第 25 卷第 2 期
2008 年 4 月

控 制 理 论 与 应 用
Control Theory & Applications

Vol. 25 No. 2
Apr. 2008

文章编号: 1000−8152(2008)02−0257−04

基于条件互信息的多维时间序列图模型

高 伟[1], 田 铮[1,2]

(1. 西北工业大学 应用数学系, 陕西 西安 710072;
2. 中国科学院 自动化研究所 模式识别国家重点实验室, 北京 100080)

摘要: 在多维时间序列的图模型中引入信息论方法, 提出了多维时间序列中各分量之间直接线性联系存在性的互信息检验. 定义了线性条件互信息图, 图中的结点表示多维时间序列的分量, 结点间的边表示各分量之间存在的直接线性相依关系. 提出了分量之间条件线性联系存在性的信息论检验方法. 图中边的存在性用基于线性条件互信息的统计量检验, 统计量的显著性用置换检验决定. 应用到实例中的结果表明本文的方法能迅速准确的捕捉各分量之间的直接线性联系.

关键词: 多维时间序列; 图模型; 互信息; 线性条件互信息图
中图分类号: O211.6　　**文献标识码**: A

Graphical models for multivariate time series based on conditional mutual information

GAO Wei[1], TIAN Zheng[1,2]

(1. Department of Applied Mathematics, Northwest Polytechnical University, Xi'an Shaanxi 710072, China;
2. National Key Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Science, Beijing 100080, China)

Abstract: The information theory is introduced for graphical models of multivariate time series. A method for testing direct linearity between two components is proposed. A class of graphical models, called linear conditional mutual information graph, is defined. The vertex set denotes the components of the series and the edges denote the direct linear dependence structure of the components. The presence of the edges is tested by a statistics based on linear conditional mutual information. The permutation procedure is used to determine the significance of the test statistics. Finally, the method is applied to a real series, and the results show that the method can efficiently capture the direct linear dependence between the components.

Key words: multivariate time series; graphical model; mutual information; linear conditional mutual information graph

1 引言(Introduction)

研究分量间存在的线性和非线性相依关系以得到系统内部的信息已成为多维时间序列在工程应用中的一个重要方面. 作为分析多元数据的一个重要工具, 图模型在研究变量间的相依关系上得到了广泛的应用. Dahlhaus[1]引入了图模型刻画多维时间序列分量间的相依结构, 建立了条件偏相关图. 序列的每一个分量用图中的一个顶点表示, 条件偏相关图的边反映了几个随机变量间的条件相依结构, 有助于区别变量间的直接相关和间接相关. Dahlhaus[1]用给定其他分量条件下两个分量间的偏谱耦合来辨识图中的边, 是一种频域的研究方法. 需要用非参数方法估计谱矩阵及其逆, 检验统计量的显著性水平基于其近似分布来确定.

信息论中的熵和互信息概念广泛用于检验时间序列的非线性联系[2∼5]. 本文用线性的信息论统计量检验时间序列各分量间的直接线性联系, 用时域的方法研究多维时间序列的相依结构, 定义了线性条件互信息图. 用线性条件互信息度量给定所有其他分量的条件下, 两个分量间的条件线性联系的存在性. 用置换检验确定检验统计量的显著性水平. 与传统的多维时间序列分析方法, 如互相关矩阵方法相比, 图模型方法更加直观的揭示分量间的相依结构, 可以区分直接联系和间接联系. 本文的方法与Dahlhaus[1]的方法相比, 主要有两个优点: 其一, 基于条件互信息的方法可用于各种残差分布情形, 如正态分布和指数分布等; 其二, 用置换检验确定检验统计量的显著性, 避免了用非参数方法估计

收稿日期: 2005−12−30; 收修改稿日期: 2007−04−11.
基金项目: 国家自然科学基金资助项目(60375003); 国家航空基础项目资助(03153059).

谱矩阵和求逆矩阵运算. 线性条件互信息的非线性对应—-广义条件互信息, 可直接用于条件独立性的检验. 最后用美国政府债券数据作为实例验证了本文提出的方法确实可行且有效.

