Bayesian modelling and Markov chain Monte Carlo

First winter school in eScience
Geilo, Friday February 2nd 2007

Pdf file available from
http://www.math.ntnu.no/∼haakont/vinterskole/

Håkon Tjelmeland
Department of Mathematical Sciences
Norwegian University of Science and Technology
Trondheim, Norway

Todays plan

- **Bayesian statistics**
- **Bayesian hierarchical modals**
- **In two examples (one small and one larger)**
  - **demonstrtate Bayesian hierarchical modelling**
  - **demonstrate how MCMC is the natural computational tool for Bayesian hierarchical settings**
- **If time:**
  - **demonstrate the flexibility of the Metropolis–Hastings setup**
  - **perfect simulation**

Bayesian statistics

- **Example (Bayes, 1763):**
  - **A billiard ball is dropped on the interval** $[0, 1]$
    - ∗ **it stops at** p
    - ∗ **assume** p **is uniformly distributed on** $[0, 1]$
  - **Drop the billiard ball** n **new times**
    - ∗ **record** $y_i = 1$ **if ball stops to the left of** p
    - ∗ $y_i = 0$ **otherwise**
    - ∗ **set** $x = \sum_{i=1}^{n} y_i$
    - ∗ **thus** $x|p \sim$ **bin**(n, p)**,**

$$\mathbf{P}(X = x|p) = \begin{pmatrix} n \\ x \end{pmatrix} p^x (1-p)^{n-x} \text{ , } x = 0, 1, \ldots, n$$

  - **want to estimate** p **from observed** x
  - **standard estimator for** p **in binomial distr.:**

$$\widehat{p} = \frac{X}{n}$$

  - **but we know** $p \sim$ **Uniform**$[0, 1]$**,**

$$\pi(p) = \begin{cases} 1 & \textbf{for } x \in [0, 1], \\ 0 & \textbf{otherwise} \end{cases}$$

Bayesian statistics (cont.)

- **Recall**

$$\pi(p) = \begin{cases} 1 & \textbf{for } x \in [0,1], \\ 0 & \textbf{otherwise} \end{cases}$$

$$\mathbf{P}(X = x|p) = \begin{pmatrix} n \\ x \end{pmatrix} p^x (1-p)^{n-x} \textbf{ , } x = 0, 1, \ldots, n$$

- **Thus**

$$\begin{aligned} \pi(p|x) &= \frac{\pi(p,x)}{\mathbf{P}(X=x)} = \frac{\pi(p)\mathbf{P}(X=x|p)}{\int_0^1 \mathbf{P}(X=x|\tilde{p})\pi(\tilde{p})\mathbf{d}\tilde{p}} \\ &= \frac{p^x(1-p)^{n-x}}{\int_0^1 \tilde{p}^x(1-\tilde{p})^{n-x}\mathbf{d}\tilde{p}} = \frac{p^x(1-p)^{n-x}}{B(x+1, n-x+1)} \end{aligned}$$

- **This is a beta-distribution, $\mathcal{B}(x+1, n-x+1)$, with**

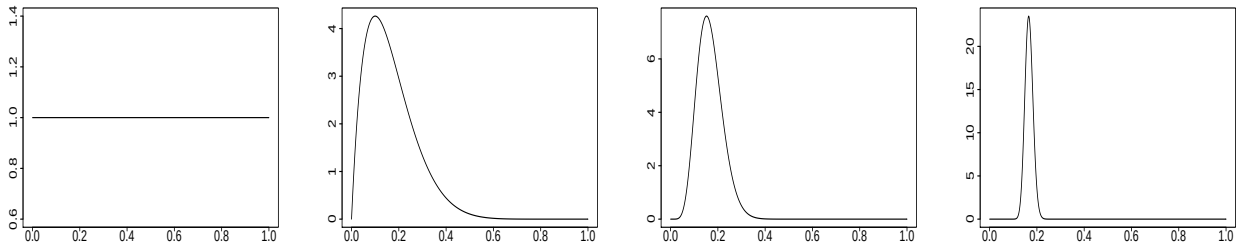
$$\mathbf{E}[p|x] = \frac{x+1}{n+2}$$

- **Natural estimator for p**

$$\widehat{p} = \frac{X+1}{n+2}$$

Bayesian statistics (cont.)

- **In example: p is a stochastic variable because it is the result of a stochastic experiment**
- **Bayesian modelling: consider parameters as stochastic variables also when their value is not the result of a stochastic experiment**
- **Another (toy) example:**
  - **I have a dice, let p: probability of getting a six**
  - **Consider p as a stochastic variable, you don't know it is a proper dice**
  - **what distribution would you assign to p?**



- **we roll the dice n times, let x: number of sixes**

$$\mathbf{P}(X = x|p) = \binom{n}{x} p^x (1-p)^{n-x} \text{ , for } x = 0, 1, \ldots, n$$

Bayesian statistics (cont.)

- **Recall**

$$\mathbf{P}(X = x|p) = \left(\begin{array}{c} n \\ x \end{array} \right) p^x (1-p)^{n-x} \text{ , for } x = 0, 1, \ldots, n$$

- **Assume** $p \sim \mathcal{B}(\alpha, \beta)$**,**

$$\pi(p) = \frac{1}{\mathbf{B}(\alpha, \beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

- **This gives**

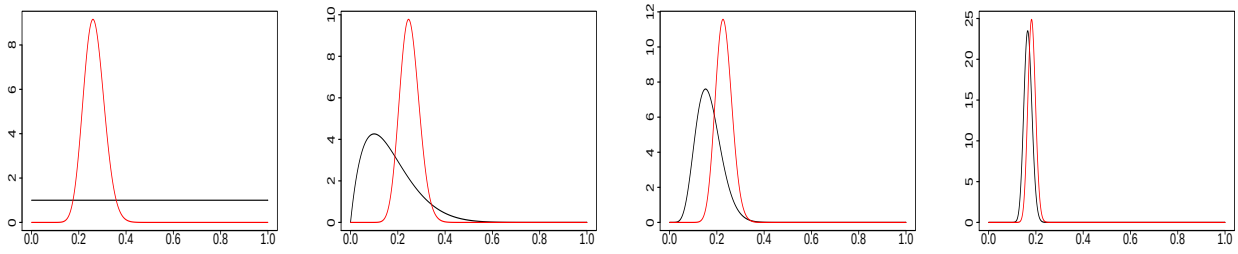
$$\begin{aligned} \pi(p|x) &= \frac{\pi(p, x)}{\mathbf{P}(X = x)} \propto \pi(p)\mathbf{P}(X = x|p) \\ &\propto p^{\alpha-1}(1-p)^{\beta-1} p^x (1-p)^{n-x} \\ &= p^{\alpha+x-1}(1-p)^{\beta+n-x-1} \end{aligned}$$

- **Thus** $p|x \sim \mathcal{B}(\alpha + x, \beta + n - x)$ **and**

$$\mathbf{E}[p|x] = \frac{\alpha + x}{\alpha + \beta + n}$$

- **Observed** $n = 100$**,** $x = 26$**:**

$\mathbf{E}[p|x] = 0.265$ $\mathbf{E}[p|x] = 0.255$ $\mathbf{E}[p|x] = 0.230$ $\mathbf{E}[p|x] = 0.183$



