南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Machine Learning and NeuroEngineering

机器学习与神经工程

Lecture 11 – Gibbs Sampling & JAGS

Quanying Liu （刘泉影）
SUSTech, BME department

Email: liuqy@sustech.edu.cn

BMEB212, Spring term, 2021

Lecture 10 – Recap

- What is MCMC?
  - Motivations: posterior distributions may not be computed
  - The Metropolis-Hastings Algorithm for MCMC
  - Estimating *single* parameter (examples from the intelligence tests)
  - Estimating *multiple* parameters (an example from the visual working memory)
- Two major problems for MCMC sampling
  - Convergence of MCMC Chains → multiple chains
  - Autocorrelation in MCMC Chains → thinning
- Approximate Bayesian Computation (ABC): a likelihood-free method
  - Likelihoods that cannot be computed
  - From simulations to estimates of the posterior
  - An Example: applying ABC in a signal-detection model

Signal Detection Theory

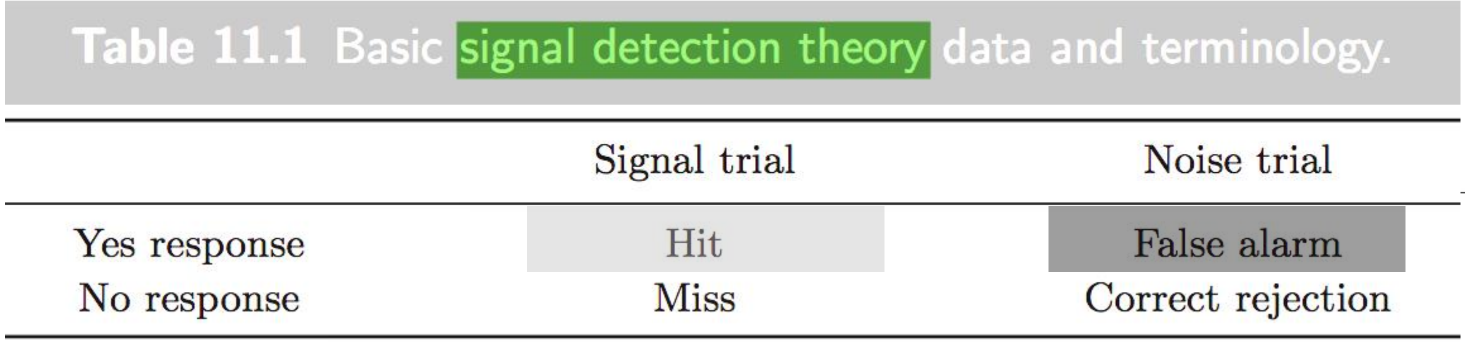


Table 11.1 Basic signal detection theory data and terminology.

	Signal trial	Noise trial
Yes response	Hit	False alarm
No response	Miss	Correct rejection

Two assumptions for signal-detection theory:

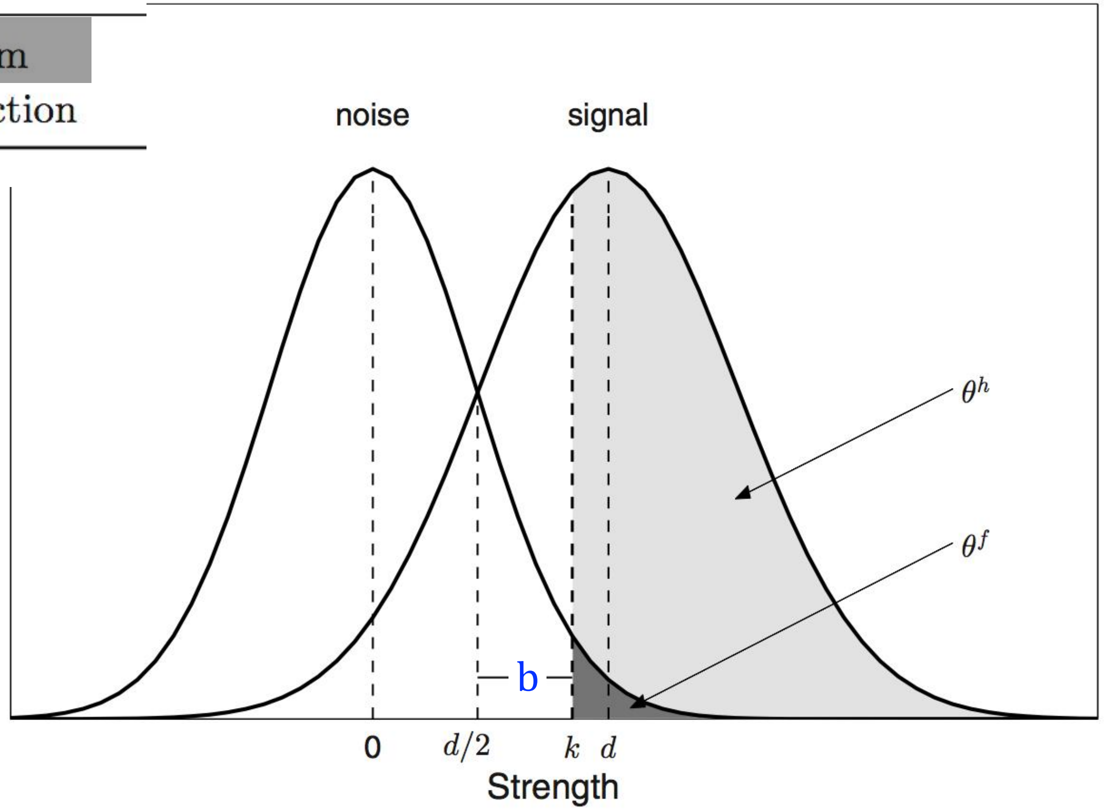
1) Signal and noise trials can be represented as values along a uni-dimensional “strength” construct.

2) Both types of trials produce strengths that vary according to a Gaussian distribution along this dimension.

d: discriminability of the signal trials from the noise trials

k: criterion for response

b: bias (b = k-d/2)



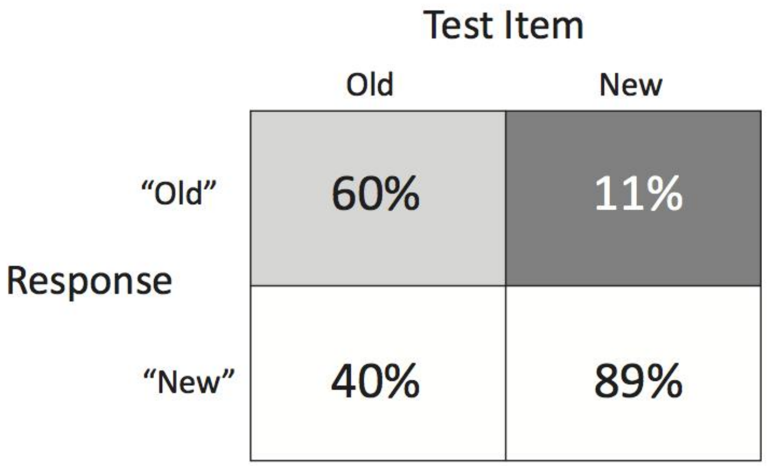
Equal-variance Gaussian signal detection theory framework.

d/2: the unbiased criterion

Applying ABC in a single-detection model

A recognition memory task

- Study session: Participants first study a list of words and are then presented with a long sequence of test items.
- Test session: Each test item is *either* a word from the study list (old) *or* an item that has not been seen before (new). Old and new items usually appear with equal probability in the test sequence. The participant *responds* to each test item by indicating “old” or “new”.



		Test Item	
		Old	New
Response	“Old”	60%	11%
	“New”	40%	89%

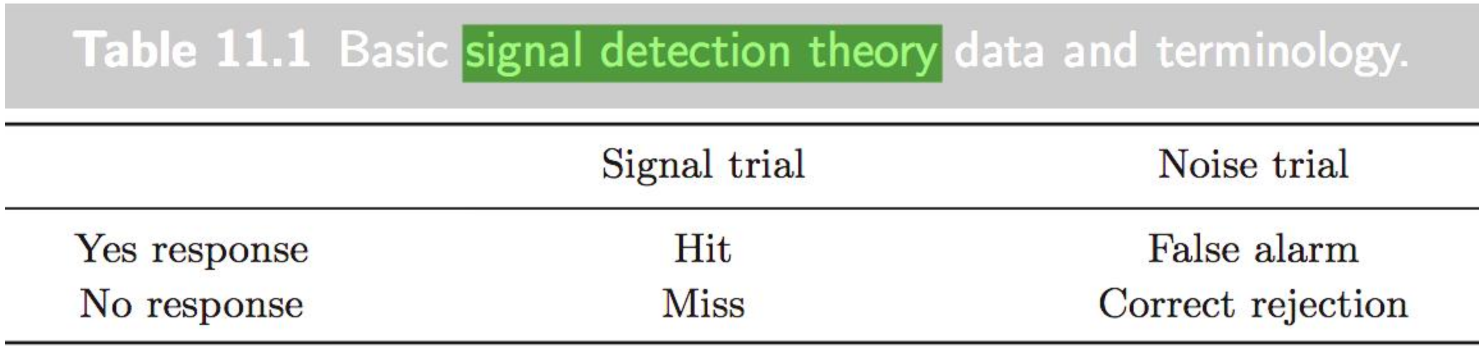
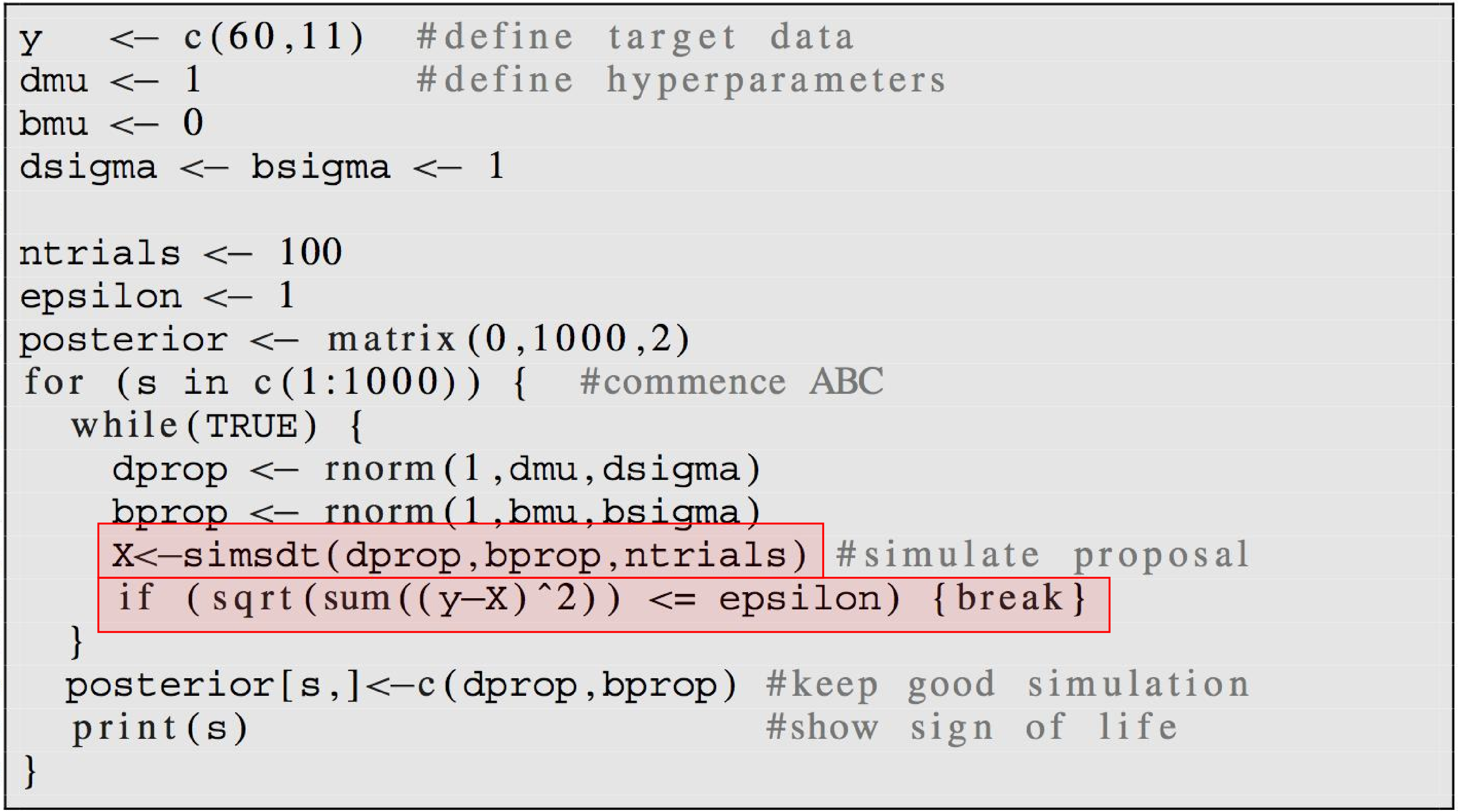


Table 11.1 Basic signal detection theory data and terminology.