2 多维时间序列的线性条件互信息图(Linear conditional mutual information graph for multivariate time series)

本节引入多维时间序列的线性条件互信息图, 其中涉及信息论中关于熵和互信息的一些基本概念, 见参考文献[4].

设$X(t)=(X_1(t),\cdots,X_k(t))^{\mathrm{T}},t\in\mathbb{Z}$为一多维严平稳时间序列. 图中的顶点表示各分量序列, 即$V=\{1,\cdots,k\}$. 边(a,b)不存在的条件为: 给定序列的其他分量$Y_{ab}(t)=(X_j(t),j\neq a,b)$的条件下, $X_a(\cdot)$和$X_b(\cdot)$是条件独立的. 令$x_a=(X_a(t);t\in\mathbb{Z})$, $y_{ab}=(Y_{ab}(t);t\in\mathbb{Z})$, 即

$$(a,b)\notin E\Leftrightarrow x_a\perp x_b|y_{ab}. \tag{1}$$

考虑到本文的主要目的是研究各分量序列之间的联系结构, 因此在建立图模型时只需要考虑两个分量之间联系的存在性. 以检验$X_a(\cdot)$和$X_b(\cdot)$之间的条件独立性为例, 研究$X_a(t)$和$X_b(t-m)$之间及$X_b(t)$和$X_a(t-m),m\in\mathbb{Z}^+$之间的条件独立性, 理论上需要去掉过去和将来所有其他随机变量的影响, 实际应用中只需要去掉其他序列从时刻$t-m$到时刻t的随机变量的影响.

设$Y_{ab}^{m+1}(t)=(Y_{ab}(t),\cdots,Y_{ab}(t-m))$, 给定所有变量$Y_{ab}^{m+1}(t)$的条件下, $X_a(t)$和$X_b(t-m)$之间的条件互信息为

$$\begin{aligned}&I(X_a(t),X_b(t-m)|Y_{ab}^{m+1}(t))=\\&-H(Y_{ab}^{m+1}(t),X_a(t),X_b(t-m))+\\&H(X_a(t),Y_{ab}^{m+1}(t))+H(X_b(t-m),\\&Y_{ab}^{m+1}(t))-H(Y_{ab}^{m+1}(t)).\end{aligned} \tag{2}$$

记

$$\begin{aligned}&Z(t)=(Y_{ab}^{m+1}(t)),\ Z_a(t)=(X_a(t),Y_{ab}^{m+1}(t)),\\&Z_b(t)=(X_b(t-m),Y_{ab}^{m+1}(t)),\\&Z_{ab}(t)=(X_a(t),X_b(t-m),Y_{ab}^{m+1}(t)),\end{aligned}$$

式(2)简化为

$$\begin{aligned}&I(X_a(t),X_b(t-m)|Z(t))=\\&-H(Z_{ab}(t))+H(Z_a(t))+H(X_b(t))-H(Z(t)).\end{aligned} \tag{3}$$

对于n维随机向量$(X_1,\cdots,X_n)$, 设均值为0, 协方差矩阵为Σ. 定义线性熵$H^l(X_1,\cdots,X_n)$为

$$H^l(X_1,\cdots,X_n)=\frac{n}{2}\ln(2\pi)+\frac{1}{2}\ln|\Sigma|. \tag{4}$$

当随机变量$X_1,\cdots,X_n$的联合分布是n维Gauss分布时, 线性熵$H^l(X_1,\cdots,X_n)$和熵$H(X_1,\cdots,X_n)$在理论上是等价的. 线性熵、线性互信息和线性条件互信息只对线性结构敏感, 可以用来度量直接和间接线性相关联系.