Interpretation of probability

- **Frequentist: Probability of event A is**

$$\mathbf{P}(A) = \lim_{n\to\infty} \frac{m}{n}$$

  where m is # times A has occurred in n identical and independent trials

- **Bayesian (subjective): Probability of event A, $\mathbf{P}(A)$, is a measure of someone's degree of belief in the occurrence of A.**
  - **different persons may have different $\mathbf{P}(A)$**

Prior and posterior distribution

- **Prior distribution:** $\pi(\theta)$
  - **a measure of our belief about the value of** θ **before we have observed the data, based on prior information/experience**

- **Observation and Likelihood:** $f(x|\theta)$
  - **observed value** x**, and its probability distribution given** θ

- **Posterior distribution:** $\pi(\theta|x)$
  - **a measure of our belief about the of value of** θ **after we have observed the data** x**, based on prior information/experience *and* the observed data** x
  - **Bayes theorem**

$$\pi(\theta|x) = \frac{\pi(\theta, x)}{\pi(x)} \propto \pi(\theta, x) = \pi(\theta) f(x|\theta)$$

Conjugate priors

- **In examples: posteriors available on closed form**
  - **this is because we have used a *conjugate* prior**
- **binomial conjugate prior**
  - $x|p \sim$ **binomial**(n, p)
  - $p \sim$ **beta**(α, β)
  - $p|x \sim$ **beta**$(\cdot, \cdot)$
- **normal (mean) conjugate prior**
  - $x_1, \ldots, x_n|\mu \sim \mathbf{N}(\mu, \sigma_0^2)$
  - $\mu \sim \mathbf{N}(\mu_0, \tau^2)$
  - $\mu|x_1, \ldots, x_n \sim \mathbf{N}(\cdot, \cdot)$
- **normal (variance) conjugate prior**
  - $x_1, \ldots, x_n|\sigma^2 \sim \mathbf{N}(\mu_0, \sigma^2)$
  - $\sigma^2 \sim (IG)(\alpha, \beta)$
  - $\sigma^2|x_1, \ldots, x_n \sim \mathbf{IG}(\cdot, \cdot)$
- **and many more**
- **Conjugate priors often used also in hierarchical Bayesian models — enable Gibbs updates**

Hierarchical Bayesian models

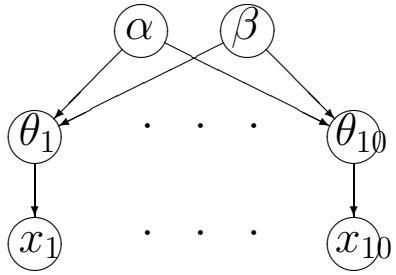
- **A simple example (from George et al., 1993)**
  - **Analysis of** 10 **power plant pumps**
  - x_i**,** t_i**: number of failures for pump** i **and length of operation time on that pump (in** 1000 **hours)**
  - **Modelling:**
    - $x_i|\theta_i \sim$ **Poisson**$(\theta_i t_i)$
    - **conjugate prior for** θ_i

$$\theta_i|\alpha,\beta \sim \textbf{Gamma}(\alpha,\beta)$$

    - **hyper-prior distribution on** α **and** β

$$\alpha \sim \textbf{Exp}(1.0)\ ,\ \beta \sim \textbf{Gamma}(0.1, 1.0)$$

  - **graphical model:**

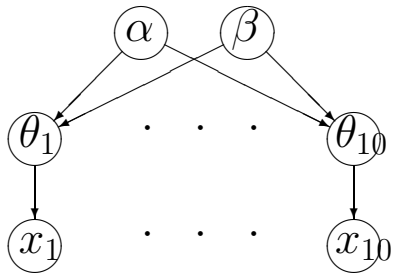


  - **observed:** $x_1, \ldots, x_n$
  - **posterior distribution of interest:**

$$\pi(\alpha, \beta, \theta_1, \ldots, \theta_{10}|x_1, \ldots, x_{10})$$

A simple example (cont.)

- **Graphical model**



- **Posterior distribution**

$$\pi(\alpha, \beta, \theta_1, \ldots, \theta_{10} | x_1, \ldots, x_{10}) \propto \pi(\alpha, \beta, \theta_1, \ldots, \theta_{10}, x_1, \ldots, x_{10})$$

$$= \pi(\alpha)\pi(\beta)\pi(\theta_1|\alpha, \beta) \cdot \ldots \cdot \pi(\theta_{10}|\alpha, \beta)\pi(x_1|\theta_1) \cdot \ldots \cdot \pi(x_{10}|\theta_{10})$$

- **Single-site Metropolis–Hastings algorithm:**
  - **for $i = 1, \ldots, 10$ update θ_i with Gibbs**

$$\pi(\theta_i | \alpha, \beta, \theta_{-i}, x_1, \ldots, x_{10}) \propto \pi(\alpha, \beta, \theta_1, \ldots, \theta_{10}, x_1, \ldots, x_{10})$$

$$\propto \pi(\theta_i|\alpha, \beta)\pi(x_i|\theta_i)$$

    - **this is a gamma distribution**
  - **update β with Gibbs**

$$\pi(\beta | \alpha, \theta_1, \ldots, \theta_{10}, x_1, \ldots, x_{10}) \propto \pi(\alpha, \beta, \theta_1, \ldots, \theta_{10}, x_1, \ldots, x_{10})$$

$$\propto \pi(\beta)\pi(\theta_1|\alpha, \beta) \cdot \ldots \cdot \pi(\theta_{10}|\alpha, \beta)$$

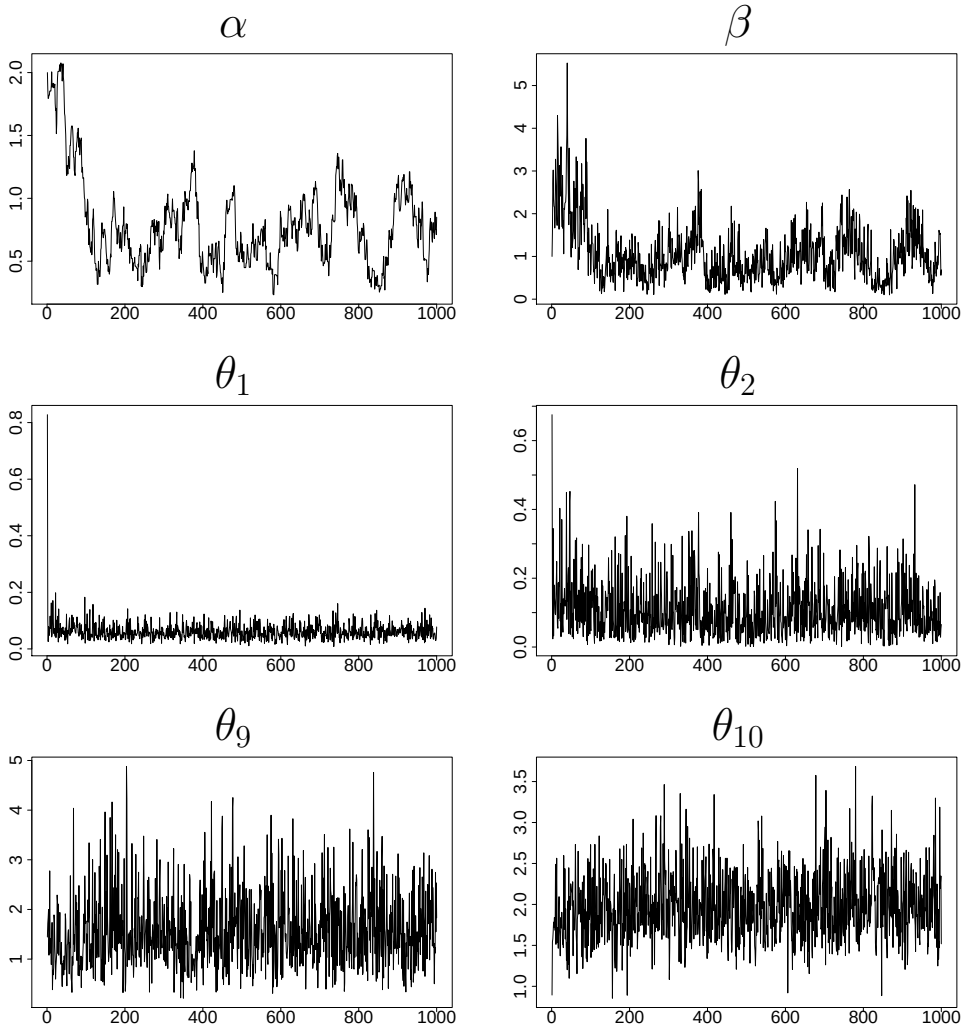
    - **this is a gamma distribution**
  - **update α with a Metropolis–Hastings proposal**
    - **for example: random walk proposal**

