Sampling

Simulate $(\boldsymbol{z}^{(j)})_{1\leq j\leq m}$ from q and estimate $\mathrm{E}_\pi[g(\boldsymbol{z})]$ by

$$\frac{\sum_{j=1}^{m} \boldsymbol{w}^{(j)} g(\boldsymbol{z}^{(j)})}{\sum_{i=1}^{n} \boldsymbol{w}^{(i)}}$$

where

$$\boldsymbol{w}^{(j)} = \pi(\boldsymbol{z}^{(j)})/q(\boldsymbol{z}^{(j)})$$

Very useful* but does not scale well to large dimensions

* Main tool in *Sequential Monte Carlo methods*

Transition Kernel

The probability distribution of a Markov chain $(\boldsymbol{z}^{(j)})_{j\geq 1}$ on Z is fully determined by its initial distribution $\nu(z)$ and its transition kernel $k(z, z')$, which are such that

$$\mathrm{P}(\boldsymbol{z}^{(1)} \in A) = \int_A \nu(z)\mathrm{d}z$$

$$\mathrm{P}(\boldsymbol{z}^{(j)} \in A | \boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(j-1)}) = \int_A k(\boldsymbol{z}^{(j-1)}, z)\mathrm{d}z$$

Chapman-Kolmogorov Equations

$$\mathrm{P}(\boldsymbol{z}^{(\boldsymbol{j}+\mathbf{1})} \in A) = \int_{z \in \mathsf{Z}} \int_{z' \in A} \nu(z) k^j(z, z') \mathrm{d}z \mathrm{d}z'$$

where

$$k^j(z, z'') = \int k^{j-1}(z, z') k(z', z'') \mathrm{d}z'$$

- $k^j(z^{(1)}, z)$ is the conditional pdf of $z^{(j+1)}$ given $z^{(1)}$

Stationary Distribution

Definition

π is stationary for k if

$$\int \pi(z) k(z, z') \mathrm{d}z = \pi(z')$$

Hence π is a stationary point of the kernel k, viewed as an operator on pdfs

- It is easily checked that this implies that if $\nu = \pi$,

$$\mathrm{P}(\boldsymbol{z}^{(j)} \in A) = \int_A \pi(z) \mathrm{d}z$$

for all $j \geq 1$

Detailed Balance Condition and Reversibility

Determining the stationary distribution(s) is hard in general, except in cases where the following stronger condition holds.

Detailed Balance Condition

$$\pi(z)k(z,z') = \pi(z')k(z',z) \qquad \text{for all } (z,z') \in Z^2$$

The chain is then said to be π-reversible and π is a stationary distribution

Proof

$$\int \pi(z)k(z,z')\mathrm{d}z = \int \pi(z')k(z',z)\mathrm{d}z = \pi(z')$$

Convergence to Stationary Distribution

If π is a stationary distribution, and under additional regularity conditions not discussed here, the following properties hold

Convergence in Distribution

$$\mathrm{E}[g(\boldsymbol{z}^{(\boldsymbol{m})})] \to \int_{\mathsf{Z}} g(z)\pi(z)\mathrm{d}z \quad \text{(irrespectively of } \nu\text{)}$$

Law of Large Numbers (Ergodic theorem)

$$\frac{1}{m}\sum_{j=1}^{m} g(\boldsymbol{z}^{(\boldsymbol{j})}) \xrightarrow{\text{as.}} \int_{\mathsf{Z}} g(z)\pi(z)\mathrm{d}z$$

Central Limit Theorem

$$\frac{\sqrt{m}}{\sigma_{\pi,k,g}}\left[\frac{1}{m}\sum_{j=1}^{m} g(\boldsymbol{z}^{(\boldsymbol{j})}) - \int_{\mathsf{Z}} g(z)\pi(z)\mathrm{d}z\right] \xrightarrow{L} \mathcal{N}(0,1)$$

Markov Chain Monte Carlo (MCMC) in a Nutshell

1. Given a target distribution π, which may be known up to a constant only, find a transition kernel k which is π-reversible, ie., such that

$$\pi(z)k(z,z') = \pi(z')k(z',z)$$

2. Simulate a (long) section $\boldsymbol{z}^{(1)},\ldots,\boldsymbol{z}^{(m)}$ of a chain with kernel k started from an arbitrary point $\boldsymbol{z}^{(1)}$ and compute the Monte Carlo estimate

$$\widehat{\pi}(g) = \frac{1}{m}\sum_{j=1}^{m} g(\boldsymbol{z}^{(j)})$$

of $\int_{Z} f(z)\pi(z)\mathrm{d}z$, perhaps discarding in the sum the very first iterations (so called burn-in period)