用Σ,Σ_a,Σ_b,Σ_{ab}分别表示前面所构造的随机向量$Z(t)$,$Z_a(t)$,$Z_b(t)$,$Z_{ab}(t)$的协方差矩阵. 则根据式(3)和(4), 给定所有变量$Y_{ab}^{m+1}(t)$的条件下, 随机变量$X_a(t)$和$X_b(t-m)$之间的线性条件互信息为

$$\begin{aligned}&I_l(X_a(t),X_b(t-m)|Y_{ab}^{m+1}(t))=\\&\frac{1}{2}\ln(\frac{|\Sigma_a||\Sigma_b|}{|\Sigma_{ab}||\Sigma|}).\end{aligned} \tag{5}$$

由条件互信息的相关性质可直接得到线性条件互信息的如下结果:

定理 1 给定所有变量$Y_{ab}^{m+1}(t)$的条件下, $X_a(t)$和$X_b(t-m)$线性无关等价于

$$I_l(X_a(t),X_b(t-m)|Y_{ab}^{m+1}(t))=0.$$

定义 1 令$X(t)=(X_1(t),\cdots,X_k(t))^{\mathrm{T}},t\in\mathbb{Z}$为一多维严平稳时间序列. 图$G=(V,E)$, 对应的顶点集为$V=\{1,\cdots,k\}$. 令$(a,b)\notin E$当且仅当$I_l(X_a(t),X_b(t-m)|y_{ab})=0$, 且$I_l(X_b(t),X_a(t-m)|y_{ab})=0,m\in\mathbb{Z}^+$成立, 则称$G=(V,E)$为多维时间序列的线性条件互信息图, 简记为LCMIG (linear conditional mutual information graph).

注意到线性条件互信息是条件相依的一个无界度量. 这里用线性条件互信息的一个变换作为检验统计量:

$$\begin{aligned}&\delta^l_{X_aX_b|Y_{ab}}(m)=\\&1-\exp(-I_l(X_a(t),X_b(t-m)|Y_{ab}^{m+1}(t))=\\&1-\sqrt{\frac{|\Sigma_{ab}||\Sigma|}{|\Sigma_a||\Sigma_b|}}.\end{aligned} \tag{6}$$

容易推出$\delta^l_{X_aX_b|Y_{ab}}(\cdot)\in[0,1)$.

定理 2 设$G=(V,E)$是多维时间序列的线性条件互信息图, 则成立

$$(a,b)\notin E\Leftrightarrow\delta^l_{X_aX_b|Y_{ab}}(\cdot)\equiv0,\ \delta^l_{X_bX_a|Y_{ab}}(\cdot)\equiv0.$$

3 多维时间序列线性条件互信息图的建立(The construction of linear conditional mutual information graph for multivariate time series)

设得到一个样本量为n的k维时间序列, 建立与之对应的图模型, 根据定理2, 两个顶点间没有边相连等价于顶点表示的两个分量时间序列不同时刻的随机变量之间不存在线性相依联系. 因此, 通过检验统计量$\delta^l_{X_aX_b|Y_{ab}}(m)$和$\delta^l_{X_bX_a|Y_{ab}}(m),m\in\mathbb{Z}^+$是否为0来建立多维时间序列的线性条件互信息图模型.

检验统计量$\delta^l_{X_aX_b|Y_{ab}}(m)$的估计可由相应随

机向量$Z(t)$,$Z_a(t)$,$Z_b(t)$,$Z_{ab}(t)$的协方差矩阵Σ,Σ_a, Σ_b, Σ_{ab}的估计$\hat{\Sigma}$,$\hat{\Sigma}_a$,$\hat{\Sigma}_b$,$\hat{\Sigma}_{ab}$计算得到

$$\hat{\delta}^l_{X_aX_b|Y_{ab}}(m)=1-\sqrt{\frac{|\hat{\Sigma}_{ab}||\hat{\Sigma}|}{|\hat{\Sigma}_a||\hat{\Sigma}_b|}}. \tag{7}$$

建立原假设:

H_0: 给定序列中所有其他变量$Y_{ab}(t)$的条件下, 随机变量序列$X_a(\cdot)$和$X_b(\cdot)$是线性无关的.

