Finance 金融, 2012, 2, 45-49
http://dx.doi.org/10.12677/fin.2012.21005 Published Online January 2012 (http://www.hanspub.org/journal/fin)

A Hidden Markov Chain Modeling of Shanghai Stock Index

Jian Gong, Chenghu Ma

School of Management, Fudan University, Shanghai
Email: {machenghu, 082025051}@fudan.edu.cn

Received: Nov. 9th, 2011; revised: Nov. 24th, 2011; accepted: Dec. 5th, 2011

Abstract: This paper develops a model of financial forecasting using hidden Markov chains. Empirical analysis was carried out with respect to Shanghai Stock Exchange Index (SSEI) using daily data for the period from 2002.12.17-2011.3.18 (2000 samples). A three dimensional hidden Markov chain is identified to fit the data in the sampling period. An altered 10-day weighted average method was proposed, and was found to be useful for out-of-sample forecasting.

Keywords: Hidden Markov Model; Index Forecast Model; Shanghai Stock Index

基于隐马尔可夫链的上证股指建模

龚　健，马成虎

复旦大学管理学院，上海
Email: {machenghu, 082025051}@fudan.edu.cn

收稿日期：2011 年 11 月 9 日；修回日期：2011 年 11 月 24 日；录用日期：2011 年 12 月 5 日

摘　要：本文引入了连续观测概率分布下的一类隐马尔可夫链模型。通过探讨该类模型的时间序列特征以及相应的模式辨识理论，提出一类新的指数预测方法。作为实证应用，我们选择上证指数 2002 年 12 月 17 日至 2011 年 3 月 18 日共 2000 个交易日的数据样本进行模型拟合和预测。结论显示，三维隐马尔可夫链模型能较好的拟合采样区间内的上证指数，而调整 10 日加权的预测方法则给出了关于样本外股指价格较为精确的预测。

关键词：隐马尔可夫链；指数预测模型；上证指数

1. 引言

证券市场的预测方法一直是技术分析与量化投资的核心。建立在经典无套利理论和一般均衡定价理论基础上的资产定价模型大多依赖先验给定的状态变量，这些变量称作定价因子。其中部分因子是可观测的经济变量，如：历史价格、交易量、利率、红利等，而另一部分定价因子却是无法直接观测到的，如：市场波动率、股民偏好等，甚至是未知变量——尽管该类因子影响价格，但建模者并没有意识到其重要性。后两类变量统称为“隐”变量。隐变量的存在，无疑对股市建模，特别是股市预测，增加了技术上的难度。一方面，需要估计状态变量的维数；另外还要对状态结构参数进行模式辨识。

隐马尔可夫链模型(Hidden Markov Model，HMM)最初是由 Baum and Egon (1967)[1]提出的参数估计和模式辨识技术。其建模过程包含一个双重随机过程的设定：一重是用隐马尔可夫链来描述的随机状态转移过程；另一重是与模型中状态相关的随机观察值输出概率分布，以描述状态和观测变量之间的统计对应关系。HMM

建模和预测的机制是：通过对观测数据的收集与统计分析对不完全可观测的(隐)状态变量的维数和结构特征进行估计和辨识，进而对将来状态和输出变量进行预测。

HMM 模型在工程领域最为人知的应用是语音识别(Jelinek *et al*., 1974; Rabiner 1989)[2,3]和生物序列分析(Durbin *et al*., 1999)[4]。其实，HMM 方法在航天、气象预报、地下勘测、医学 CT 成像诸多领域都有很多成功的应用，文献之多此处不一一列举。相比之下，马尔可夫模型在金融建模中的应用大多基于可观测的(显)马尔可夫机制转换模型(Markov Switching Model)。比如：在 Schaller and Van Norden (1997)[5]对金融市场收益的波动性分析，Bhar and Hamori (2004)[6]关于证券市场收益的长期和短期效应，以及 Gray (1996)[7]的利率期限结构分析中，都回避了关于隐状态变量的结构辨识问题。

本文提出的应用 HMM 对股市动态建模方法是该模型在金融建模中的一种新尝试。利用上证指数在 2002 年 12 月 17 日至 2011 年 3 月 18 日期间共 2000 个交易日数据，对上证指数的价格动态变化特征进行了建模。一方面，我们假设影响股市动态的是一组服从连续时间分布的隐马尔可夫链状态过程；另外，还有一组变量则代表适应于状态向量、可观测的经济指标，特别包括我们最为关心的指数价格过程。利用样本区间的数据，我们对隐状态变量的维数和转移矩阵进行了辨识，并在此基础上提出了一类指数预测的方法，并验证了该类方法的有效性。