Part III

Markov Chain Monte Carlo Methods

7 MCMC Basics

- Why Do We Need MCMC?
- (Minimal) Markov Chain Theory
- MCMC Essentials

8 The Gibbs Sampler

9 Rao-Blackwellization

10 Auxiliary Targets

Partial Updates

In most cases of interest $\boldsymbol{z} = (\boldsymbol{z_1}, \ldots, \boldsymbol{z_d})$, and the individual moves only update some components of $\boldsymbol{z}$:

- The i–th component $\boldsymbol{z_i}$ is updated from the kernel $k(z, z_i')$
- The remaining components $\boldsymbol{z_{-i}}$ are left unchanged

The detailed balance condition becomes

$$\pi(z_i|z_{-i})\,k(z, z_i') = \pi(z_i'|z_{-i})\,k(z', z_i)$$

✎

To ensure *irreducibility*, all components need to be updated in turn either systematically[*] or in a random scanning order

[*] Prevent the complete chain to be reversible

The Gibbs sampler is based on the choice $k(z, z_i') = \pi(z_i'|z_{-i})$

Gibbs Sampler

Starting from an initial arbitrary state $\boldsymbol{z}^{(1)}$, update the current state $\boldsymbol{z}^{(j)} = (\boldsymbol{z}_1^{(j)}, \ldots, \boldsymbol{z}_d^{(j)})$ to a new state $\boldsymbol{z}^{(j+1)}$ as follows.

For $i = 1, 2, \ldots, d$: Simulate $\boldsymbol{z}_i^{(j+1)}$ from

$$\pi(z_i|z_1^{(j+1)}, \ldots, z_{i-1}^{(j+1)}, z_{(i+1)}^{j}, \ldots, z_d^{(j)})$$

The above is the systematic scan Gibbs sampler; one may also use the random scan Gibbs sampler by choosing at random the index i of the component to be updated

Gaussian Posterior

In the Gaussian model $\boldsymbol{y_1}, \ldots, \boldsymbol{y_n} \sim_{iid} \mathcal{N}(y|\mu, v)$ with constant (improper) priors for both μ and v^*, we have

$$\begin{aligned} \boldsymbol{\mu}|y_{1:n}, v &\sim \mathcal{N}\left(\mu \,\middle|\, \tfrac{1}{n}\textstyle\sum_{i=1}^{n} y_i, \tfrac{v}{n}\right) \\ \boldsymbol{v}|y_{1:n}, \mu &\sim \text{Inv-Gamma}\left(v \,\middle|\, \tfrac{n}{2} - 1, \tfrac{1}{2}\textstyle\sum_{i=1}^{n}(y_i - \mu)^2\right) \end{aligned}$$

- This suggests a simple Gibbs sampler for simulating from the posterior of (μ, v)

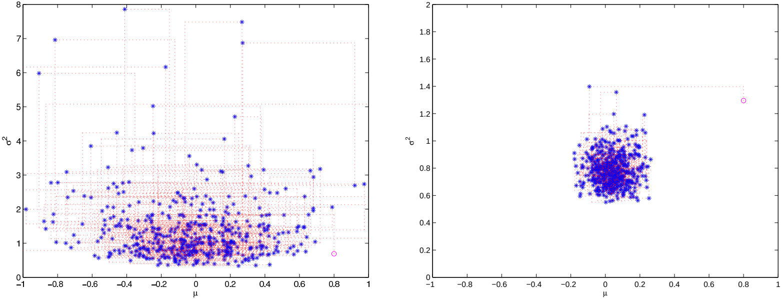
[*] Arguably not the best choice for v

The Systematic Gibbs Sampler for Gaussian Observations (in MATLAB/OCTAVE)[*]

```
n = length(Y);
S = sum(Y);
mu = S/n;
for i = 1:500        % Small number of iterations
   S2 = sum((Y-mu).^2);
   v = 1/gamrnd(n/2-1,2/S2);
   mu = S/n + sqrt(v/n)*randn;
end
```

[*] Observe that for loops are unavoidable and hence that MATLAB/OCTAVE is not very MCMC-friendly

Example of Results with, Left $n = 10$ Observations; Right, $n = 100$ Observations from the $\mathcal{N}(0,1)$ Distribution