在原假设成立的条件下, $\delta^l_{X_aX_b|Y_{ab}}(m)=0$和$\delta^l_{X_bX_a|Y_{ab}}(m)=0$,对所有的$m\in\mathbb{Z}^+$成立, 由于笔者主要考虑的是$X_a(\cdot)$和$X_b(\cdot)$之间的联系, 因此在各分量独立的条件下, 对各分量的样本分别进行置换所得到的样本对应分量之间仍是独立的, 这里用置换检验来判定原假设是否成立. 由于样本量的局限性, 滞后值m从集合$m\in\{0,1,\cdots,M-1,M\}$中取得, 其中M为事先设定的正整数. 在实际应用中, 可通过分析样本协方差矩阵初步获得.

检验步骤:

1) 计算多维时间序列$X(t)$的$\hat{\delta}^l_{X_aX_b|Y_{ab}}(m)$, $m\in\{0,1,\cdots,M-1,M\}$, $a,b\in\{1,\cdots,k\}$.

2) 分别随机置换时间序列$X_1(t),\cdots,X_k(t)$. 得到$\tilde{X}(t)=(\tilde{X}_1(t),\cdots,\tilde{X}_k(t))'$.

3) 计算序列$\tilde{X}(t)$的检验统计量值, 记为$\tilde{\delta}^l_{\tilde{X}_a\tilde{X}_b|\tilde{Y}_{ab}}(m)$.

4) 重复2)3)步B次.

5) 计算单边bootstrap p值:

$$\hat{p}_{X_aX_b|Y_{ab}}(m)=$$
$$(1+\#[\tilde{\delta}_{\tilde{X}_a\tilde{X}_b|\tilde{Y}_{ab}}(m)\geqslant\hat{\delta}_{X_aX_b|Y_{ab}}(m)])/(1+B).$$

6) 对于选定的显著性水平α, 如果$\hat{p}_{X_aX_b|Y_{ab}}(m)\leqslant\alpha$, 则拒绝$X_a(t)$和$X_b(t-m)$之间线性无关的原假设.

7) 对于多维时间序列$X(t)$, 计算出所有的$\hat{p}_{X_iX_j|Y_{ij}}(m)$, $i,j\in V, i\neq j$, 则在图$G=(V,E)$中:

$$(i,j)\in E\Leftrightarrow\exists m,\hat{p}_{X_iX_j|Y_{ij}}(m)\leqslant\alpha$$
$$\text{或 }\hat{p}_{X_jX_i|Y_{ij}}(m)\leqslant\alpha.$$

4　实例分析(Example analysis)

为了检验本文提出的多维时间序列的线性条件互信息图模型方法可以有效的表示各类多维时间序列分量之间的直接线性联系, 对实际数据进行了验证, 结果表明本文的方法确实可以准确描述多维时间序列各分量间的相依结构.