文章由 5 部分组成：第 2 节将系统介绍 HMM 模型设定的理论与方法；第 3 节将介绍基于 HMM 的不同预测方法；第 4 节则针对上证指数进行 HMM 实证建模，并对上证指数进行样本外的预测检验；第 5 节概括了文章的主要结论和局限性。

2. HMM 模型设定

本节内容由两部分组成：2.1 节详细介绍 HMM 参数模型的设置；2.2 节是关于 HMM 模型的状态个数和观测值概率分布的参数估计方法，利用上证指数在 2002 年 12 月 17 日至 2011 年 3 月 18 日期间(共 2000 个样本)的日交数据对基于上证指数的 HMM 模型进行了似然估计。

2.1. 基于连续观测概率分布的 HMM 模型

HMM 模型可用五元组 $\lambda=(N,M,\pi,A,B)$ 表示，具体描述如下：

1) N 表示隐马尔可夫链所包含的状态的数目，q_t 表示 t 时刻的隐状态，

$$q_t=S_i,\ 1\le i\le N\ .$$

2) M 表示构成随机观察值输出概率函数的混合成分的数目；

3) π 表示隐马尔可夫链初始状态的概率分布，

$$\pi=\{\pi_i\};\ \pi_i=P\{q_0=S_i\},\ 1\le i\le N\ .$$

4) A 表示隐马尔可夫链的状态转移概率分布，

$$A=\{a_{ij}\};\ a_{ij}=P\{q_{t+1}=S_j\,|\,q_t=S_i\},\ 1\le i,j\le N\ .$$

5) $\boldsymbol{B}$ 表示随机观察值输出概率矩阵，$b_j(O_t)$ 表示给定隐状态 $S_t=j$，产生随机观察值 O_t 的概率，记作

$$\boldsymbol{B}=\{b_j(O_t)\};\ b_j(O_t)=P\{O_t\,|\,q_t=S_j\},\ 1\le j\le N\ .$$

在显性可观测过程为连续值的情况下，从给定隐状态 $S_t=j$ 产生随机观测值 O_t 的概率，$b_j(O_t)$，应表示为连续概率密度函数或者混合连续概率密度函数，即“发射函数”。

如果用高斯分布去近似地表达这种关系，则发射函数可以表示成(k 个混合)高斯分布的组合。假设时间点 t 时，隐状态为 q_t，观测值（显性链状态）为 O_t，用(k 个混合)高斯分布 $M(t)$ 作为发射函数。即发射函数 $b_j(O_t)$ 的形式是连续概率密度函数或者数个连续概率密度函数的混合：

$$b_j(O_t)=\sum_{k=1}^{M}w_{jk}b_{jk}(O_t),\ j=1,\cdots,N$$

M 是混合成分的个数，w_{jk} 表示隐状态为 $q_t=j$ 时第 k 个混合成分的权重，且有：

$$\sum_{k=1}^{M}w_{jk}=1,\ j=1,\cdots,N;\ k=1,\cdots,M$$

如果混合成分 $b_{jk}(O_t)$ 为 D 维高斯分布（即显性观测序列为 D 维序列），具有均值则 μ_{jk} 和协方差阵 $\sum_{jk}$，则：

$$b_{jk}(O_t)=N(O_t,\mu_{jk},\sum_{jk})$$

当隐状态 $q_t=j$ 时，由第 k 个高斯分布 $(M(t)=k)$ 所产生的显状态的统计特征为：

$$\mu_{jk} = E\left[O_t \mid q_t = j, M(t) = k\right]$$

$$\sum_{jk} = Cov\left[O_t \mid q_t = j, M(t) = k\right]$$

此时，多个混合高斯分布的概率密度函数为：

$$b_{jk}(O_t) = N\left(O_t, \mu_{jk}, \Sigma_{jk}\right) = \frac{1}{(2\pi)^{D/2}\left|\Sigma_{jk}\right|^{1/2}} \cdot \exp\left\{-\frac{1}{2}\left(O_t - \mu_{jk}\right)^T \left(\Sigma_{jk}\right)^{-1}\left(O_t - \mu_{jk}\right)\right\}$$