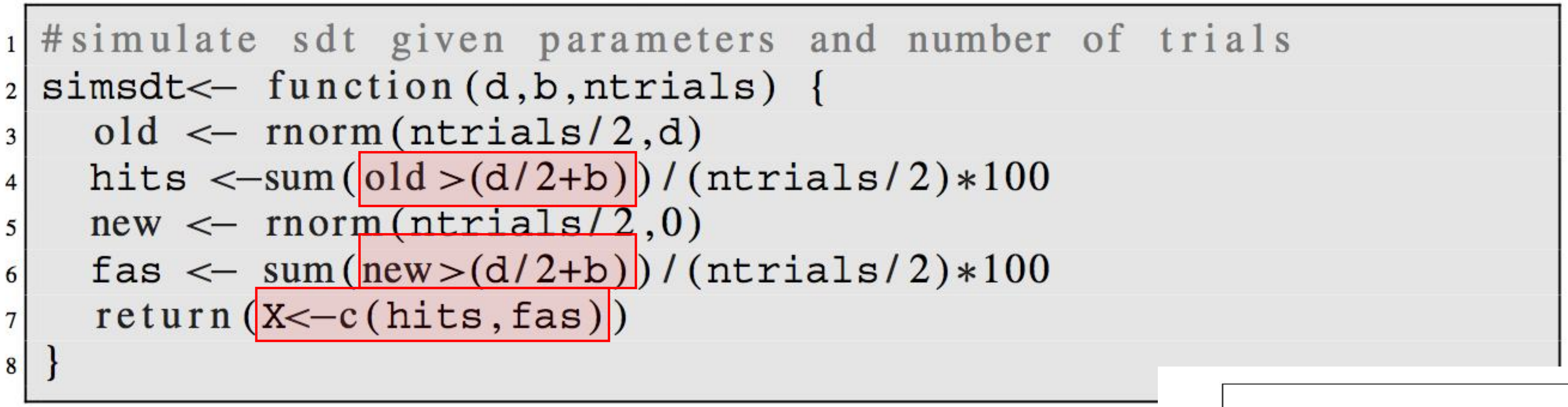
	Signal trial	Noise trial
Yes response	Hit	False alarm
No response	Miss	Correct rejection

SEE CODE: Lecture9_5_abcsdt.R



```r
y   <- c(60,11)  #define target data
dmu <- 1         #define hyperparameters
bmu <- 0
dsigma <- bsigma <- 1

ntrials <- 100
epsilon <- 1
posterior <- matrix(0,1000,2)
for (s in c(1:1000)) {  #commence ABC
  while(TRUE) {
    dprop <- rnorm(1,dmu,dsigma)
    bprop <- rnorm(1,bmu,bsigma)
    X<-simsdt(dprop,bprop,ntrials) #simulate proposal
    if (sqrt(sum((y-X)^2)) <= epsilon) {break}
  }
  posterior[s,]<-c(dprop,bprop) #keep good simulation
  print(s)                      #show sign of life
}
```

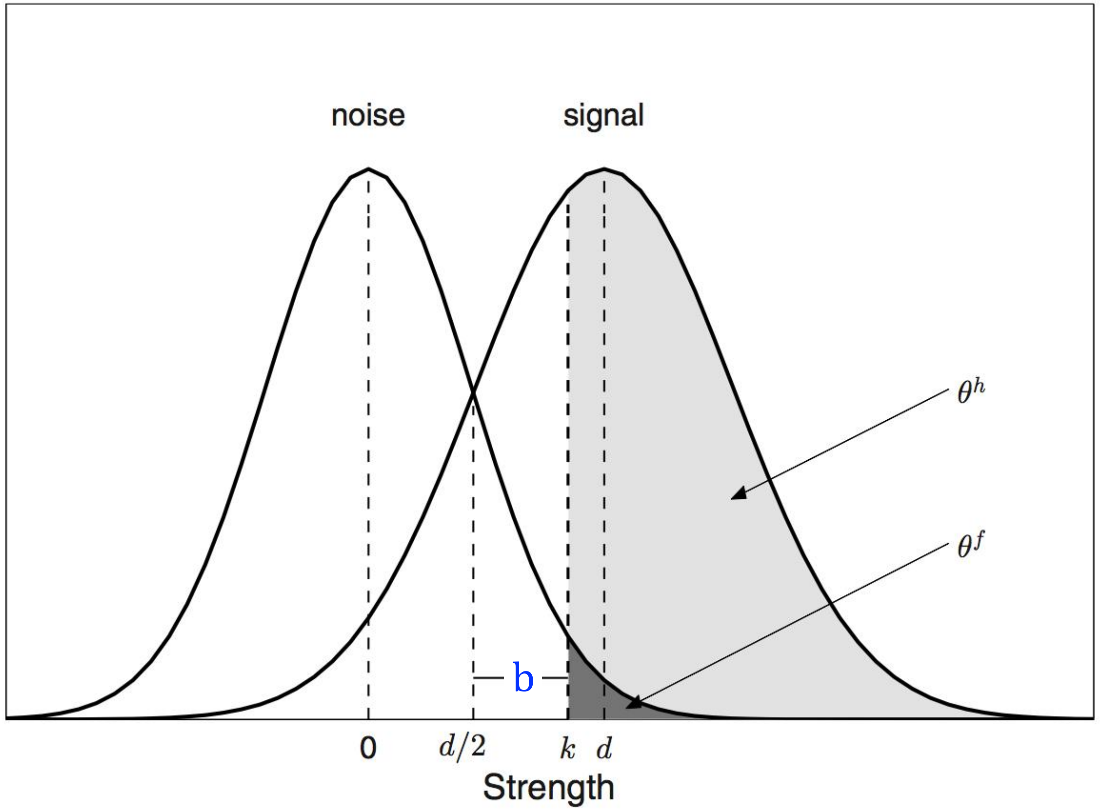


```
#simulate sdt given parameters and number of trials
simsdt<- function(d,b,ntrials) {
  old <- rnorm(ntrials/2,d)
  hits <-sum(old >(d/2+b))/(ntrials/2)*100
  new <- rnorm(ntrials/2,0)
  fas <- sum(new>(d/2+b))/(ntrials/2)*100
  return(X<-c(hits,fas))
}
```

Criterion: in the simulated trials, pro > d/2 + b

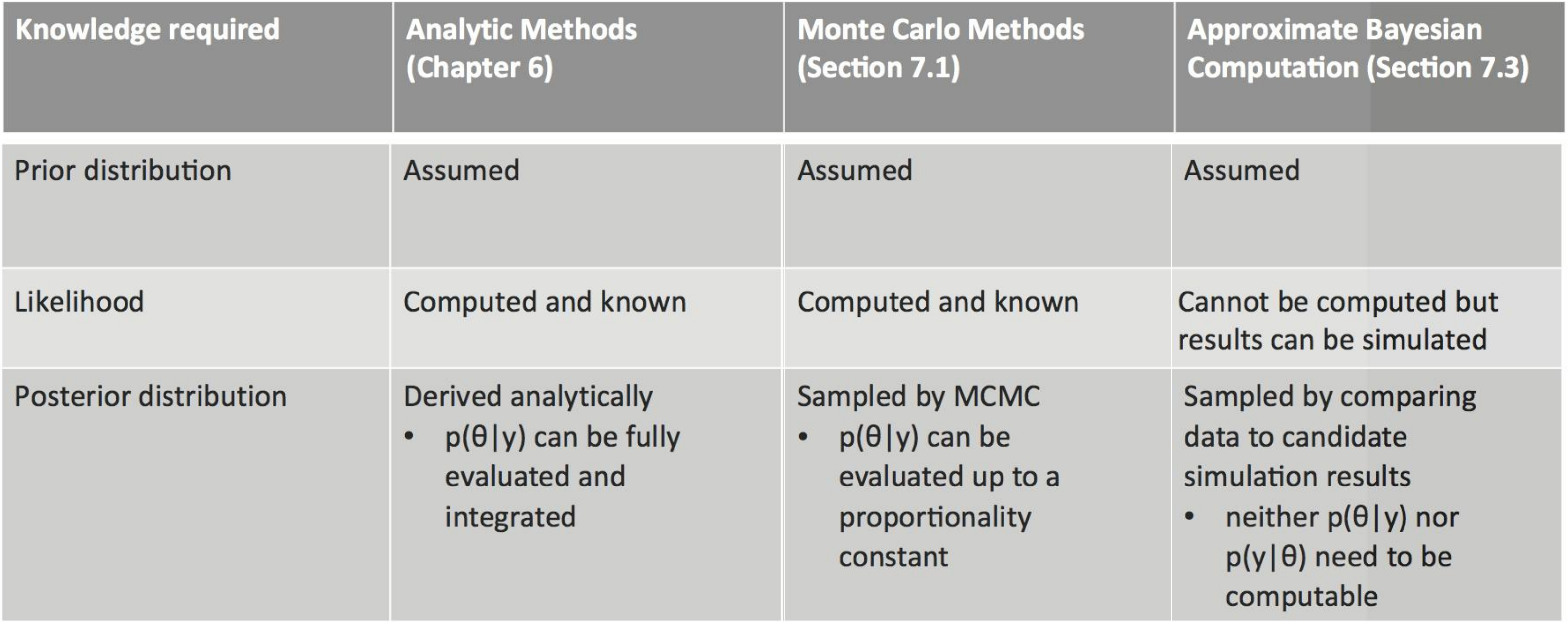
Return X: Percentage of hits and false alarm

X is simulated from the SDT model



Equal-variance Gaussian signal detection theory framework.

Summary for Bayesian Parameter Estimation



Knowledge required	Analytic Methods (Chapter 6)	Monte Carlo Methods (Section 7.1)	Approximate Bayesian Computation (Section 7.3)
Prior distribution	Assumed	Assumed	Assumed
Likelihood	Computed and known	Computed and known	Cannot be computed but results can be simulated
Posterior distribution	Derived analytically • p(θ|y) can be fully evaluated and integrated	Sampled by MCMC • p(θ|y) can be evaluated up to a proportionality constant	Sampled by comparing data to candidate simulation results • neither p(θ|y) nor p(y|θ) need to be computable

Lecture 11 – Gibbs Sampling & JAGS

- What is Gibbs Sampling?
  - Motivations
  - Gibbs Sampling: a Bivariate Example
  - Gibbs Sampling vs Metropolis-Hasting Sampling
- JAGS
  - Installation
  - Scripting for JAGS
- Examples of JAGS: Revisiting Some Known Models
  - Bayesian Modeling of Signal-Detection Theory (SDT)
  - A Bayesian Approach to a High-Threshold Model (1HT)
  - A Bayesian Approach to Multinomial Processing Tree (MPT)