A simple example (cont.)

- **Data:**

Pump	1	2	3	4	5	6	7	8	9	10
t_i	94.3	15.7	62.9	126	5.24	31.4	1.05	1.05	2.1	10.5
x_i	5	1	5	14	3	19	1	1	4	22

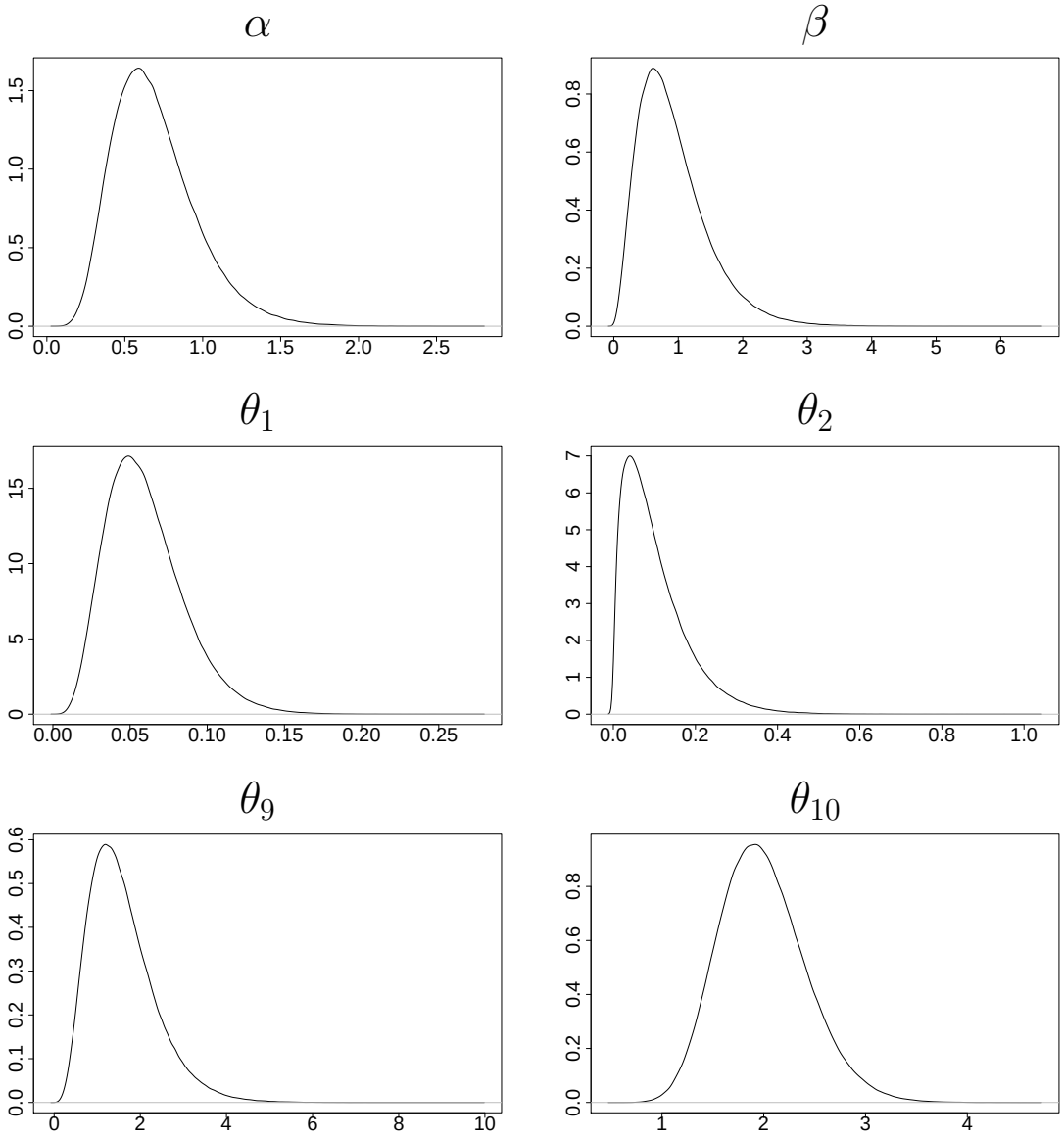
- **Trace plots**



- **Convergence in less than 500 iterations**

A simple example (cont.)

- **Posterior density plots**



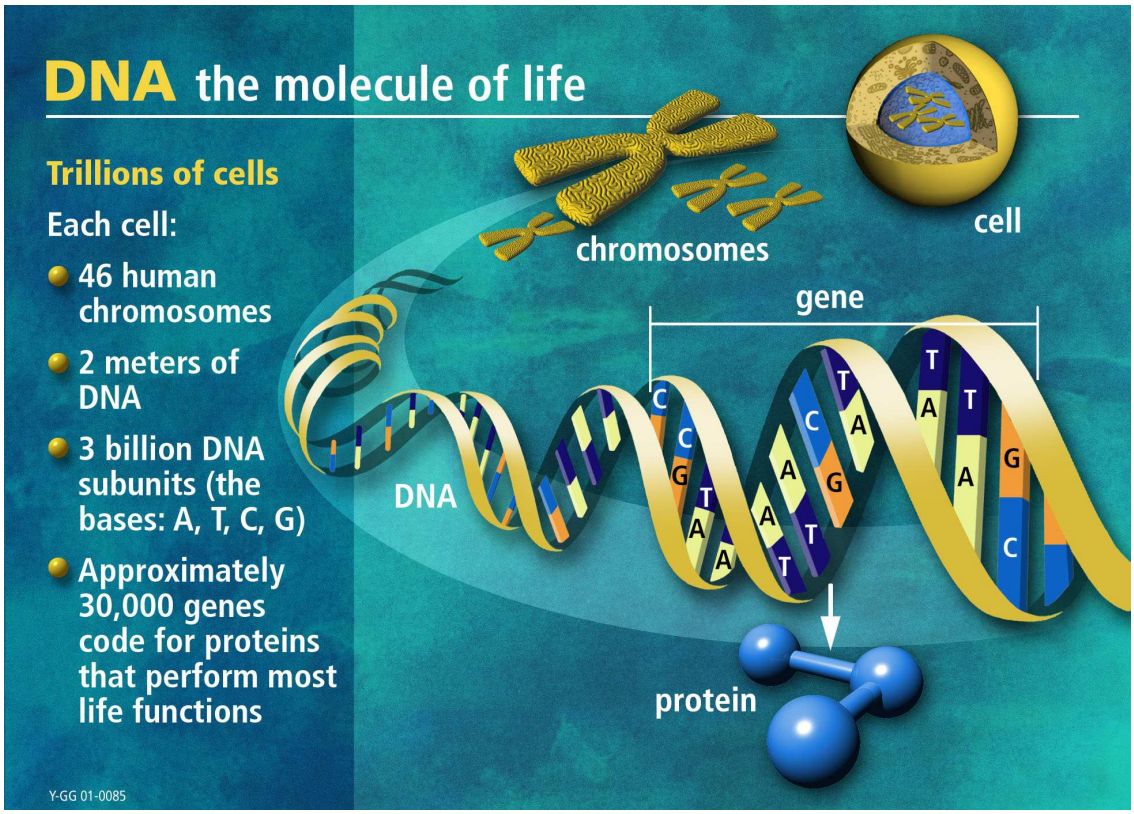
A simple example (cont.)