2.2. 模型设定标准及拟合结果

在应用模型对证券价格指数进行预测之前，需要解决的一个首要问题就是确定 HMM 模型中隐状态的个数以及混合高斯分布的数量。有很多学者对其进行研究，如 McLachlan and Peel (2000)[8]，我们选择的检验标准如下：

2.2.1. BIC 准则

BIC 准则(Bayesian information criterion)是惩罚性似然值检验标准的一种。通过引入一项对模型参数数量的惩罚项，解决由增加参数数量所导致的模型过度适合问题(Posada & Buckley, 2004)[9]。

$$-2In\,p\left(O\middle|k\right) \approx BIC = -2 \cdot InL + k \cdot In(T)$$

其中，O 为观测序列，k 为参数个数，$p\left(O\middle|k\right)$ 为在给定模型下获得观测序列的概率，L 为似然值函数的最大值，T 为样本量。

2.2.2. 交叉检验似然值准则

根据 Celeux and Durand (2008)[10]，如果原序列服从齐次 HMM 模型，可将原始数据集分成奇数集和偶数集，这些数据集仍然服从 HMM 模型。由此产生的检验方法被称作 OEHS (Odd-Even Half-Sampling)标准。

在 OEHS 方法下，将原始数据分成奇数组及偶数组，对每组序列分别产生 10 组随机的初始分布和转移矩阵及混合高斯分布(作为观测概率函数)，使用 EM 算法训练模型参数，当迭代次数达到 50 次或者对数似然值增幅小于 0.1 时，停止计算。

使用 2002.12.17~2011.3.18 (2000 个样本)[1] 的上证指数历史收益率数据为样本，分别选取隐状态为二状态、三状态、四状态、五状态进行模型检验，检验结果汇总于表 1。

表 1 可见，基于 BIC 标准及交叉检验似然值的 OEHS 标准，最稳定的 HMM 模型是三维状态 HMM 模型。

3. 基于 HMM 的预测模型

3.1. 预测机制

以 2010.12.14 的测试样本为例。利用历史数据样本得出训练模型 $\lambda = (N, M, \pi, A, B)$ 后，将 2010.12.14 的观测值计算得到似然值–19.8913，通过在历史数据中寻找，我们锁定最接近的一个似然值为 2010.11.2，似然值为–19.8720。

图 1 分别显示 2010.12.14 与 2010.11.2 两个交易日的当日观测向量：从图 1 中可以清楚的看到，2010.12.14 与 2010.11.2 的行为模式具有很大的相似性。由此我们预测，2010.12.14 的下一个交易日的收盘价变化会与 2010.11.2 的下一个交易日的收盘价变化有类似的模式。

Table 1. State number analysis of HMM
表 1. HMM 模型的状态参数分析

Model	P	–logL	$logL_1$	$logL_2$	BIC	OEHS[2]
2-state HMM	8	–3784	–2560.6	–2595.2	7594	0.67%
3-state HMM	15	–3769	–2664.2	–2654.6	7587	0.18%
4-state HMM	24	–3756	–2953.6	–3060.6	7592	1.78%
5-state HMM	35	–3745	–3594.1	–3289.9	7606	4.42%

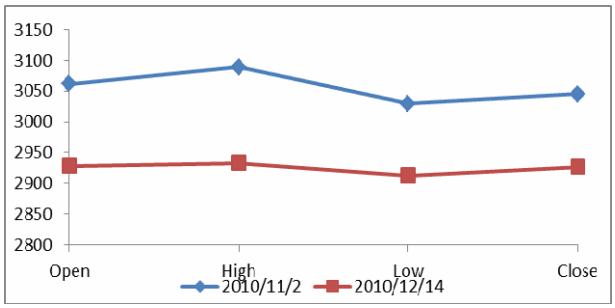

Figure 1. Pattern matching of observations
图 1. 观测值的模式匹配

[1] 数据来源：wind 数据；本文的实证数据均采用此数据区间。
[2] 此处变量的计算方式为：$\text{abs}(\text{logL}_1 - \text{logL}_2)/(\text{logL}_1 + \text{logL}_2) \times 100\%$ 。

3.2. 预测方法

这里提出的基于隐马尔可夫链模型的预测方法包括如下几个步骤：

1) 寻找观测时间序列$\left(O_t=\left[O_{1t},O_{2t},\cdots\right]'\right)$，如开盘价，收盘价等，认为这些时间序列间的变化符合某些模式；