Motivations for Gibbs Sampling

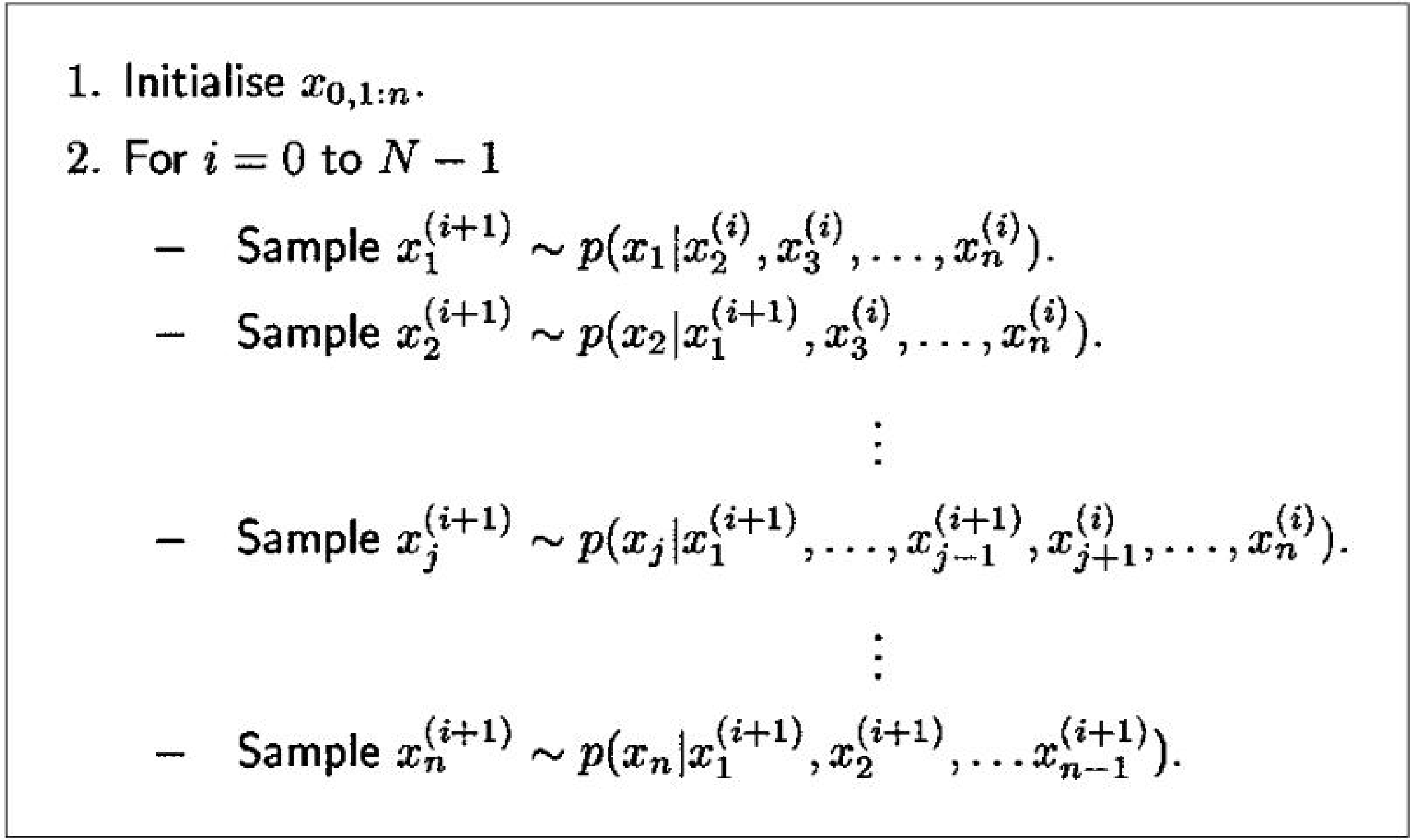
Sampling from the joint posterior for all the parameters may be unachievable in many situations;

P(x1, x2, … xn)

We can often easily sample from the posterior for one parameter given knowledge of the other parameters values.

- *Gibbs Sampler* samples from conditional distributions;
- By iterating through the parameters, sampling each conditional upon the others being constant, the Gibbs sampler manages to provide us with posterior distributions for each parameter.

Gibbs Sampler



```
1. Initialise $x_{0,1:n}$.
2. For $i = 0$ to $N - 1$
    – Sample $x_1^{(i+1)} \sim p(x_1|x_2^{(i)}, x_3^{(i)}, \ldots, x_n^{(i)})$.
    – Sample $x_2^{(i+1)} \sim p(x_2|x_1^{(i+1)}, x_3^{(i)}, \ldots, x_n^{(i)})$.
                                $\vdots$
    – Sample $x_j^{(i+1)} \sim p(x_j|x_1^{(i+1)}, \ldots, x_{j-1}^{(i+1)}, x_{j+1}^{(i)}, \ldots, x_n^{(i)})$.
                                $\vdots$
    – Sample $x_n^{(i+1)} \sim p(x_n|x_1^{(i+1)}, x_2^{(i+1)}, \ldots x_{n-1}^{(i+1)})$.
```

C Andrieu et al (2003) An Introduction to MCMC for Machine Learning

An example of Gibbs Sampling: Bivariate Gaussian

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) \triangleq \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$

Let us consider a 2d example.

$$\mu_x = \mu_y = 0 \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

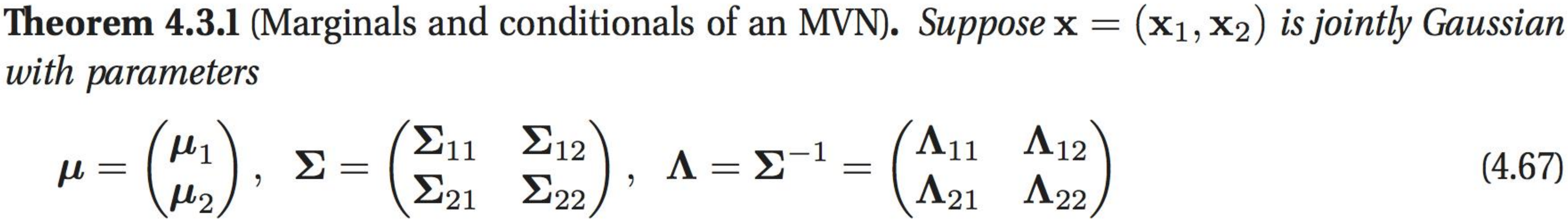
$$x^{(i+1)} \sim f(x|y^{(i)}) = N\left(\rho\left(\frac{\sigma_x}{\sigma_y}\right)y^{(i)}, \sqrt{\sigma_x^2\,(1-\rho^2)}\right)$$

$$y^{(i+1)} \sim f(y|x^{(i+1)}) = N\left(\rho\left(\frac{\sigma_y}{\sigma_x}\right)x^{(i+1)}, \sqrt{\sigma_y^2\,(1-\rho^2)}\right)$$

```
#gibbs sampling
sxt1mr <- sqrt(sigx^2*(1-rho^2))
syt1mr <- sqrt(sigy^2*(1-rho^2))
rxy <- rho*(sigx/sigy)
ryx <- rho*(sigy/sigx)
xsamp <- ysamp <- rep(0,nsamples)
xsamp[1] <- -2
ysamp[1] <- 2
for (i in c(1:(nsamples-1))) {
   xsamp[i+1] <- rnorm(1, mean=rxy*ysamp[i], sd=sxt1mr)
   ysamp[i+1] <- rnorm(1, mean=ryx*xsamp[i+1], sd=syt1mr)
}
```

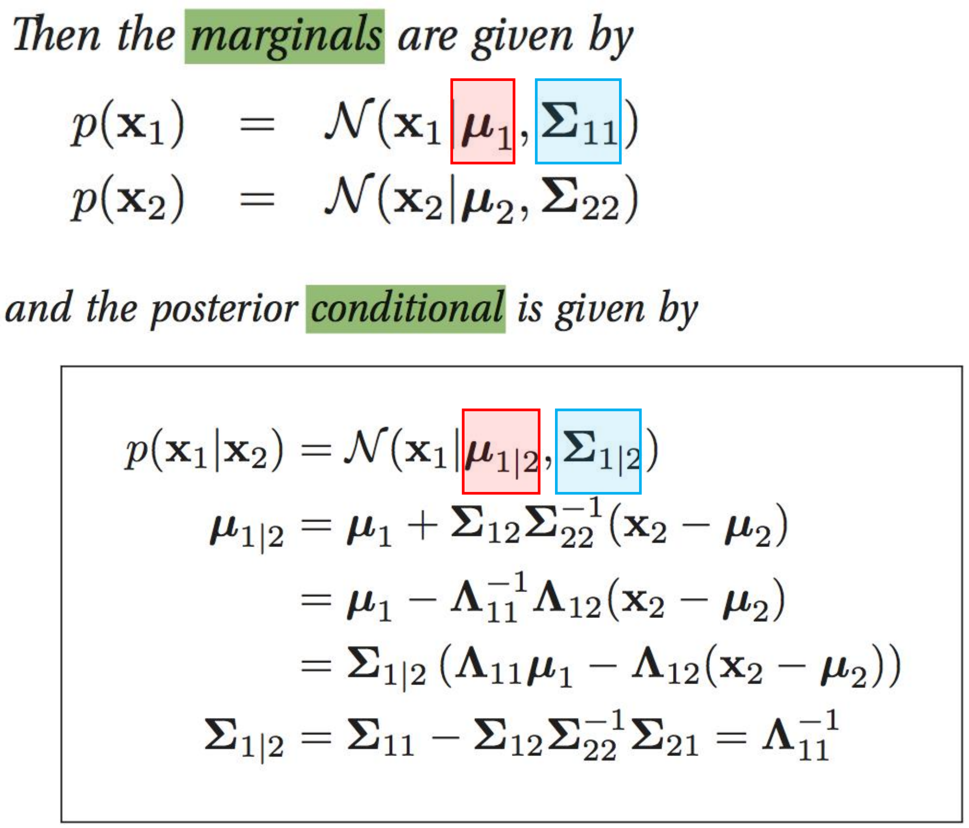
SEE CODE: Lecture11_1_gibbs.R

Multivariate Gaussian



Theorem 4.3.1 (Marginals and conditionals of an MVN). *Suppose* $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$ *is jointly Gaussian with parameters*

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}, \quad \boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1} = \begin{pmatrix} \boldsymbol{\Lambda}_{11} & \boldsymbol{\Lambda}_{12} \\ \boldsymbol{\Lambda}_{21} & \boldsymbol{\Lambda}_{22} \end{pmatrix} \tag{4.67}$$



Then the marginals are given by

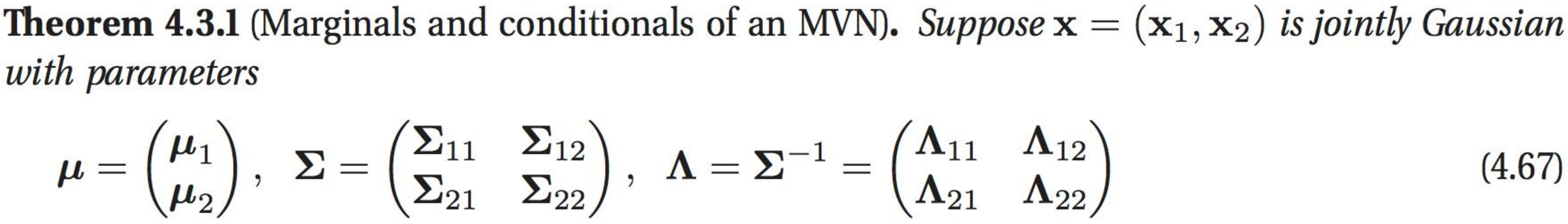
$$\begin{aligned} p(\mathbf{x}_1) &= \mathcal{N}(\mathbf{x}_1 | \boxed{\boldsymbol{\mu}_1}, \boxed{\boldsymbol{\Sigma}_{11}}) \\ p(\mathbf{x}_2) &= \mathcal{N}(\mathbf{x}_2 | \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22}) \end{aligned}$$

and the posterior conditional is given by

$$\begin{aligned} p(\mathbf{x}_1 | \mathbf{x}_2) &= \mathcal{N}(\mathbf{x}_1 | \boxed{\boldsymbol{\mu}_{1|2}}, \boxed{\boldsymbol{\Sigma}_{1|2}}) \\ \boldsymbol{\mu}_{1|2} &= \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2) \\ &= \boldsymbol{\mu}_1 - \boldsymbol{\Lambda}_{11}^{-1}\boldsymbol{\Lambda}_{12}(\mathbf{x}_2 - \boldsymbol{\mu}_2) \\ &= \boldsymbol{\Sigma}_{1|2}\left(\boldsymbol{\Lambda}_{11}\boldsymbol{\mu}_1 - \boldsymbol{\Lambda}_{12}(\mathbf{x}_2 - \boldsymbol{\mu}_2)\right) \\ \boldsymbol{\Sigma}_{1|2} &= \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} = \boldsymbol{\Lambda}_{11}^{-1} \end{aligned}$$

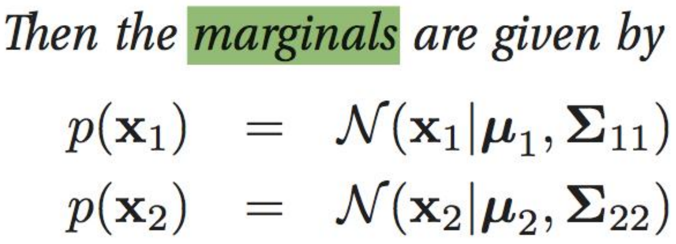
See Proof in

- MLaPP (pp116-pp119)
- 白板推导机器学习 P5



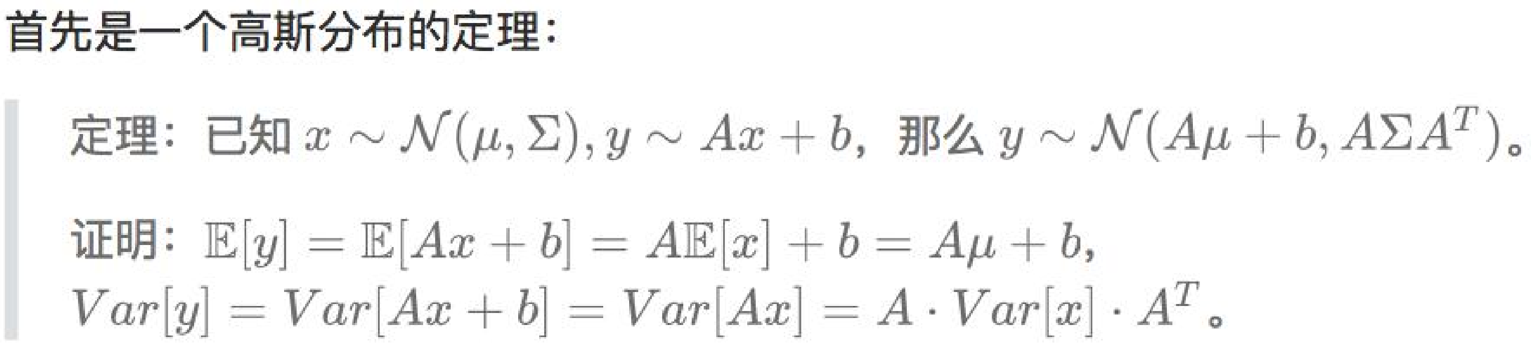
Theorem 4.3.1 (Marginals and conditionals of an MVN). *Suppose* $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$ *is jointly Gaussian with parameters*

