隐马尔科夫链模型

理论及其在R中的应用

陆银波

中国人民大学统计学院

第5届R语言会议，2012

大纲

Outline

从伯努利到隐马氏

- 伯努利模型：抛一枚硬币，记下结果，重复实验，估计参数p
- 混合模型：从若干枚硬币中选出一枚，抛出，记下结果，重复实验，估计参数p 和 π
- 隐马氏模型：从若干枚硬币中，以马尔可夫过程(未知)选择一枚，抛出...，估计参数...

从伯努利到隐马氏

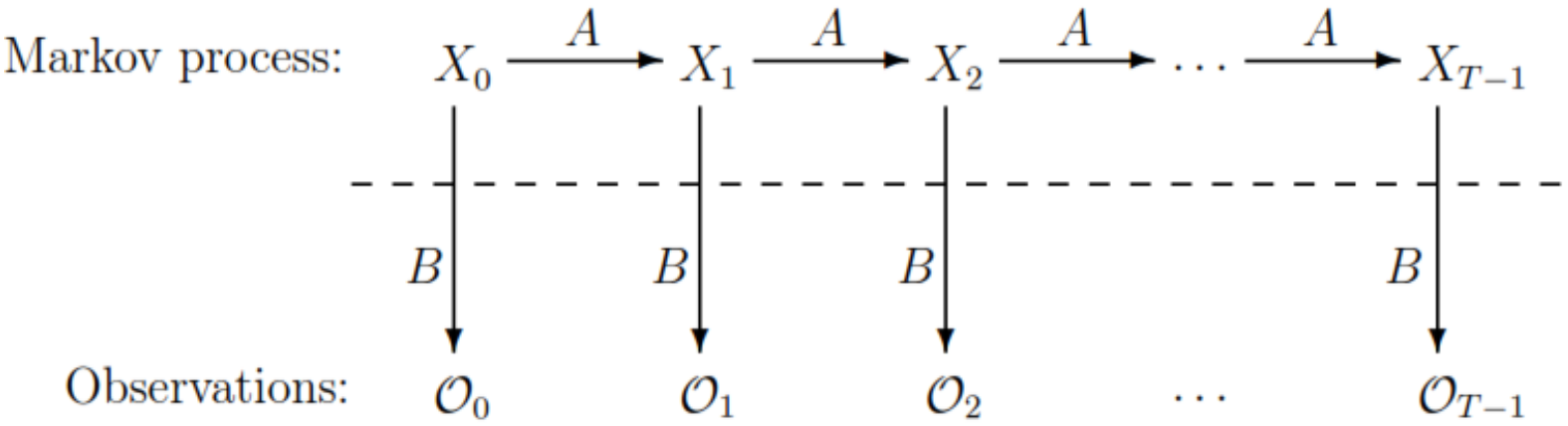
- 伯努利模型：抛一枚硬币，记下结果，重复实验，估计参数p
- 混合模型：从若干枚硬币中选出一枚，抛出，记下结果，重复实验，估计参数p 和 π
- 隐马氏模型：从若干枚硬币中，以马尔可夫过程(未知)选择一枚，抛出...，估计参数...

从伯努利到隐马氏

- 伯努利模型：抛一枚硬币，记下结果，重复实验，估计参数p
- 混合模型：从若干枚硬币中选出一枚，抛出，记下结果，重复实验，估计参数p 和 π
- 隐马氏模型：从若干枚硬币中，以马尔可夫过程(未知)选择一枚，抛出...，估计参数...