2) 利用所选择时间序列的数据样本训练 HMM 模型参数，得到 HMM 模型：$\lambda=\left(N,M,\pi,A,B\right)$；

3) 根据模型$\lambda=\left(N,M,\pi,A,B\right)$计算目标序列$\left(T_t=\left[T_{1t},T_{2t},\cdots\right]'\right)$出现的似然值：$LL\left(T\right)$；

4) 在历史序列中寻找与$LL\left(T\right)$最接近 d 个似然值$\left\{LL\left(O_{t1}\right),LL\left(O_{t2}\right),\cdots,LL\left(O_{td}\right)\right\}$，则$\sum_{i=1}^{d}w_i\left(O_{t+1}-O_t\right)$构成对 T 期到 $T+1$ 期变化量的预测值。

根据 d 值和预测权重 w 的选取可以产生不同的预测方法。

3.2.1. 单日预测方法

单日预测$\left(d=1\right)$假定相似的走势会使得下一交易日的走势也相似。通过找到历史数据中与目标走势相近似的交易日，下一个交易日的收盘价变化就是对目标走势下一个交易日的预测。即：

$$T_{0+1}-T_0=O_{t+1}-O_t$$

其中，T_0是目标走势，T_{0+1}是其下一交易日的预测值；$O_{t+1}-O_t$是根据 HMM 模型从历史数据中得到的与T_0走势最为接近的交易日走势。

3.2.2. d 日加权预测法

考虑到单日交易的波动性过大，可以选取 d 个走势相近的历史交易日，并对其下一日的涨跌幅进行加权，以此构成对目标走势的预测：

$$T_{0+1}-T_0=\sum_{i=1}^{d}w_i\left(O_{i+1}-O_i\right)$$

其中，$O_i\left(1\le i\le d\right)$表示历史数据中与$T_0$走势第 i 接近的交易日，w_i则表示$O_{i+1}-O_i$在预测中所占的权重。这里，我们根据样本似然值与目标似然值的差异大小选取权重：

$$w_i=\frac{1/\left(LL_i-LL_0\right)}{\sum_{i=1}^{d}1/\left(LL_i-LL_0\right)},\ \sum_{i=1}^{d}w_i=1$$

这里，LL_i-LL_0代表似然值的差异，w_i表示O_i与T_0的走势差异占所有 d 个参照样本的比重。可见，似然值差异越小则所占预测部分的权重越大。

3.2.3. 调整 d 日加权预测法

考虑时间因素对权重选取的影响。理论上看，近期的走势应对预测结果产生更大的影响。如果加入一项时间权重参数，可能使模型的预测能力更为完善。根据马成虎(2010)[11]，可考虑采用如下权重取法：

$$w_i=\frac{1/\left[\left(LL_i-LL_0\right)e\left(t-t_i\right)\right]}{\sum_{i=1}^{d}1/\left[\left(LL_i-LL_0\right)e\left(t-t_i\right)\right]},\ \sum_{i=1}^{d}w_i=1$$

其中，$e\left(t-t_i\right)$[3]为指数分布函数。参考日离目标日的距离越近，其所占预测权重则越大。

4. 模型实证结果

模型选取的观测变量序列包括：当日开盘价O_{1t}，最高价O_{2t}，最低价O_{3t}，收盘价O_{4t}，组成显性观测序列O_t。即：

$$O_t=\begin{Bmatrix}\text{开盘价}\\ \text{最高价}\\ \text{最低价}\\ \text{收盘价}\end{Bmatrix}=\begin{Bmatrix}O_{1t}\\ O_{2t}\\ O_{3t}\\ O_{4t}\end{Bmatrix}$$

为保持模型参数的一致性，使用上一节的数据集作为训练样本(2000 个交易日)，使用 2010.12.13 至 2011.3.14 的数据集(60 个交易日)作为测试样本。

为使实证结果尽可能少的对初始状态的依赖，首先产生 10 组随机的初始分布和转移矩阵及三维混合高斯分布，使用 EM 算法训练模型参数，选取 50 次迭代之后对数似然值最大的模型参数作为初始值，进行第二阶段的模型估计；当第二阶段的迭代次数达到 50 次或者对数似然值增幅小于 0.1 时，停止计算。在 d 日加权预测方法中，分别取 d 为 5 日，10 日，20 日，30 日，并对比调整 d 日加权预测法。所得预测结果分别汇总于如表 2、表 3 中。