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}, \quad \boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1} = \begin{pmatrix} \boldsymbol{\Lambda}_{11} & \boldsymbol{\Lambda}_{12} \\ \boldsymbol{\Lambda}_{21} & \boldsymbol{\Lambda}_{22} \end{pmatrix} \tag{4.67}$$



Then the marginals are given by

$$\begin{aligned} p(\mathbf{x}_1) &= \mathcal{N}(\mathbf{x}_1 | \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}) \\ p(\mathbf{x}_2) &= \mathcal{N}(\mathbf{x}_2 | \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22}) \end{aligned}$$



首先是一个高斯分布的定理：

> 定理：已知 $x \sim \mathcal{N}(\mu, \Sigma), y \sim Ax + b$，那么 $y \sim \mathcal{N}(A\mu + b, A\Sigma A^T)$。
>
> 证明：$\mathbb{E}[y] = \mathbb{E}[Ax + b] = A\mathbb{E}[x] + b = A\mu + b$,
> $Var[y] = Var[Ax + b] = Var[Ax] = A \cdot Var[x] \cdot A^T$。

构造 $x_a = \begin{pmatrix} \mathbb{I}_{m\times m} & \mathbb{O}_{m\times n} \end{pmatrix} \begin{pmatrix} x_a \\ x_b \end{pmatrix}$

$$\mathbb{E}[x_a] = \begin{pmatrix} \mathbb{I} & \mathbb{O} \end{pmatrix} \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix} = \mu_a$$

$$Var[x_a] = \begin{pmatrix} \mathbb{I} & \mathbb{O} \end{pmatrix} \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix} \begin{pmatrix} \mathbb{I} \\ \mathbb{O} \end{pmatrix} = \Sigma_{aa}$$

Theorem 4.3.1 (Marginals and conditionals of an MVN). *Suppose* $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$ *is jointly Gaussian with parameters*

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}, \quad \boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1} = \begin{pmatrix} \boldsymbol{\Lambda}_{11} & \boldsymbol{\Lambda}_{12} \\ \boldsymbol{\Lambda}_{21} & \boldsymbol{\Lambda}_{22} \end{pmatrix} \tag{4.67}$$

Then the marginals are given by

$$\begin{aligned} p(\mathbf{x}_1) &= \mathcal{N}(\mathbf{x}_1|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}) \\ p(\mathbf{x}_2) &= \mathcal{N}(\mathbf{x}_2|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22}) \end{aligned}$$

and the posterior conditional is given by

$$\begin{aligned} p(\mathbf{x}_1|\mathbf{x}_2) &= \mathcal{N}(\mathbf{x}_1|\boxed{\boldsymbol{\mu}_{1|2}}, \boxed{\boldsymbol{\Sigma}_{1|2}}) \\ \boldsymbol{\mu}_{1|2} &= \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2) \\ &= \boldsymbol{\mu}_1 - \boldsymbol{\Lambda}_{11}^{-1}\boldsymbol{\Lambda}_{12}(\mathbf{x}_2 - \boldsymbol{\mu}_2) \\ &= \boldsymbol{\Sigma}_{1|2}\left(\boldsymbol{\Lambda}_{11}\boldsymbol{\mu}_1 - \boldsymbol{\Lambda}_{12}(\mathbf{x}_2 - \boldsymbol{\mu}_2)\right) \\ \boldsymbol{\Sigma}_{1|2} &= \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} = \boldsymbol{\Lambda}_{11}^{-1} \end{aligned}$$

构造 $x_{b\cdot a} = x_b - \Sigma_{ba}\Sigma_{aa}^{-1}x_a$

$$x_{b\cdot a} = \begin{pmatrix} -\Sigma_{ba}\Sigma_{aa}^{-1} & \mathbb{I}_{n\times n} \end{pmatrix} \begin{pmatrix} x_a \\ x_b \end{pmatrix}$$

$$\mathbb{E}[x_{b\cdot a}] = \begin{pmatrix} -\Sigma_{ba}\Sigma_{aa}^{-1} & \mathbb{I}_{n\times n} \end{pmatrix} \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix} = \mu_{b\cdot a}$$

$$Var[x_{b\cdot a}] = \begin{pmatrix} -\Sigma_{ba}\Sigma_{aa}^{-1} & \mathbb{I}_{n\times n} \end{pmatrix} \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix} \begin{pmatrix} -\Sigma_{aa}^{-1}\Sigma_{ba}^T \\ \mathbb{I}_{n\times n} \end{pmatrix} = \Sigma_{bb\cdot a}$$

$$x_b = x_{b\cdot a} + \Sigma_{ba}\Sigma_{aa}^{-1}x_a$$

$$\mathbb{E}[x_b|x_a] = \mu_{b\cdot a} + \Sigma_{ba}\Sigma_{aa}^{-1}x_a$$

$$Var[x_b|x_a] = \Sigma_{bb\cdot a}$$

Quiz

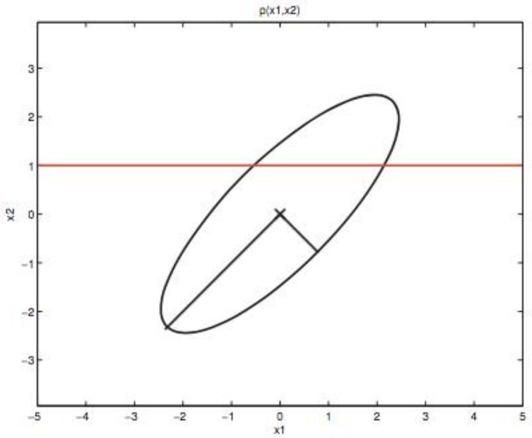
$$
\begin{aligned}
p(\mathbf{x}_1|\mathbf{x}_2) &= \mathcal{N}(\mathbf{x}_1|\boldsymbol{\mu}_{1|2}, \boldsymbol{\Sigma}_{1|2}) \\
\boldsymbol{\mu}_{1|2} &= \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2) \\
&= \boldsymbol{\mu}_1 - \boldsymbol{\Lambda}_{11}^{-1}\boldsymbol{\Lambda}_{12}(\mathbf{x}_2 - \boldsymbol{\mu}_2) \\
&= \boldsymbol{\Sigma}_{1|2}\left(\boldsymbol{\Lambda}_{11}\boldsymbol{\mu}_1 - \boldsymbol{\Lambda}_{12}(\mathbf{x}_2 - \boldsymbol{\mu}_2)\right) \\
\boldsymbol{\Sigma}_{1|2} &= \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} = \boldsymbol{\Lambda}_{11}^{-1}
\end{aligned}
$$

Let us consider a 2d example.

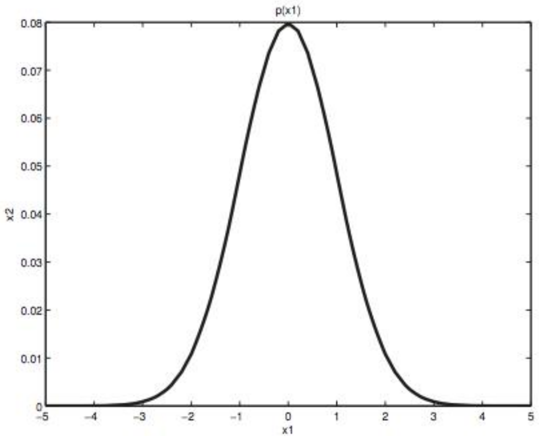
$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

$\rho = 0.8,\ \sigma_1 = \sigma_2 = 1,\ \boldsymbol{\mu} = \mathbf{0}$ and $x_2 = 1$

Suppose we observe $X_2 = x_2$; the conditional $p(x_1|x_2)$ is ?



p(x1,x2)



$p(x_1) = \mathcal{N}(x_1|\mu_1, \sigma_1^2)$

?

p(x1|x2)

Quiz

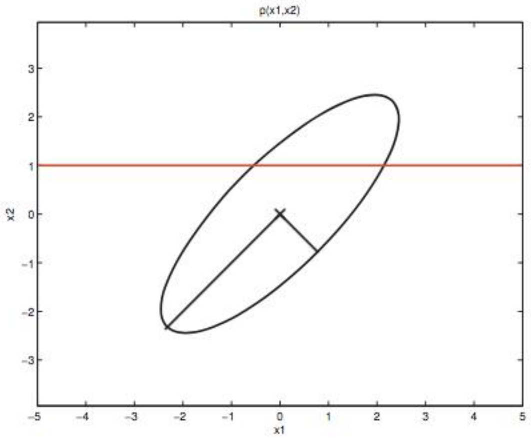
$$
\begin{aligned}
p(\mathbf{x}_1|\mathbf{x}_2) &= \mathcal{N}(\mathbf{x}_1|\boldsymbol{\mu}_{1|2}, \boldsymbol{\Sigma}_{1|2}) \\
\boldsymbol{\mu}_{1|2} &= \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2) \\
&= \boldsymbol{\mu}_1 - \boldsymbol{\Lambda}_{11}^{-1}\boldsymbol{\Lambda}_{12}(\mathbf{x}_2 - \boldsymbol{\mu}_2) \\
&= \boldsymbol{\Sigma}_{1|2}\left(\boldsymbol{\Lambda}_{11}\boldsymbol{\mu}_1 - \boldsymbol{\Lambda}_{12}(\mathbf{x}_2 - \boldsymbol{\mu}_2)\right) \\
\boldsymbol{\Sigma}_{1|2} &= \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} = \boldsymbol{\Lambda}_{11}^{-1}
\end{aligned}
$$

Let us consider a 2d example.

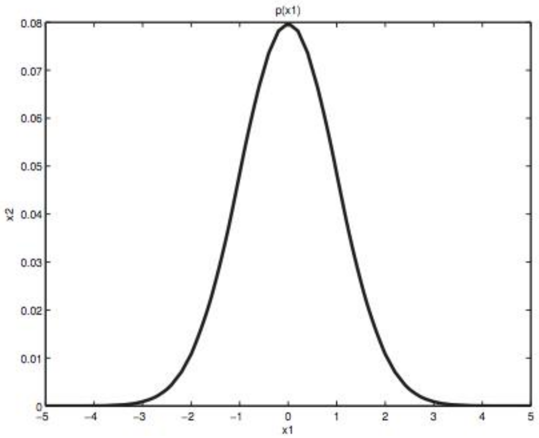
$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

$\rho = 0.8,\ \sigma_1 = \sigma_2 = 1,\ \boldsymbol{\mu} = \mathbf{0}$ and $x_2 = 1.$

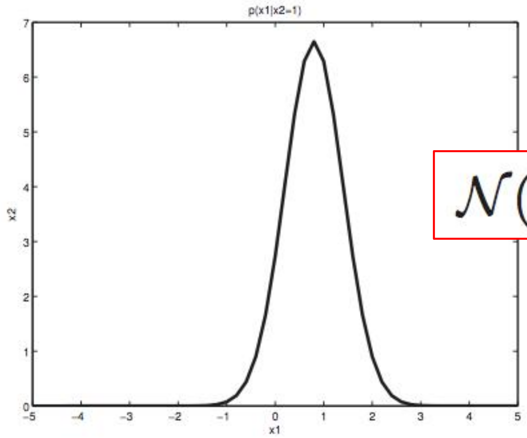
$$p(x_1|x_2) \quad = \quad \mathcal{N}\left(x_1|\mu_1 + \frac{\rho\sigma_1\sigma_2}{\sigma_2^2}(x_2 - \mu_2),\ \sigma_1^2 - \frac{(\rho\sigma_1\sigma_2)^2}{\sigma_2^2}\right)$$



p(x1,x2)



$p(x_1) = \mathcal{N}(x_1|\mu_1, \sigma_1^2)$



$\mathcal{N}(x_1|0.8, 0.36)$.