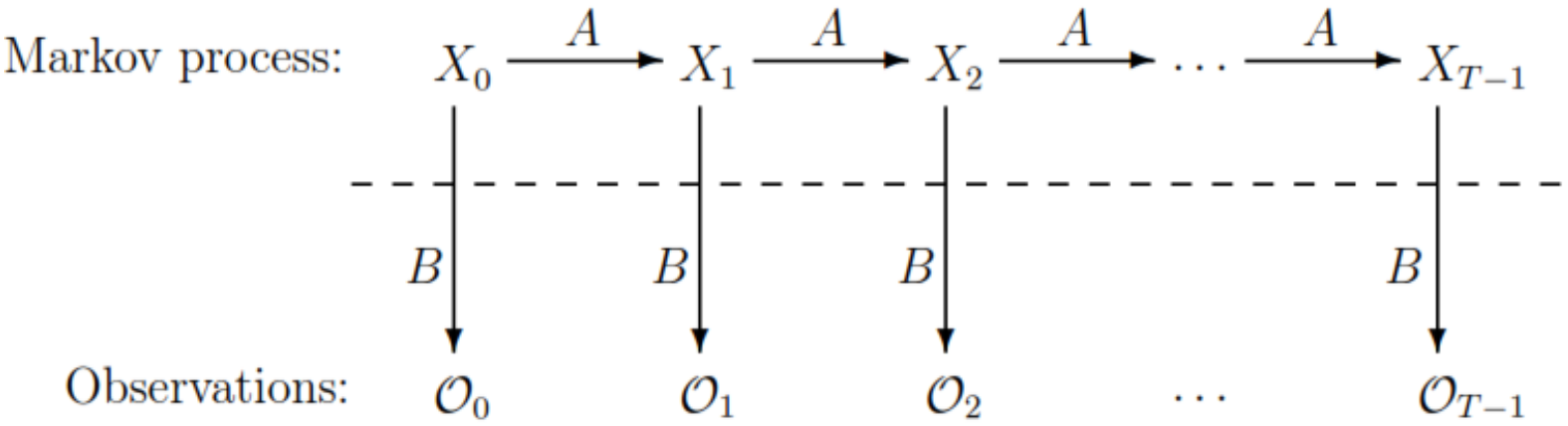
Outline

示意图



- X_i均取值于离散状态空间$\{1, 2, \ldots k\}$k 事先给定
- X_0由初始状态分布(多项式分布)产生，之后以状态转移矩阵A 产生$X_1 \ldots X_{T-1}$
- B表示不同状态下，观测值所服从的分布的参数
- 通常假定马氏过程具有平稳分布：$\pi A = \pi$

示意图



- X_i均取值于离散状态空间$\{1, 2, \ldots k\}$k 事先给定
- X_0由初始状态分布(多项式分布)产生，之后以状态转移矩阵A 产生$X_1 \ldots X_{T-1}$
- B表示不同状态下，观测值所服从的分布的参数
- 通常假定马氏过程具有平稳分布：$\pi A = \pi$

小结

一个简单的隐马氏模型由以下5个要素定义

- k 状态数
- O观测序列
- π初始状态分布
- A状态转移矩阵
- B依赖于状态的分布参数

k ,O 已知。π, A, B, 需要估计, 被称为模型参数

Outline

3个问题

问题1

已知模型参数，计算似然函数

问题2

估计最优模型参数(最大似然)

问题3

已知模型参数，估计最优状态序列X_i

3个问题

问题1

已知模型参数，计算似然函数

问题2

估计最优模型参数(最大似然)

问题3

已知模型参数，估计最优状态序列X_i

3个问题

问题1

已知模型参数，计算似然函数

问题2

估计最优模型参数(最大似然)

问题3

已知模型参数，估计最优状态序列X_i

算法

- 前向行算法；递推过程
- BaumWelch算法;即EM算法
- Viterbi算法；动态规划过程

Outline

Fact

在平稳分布的假定下，观测值边际分布为混合分布；混合比例为初始状态分布均值，方差等各级矩均易求得。

Fact

$$\begin{aligned}&\mathrm{cov}(g(O_t), g(O_{t+k})) = \\ &\mathrm{cov}(g(O_t), g(O_{t+k})|X_t, X_{t+k})\Pr(X_t)\Pr(X_{t+k}|X_t) = \\ &E(g(O_t)|X_t)E(g(O_{t+k})|X_{t+k})\Pr(X_t)A^k_{X_tX_{t+k}} - \dots\end{aligned}$$

Fact

在平稳分布的假定下，观测值边际分布为混合分布；混合比例为初始状态分布均值，方差等各级矩均易求得。

Fact

$$\begin{aligned}
&\mathrm{cov}(g(O_t), g(O_{t+k})) = \\
&\mathrm{cov}(g(O_t), g(O_{t+k})|X_t, X_{t+k})\Pr(X_t)\Pr(X_{t+k}|X_t) = \\
&E(g(O_t)|X_t)E(g(O_{t+k})|X_{t+k})\Pr(X_t)A^k_{X_tX_{t+k}} - \ldots
\end{aligned}$$