Number of Iterations 1, 2, 3, 4, 5, 10, 25, 50, 100, 500

Part III

Markov Chain Monte Carlo Methods

7 MCMC Basics
- Why Do We Need MCMC?
- (Minimal) Markov Chain Theory
- MCMC Essentials

8 The Gibbs Sampler

9 Rao-Blackwellization

10 Auxiliary Targets

Rao-Blackwellization

If we can find $(\boldsymbol{z}, t)$ such that $\boldsymbol{z} \sim \pi$, $\boldsymbol{t} \sim \nu$ and $\mathrm{E}\left[g(\boldsymbol{z})|\boldsymbol{t}\right]$ may be computed in closed-form,
MCMC simulation $\boldsymbol{t}^{(1)}, \ldots, \boldsymbol{t}^{(n)}$ are performed using ν as target pdf and the Rao-Blackwellized estimator

$$\widehat{\boldsymbol{\pi}^{RB}}(\boldsymbol{g}) = \frac{1}{m}\sum_{j=1}^{m} \mathrm{E}\left[\, g(\boldsymbol{z}) \middle| \, \boldsymbol{t}^{(j)} \right]$$

is used instead of $\hat{\boldsymbol{\pi}}(\boldsymbol{g}) = \frac{1}{m}\sum_{j=1}^{m} g(\boldsymbol{z}^{(j)})$

For independent simulations, the Rao-Blackwell Theorem[*] shows that

$$\mathrm{Var}\left(\widehat{\boldsymbol{\pi}^{RB}}(\boldsymbol{g})\right) \leq \mathrm{Var}\left(\hat{\boldsymbol{\pi}}(\boldsymbol{g})\right)$$

This does not necessarily hold true for MCMC simulations, but empirically it does in most settings

[*] $\mathrm{Var}(\mathrm{E}[g(\boldsymbol{z})|t]) + \mathrm{E}(\mathrm{Var}[g(\boldsymbol{z})|t]) = \mathrm{Var}[g(\boldsymbol{z})]$

The term *Rao-Blackwellization* is used loosely in MCMC to describe approaches in which explicit marginalizations replace simulations

A first common use of the idea is to run the MCMC simulation on an augmented target and to use Rao-Blackwellization as a post-processing for computing estimates

For instance, when using the Gibbs sampler a natural Rao-Blackwellized estimator of the marginal pdf of $\boldsymbol{z_i}$ is

$$\frac{1}{m}\sum_{j=1}^{m}\pi(z_i|\boldsymbol{z}_{-\boldsymbol{i}}^{(\boldsymbol{j})})$$

Another option is to take profit of Rao-Blackwellization during the simulations resulting in marginalized or collapsed Gibbs samplers

In a Bayesian latent variable model, a typical scheme for the Gibbs sampler is to alternate

$$\begin{cases} \boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta} \\ \boldsymbol{\theta}|\boldsymbol{x},\boldsymbol{y} \end{cases}$$

But for an exponential family complete-data model with conjugate prior, the pdf $\pi(\theta|x,y)$ is available in closed-form, thus allowing for Rao-Blackwellization

Recall that for an exponential family complete-data model with conjugate prior

$$\pi(x, \theta | y) \propto h(x, y) \exp\left[\langle s(x, y), \psi(\theta)\rangle - nB(\theta)\right]$$
$$Z_0^{-1}(\mu_0, \lambda_0) \exp\left[\langle \mu_0, \psi(\theta)\rangle - \lambda_0 B(\theta)\right]$$

and hence

$$\pi(\theta | x, y) = Z_0^{-1}(s(x, y) + \mu_0, n + \lambda_0)$$
$$\exp\left[\langle s(x, y) + \mu_0, \psi(\theta)\rangle - (n + \lambda_0) B(\theta)\right]$$

Thus a Rao-Blackwellized estimator of the posterior mean of $\boldsymbol{\theta}$, may be computed as

$$\frac{1}{m} \sum_{j=1}^{m} \mathrm{E}(\boldsymbol{\theta} | \boldsymbol{x}^{(j)}, \boldsymbol{y})$$

But as the normalizing constant Z_0 is known explicitly, it is also possible to integrate out θ to obtain a closed-form expression of $\pi(x|y)$:

$$\pi(x|y) \propto h(x,y) \frac{Z_0(s(x,y)+\mu_0, n+\lambda_0)}{Z_0(\mu_0,\lambda_0)}$$