p(x1|x2)

An example of Gibbs Sampling: Bivariate Gaussian

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) \triangleq \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$

Let us consider a 2d example.

$$\mu_x = \mu_y = 0 \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

$$p(x_1|x_2) \quad = \quad \mathcal{N}\left(x_1|\mu_1 + \frac{\rho\sigma_1\sigma_2}{\sigma_2^2}(x_2-\mu_2),\ \sigma_1^2 - \frac{(\rho\sigma_1\sigma_2)^2}{\sigma_2^2}\right)$$

$$x^{(i+1)} \sim f(x|y^{(i)}) = N\left(\rho\left(\frac{\sigma_x}{\sigma_y}\right)y^{(i)}, \sqrt{\sigma_x^2\,(1-\rho^2)}\right)$$

$$y^{(i+1)} \sim f(y|x^{(i+1)}) = N\left(\rho\left(\frac{\sigma_y}{\sigma_x}\right)x^{(i+1)}, \sqrt{\sigma_y^2\,(1-\rho^2)}\right)$$

$$\mu_x = \mu_y = 0$$

An example of Gibbs Sampling: Bivariate Gaussian

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) \triangleq \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$

Let us consider a 2d example.

$$\mu_x = \mu_y = 0 \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

$$p(x_1|x_2) \quad = \quad \mathcal{N}\left(x_1|\mu_1 + \frac{\rho\sigma_1\sigma_2}{\sigma_2^2}(x_2 -\right.$$

$$x^{(i+1)} \sim f(x|y^{(i)}) = N\left(\rho\left(\frac{\sigma_x}{\sigma_y}\right)y^{(i)}, \sqrt{\sigma_x^2\left(1-\rho^2\right)}\right)$$

$$y^{(i+1)} \sim f(y|x^{(i+1)}) = N\left(\rho\left(\frac{\sigma_y}{\sigma_x}\right)x^{(i+1)}, \sqrt{\sigma_y^2\left(1-\rho^2\right)}\right)$$

```
#gibbs sampling
sxt1mr <- sqrt(sigx^2*(1-rho^2))
syt1mr <- sqrt(sigy^2*(1-rho^2))
rxy <- rho*(sigx/sigy)
ryx <- rho*(sigy/sigx)
xsamp <- ysamp <- rep(0,nsamples)
xsamp[1] <- -2
ysamp[1] <- 2
for (i in c(1:(nsamples-1))) {
   xsamp[i+1] <- rnorm(1, mean=rxy*ysamp[i], sd=sxt1mr)
   ysamp[i+1] <- rnorm(1, mean=ryx*xsamp[i+1], sd=syt1mr)
}
```

Gibbs Sampling vs Metropolis-Hastings Sampling

Gibbs sampling only applies to:

1) multiple variables
2) the conditional distributions for each single variable are known

M-H sampling can **also** operate in univariate situations

C Andrieu et al (2003)

Gibbs sampling do not reject samples. Acceptance probability is 1.

M-H sampling reject the proposals with certain probability. Acceptance probability is

$$\min\left\{1, \frac{p(x^{\star})q\left(x^{(i)} \mid x^{\star}\right)}{p\left(x^{(i)}\right)q\left(x^{\star}|x^{(i)}\right)}\right\}$$

JAGS - installation

JAGS stands for ‘Just Another Gibbs Sampler’.

To install JAGS:

Download the relevant files from this website:

http://mcmc-jags.sourceforge.net/

Run the installer

To call JAGS using R:

```
install.packages("rjags")
require(rjags) or library(rjags)
```

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) \triangleq \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$

$$\begin{aligned} p(\mathbf{x}_1|\mathbf{x}_2) &= \mathcal{N}(\mathbf{x}_1|\boldsymbol{\mu}_{1|2},\boldsymbol{\Sigma}_{1|2}) \\ \boldsymbol{\mu}_{1|2} &= \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2-\boldsymbol{\mu}_2) \\ &= \boldsymbol{\mu}_1 - \boldsymbol{\Lambda}_{11}^{-1}\boldsymbol{\Lambda}_{12}(\mathbf{x}_2-\boldsymbol{\mu}_2) \\ &= \boldsymbol{\Sigma}_{1|2}\left(\boldsymbol{\Lambda}_{11}\boldsymbol{\mu}_1 - \boldsymbol{\Lambda}_{12}(\mathbf{x}_2-\boldsymbol{\mu}_2)\right) \\ \boldsymbol{\Sigma}_{1|2} &= \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} = \boldsymbol{\Lambda}_{11}^{-1} \end{aligned}$$

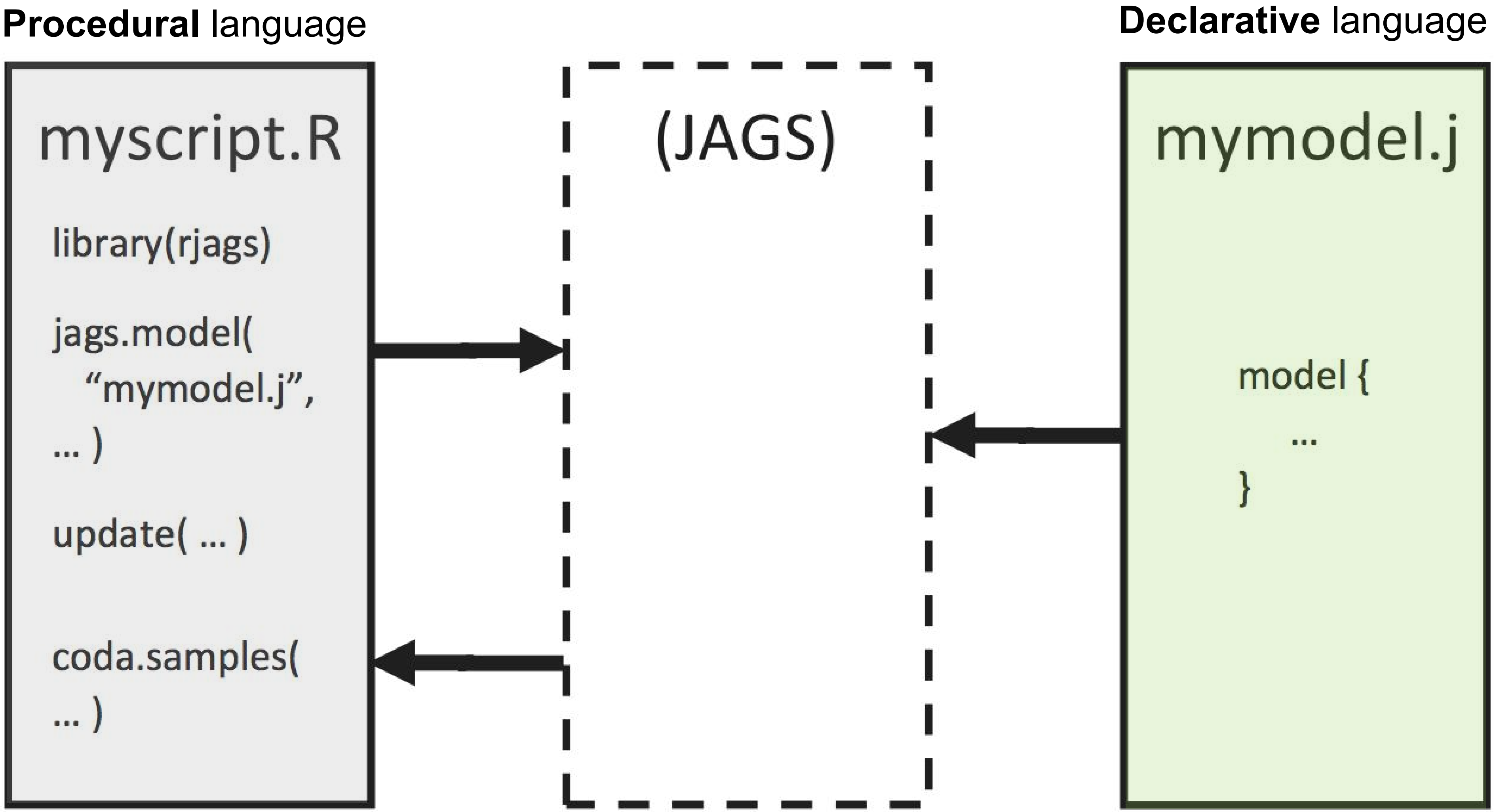
$$x^{(i+1)} \sim f(x|y^{(i)}) = N\left(\rho\left(\frac{\sigma_x}{\sigma_y}\right)y^{(i)}, \sqrt{\sigma_x^2\,(1-\rho^2)}\right)$$

$$y^{(i+1)} \sim f(y|x^{(i+1)}) = N\left(\rho\left(\frac{\sigma_y}{\sigma_x}\right)x^{(i+1)}, \sqrt{\sigma_y^2\,(1-\rho^2)}\right)$$

```
#gibbs sampling
sxt1mr <- sqrt(sigx^2*(1-rho^2))
syt1mr <- sqrt(sigy^2*(1-rho^2))
rxy <- rho*(sigx/sigy)
ryx <- rho*(sigy/sigx)
xsamp <- ysamp <- rep(0,nsamples)
xsamp[1] <- -2
ysamp[1] <- 2
for (i in c(1:(nsamples-1))) {
   xsamp[i+1] <- rnorm(1, mean=rxy*ysamp[i], sd=sxt1mr)
   ysamp[i+1] <- rnorm(1, mean=ryx*xsamp[i+1], sd=syt1mr)
}
```

→ Enveloped In JAGS

JAGS - Scripting



Procedural language

Declarative language

```
library(rjags)

jags.model(
  “mymodel.j”,
… )

update( … )

coda.samples(
… )
```

```
model {
    …
}
```

JAGS - Scripting

myscript.R

```
require(rjags)

N <- 1000
x <- rnorm(N, 0, 2)

myj <- jags.model("mymodel.j",
                  data = list("xx" = x, "N" = N))
update(myj, n.iter=1000)
mcmcfin<-coda.samples(myj, c("mu", "tau"),5000)

summary(mcmcfin)
plot(mcmcfin)
```

mymodel.j

```
#Gaussian
model {
   #model the data
         for (i in 1:N) {
                  xx[i] ~ dnorm(mu, tau)
         }

         #priors for parameters
         mu ~ dunif(-100,100)
         tau <- pow(sigma, -2)
         sigma ~ dunif(0, 100)
}
```