Outline

对缺失值的估计

$$Pr(0_t|0^{(-t)}) = Pr(0|\text{模型参数})/pr(0^{(-t)}|\text{模型参数})$$

对未来值的预测

$$Pr(0_{T+k}|0)=Pr(0_{T+k}|\text{模型参数})/Pr(0|\text{模型参数})$$

对缺失值的估计

$$Pr(0_t|0^{(-t)}) = Pr(0|\text{模型参数})/pr(0^{(-t)}|\text{模型参数})$$

对未来值的预测

$$Pr(0_{T+k}|0)=Pr(0_{T+k}|\text{模型参数})/Pr(0|\text{模型参数})$$

Outline

检验

Theorem

$\Phi^{-1}(F(X)) \sim N(0,1)$

$\text{residual}=\Phi^{-1}(F_t(X_t \leq x_t))$

Outline

数据

数据：1900—2005年世界范围内震级7级以上地震发生数
数据来源： http://neic.usgs.gov/neis/eqlists

R包： HiddenMarkov

特点:简单，直观，用于处理简单的隐马氏模型；
观测序列为单变量；
边际分布为一些常见的混合分布,如泊松,伯努利,高斯,指数等；
状态转移矩阵不随时间变化；

数据

数据：1900—2005年世界范围内震级7级以上地震发生数
数据来源： http://neic.usgs.gov/neis/eqlists

R包： HiddenMarkov

特点:简单，直观，用于处理简单的隐马氏模型；
观测序列为单变量；
边际分布为一些常见的混合分布,如泊松,伯努利,高斯,指数等；
状态转移矩阵不随时间变化；

主要的函数

- dthmm():设定模型参数的初始值，数据
- BaumWelch():估计模型参数
- forward():前向型算法，预测时需要用到一些中间值
- simulate():给出数据拟合值
- residuals()：计算残差

主要的函数

- dthmm():设定模型参数的初始值，数据
- BaumWelch():估计模型参数
- forward():前向型算法，预测时需要用到一些中间值
- simulate():给出数据拟合值
- residuals()：计算残差

主要的函数

- dthmm():设定模型参数的初始值，数据
- BaumWelch():估计模型参数
- forward():前向型算法，预测时需要用到一些中间值
- simulate():给出数据拟合值
- residuals()：计算残差

最优状态数

	均值	方差	AIC
样本	19.364	51.573	--
1状态	19.364	19.364	785.8
2状态	19.086	44.523	692.6
3状态	18.322	50.709	676.9
4状态	18.021	49.837	687.7
5状态	18.011	48.956	701.5

Table: 选择合适的状态空间

状态数选为3

预测

	2007	2008	2009	2010	2011
$\Pr([15,20])$	0. 247	0. 251	0. 255	0. 258	0. 260
真实值	18	12	17	24	20

