文章编号:1001 – 2486(2005)05 – 0078 – 04

基于时间序列预测的独立分量排序*

王　刚[1,2],胡德文[1]

(1.国防科技大学 机电工程与自动化学院,湖南 长沙　410073; 2.空军工程大学 电讯工程学院,陕西 西安　710077)

摘　要:独立分量排序是独立分量分析的热点问题,是提高特征空间鲁棒性和减少计算复杂度的必要前提。结合 ICA 在时间序列预测的应用,给出了基于一阶差分和最小方差误差的多分量联合重构预测排序准则。为了避免联合优化中出现的海量计算问题,提出了添加 – 测试 – 接受机制(ATA)的次优搜索方法。实验结果表明,和传统排序方法比较,新方法具有优异的预测能力和搜索效率。

关键词:独立分量分析;时间序列;预测;添加 – 检测 – 接受

中图分类号:TP183　　文献标识码:A

Independent Component Ordering in Time Series Forecasting

WANG Gang[1,2], HU De-wen[1]

(1. College of Mechatronics Engineering and Automation, National Univ. of Defense Technology, Changsha, 410073, China;
2. The Telecommunication Engineering Institute, Air Force Engineering University, Xi'an 710077, China)

Abstract: The ordering of independent components is a hot issue in independent component analysis (ICA), and the critical step for the robustness and computing complexity of independent feature space. In time series forecasting, a novel criterion has been presented based on the mechanism of first-order differential and minimum variance error under multiple components reconstruction. To avoid the exhaustive search in combinatorial optimization, a sub-optimum approach named Adding-Testing-Acceptance (ATA) is proposed. Experimental results show that the proposed method has a better forecasting ability and more efficient in comparison with the existing ones.

Key words: independent component analysis; time series; ordering; adding-testing-acceptance (ATA)

独立分量分析(ICA)是一种基于高阶统计信息的信号处理和分析方法,由于具有良好的特征表示和辨识能力,目前已经广泛应用于盲源分离、特征提取和盲解卷等领域。在基本模型的估计中,ICA 可用于提取潜在独立分量的波形特征,而无法给出幅度、方向和分量的先后排序等信息[1~4]。对盲源分离而言,独立分量顺序的不确定性使得估计结果和源信号之间无法直接一一对应,要求引入其它先验知识,如频谱信息[3]。在模式识别和数据特征分析中,为了减少计算的复杂度并保证特征空间对样本的鲁棒性,要求先对独立分量集中的分量按照某一标准进行排序,然后再选择前面的少数或者部分分量组成新的特征集[5,6]。独立分量排序是基于 ICA 的数据分析的重要环节。

目前独立分量排序主要包括:(1)基于分量的自身特性(如非高斯性)的方法[4,6];(2)基于对不同观测变量的贡献率的方法[7];(3)基于对观测数据的重构误差的方法[4];(4)基于与观测数据互信息的方法[5]。此外还有基于模式聚类的方法[7,8]。研究表明,独立分量排序不仅与分量自身的特点有关,关键取决于应用背景[4,7]。在基于 ICA 的时间序列预测中,Cheung 等首次提出了多分量联合优化的思想,给出了基于相对汉明码距的目标函数准则和检测 – 接受(Testing-and-Acceptance, TnA)机制的次优搜索算法[5]。分析表明,采用相对汉明码距虽然突出了预测的方向信息,却丢弃了变化量信息;其次,TnA 本质上是一种逐步剔除的策略,对于高维分量空间且仅要求选择极少独立分量的情况,计算量仍然太大。

* 收稿日期:2005 – 05 – 20
基金项目:国家自然科学基金项目(30370416);国家杰出青年科学基金项目(60225015);高等学校优秀教师教学科研奖励计划项目
作者简介:王　刚(1976—),男,讲师,博士生。