Examples of JAGS - SDT

Revisit SDT we have learned in Lecture 9

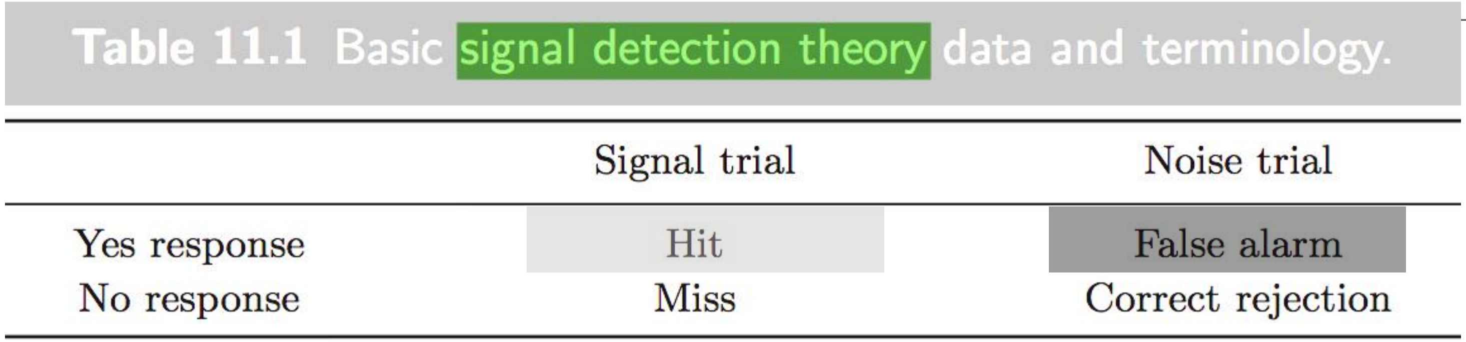
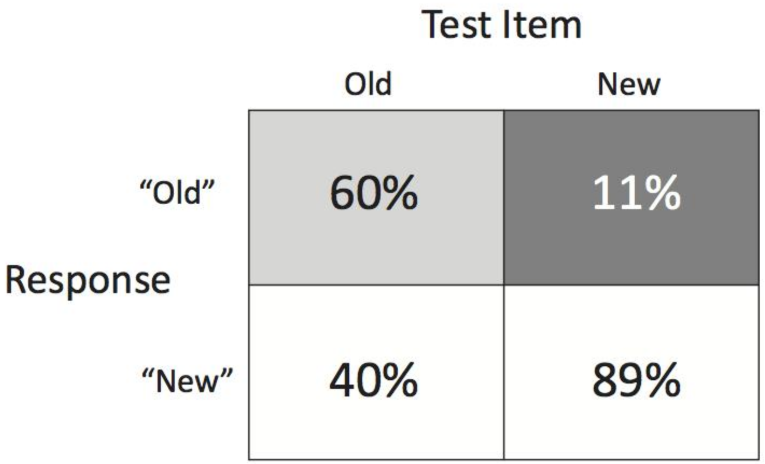
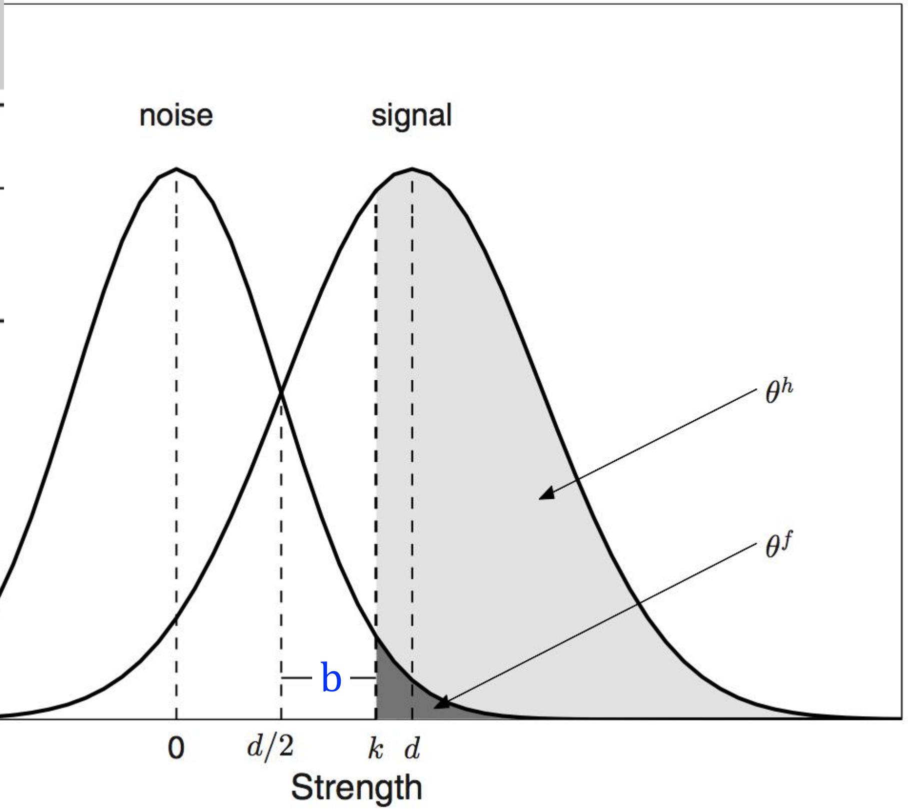


Table 11.1 Basic signal detection theory data and terminology.

	Signal trial	Noise trial
Yes response	Hit	False alarm
No response	Miss	Correct rejection



		Test Item	
		Old	New
Response	"Old"	60%	11%
	"New"	40%	89%



Equal-variance Gaussian signal detection theory framework.

Examples of JAGS – Scripting for SDT

SDT.R

```
library(rjags)
h <- 60
f <- 11
sigtrials <- noistrials <- 100
oneinit <- list(d=0, b=0)
myinits <- list(oneinit)[rep(1,4)]
myinits <- lapply(myinits,FUN=function(x) lapply(x,
FUN=function(y) y+rnorm(1,0,.1)))
sdtj <- jags.model("SDT.j", data = list("h"=h, "f"=f,
                              "sigtrials"=sigtrials,"noistrials"=noistrials),
                    inits=myinits, n.chains=4)
update(sdtj,n.iter=1000)
parameters <- c("d", "b", "phih", "phif")
mcmcfin<-coda.samples(sdtj, parameters, 5000)

summary(mcmcfin)
plot(mcmcfin)
gelman.plot(mcmcfin)
```

SDT.j

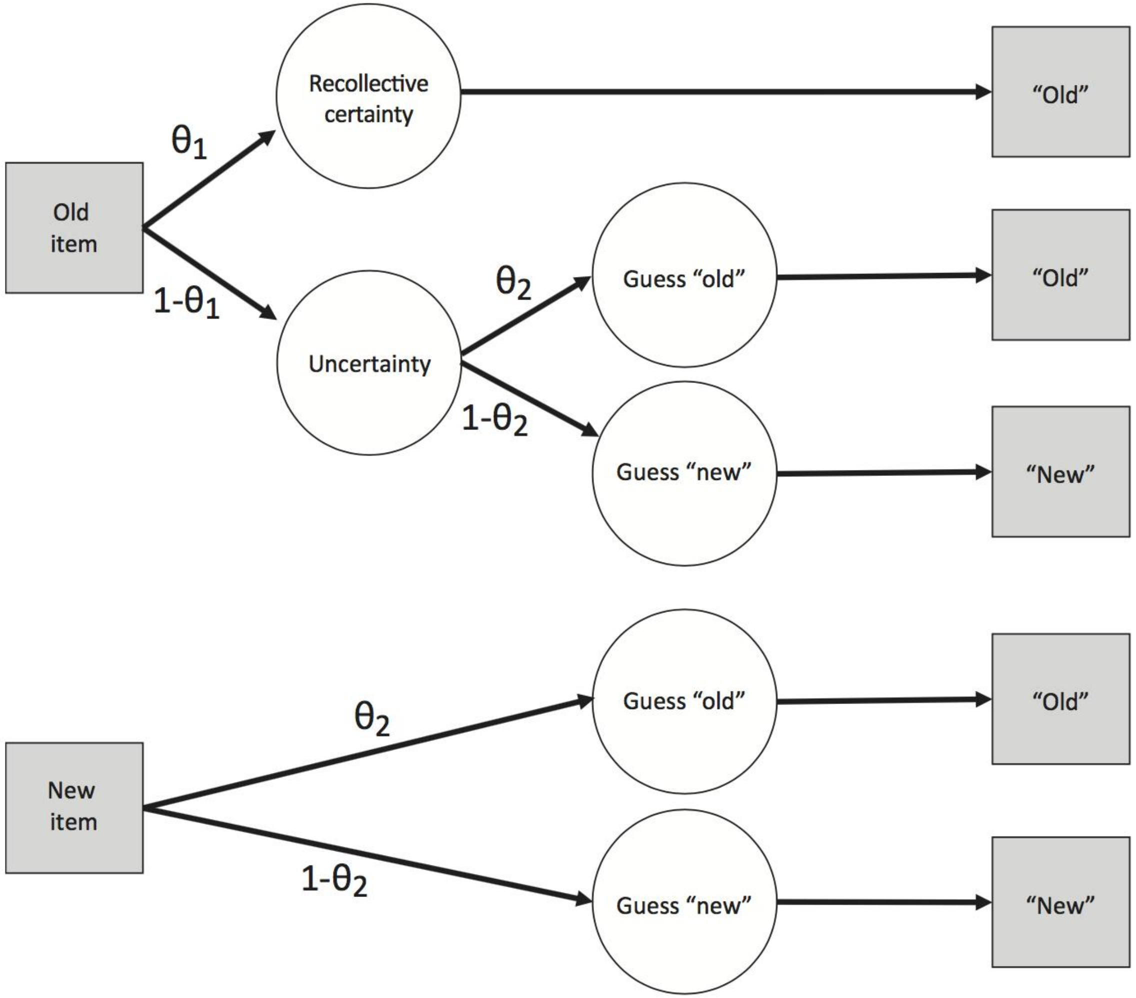
```
model{
# priors for discriminability and bias
    d ~ dnorm(1,1)
    b ~ dnorm(0,1)

    # express as areas under curves
    phih <- phi(d/2-b)   #normal cdf
    phif <- phi(-d/2-b)

    # Observed hits and false alarms
    h ~ dbin(phih, sigtrials)
    f ~ dbin(phif, noistrials)
}
```

Examples of JAGS – A High-Threshold Model

The high-threshold(1HT) model of recognition memory



Unlike signal-detection theory, the 1HT model does not contain a response criterion (d/2+b).

1HT model relies on two parameters:

1. the probability of being in the certain state, captured by the parameter θ1,
2. the probability of guessing "old", described by θ2, when in the state of uncertainty.

$$p(hit) = \theta_1 + (1 - \theta_1)\theta_2,$$
$$p(FA) = \theta_2,$$

Examples of JAGS – Scripting for 1HT

1HT.R

```
library(rjags)
#provide data from experiment
h <- 60
f <- 11
sigtrials <- noistrials <- 100

#define JAGS model
onehtj <- jags.model("1HT.j",
                  data = list("h"=h, "f"=f,
                              "sigtrials"=sigtrials,
                              "noistrials"=noistrials),
                   n.chains=4)

# burnin
update(onehtj,n.iter=1000)

# perform MCMC
parameters <- c("th1", "th2", "predh", "predf")
mcmcfin<-coda.samples(onehtj, parameters, 5000)
```

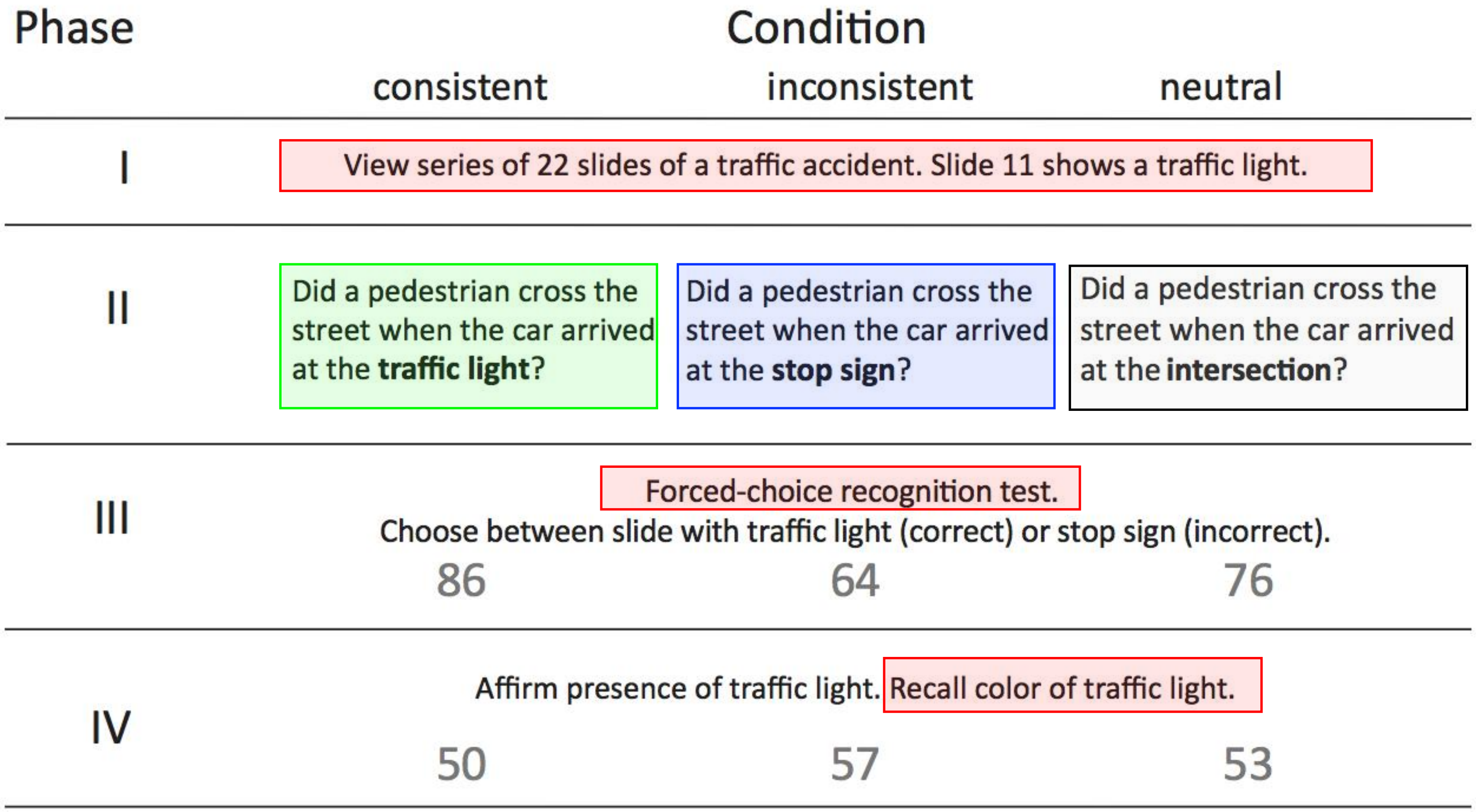
1HT.j

```
# High-threshold model
model{
    # priors for MPT parameters
    th1 ~ dbeta(1,1)
    th2 ~ dbeta(1,1)

    # predictions for responses
    predh <- th1+(1-th1)*th2
    predf <- th2

    # Observed responses
    h ~ dbin(predh, sigtrials)
    f ~ dbin(predf, noistrials)
}
```