- **Posterior mean for θ_i compared to x_i/t_i**

parameter	**posterior mean**	x_i/t_i
θ_1	0.0598	0.0530
θ_2	0.1017	0.0636
θ_3	0.0892	0.0795
θ_4	0.1157	0.1111
θ_5	0.6011	0.5725
θ_6	0.6095	0.6051
θ_7	0.8910	0.9524
θ_8	0.8928	0.9524
θ_9	1.5867	1.9047
θ_{10}	1.9901	2.0952

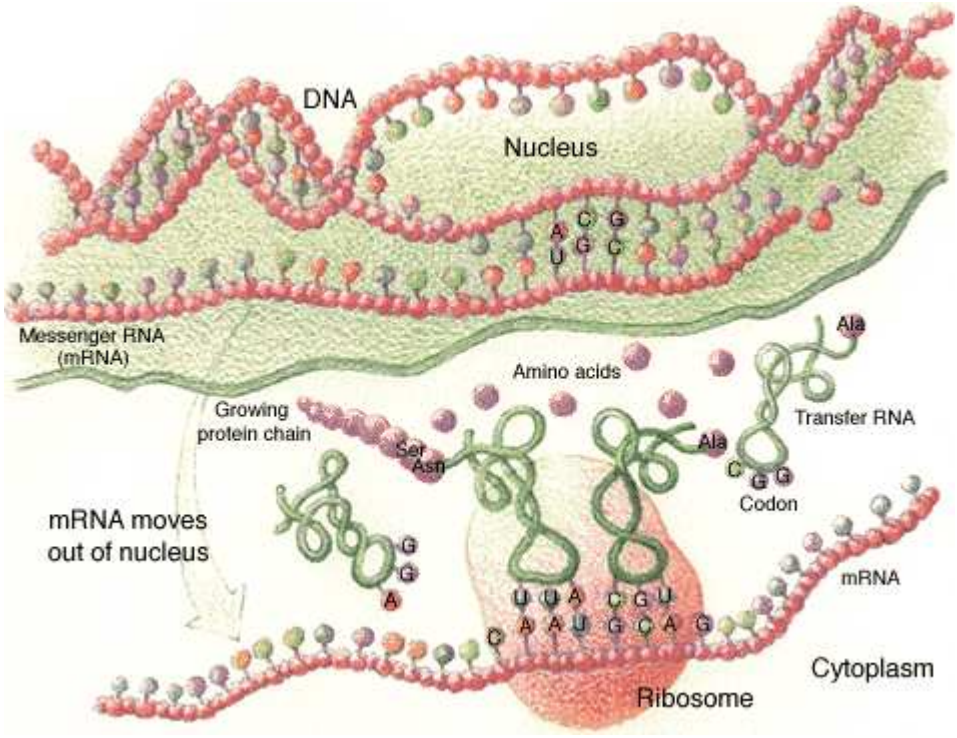
Microarray data example

- Joint work with Rob Scharpf, Giovanni Parmigiani and Andrew Nobel
- Example include
  - problem description
  - Bayesian model formulation
  - Metropolis–Hastings algorithm
  - convergence analysis
  - presentation of results
- DNA contains genes (about $30,000$ in humans)



Microarray data example (cont.)

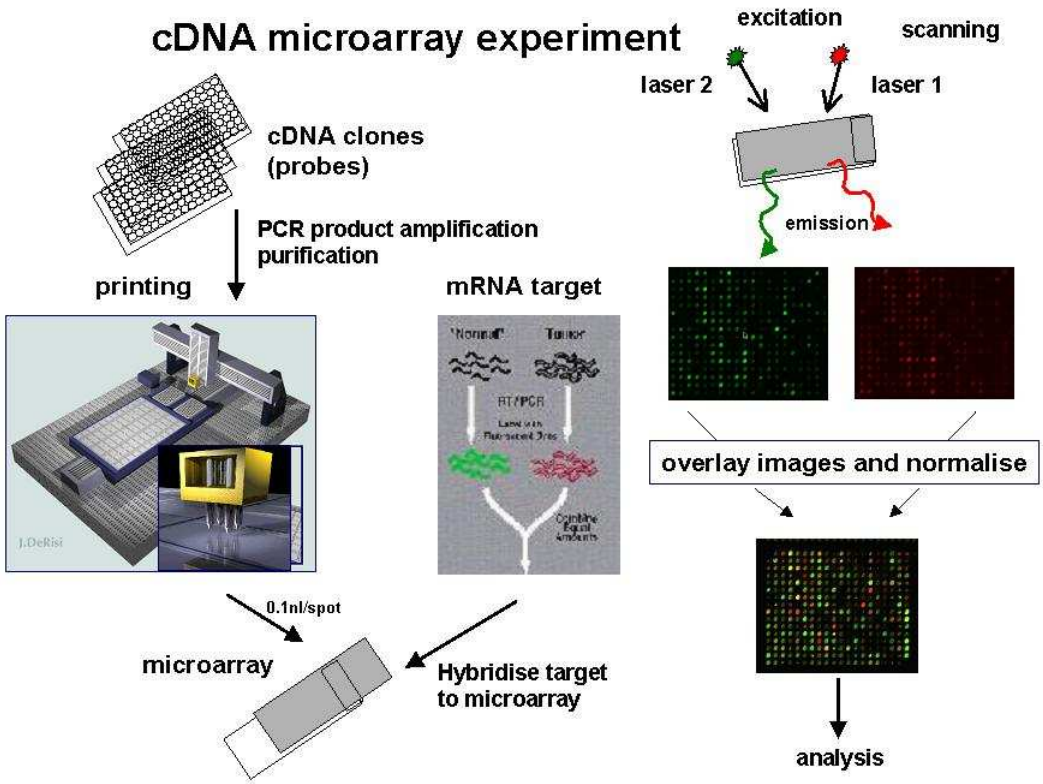
- **A gene can be more or less turned on, *expressed*, in a cell**



- **Gene expression: the process by which a gene's coded information is converted into the structures present and operating in the cell**
- **Can measure the amount of mRNA**
- **One goal: Find genes that are differentially expressed in (for example) breast cancer cells and healthy breast cells, or in breast cancer cells of two different (sub)types of cancer**

Microarray data example (cont.)