1　独立分量分析和排序

考虑具有时间结构信息的 ICA 模型[2]

$$x(t) = As(t), 1 \leqslant t \leqslant N \tag{1}$$

其中,$x(t) = [x_1(t), \cdots, x_n(t)]^{\mathrm{T}}$ 为观测时间序列,$s(t) = [s_1(t), \cdots, s_n(t)]^{\mathrm{T}}$ 为潜在的独立分量,A 表示未知的 $n \times n$ 维的混合矩阵。通过最大非高斯估计、Infomax 等方法可以实现潜在分量的估计,设 $y(t)$ 是 $s(t)$ 的一种实现,则

$$y(t) = Wx(t) = P \wedge s(t), 1 \leqslant t \leqslant N \tag{2}$$

其中,$y(t) = [y_1(t), \cdots, y_n(t)]^{\mathrm{T}}$,$y_i(t)$ 对应 $s(t)$ 中某一元素的保形映射[2,3]。

式(2)表明,通常 ICA 方法无法确定独立分量的先后顺序。而在应用中,为了提高特征空间对样本对象的鲁棒性,并减少计算和分析的复杂度,通常要求对独立分量集 $\phi_n = \{y_i(t)\}_{i=1}^{n}$ 中的分量进行排序,然后根据需要(如某种标准)选择合适的子空间,独立分量的排序是选择特征子空间的关键[4,5,7,8]。

2　目标函数和优化

2.1　目标函数

引入 Cheung 在文献[5]中提出的联合优化思想。首先考虑独立分量 $y_j(t)$ 对观测数据 $x_i(t)$ 的重构

$$u_{ij}(t) = \hat{A}_{ij} y_j(t), 1 \leqslant j \leqslant k \tag{3}$$

其中 $\hat{A}_{ij}$ 为混合阵 A 的估计 $\hat{A}$ 的第 (i,j) 元素。ϕ_n(下标 n 表示集合中元素的个数)中前 m 个独立分量对观测数据 $x_i(t)$ 的重构可以表示为

$$u_{i,1:m}(t) = \sum_{r=1}^{m} \hat{A}_{ir} y_r(t), 1 \leqslant r \leqslant m \leqslant k \tag{4}$$

鉴于相对汉明距的缺陷,以下直接引入基于一阶差分和最小方差误差的预测准则。分别定义 $\tilde{x}_i(t)$ 和 $\tilde{u}_{i,1:m}(t)$ 为 $x_i(t)$ 和 $u_{i,1:m}(t)$ 的一阶差分,用来表示时间序列的变化趋势:

$$\tilde{x}_i(t) = x_i(t) - x_i(t-1), \tilde{u}_{i,1:m}(t) = u_{i,1:m}(t) - u_{i,1:m}(t-1) \tag{5}$$

定义 $J_m(x_i, u_{i,1:m})$ 为 ϕ_n 中前 m 个分量的重构数据 $u_{i,1:m}(t)$ 对时间序列 $x_i(t)$ 的预测误差

$$J_m(x_i, u_{i,1:m}) = (\tilde{x}_i(t) - \tilde{u}_{i,1:m}(t))^2 \tag{6}$$

改变 ϕ_n 中各分量的顺序,可得到多种组合,包含 m 个分量的集合 ϕ_m 中联合预测能力最优的集合 ϕ_m^* 为

$$\phi_m^* = \arg\min J_m(x_j, u_{i,1:m}) \tag{7}$$

2.2　次优搜索算法

以上确定全部分量最优排序的搜索次数为 $n!$[5],在 n 较大时计算量非常大,显然难以接受。针对此问题,文献[5]提出了 TnA 次优搜索算法,其核心思想是将依次从包含 $k(1 < k \leqslant n)$ 个独立分量的集合 ϕ_k 中选出联合预测最优的 $k-1$ 个分量集,并将剩下的一个分量置末。本质上这是一种逐步剔除的策略。以下提出 ATA(Adding-Testing-Acceptance)的次优搜索算法,通过逐步增加分量的方法实现优化排序。具体而言,首先从包含 n 个分量的集合 ϕ_n 中选择满足(7)式的分量 y_l,定义 $\hat{y}_1 = y_l$ 添加在空集合 φ 中,并从 ϕ_n 中剔除 y_l 对 ϕ_{n-1} 中各分量 $y_p(1 \leqslant p \leqslant n-1)$,应用 y_p 和 φ 中的分量 $\hat{y}_1$ 对 $x_i(t)$ 进行重构预测分析,满足(7)式的分量 y_p 定义为 $\hat{y}_2$,……,依次可以得到前 m 个独立分量和排序,也可以实现完全排序。