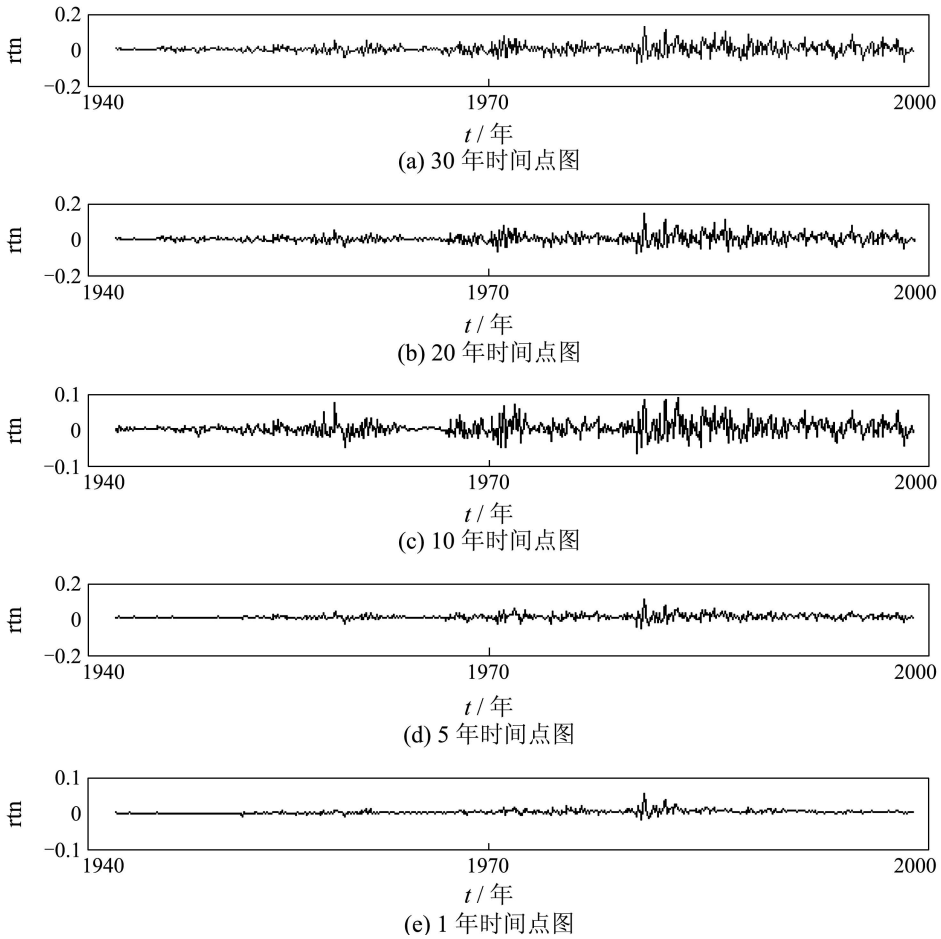

图 1　到期日分别为30年、20年、10年、5年和1年的美国政府债券月简单收益序列的时间点图

Fig. 1　Time plots of monthly simple returns of five indexes of U.S. government bonds with maturities in 30 years, 20 years, 10 years, 5 years, and 1 year

考虑30年、20年、10年、5年和1年共5个不同到期日的美国政府债券从1942年1月到1999年12月的月简单收益率数据[6], 共696个观测值. 令$X_t = (X_{1t}, \cdots, X_{5t})^{\mathrm{T}}$为按到期日递减顺序排列的收益序列. 图1为在相同尺度上X_t的的时间点图.

根据上节提出的检验方法, 滞后选择$m = 0, 1, \cdots, 10$, 得到$B = 199$组置换样本, 计算出各分量之间的$p_{X_iX_j|Y_{ij}}(m)$, $i, j = 1, \cdots, 5, i \neq j$. 虽然统计量$\delta^l_{X_aX_b|Y_{ab}}(m), m \in \mathbb{Z}^+$不是对称的, 即$\delta^l_{X_aX_b|Y_{ab}}(m) \neq \delta^l_{X_bX_a|Y_{ab}}$. 但$\delta^l_{X_aX_b|Y_{ab}}(m)$和$\delta^l_{X_bX_a|Y_{ab}}$之间成立关系式: $\delta^l_{X_aX_b|Y_{ab}}(m) = \delta^l_{X_bX_a|Y_{ab}}(-m), m \in \mathbb{Z}^+$. 因此为方便起见, 把$p_{X_iX_j|Y_{ij}}(m)$和$p_{X_jX_i|Y_{ij}}(m)$画到一个图上, 即$p_{X_iX_j|Y_{ij}}(-m) = p_{X_jX_i|Y_{ij}}(m)$. 各收益序列之间联系存在性的检验结果见图2, 其中虚线表示显著性水平$\alpha = 0.05$.