- **DNA microarrays: A high throughput technology for measuring the gene expression of thousands of genes for tissue samples**

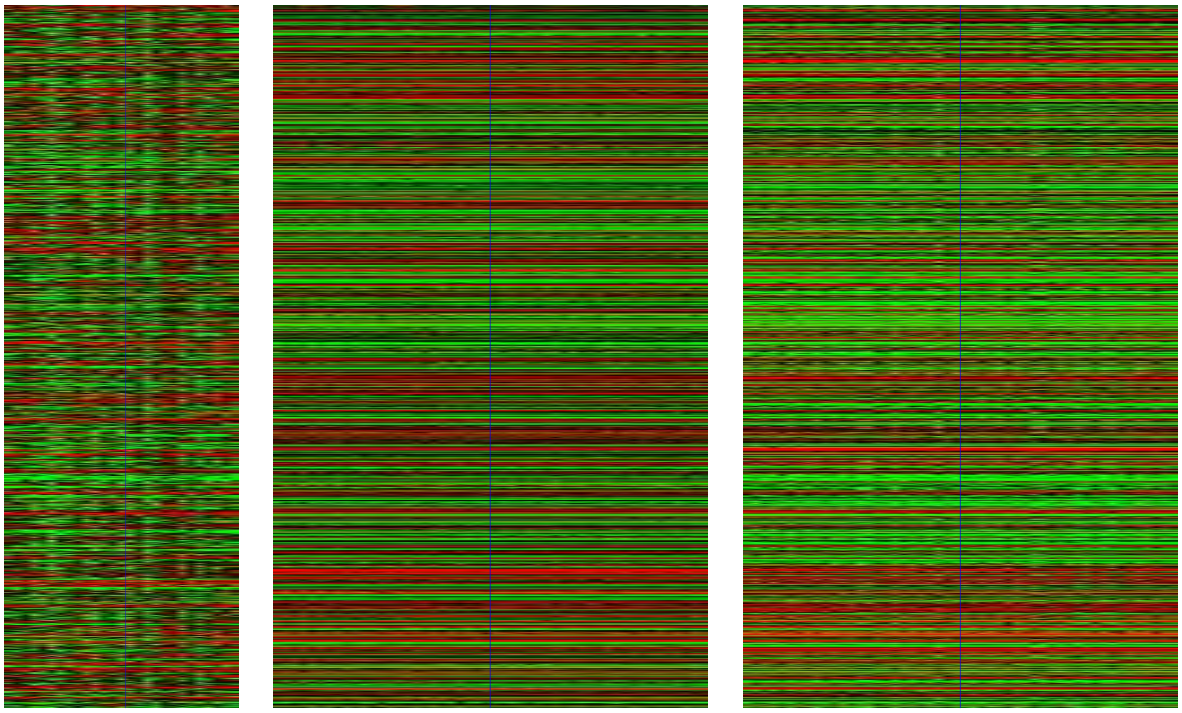


Copied from talk by Terry Speed at http://www.ipam.ucla.edu/programs/fg2000/fgt_tspeed7.ppt

- **Different technologies exist — even with the same technology measurements from different labs are not comparable**
  - **measure difference to a reference tissue**

Microarray data example (cont.)

- **Focus here: Use microarray data from several studies to find genes that are differentially expressed**



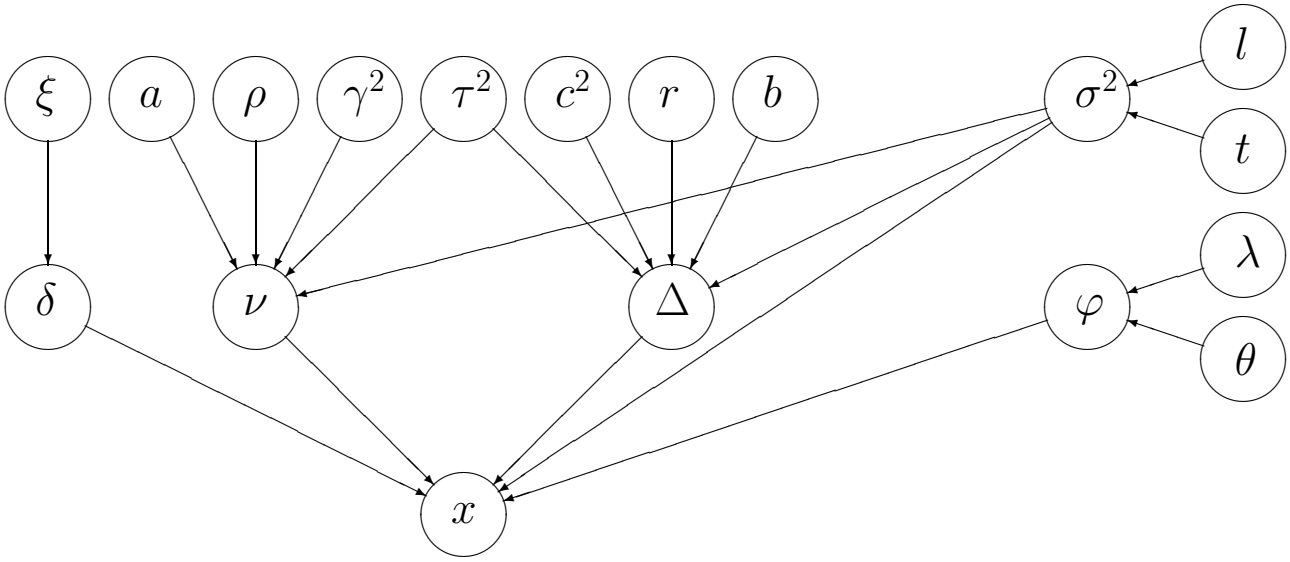
- **Sources of variation in data**
  - **biology**
  - **technology**
  - **observation noise**
- **Note:**
  - **# genes is large, thousands**
  - **# samples is small,** $10 - 200$

Microarray data example (cont.)

- **Different approaches to merge information from several studies**
  - **normalise and combine studies**
    - **then analyse as one data set**
  - **meta-analysis**
    - **combine information from primary statistics**
    - **for example t-statistics**
  - **joint model for data from all studies**
    - **model variation caused from both biology and technology**
- **Notation:**
  - $p = 1, \ldots, P$**: study (or platform)**
  - $g = 1, \ldots, G$**: gene**
  - $s = 1, \ldots, S_p$**: sample**
  - x_{gsp}**: expression value**
  - $\psi_{sp} \in \{0, 1\}$**: two possible conditions**

Microarray data example (cont.)

- **Graphical model**



- $\delta_g \in \{0, 1\}$**: indicator for differential expression.**
- **Likelihood:**
  - **if** $\delta_g = 0$