Table: 预测2007-2011年发生15-20次的地震的概率

自相关性

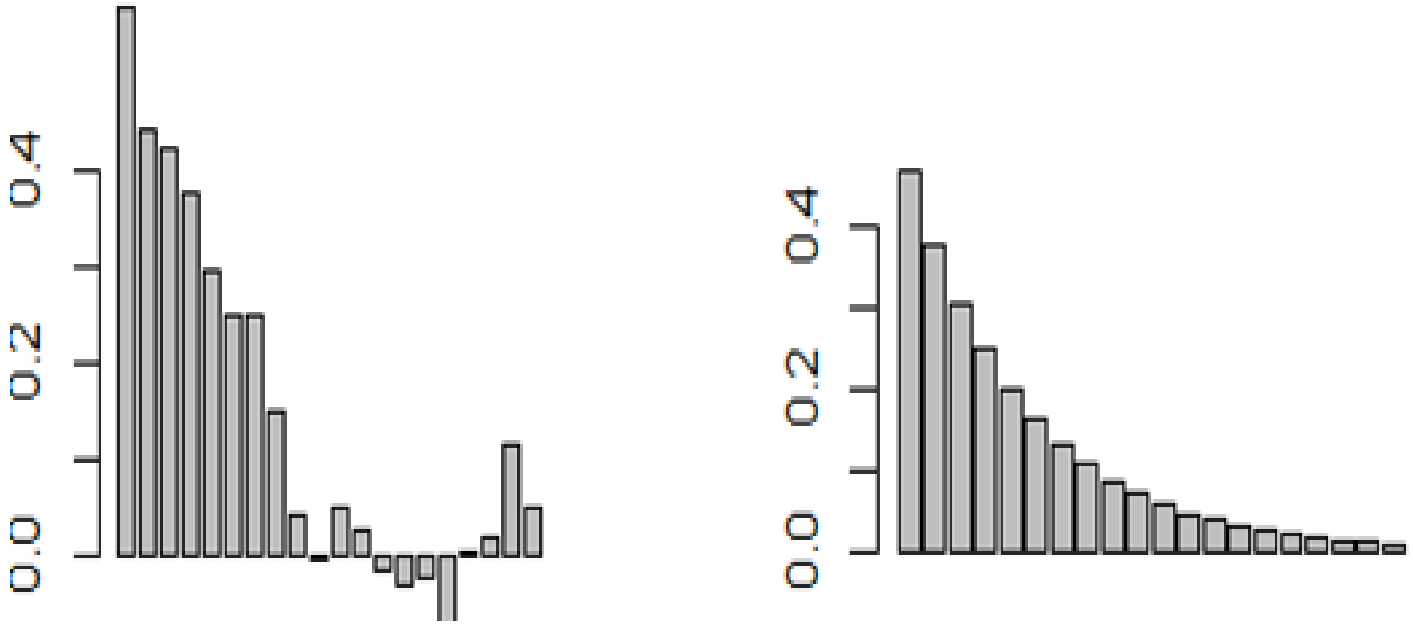


Figure: 自相关性比较

残差检验

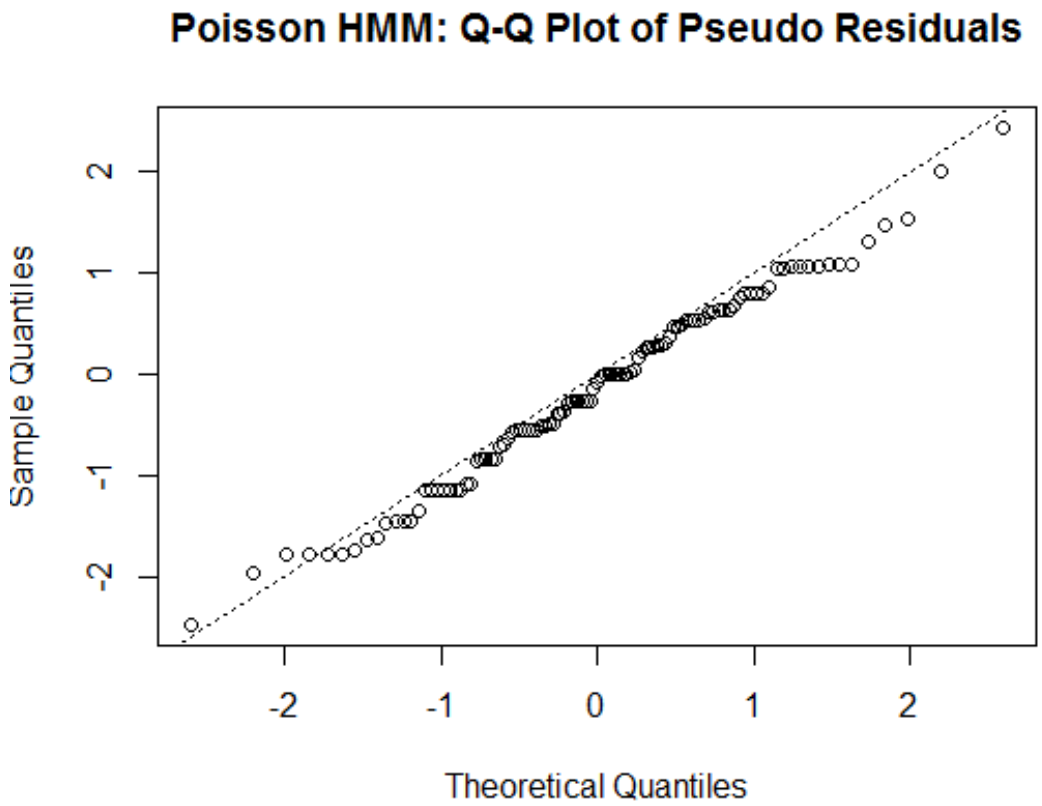
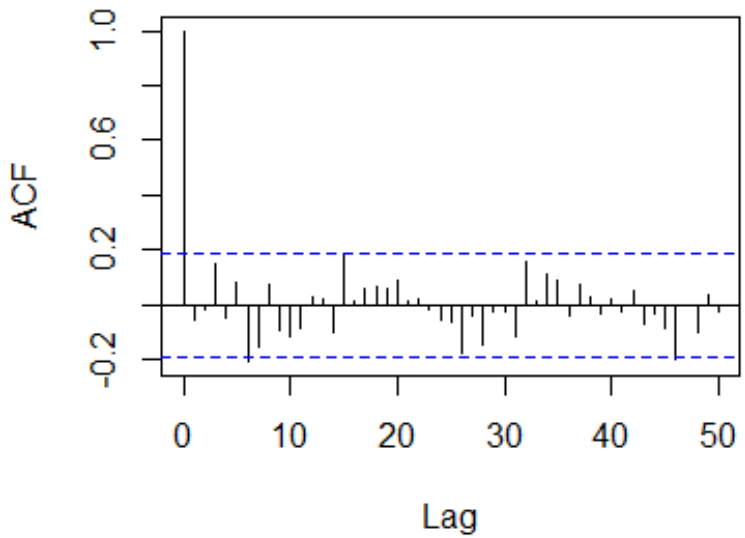