The resulting marginal is usually complex but amenable to single site Gibbs sampling on the components of $x=(x_1,\ldots,x_n)$, especially when these are discrete variables

This is the preferred sampling method for LDA and related models [Griffiths and Steyvers, 2002; Rigouste et al., 2007]

Mixture of Gaussian Example

Assume that we observe $\boldsymbol{y_1}, \ldots, \boldsymbol{y_n}$ in the Gaussian mixture model $\sum_{i=1}^{r} \alpha_i \mathcal{N}(y|\theta_i, v_i)$. To make derivations simpler, α and v are treated as fixed parameters and we use an improper flat prior on the vector of means θ

$$\pi(x_k, \theta | y_k) \propto \exp\left[\sum_{i=1}^{r}\left(\log \alpha_i - \frac{(y_k - \theta_i)^2}{2v_i}\right)\mathbb{1}\{x_k = i\}\right]$$

and hence

$$\pi(x_{1:n}, \theta | y_{1:n}) \propto \exp\left[\sum_{i=1}^{r} \log \alpha_i n_i - \frac{\theta_i^2 n_i}{2v_i} + \frac{\theta_i s_i}{v_i}\right]$$

where

$$\begin{cases} n_i = \sum_{k=1}^{n} \mathbb{1}\{x_k = i\} \\ s_i = \sum_{k=1}^{n} y_k \mathbb{1}\{x_k = i\} \end{cases}$$

Upon completing the square,

$$\pi(x_{1:n},\theta|y_{1:n}) \propto \exp\left[\sum_{i=1}^{r} \log \alpha_i n_i + \frac{s_i^2}{2n_i v_i}\right] \exp\left[\sum_{i=1}^{r} -\frac{1}{2v_i/n_i}\left(\theta_i - \frac{s_i}{n_i}\right)^2\right]$$

and

$$\pi(x_{1:n}|y_{1:n}) \propto \prod_{i=1}^{r} \alpha_i^{n_i} \sqrt{\frac{v_i}{n_i}} \exp\left[\frac{s_i^2}{2n_i v_i}\right]$$

Finally,

$$\pi(\boldsymbol{x_k} = i|x_{-k}, y_{1:n}) \propto \alpha_i^{n_{i,-k}+1} \sqrt{\frac{v_i}{n_{i,-k}+1}} \exp\left[\frac{(s_{i,-k}+y_k)^2}{2(n_{i,-k}+1)v_i}\right]$$

where

$$\begin{cases} n_{i,-k} = \sum_{j\neq k} \mathbb{1}\{x_j = i\} & = n_i - \mathbb{1}\{x_k = i\} \\ s_{i,-k} = \sum_{j\neq k} y_j \mathbb{1}\{x_j = i\} & = s_i - y_k \mathbb{1}\{x_k = i\} \end{cases}$$

To run the collapsed (or marginalized) single site Gibbs sampler

- Repeatedly simulate from the conditionals $\pi(x_k|x_{-k}, y_{1:n})$
- keeping track of the accumulated component statistics $(n_i, s_i)_{1\leq i\leq r}$

The idea can be extended to mixture models with an unknown number of components [Nobile and Fearnside, 2007][*]

[*] Recall that this would require using a proper prior on θ

State-Space models

In models with continuous state variables,

$$\begin{cases} \boldsymbol{x_i} | \boldsymbol{x_{-i}}, \boldsymbol{y}, \boldsymbol{\theta} & 1 \le i \le n \\ \boldsymbol{\theta} | \boldsymbol{x}, \boldsymbol{y} \end{cases}$$

is often the only option

There exists variants of the collapsed Gibbs sampler for important classes of models, in particular for conditionally Gaussian state-space models [Carter and Kohn, 1996; Doucet and Andrieu, 2001; Cappé et al., 2005]

Part III

Markov Chain Monte Carlo Methods

7 MCMC Basics

- Why Do We Need MCMC?
- (Minimal) Markov Chain Theory
- MCMC Essentials

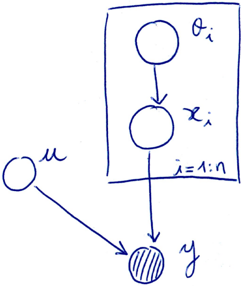
8 The Gibbs Sampler

9 Rao-Blackwellization

10 Auxiliary Targets

Auxiliary Targets

Apart from Rao-Blackwellization, the other common design trick is to use cleverly chosen auxiliary targets



[Doucet et al., 2005]