$$x_{gsp} = \nu_{gp} + \varepsilon_{gsp}$$

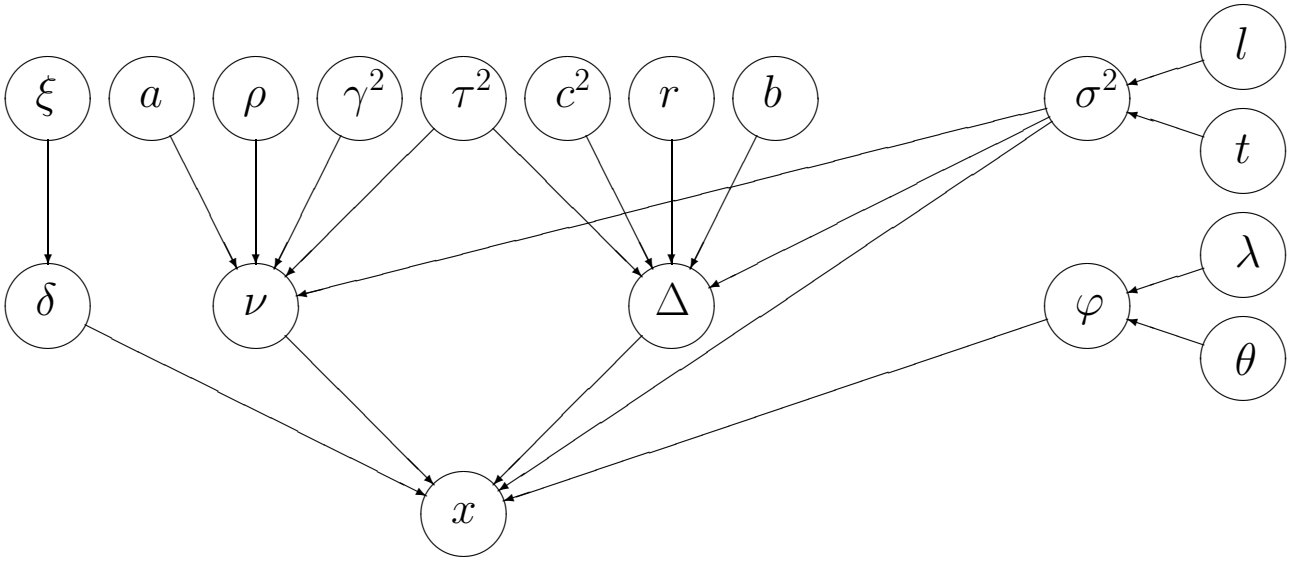
  - **if** $\delta_g = 1$

$$x_{gsp} = \begin{cases} \nu_{gp} - \Delta_{gp} + \varepsilon_{gsp} & \textbf{if } \psi_{sp} = 0, \\ \nu_{gp} + \Delta_{gp} + \varepsilon_{gsp} & \textbf{if } \psi_{sp} = 1. \end{cases}$$

  - **different variance for** $\psi_{sp} = 0$ **and** $\psi_{sp} = 1$

$$\mathbf{Var}[\varepsilon_{gsp}] = \begin{cases} \sigma^2_{gp} \cdot \varphi_{gp} & \textbf{if } \psi_{sp} = 0, \\ \frac{\sigma^2_{gp}}{\varphi_{gp}} & \textbf{if } \psi_{sp} = 1. \end{cases}$$

Microarray data example (cont.)

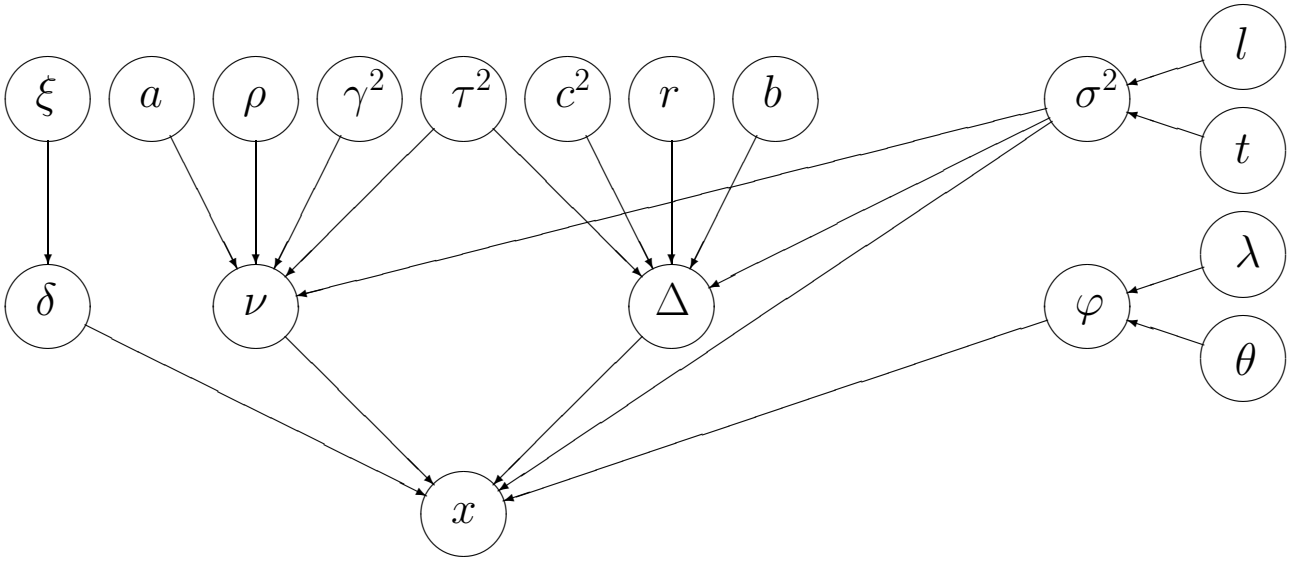


- **Prior for δ_g**
  - **assume $\delta_1, \ldots, \delta_G$ apriori independent given ξ.**

$$\mathbf{P}(\delta_g = 1|\xi) = \xi.$$

  - **apriori $\xi \sim$ Uniform$[0, 1]$.**

Microarray data example (cont.)



- **Priors for $\boldsymbol{\nu}_g = (\nu_{g1}, \ldots, \nu_{gP})^T$ and $\boldsymbol{\Delta}_g = (\Delta_{g1,\ldots,gP})^T$**

$$\boldsymbol{\nu}_g | \text{ hyper-parameters } \sim \mathbf{N}(0, \Sigma_g)$$

$$\boldsymbol{\Delta}_g | \text{ hyper-parameters } \sim \mathbf{N}(0, R_g)$$

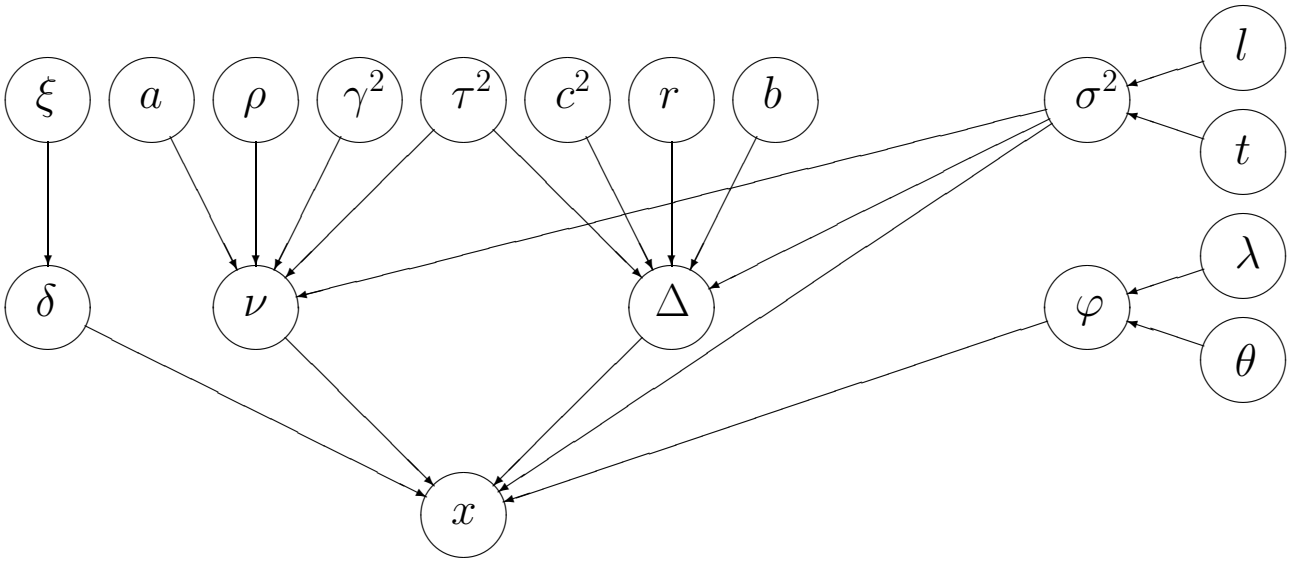
  – **model variances and correlations separately**