图 2 显示，通过对上证指数进行 HMM 建模后，使用模式识别的方法能有效地预测该指数的走势。相

[3]此处，$e\left(t-t_i\right)$中的时间为连续时间。在实证分析中，为避免极大极小值得出现，我们选择单位化的时间，将时间间隔定义为占样本量的比：$t=T/n,\ t_i=T_i/n$。

Table 2. Comparative analysis of forecast model
表 2. 预测模型对比分析

	单日预测	5 日预测	10 日预测	20 日预测	30 日预测
平均错误预测	45.7	38.7	32.5	30.2	29.3
MAPE	0.0163	0.0142	0.0136	0.0135	0.0135
R^2	0.6415	0.7596	0.7798	0.8006	0.8184

Table 3. Comparative analysis of forecast model
表 3. 预测模型对比分析

	调整 10 预测	调整 20 预测	调整 30 预测
平均错误预测	28.78	28.54	28.01
MAPE	0.0104	0.0103	0.0101
R^2	0.8055	0.8238	0.8416

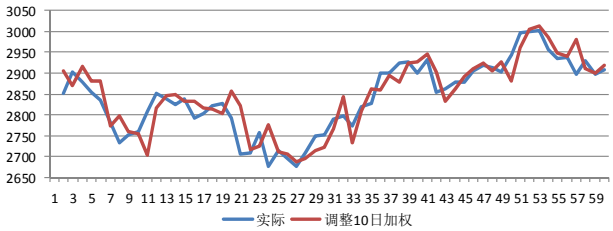

Figure 2. Altered 10 day weighted forecast result
图 2. 调整 10 日加权平均预测结果

对单日预测而言，加权平均能够进行更为准确和平滑的预测。对比 *d* 日加权平均及调整 *d* 日加权平均两种方法，后者由于考虑时间因素，其预测显得更为有利。在参数 *d* 的取值上，本文认为，调整 10 日加权平均预测方法已经能够较好的给出预测结果。

5. 结论

我们发现 HMM 为多变量金融时间序列的建模和预测提供了一种强有力的概率分析框架。在 HMM 的基础上，选择适当的观测向量、隐状态数量、发射概率密度函数，及对目标时间序列变动模式的识别，通过在历史序列中寻找相匹配的模式序列，能够产生具有概率统计支持的预测结果。本文通过对中国上证指数进行 HMM 建模，引入模式识别的预测机制，提出了 *d* 日加权平均及调整 *d* 日加权平均两种方法并进行实证检验，使得结论具有很强的说服力。

6. 致谢

本文得到了《国家自然科学基金》面上项目(#70871100)的赞助，在此表示感谢。

参考文献 (References)

[1] L. E. Baum, J. A. Egon. An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology. Bulletin of the American Mathematical Society, 1967, 73(3): 360-363.
[2] F. Jelinek, L. Bahl and R. Mercer. Optimal decoding of linear codes for minimizing symbol error rate. IEEE Transactions on Information Theory, 1974, 20(2): 284-287.
[3] L. R. Rabiner. A tutorial on Hidden Markov Models and selected applications in speech recognition. Proceedings of the IEEE, 1989, 77(2): 257-286.
[4] R. Durbin, S. R. Eddy, A. Krogh and G. Mitchison. Biological sequence analysis: Probabilistic models of proteins and nucleic acids. Cambridge: Cambridge University Press, 1999.
[5] H. Schaller, S. Van Norden. Regime switching in stock market returns. Applied Financial Economics, 1997, 7(2): 177-191.
[6] R. Bhar, S. Hamori. Empirical characteristics of the permanent and transitory component of stock return: Analysis in a Markov switching heteroskedasticity framework. Economics Letters, 2004, 82(2): 157-165.
[7] S. Gray. Modeling the conditional distribution of interest rates as a regime switching process. Journal of Financial Economics, 1996, 42(1): 27-62.
[8] G. J. McLachlan, D. Peel. Finite mixture models. Wiley Series in Probability and Statistics, 2000: 175-195.
[9] D. Posada, T. R. Buckley. Model selection and model averaging in phylogenetics: Advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. Systematic Biology, 2004, 53(5): 793-808.
[10] G. Celeux, J. B. Durand. Selecting hidden Markov model state number with cross-validated likelihood. Computational Statistics, 2008, 23(4): 541-564.
[11] 马成虎. 高级资产定价理论[M]. 北京: 中国人民大学出版社，2010: 286-287.