If $p(x_i|\theta_{1:n}, y)$ is available, scan all components, alternating between

1. $\boldsymbol{x_i} \sim p(x_i|\theta_{1:n}, y)$
2. $\boldsymbol{\theta_i} \sim q(\theta_i|x_i)$

Proving that the previous algorithm is correct can be challenging

⚠#1: The Target pdf

The algorithm is not simulating under $\pi(x_{1:n}, \theta_{1:n}|y)$

Reverse-Engineering Solution If you believe that the algorithm is marginally correct for $\theta_{1:n}$ then the target pdf must be

$$p_{\text{aux}}(\theta_{1:n}, x_{1:n}) = \pi(\theta_{1:n}|y) \prod_{i=1}^{n} p(x_i|\theta_{1:n}, y)$$

⚠ #2: The Updating Scheme

The algorithm is not alternating between the full conditionals

$$\begin{cases} \boldsymbol{x_i} \sim p_{\text{aux}}(x_i|\theta_{1:n}, x_{-i}) \\ \boldsymbol{\theta_i} \sim p_{\text{aux}}(\theta_i|\theta_{-i}, x_{1:n}) \end{cases}$$

The second update is indeed a draw from[*]

$$p_{\text{aux}}(\theta_i, x_{-i}|\theta_{-i}, x_i)$$

$$\begin{aligned} p_{\text{aux}}(\theta_i, x_{-i}|\theta_{-i}, x_i) &= \frac{\pi(\theta_{1:n}|y)\, p(x_i|\theta, y) \prod_{j\neq i} p(x_j|\theta, y)}{\int_\Theta \pi(\theta_{1:n}|y)\, p(x_i|\theta, y)\,\mathrm{d}\theta_i} \\ &= \underbrace{\pi(\theta_i|x_i, y, \theta_{-i})}_{q(\theta_i|x_i)} \prod_{j\neq i} p(x_j|\theta, y) \end{aligned}$$

[*] The x_{-i} part is not required and can be “discarded” (in practice, it is not even simulated)

The particle Gibbs sampler of [Andrieu et al., 2010] is another striking example where one simulates a population of “particles” $\boldsymbol{z}^{(j)} = \boldsymbol{z}^{(j)}_{\boldsymbol{1:d}}$ and an index $\boldsymbol{k}^{(j)}$ such that only $\boldsymbol{z}^{(j)}_{\boldsymbol{k}^{(j)}}$ is converging to $\pi(z)$

Here, the auxiliary target is

$$p_{\text{aux}}(k, z_{1:d}) = \frac{1}{d}\pi(z_k) \prod_{i \neq k} q(z_i)$$

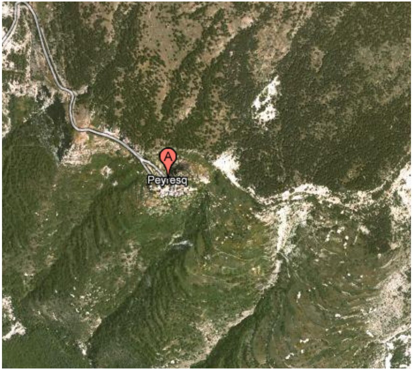
And the update rule

- $\boldsymbol{z}_{-\boldsymbol{k}} | k, z_k \sim \prod_{i \neq k} q(x_i)$
- $\boldsymbol{k} | z_{1:d} \sim p_{\text{aux}}(k | z_{1:d}) = \frac{\pi(z_k)/q(z_k)}{\sum_{i=1}^{d} \pi(z_i)/q(z_i)}$

Important things that have not been discussed here

- Bayesian nonparametric models
- Advanced variational methods [Wainwright and Jordan, 2008]
- Reversible jump MCMC [Green, 1995]
- Sequential Monte Carlo methods [Doucet et al., 2001; Cappé et al., 2005, 2007] and their applications for static inference [Andrieu et al., 2010]
- Techniques specific to the case of state-space models [Cappé et al., 2005] and applications, eg., to changepoint models [Fearnhead, 2006]

Thank you for your attention!



References I

A subjective choice of basic references

- Bayesian statistics [Gelman et al., 1995; Robert, 2001]
- MCMC [Robert and Casella, 2004], Chapters 6 and 13 of [Cappé et al., 2005] (for people specifically interested in state-space models), [Andrieu et al., 2003] for a shorter introduction

C. Andrieu and A. Doucet. Joint bayesian detection and estimation of noisy sinusoids via reversible jump MCMC. *IEEE Trans. Signal Process.*, 47(10):2667–2676, 1999.