$$(\Sigma_g)_{pp} = \gamma^2 \tau_p^2 (\sigma_{gp}^2)^{a_p} \quad \textbf{and} \quad (R_g)_{pp} = c^2 \tau_p^2 (\sigma_{gp}^2)^{b_p}$$

$$\frac{(\Sigma_g)_{pq}}{\sqrt{(\Sigma_g)_{pp}(\Sigma_g)_{qq}}} = \rho_{pq} \quad \textbf{and} \quad \frac{(R_g)_{pq}}{\sqrt{(R_g)_{pp}(R_g)_{qq}}} = r_{pq}$$

  * τ_p^2**: relative scale for study** p**;** $\tau_1^2 \cdot \ldots \cdot \tau_P^2 = 1$
  * $a_p, b_p \in [0, 1]$
  * **hyper-priors on** a**,** ρ**,** γ^2**,** τ^2**,** c^2**,** r**,** b**,** σ^2 **and** φ

Microarray data example (cont.)

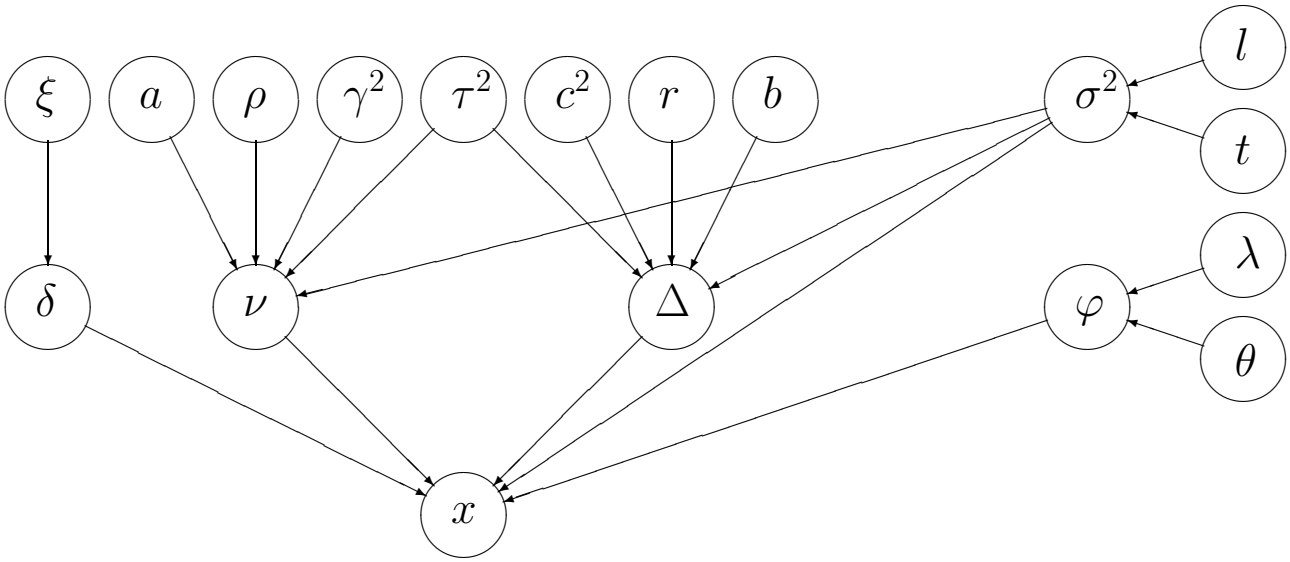


- **Gibbs updates possible for many parameters**
  - ξ, δ_g, γ^2, c^2, ν_{gp}, Δ_{gp}
- **First try: update each parameter separately**
  - **gives very slow convergence/mixing**
  - **strong dependence between some parameters**
- **Next try: introduce block updates**
  - **correlation matrix** $[\rho_{pq}]$ **and** γ^2
  - **propose new** ρ_{pq} **by**

$$\tilde{\rho}_{pq} = (1 - \varepsilon)\rho_{pq} + \varepsilon T_{pq}.$$

  - **propose** γ^2 **from full conditional.**
  - **accept/reject** $[\tilde{\rho}_{pq}]$ **and** $\tilde{\gamma}^2$ **jointly**
- **Similar block update for** $[r_{pq}]$ **and** c^2

Microarray data example (cont.)



- **More block updates**
  - δ_g **and** $\mathbf{\Delta}_g$
  - **propose to change value for** δ_g

$$\tilde{\delta}_g = 1 - \delta_g.$$

  - **propose** $\mathbf{\Delta}_g$ **from full conditional**
  - **accept/reject** $\tilde{\delta}_g$ **and** $\tilde{\mathbf{\Delta}}_g$ **jointly**
- **Last block update**
  - c^2 **and** $\mathbf{\Delta}_g$ **for genes with** $\delta_g = 0$
  - **block Gibbs update for these parameters**
- **Resulting algorithm seems to have good convergence/mixing properties**
- **Algorithm contains several tuning parameters, performance not very sensitive to the values of these**

Microarray data example (cont.)

- **Alternative methods**
  - **(estimated) posterior probability for differential expression**

$$S_g = \frac{1}{n}\sum_{i=1}^{n} \delta_g^{(i)}$$

  - **t-score: combine t-statistics**
  - **SAM-score: combine SAM-statistics**
  - **Choi**

- **Test for differential expression: For statistic S and a threshold $t > 0$, use $|S_g| > t$ as a test for $\theta_g = 1$.**

- **Summing over $g = 1, \ldots, G$ gives a 2×2 table**

	$\theta = 0$	$\theta = 1$
$\lvert S\rvert \geq t$	$\mathbf{FP}(t)$	$\mathbf{TP}(t)$
$\lvert S\rvert < t$	$\mathbf{TN}(t)$	$\mathbf{FN}(t)$

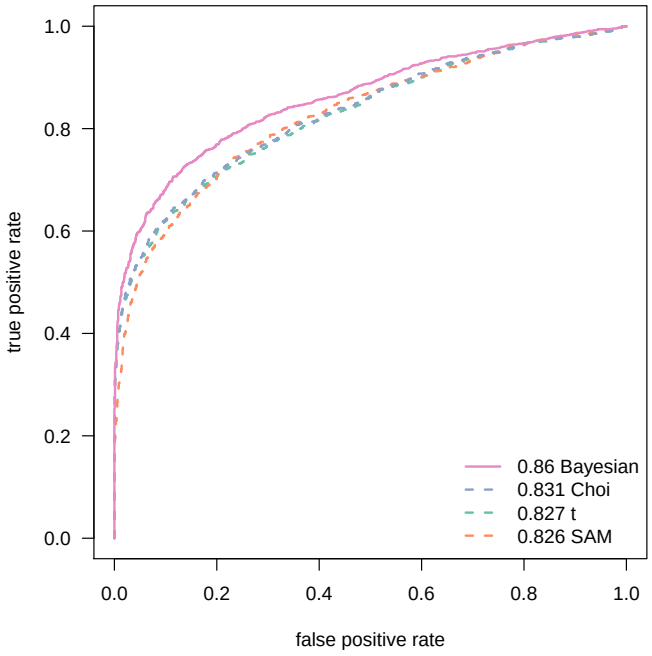
- **ROC curve:**

$$\mathbf{FPR}(t) = \frac{\mathbf{FP}(t)}{\mathbf{TN}(t) + \mathbf{FP}(t)} \quad \textbf{vs.} \quad \mathbf{TPR}(t) = \frac{\mathbf{TP}(t)}{\mathbf{FN}(t) + \mathbf{TP}(t)}$$