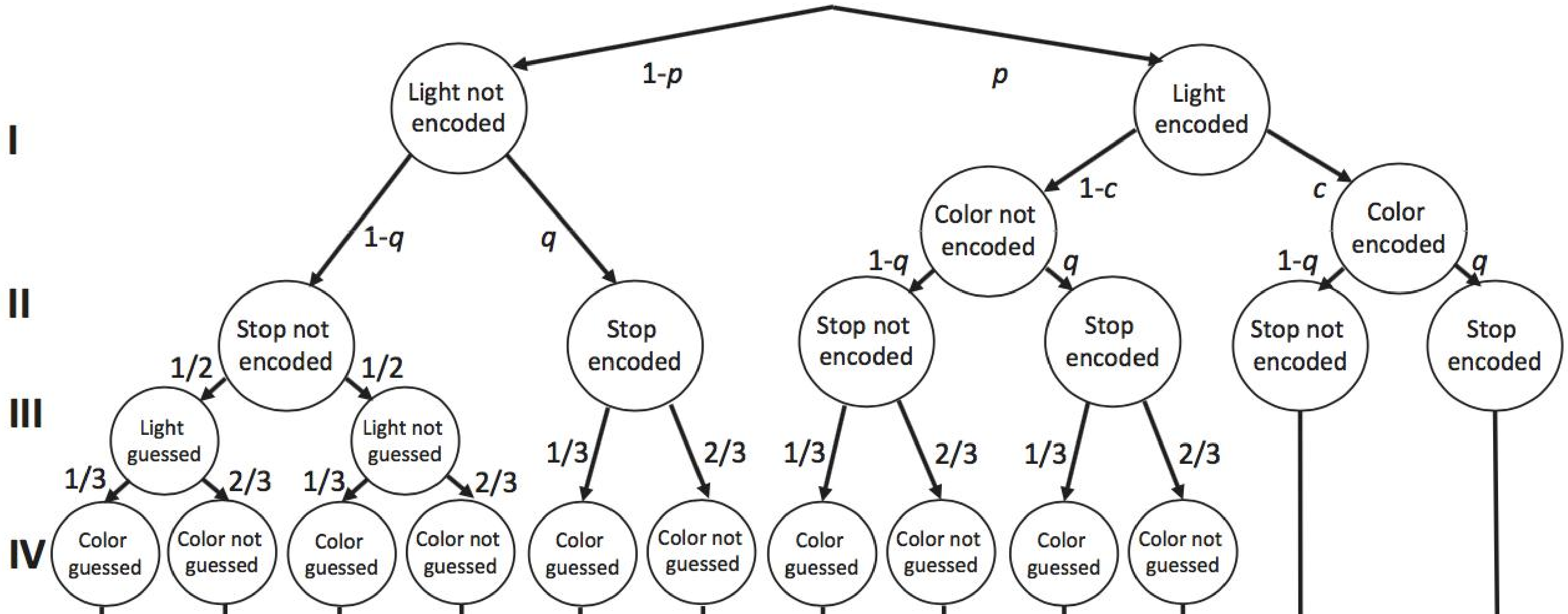
Examples of JAGS - Multinomial Processing Tree (MPT)



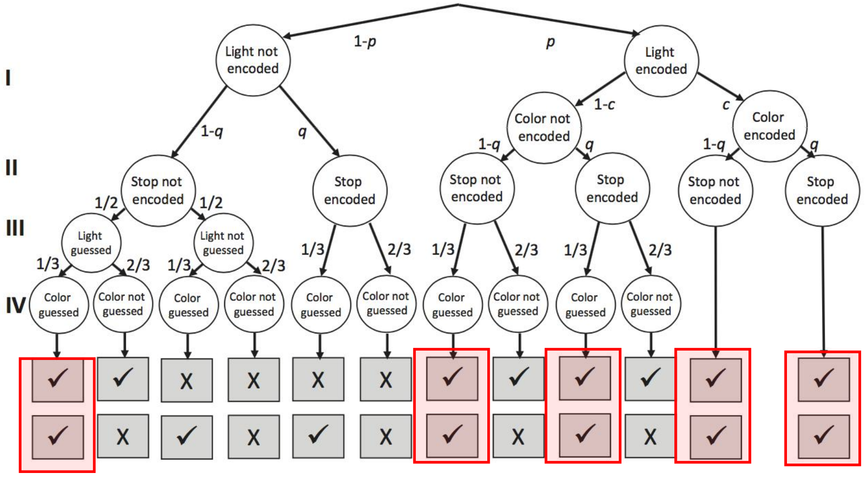
Phase	Condition		
	consistent	inconsistent	neutral
I	View series of 22 slides of a traffic accident. Slide 11 shows a traffic light.		
II	Did a pedestrian cross the street when the car arrived at the **traffic light**?	Did a pedestrian cross the street when the car arrived at the **stop sign**?	Did a pedestrian cross the street when the car arrived at the **intersection**?
III	Forced-choice recognition test. Choose between slide with traffic light (correct) or stop sign (incorrect).		
	86	64	76
IV	Affirm presence of traffic light. Recall color of traffic light.		
	50	57	53

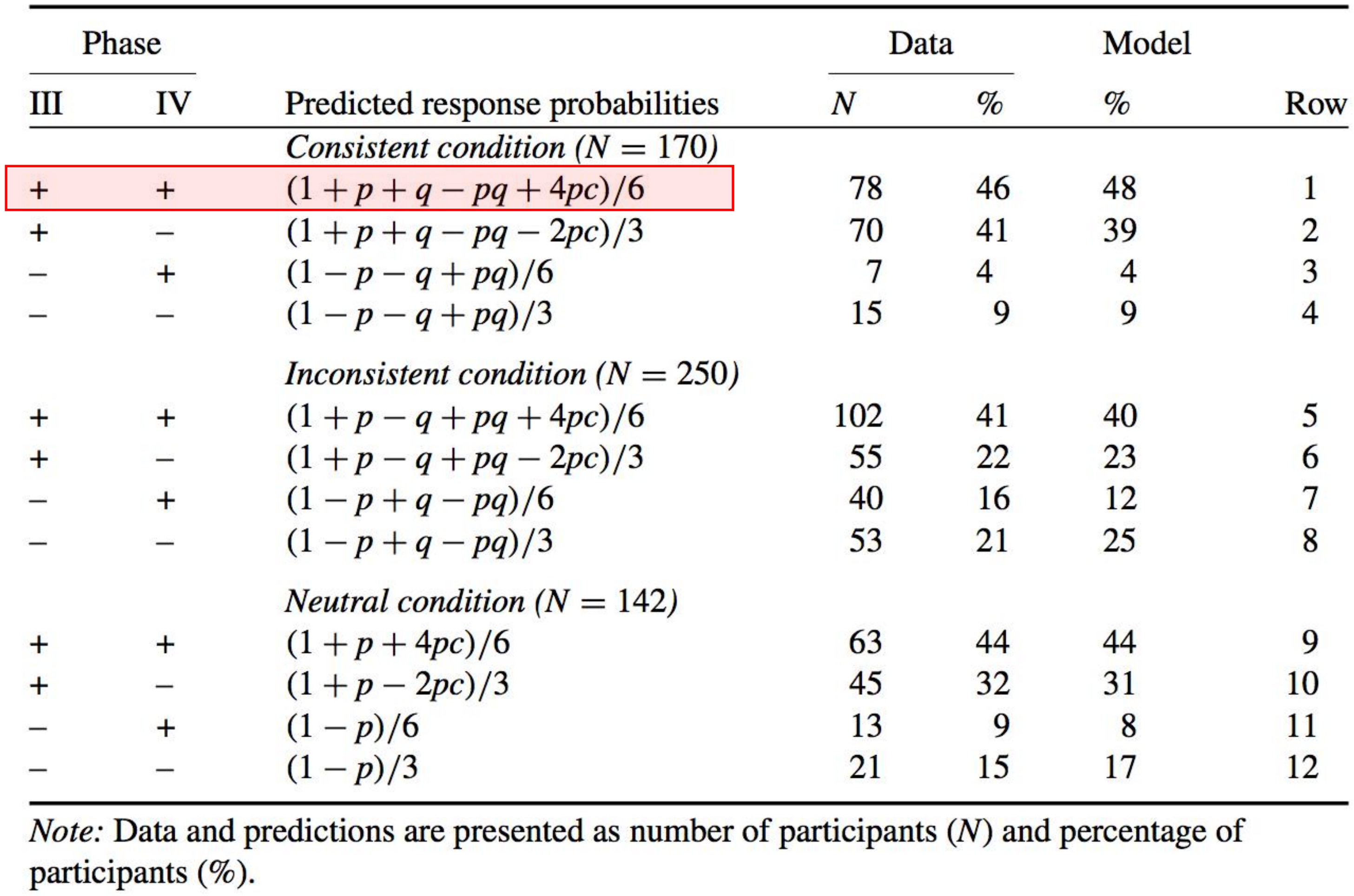
Wagenaar and Boer (1987)

Examples of JAGS – MPT



Decision in Phase III	✓	✓	X	X	X	X	✓	✓	✓	✓	✓	✓
Decision in Phase IV	✓	X	✓	X	✓	X	✓	X	✓	X	✓	✓





Phase			Data		Model	
III	IV	Predicted response probabilities	N	%	%	Row
		Consistent condition ($N = 170$)				
+	+	$(1 + p + q - pq + 4pc)/6$	78	46	48	1
+	–	$(1 + p + q - pq - 2pc)/3$	70	41	39	2
–	+	$(1 - p - q + pq)/6$	7	4	4	3
–	–	$(1 - p - q + pq)/3$	15	9	9	4
		Inconsistent condition ($N = 250$)				
+	+	$(1 + p - q + pq + 4pc)/6$	102	41	40	5
+	–	$(1 + p - q + pq - 2pc)/3$	55	22	23	6
–	+	$(1 - p + q - pq)/6$	40	16	12	7
–	–	$(1 - p + q - pq)/3$	53	21	25	8
		Neutral condition ($N = 142$)				
+	+	$(1 + p + 4pc)/6$	63	44	44	9
+	–	$(1 + p - 2pc)/3$	45	32	31	10
–	+	$(1 - p)/6$	13	9	8	11
–	–	$(1 - p)/3$	21	15	17	12

Note: Data and predictions are presented as number of participants (N) and percentage of participants (%).

SEE CODE:
MPT.R
MPT.j

Tips for JAGS

1. Run MCMC to achieve effective sample size (ESS) of 10,000
    - ESS is computed in R by the effectiveSize function

      effectiveSize(mcmcfin)
    - check the ESS of *every* parameter of interest, and the ESS of any interesting *parameter combinations*

2. Compose JAGS model statements for **human readability**
    - JAGS is a declarative language. The orders of script do not matter.
    - Start with the data, write the likelihood function, then write any dependencies among parameters, and finish with the prior distribution on the parameters

3. Make model diagrams for human comprehension and ease of programming.

   → We will learn the Graphical model in **Lecture 12**

Lecture 11 – Summary

- What is Gibbs Sampling?
  - Motivations
  - Gibbs Sampling: a Bivariate Gaussian Example
  - Gibbs Sampling vs Metropolis-Hasting Sampling
- JAGS
  - Installation
  - Scripting for JAGS
- Examples of JAGS: Revisiting Some Known Models
  - Bayesian Modeling of Signal-Detection Theory (SDT)
  - A Bayesian Approach to a High-Threshold Model (1HT)
  - A Bayesian Approach to Multinomial Processing Tree (MPT)

Homework 3

DDL: April 12, 2021，上课前
把code和结果图 做成pdf发邮件给 曲由之
12031145@mail.sustech.edu.cn

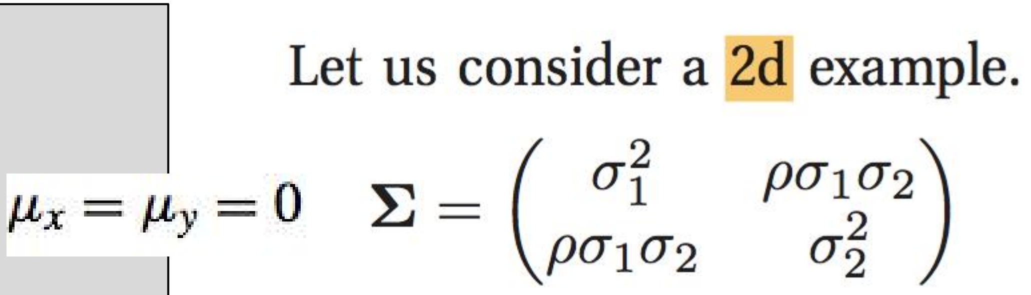
Fit data z to estimate the parameters (sig1, sig2, rho**) in Bivariate Gaussian model using JAGS.**