C. Andrieu, N. de Freitas, A. Doucet, and M. I. Jordan. An introduction to MCMC for machine learning. *Machine Learning*, 50:5–43, 2003.

C. Andrieu, A. Doucet, and R. Holenstein. Particle markov chain Monte Carlo methods (with discussion). *J. Roy. Statist. Soc. B*, 72(3):269–342, 2010.

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. In *Advances in Neural Information Processing Systems (NIPS)*, volume 14, pages 601–608, 2002.

W. Buntine and A. Jakulin. Discrete component analysis. In *Proceedings of the Subspace, Latent Structure and Feature Selection: Statistical and Optimization Perspectives Workshop (SLSFS)*, pages 1–33, 2006.

O. Cappé, E. Moulines, and T. Rydén. *Inference in Hidden Markov Models*. Springer, 2005.

References II

O. Cappé, S. J. Godsill, and E. Moulines. An overview of existing methods and recent advances in sequential Monte Carlo. *IEEE Proceedings*, 95(5):899–924, 2007. doi: 10.1109/JPROC.2007.893250.

C. K. Carter and R. Kohn. Markov chain Monte Carlo in conditionnaly Gaussian state space models. *Biometrika*, 83(3):589–601, 1996.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. B*, 39(1):1–38 (with discussion), 1977.

A. Doucet and C. Andrieu. Iterative algorithms for state estimation of jump Markov linear systems. *IEEE Trans. Signal Process.*, 49(6):1216–1227, 2001.

A. Doucet, N. De Freitas, and N. Gordon, editors. *Sequential Monte Carlo Methods in Practice*. Springer, New York, 2001.

A. Doucet, S. Senecal, and T. Matsui. Space alternating data augmentation: Application to finite mixture of gaussians and speaker recognition. In *Proc. ICASSP*, pages IV–713–716, 2005.

P. Fearnhead. Exact and efficient bayesian inference for multiple changepoint problems. *Stat. Comput.*, 16:203–213, 2006.

J. A. Fessler and A. O. Hero. Penalized maximum-likelihood image reconstruction using space-alternating generalized em algorithms. *IEEE Trans. Image Process.*, 4 (10):1417–1429, 1995.

References III

A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman, New York, 1995.

S. J. Godsill and P. J. W. Rayner. *Digital Audio Restoration: A Statistical Model-Based Approach*. Springer, 1998.

P. J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82:711–732, 1995.

T. L. Griffiths and M. Steyvers. A probabilistic approach to semantic representation. In *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, 2002.

K. A. Heller, S. Williamson, and Z. Ghahramani. Statistical models for partial membership. In *Proceed. International Conference on Machine Learning (ICML)*, Helsinki, Finland, 2008.

D. R. Hunter and K. Lange. A tutorial on MM algorithms. *Amer. Statist.*, 58(1): 30–37, 2004.

M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37:183–233, 1999.

R. M. Neal and G. E. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In M. I. Jordan, editor, *Learning in graphical models*, pages 355–368. MIT Press, Cambridge, MA, USA, 1999.

A. Nobile and A. T. Fearnside. Bayesian finite mixtures with an unknown number of components: The allocation sampler. *Statistics and Computing*, 17(2):147–162, 2007.

References IV

J. J. K. Ó Ruanaidh and W. J. Fitzgerald. *Numerical Bayesian Methods Applied to Signal Processing*. Springer, New York, 1996.

D. Peel and G. McLachlan. Robust mixture modelling using the t distribution. *Statistics and Computing*, 10:339–348, 2000.

L. Rigouste, O. Cappé, and F. Yvon. Inference and evaluation of the multinomial mixture model for text clustering. *Information Processing & Management*, 43(5): 1260–ŋ1280, 2007. doi: 10.1016/j.ipm.2006.11.001.

C. P. Robert. *The Bayesian Choice*. Springer, New York, 2nd edition, 2001.

C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, New York, 2nd edition, 2004.

M. Tipping and C. Bishop. Probabilistic principal component analysis. *J. Roy. Statist. Soc. B*, 6(3):611–622., 1999.

M. E. Tipping. Sparse Bayesian learning and the relevance vector machine. *J. Machine Learning Research*, 1:211–214, 2001.

M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2008.

C. F. J. Wu. On the convergence properties of the EM algorithm. *Ann. Statist.*, 11: 95–103, 1983.