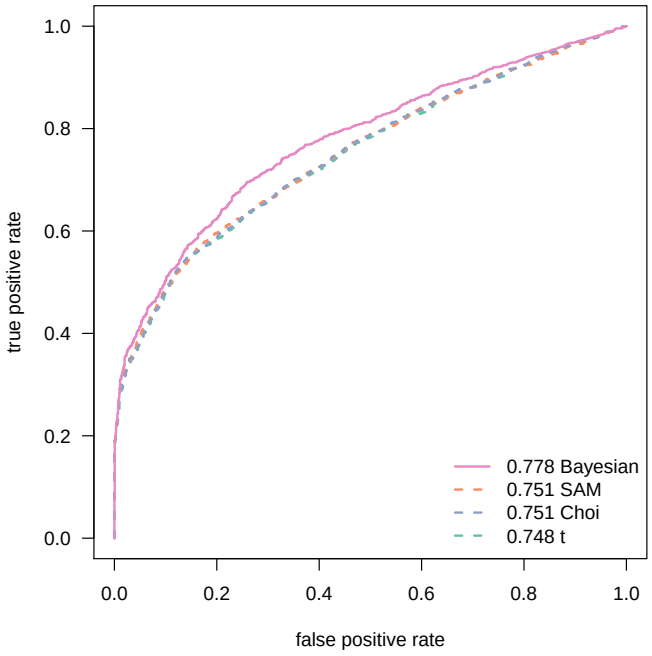
Microarray data example (cont.)

- **Simulation study: simulate data from model**

50 samples

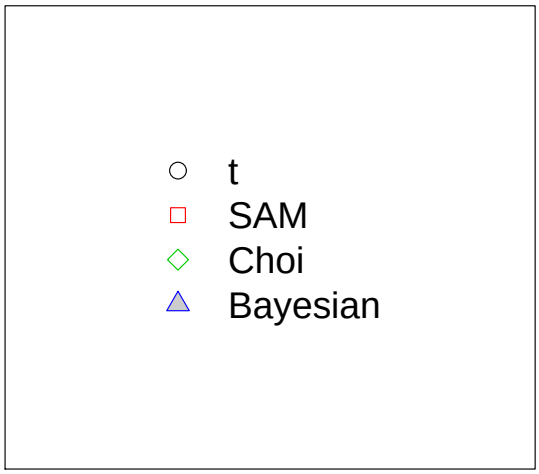


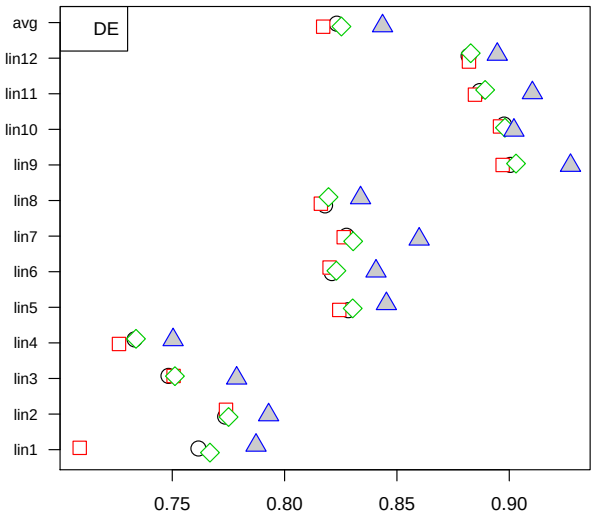
25 samples



Microarray data example (cont.)

- **Area under curve (AUC)**
  - **for different number of samples and parameter values**

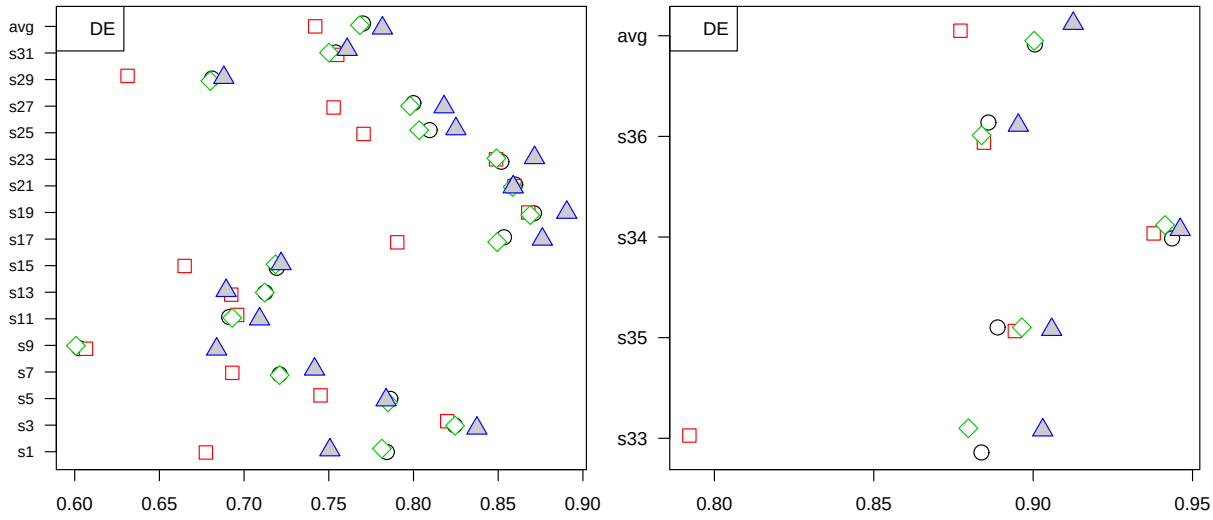




Microarray data example (cont.)

- **Real data from three studies**
  - $3,171$ **common genes.**
  - **Use adenocarcinoma samples.**
  - **For each study: split samples (or subset) in two at random.**
  - **Simulate** $\delta_g, g = 1, \ldots, G$**.**
  - **Simulate offsets** $\Delta_{g1}, \ldots, \Delta_{gP}$
  - **Make simulated data set by adding/subtracting the** Δ**'s from the observed values.**

26 and 50 samples **8 samples**



Microarray example — closing remarks

- **Algorithm specification an iterative process**
  - **tuning parameters**
  - **update types**
- **Model specification may be iterative process**
- **Model dependencies via the hierarchical model**
- **Bayesian hierarchical models and MCMC are modular — ideal for object oriented programming**
  - **one object for each node in graphical model**
  - **one object for each update type**
- **Note:**
  - **the probabilities/densities in the acceptance probability may be very small/large**
  - **all probability calculations should be done on a log scale to avoid numerical problems**
  - $U(x) = -\ln(\pi(x))$**: potential, energy**