Figure: 模性拟合优度检验

残差检验



Figure： 自相关性检验

Outline

数据

数据：ALV,DBK,DCX,SIE股票日收益率
从2003年3月4日至2005年2月17日
数据来源： http://finance.yahoo.com/q?s=^gdaxi

R包： depmixS4

特点:功能更为强大;不够直观
观测序列可以为多变量(例2)；
边际分布可以为一些不常见的混合分布,如广义t分布,指数高斯分布等；
状态转移矩阵可以随时间变化(例3)；

数据

数据： ALV, DBK, DCX, SIE股票日收益率
从2003年3月4日至2005年2月17日
数据来源： http://finance.yahoo.com/q?s=^gdaxi

R包： depmixS4

特点:功能更为强大;不够直观
观测序列可以为多变量(例2)；
边际分布可以为一些不常见的混合分布, 如广义t分布, 指数高斯分布等；
状态转移矩阵可以随时间变化(例3)；

主要的函数

- depmix():设定初始模型
- fit():估计参数

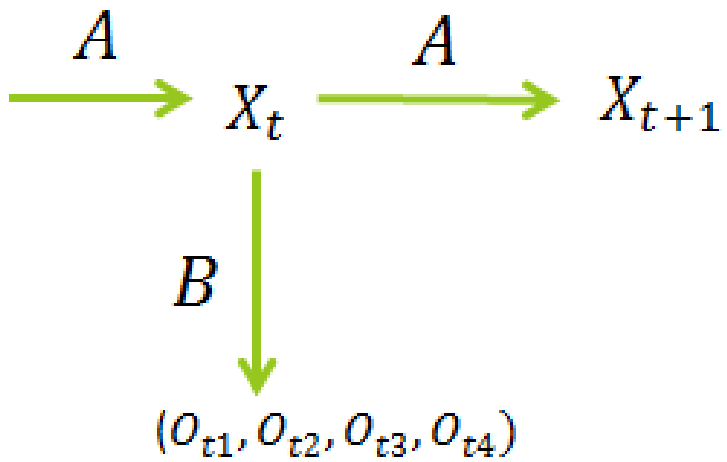
- transInit:设置转移概率矩阵及初始状态分布
- response:设置边际分布
- makeDepmix():设定初始模型
- fit():估计参数