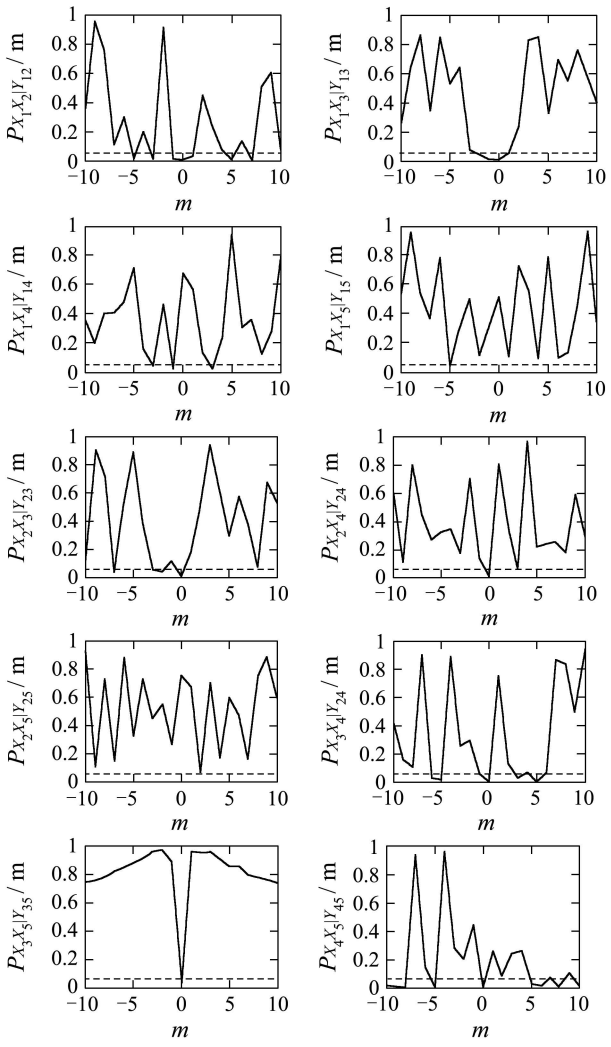

图 2 5个美国政府债券月简单收益数据各分量的p值

Fig. 2 The values of p for the components of monthly simple returns of five indexes of U.S. government bonds

尽管在一些滞后值上稍微超出了检验界, 检验表明一些序列在给定其他序列的条件下是条件无关的. 从图2得到条件互信息图3是很合理的. 图中的顶点(1, 2, 3, 4, 5)分别表示分量序列$(X_{1t}, X_{2t}, X_{3t}, X_{4t}, X_{5t})$.

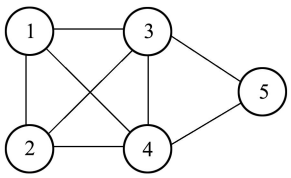

图 3 根据图2得到的LCMIG

Fig. 3 LCMIG from Fig.2

图3正确反映了30年、20年、10年和5年到期债券之间及10年、5年和1年到期债券之间存在直接相关关系, 证实了不同期限(中长期和短期)债券收益分别受到各种因素的不同程度的影响.

通过序列的同期互相关矩阵[6]:

$$\hat{\rho}_0 = \begin{bmatrix} 1.00 & 0.98 & 0.92 & 0.85 & 0.63 \\ 0.98 & 1.00 & 0.91 & 0.86 & 0.64 \\ 0.92 & 0.91 & 1.00 & 0.90 & 0.68 \\ 0.85 & 0.86 & 0.90 & 1.00 & 0.82 \\ 0.63 & 0.64 & 0.68 & 0.82 & 1.00 \end{bmatrix}$$

得到的序列之间相依联系结构图为图4.

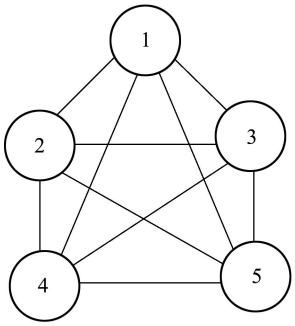

图 4 根据互相关矩阵得到的图

Fig. 4 Graph from cross correlation matrix

比较图3和图4: 图3中的边反映了几个分量之间的条件相依结构, 特别是可以区别直接和间接相关. X_{1t}和X_{5t},X_{2t}和X_{5t}之间没有边, 但它们分别与X_{3t}和X_{4t}之间都有直接联系, 这表明X_{1t}和X_{5t},X_{2t}和X_{5t}通过X_{3t}和X_{4t}发生间接联. 而直接相关分析反映出相反的结论, 因为X_{1t}和X_{5t}, X_{2t}和X_{5t}之间的相关系数很大, 分别为0.63和0.64.