1) **When** *sig1* **and** *sig2* **have prior** unif(0, 10)**, and** *rho* **has prior** unif(0, 1)
2) **When** *sig1* **and** *sig2* **have prior** inverse gamma(0.5, 0.5)**, and** *rho* **has prior** unif(0, 1)

Hint:

- Data (z) is from Lecture11_1_gibbs.R;
- *mux = muy = 0*

```
#gibbs sampling
sxt1mr <- sqrt(sigx^2*(1-rho^2))
syt1mr <- sqrt(sigy^2*(1-rho^2))
rxy <- rho*(sigx/sigy)
ryx <- rho*(sigy/sigx)
xsamp <- ysamp <- rep(0,nsamples)
xsamp[1] <- -2
ysamp[1] <- 2
for (i in c(1:(nsamples-1))) {
    xsamp[i+1] <- rnorm(1, mean=rxy*ysamp[i], sd=sxt1mr)
    ysamp[i+1] <- rnorm(1, mean=ryx*xsamp[i+1], sd=syt1mr)
}
```



Let us consider a 2d example.

$$\mu_x = \mu_y = 0 \qquad \Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

Enveloped In JAGS

Reading materials

Textbooks (must read)

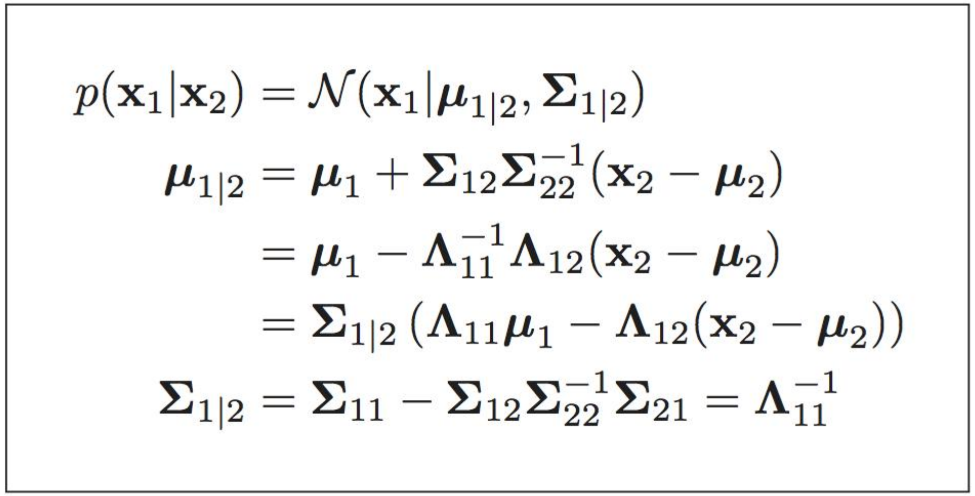
- Chapter 8 (The JAGS Language)

Papers (optional)

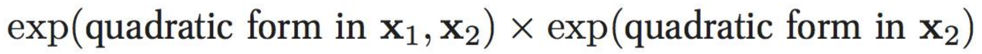
- C Andrieu, N de Freitas, A Doucet, M Jordan, (2003) An Introduction to MCMC for Machine Learning, Machine Learning

- **Extra readings:**
    - MLaPP (pp112-pp119)
    - 白板推导机器学习 (P5)

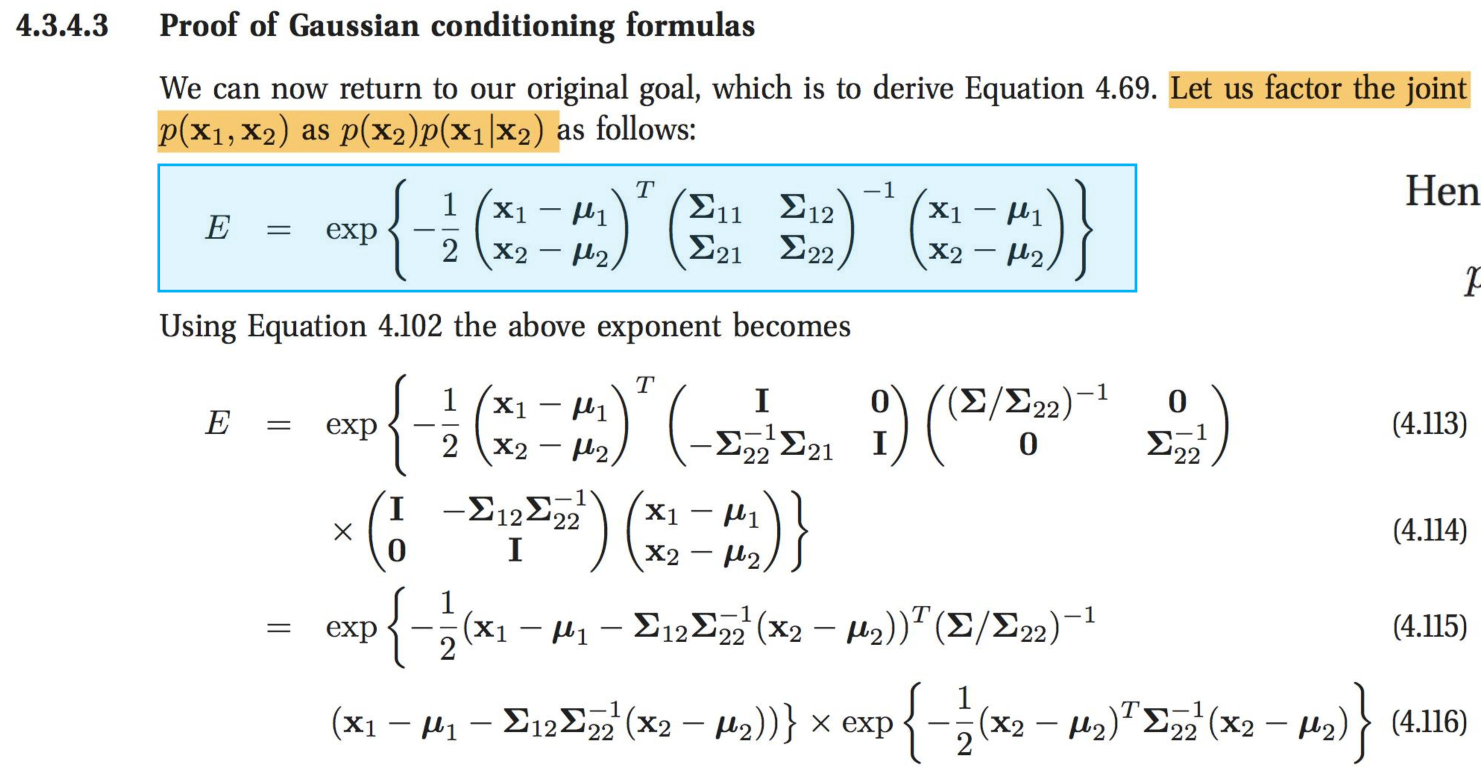
Proof in MLaPP (pp116-pp119)



$$
\begin{aligned}
p(\mathbf{x}_1|\mathbf{x}_2) &= \mathcal{N}(\mathbf{x}_1|\boldsymbol{\mu}_{1|2}, \boldsymbol{\Sigma}_{1|2}) \\
\boldsymbol{\mu}_{1|2} &= \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2) \\
&= \boldsymbol{\mu}_1 - \boldsymbol{\Lambda}_{11}^{-1}\boldsymbol{\Lambda}_{12}(\mathbf{x}_2 - \boldsymbol{\mu}_2) \\
&= \boldsymbol{\Sigma}_{1|2}\left(\boldsymbol{\Lambda}_{11}\boldsymbol{\mu}_1 - \boldsymbol{\Lambda}_{12}(\mathbf{x}_2 - \boldsymbol{\mu}_2)\right) \\
\boldsymbol{\Sigma}_{1|2} &= \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} = \boldsymbol{\Lambda}_{11}^{-1}
\end{aligned}
$$



$$\exp(\text{quadratic form in } \mathbf{x}_1, \mathbf{x}_2) \times \exp(\text{quadratic form in } \mathbf{x}_2)$$



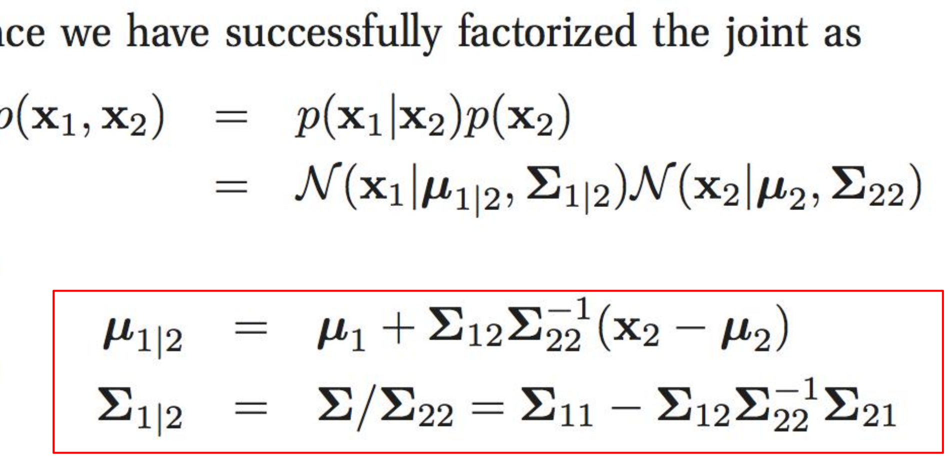
4.3.4.3 Proof of Gaussian conditioning formulas

We can now return to our original goal, which is to derive Equation 4.69. Let us factor the joint $p(\mathbf{x}_1, \mathbf{x}_2)$ as $p(\mathbf{x}_2)p(\mathbf{x}_1|\mathbf{x}_2)$ as follows:

$$E = \exp\left\{-\frac{1}{2}\begin{pmatrix}\mathbf{x}_1 - \boldsymbol{\mu}_1 \\ \mathbf{x}_2 - \boldsymbol{\mu}_2\end{pmatrix}^T \begin{pmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{pmatrix}^{-1} \begin{pmatrix}\mathbf{x}_1 - \boldsymbol{\mu}_1 \\ \mathbf{x}_2 - \boldsymbol{\mu}_2\end{pmatrix}\right\}$$

Using Equation 4.102 the above exponent becomes

$$
\begin{aligned}
E &= \exp\left\{-\frac{1}{2}\begin{pmatrix}\mathbf{x}_1 - \boldsymbol{\mu}_1 \\ \mathbf{x}_2 - \boldsymbol{\mu}_2\end{pmatrix}^T \begin{pmatrix}\mathbf{I} & \mathbf{0} \\ -\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} & \mathbf{I}\end{pmatrix} \begin{pmatrix}(\boldsymbol{\Sigma}/\boldsymbol{\Sigma}_{22})^{-1} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{22}^{-1}\end{pmatrix}\right. && (4.113) \\
&\quad \left. \times \begin{pmatrix}\mathbf{I} & -\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \\ \mathbf{0} & \mathbf{I}\end{pmatrix} \begin{pmatrix}\mathbf{x}_1 - \boldsymbol{\mu}_1 \\ \mathbf{x}_2 - \boldsymbol{\mu}_2\end{pmatrix}\right\} && (4.114) \\
&= \exp\left\{-\frac{1}{2}(\mathbf{x}_1 - \boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2))^T(\boldsymbol{\Sigma}/\boldsymbol{\Sigma}_{22})^{-1}\right. && (4.115) \\
&\quad \left.(\mathbf{x}_1 - \boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2))\right\} \times \exp\left\{-\frac{1}{2}(\mathbf{x}_2 - \boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2)\right\} && (4.116)
\end{aligned}
$$



Hence we have successfully factorized the joint as

$$
\begin{aligned}
p(\mathbf{x}_1, \mathbf{x}_2) &= p(\mathbf{x}_1|\mathbf{x}_2)p(\mathbf{x}_2) \\
&= \mathcal{N}(\mathbf{x}_1|\boldsymbol{\mu}_{1|2}, \boldsymbol{\Sigma}_{1|2})\mathcal{N}(\mathbf{x}_2|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})
\end{aligned}
$$

$$
\begin{aligned}
\boldsymbol{\mu}_{1|2} &= \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2) \\
\boldsymbol{\Sigma}_{1|2} &= \boldsymbol{\Sigma}/\boldsymbol{\Sigma}_{22} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}
\end{aligned}
$$