主要的函数

- depmix():设定初始模型
- fit():估计参数

- transInit:设置转移概率矩阵及初始状态分布
- response:设置边际分布
- makeDepmix():设定初始模型
- fit():估计参数

模型



Figure： 多变量隐马氏图示

给定状态下，观测值服从多元正太分布

	ALV	DBK	DCX	SIE
状态1下均值	0.05	0.03	0.01	0.02
状态2下均值	0.12	0.25	0.16	0.22

Table: 均值比较

	ALV	DBK	DCX	SIE
状态1下方差	1.77	1.38	1.28	1.45
状态2下方差	9.43	5.62	5.04	4.53

Table: 方差比较

协方差

1.0	0.74	0.73	0.77
0.74	1.0	0.71	0.72
0.73	0.71	1.0	0.74
0.77	0.72	0.74	1.0

Table: 状态1下协方差在0.71-0.77之间

1.0	0.61	0.67	0.69
0.61	1.0	0.62	0.68
0.67	0.62	1.0	0.70
0.69	0.68	0.70	1.0

Table: 状态2下协方差在0.61-0.70之间

Outline

数据

实验数据

每次判定一串字母是否为英文单词;
每次要求不一，可能要求以最短的时间，也可能要求以最高的准确度
数据变量：rt(反应时间)；corr(0-1,答对与否)；Pacc(时间与准确度的权衡)

目的

当对时间的要求越来越高时，是否有一个临界点，使得准确度从思考水平到猜谜水平

R包：depmixS4

运用协变量

数据

实验数据

每次判定一串字母是否为英文单词;
每次要求不一，可能要求以最短的时间，也可能要求以最高的准确度
数据变量：rt(反应时间)；corr(0-1,答对与否)；Pacc(时间与准确度的权衡)

目的

当对时间的要求越来越高时，是否有一个临界点，使得准确度从思考水平到猜谜水平

R包：depmixS4

运用协变量

数据

实验数据

每次判定一串字母是否为英文单词;
每次要求不一，可能要求以最短的时间，也可能要求以最高的准确度
数据变量：rt(反应时间)；corr(0-1,答对与否)；Pacc(时间与准确度的权衡)

目的

当对时间的要求越来越高时，是否有一个临界点，使得准确度从思考水平到猜谜水平

R包：depmixS4

运用协变量

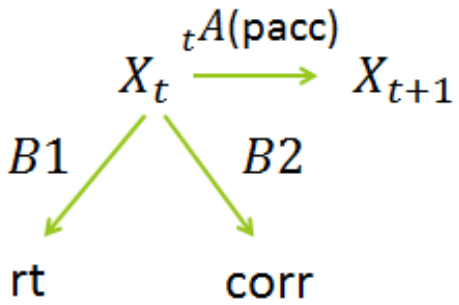


Figure: 图示：2观测变量条件独立,转移矩阵与变量pacc有关

Fact

$$
{}_t\mathrm{A} = \begin{bmatrix} {}_t\mathrm{a}_{11} & {}_t\mathrm{a}_{12} \\ {}_t\mathrm{a}_{21} & {}_t\mathrm{a}_{22} \end{bmatrix} = \begin{bmatrix} \frac{1}{1+\exp(\mathrm{a1}+\mathrm{b1}*\mathrm{y_t})} & \frac{\exp(\mathrm{a1}+\mathrm{b1}*\mathrm{y_t})}{1+\exp(\mathrm{a1}+\mathrm{b1}*\mathrm{y_t})} \\ \frac{1}{1+\exp(\mathrm{a2}+\mathrm{b2}*\mathrm{y_t})} & \frac{\exp(\mathrm{a2}+\mathrm{b2}*\mathrm{y_t})}{1+\exp(\mathrm{a2}+\mathrm{b2}*\mathrm{y_t})} \end{bmatrix}
$$

Fact

给定t时刻状态；rt服从正太分布，corr服从0-1分布；rt与corr独立

	rt均值	rt标准差	答错概率	答对概率
状态1	5.51	0.19	0.47	0.53
状态2	6.39	0.23	0.09	0.91

Table: 结果展示

思考状态的时间区间(2倍标准差)大概为[5.93, 6.85]
猜谜状态的时间区间(2倍标准差)大概为[5.13, 5.89]

总结

- 隐马氏模型的模型参数：$\tau = \{\pi, \mathrm{A}, \mathrm{B}\}$;
- forward, BaumWelch算法
- 简单隐马氏模型可以用HiddenMarkov包；复杂的用depmixS4

参考资料：

- A Revealing introduction to Hidden Markov Models .Mark Stamp ,2003
- Hidden Markov Models for Time Series An Introduction Using R ,Walter Zucchini &Iain L. MacDonald ,2009.
- depmixS4_A R package for hidden markov model ,Ingmar Visser,2010
- Hidden Markov Models application to Financial Economics.
- Hidden Markov models for bioinformatics.
- Image segmentation and compression using hidden Markov models.