(下转第267页)

[4] LIU R, ALLEYNE A. Nonlinear force/pressure tracking of an electro-hydraulic actuator[J]. *ASME Journal Dynam Syst Meas Contr*, 2000, 122(3): 232 – 237.

[5] YAO B, BU F, CHIU G. Nonlinear adaptive robust control of electro-hydraulic servo systems with discontinuous projections[C]//*Proceedings of the IEEE Decision and Control*. Piscataway, USA: IEEE Press, 1998: 2265 – 2270.

[6] YAO B, BU F, REEDY J, et al. Adaptive robust motion control of single-rod hydraulic actuators: Theory and experiments [J]. *IEEE/ASME Transactions Mechatronics*, 2000, 5(1)：79 – 91.

[7] BU F, YAO B. Desired compensation adaptive robust control of single-rod electro-hydraulic actuator[C] *//Proceedings of the American Control Conference*. Piscataway, USA: IEEE Press, 2001: 3927 – 3931.

[8] LIU S, YAO B. Indirect adaptive robust control of electro-hydraulic systems driven by single-rod hydraulic actuator [C] *//Proceedings of the IEEE/ASME international Conference on Advanced Intelligent Mechatronics*. Piscataway, USA: IEEE Press, 2003: 296 – 301.

[9] DURAISWAMY S, CHIU G, REEDY J. Nonlinear adaptive nonsmooth dynamic surface control of electro-hydraulic systems[C] *//Proceedings of the American Control Conference*. Piscataway, USA: IEEE Press, 2003: 3287 – 3292.

[10] ZHU W, PIEDBOEUF J. Adaptive output force tracking control of hydraulic cylinders with applications to robot manipulators[J]. *ASME Journal of Dynamic Systems Measurement and Control*, 2005, 127(20): 206 – 217.

作者简介:

管　成　(1968—), 男, 博士, 主要从事电液系统控制、非线性控制理论的研究, E-mail: gchlsq@163.com;

潘双夏　(1963—), 男, 教授, 博士生导师, 主要从事机电控制的研究.

(上接第260页)

6　结论(Conclusions)

本文定义了描述多维时间序列分量间直接线性相依结构的线性条件互信息图, 提出了检验边存在性的信息论统计量和置换检验方法. 实例分析结果表明, 对于各种不同结构的线性模型, 检验方法可以准确揭示分量之间的相依联系, 建立描述多维序列相依结构的图模型. 并且模型的残差分布不必限制为正态分布. 由于定义的局限, 线性条件互信息图只反映了模型的线性联系. 但是, 基于信息论的线性检验对于用广义互信息建立研究多维时间序列分量间非线性相依结构的图模型提供了基本方法, 这也是今后研究的一个内容.

参考文献(References):

[1] DAHLHAUS R. Graphical interaction models for multivariate time series[J]. *Metrika*, 2000, 51(2): 157 – 172.

[2] PALUS M. Testing for nonlinearity using redundancies: quantitative and qualitative aspects[J]. *Physica D*, 1995, 80(1): 186 – 205.

[3] PALUS M. Detecting nonlinearity in multivariate time series[J]. *Physics Latters A*, 1996, 213(3): 138 – 147.

[4] GRANGER C, LIN J L. Using the mutual information coefficient to identify lags in nonlinear models[J]. *Journal of Time Series Analysis*, 1994, 15(4): 371 – 384.

[5] DIKS C G H, MANZAN S. Test for serial independence and linearity based on correlation integrals[J]. *Studies in Nonlinear Dynamics & Econometricsl*, 2002, 6(2): 1 – 20.

[6] TSAY R S. *Analysis of Financial Time Series*[M]. New York: Wiley & Sons, 2002.

作者简介:

高　伟　(1978—), 女, 西北工业大学理学院应用数学系博士研究生, 主要研究方向为非线性时间序列分析的理论与应用, E-mail: gaoww525@tom.com;

田　铮　(1958—), 女, 教授, 西北工业大学理学院应用数学系博士生导师, 计算机科学与工程系博士生导师, 主要从事非线性时间序列与信息处理、多尺度随机模型与图像处理等方面的研究, E-mail: zhtian@nwpu.edcu.cn.