ATA 算法的求解过程归纳如下:(1)定义集合 $\phi_n = \{y_j\}_{j=1}^{n}$,空集 φ_0,令 $k=1$;(2)对任意的 $y_l \in \phi_{n+1-k}$,令 $\varphi' = \varphi_{k-1} + \{y_l\}$,依式(1)~(7),求解满足 φ' 预测能力最佳的 y_l;(3)令 $\hat{y}_k = y_l$,$\varphi_k = \varphi_{k-1} + \{y_k\}$,$\phi_{n-k} = \phi_{n+1-k} - \{y_l\}$;(4)如果 $k < n$,令 $k = k+1$,进入步骤(2);否则中止。

φ_k 中元素 $\hat{y}_k$ 的下标 k 表示分量新的排序。以上算法对应完全排序,对于不完全排序,如仅关注前 m 个独立分量,修改步骤(4)中 $k < n$ 为 $k < n-m$;对于给定预测指标 J^*,修改步骤(4)中的收敛条件。

表 1 给出了最优、TnA、ATA 三种搜索的完全排序、不完全排序、$m \ll n$ 时的搜索次数比较。显然,在

$m<0.5n$ 时，ATA 方法的搜索效率优于最优搜索和 TnA 方法。在模式识别等应用中，考虑特征空间的鲁棒性、计算的复杂度，通常要求 $m\ll n$，此时 ATA 的优势更加明显。

表 1 最优算法、TnA 和 ATA 三种搜索复杂度比较

Tab.1 Complexity comparison of optimal approach, TnA and ATA

搜索方法	完全搜索	不完全搜索	$m\ll n$
最优	$n!$	$n!/(n-m)!$	$\approx n^m$
TnA	$n(n+1)/2-1$	$(n+m+1)(n-m)/2$	$\approx n^2/2$
ATA	$n(n+1)/2-1$	$(2n-m+1)m/2$	$\approx mn$

3 实验验证

选用文献[9]提供的 1979.12.31 ~ 1984.10.09 之间共计 1200 个观测数据的 8 种外汇交易数据，分别为美元对日元、澳元、加元、德国马克、荷兰先令、法郎、日元和对瑞士法郎的汇率。将这些数据进行标准化处理，得到零均值单位方差的实验数据。选择美元/日元的金融序列 $x_7(t)$ 作为分析对象。首先应用 ICA（如 FastICA 算法[2]）提取 8 个潜在的独立分量，然后按照非高斯性方法、L_∞ 范数和给定预测准则的 TnA 和 ATA 两种次优搜索算法进行排序。表 2 给出了四种方法（算法）的排序结果和预测误差值。其中定义按照分量的非高斯性强弱排序记为 $y_1, y_2, \cdots, y_8$。“排序/预测误差”中的第 i 列中‘排序’数目表示该独立分量对应非高斯性排序的独立分量顺序，而‘预测误差’表示该分量和前面的低位排序分量联合重构对 $x_7(t)$ 的预测误差值。如 TnA 中 $m=2$ 时的‘1/7.8075’表示按照 TnA 算法，y_1 排在第二位，它与前面已经排定的分量 y_5 对 $x_7(t)$ 的联合预测误差为 7.8075。

表 2 四种方法（算法）的排序结果和预测误差值

Tab.2 IC ordering and forecasting error

方 法	排序/预测误差							
	1	2	3	4	5	6	7	8
非高斯性	1/15.6567	2/14.3830	3/18.3906	4/17.8969	5/9.4948	6/4.8199	7/0.1431	8/0.0000
L_∞ 范数	1/15.6567	2/14.3830	3/18.3906	5/ 7.9113	4/9.4948	6/4.8199	8/4.1688	7/0.0000
TnA	5/9.1905	1/7.8075	3/6.9729	7/4.6621	2/2.8557	6/0.9044	4/0.1431	8/0.0000
ATA	5/9.1905	7/7.0199	3/4.2264	4/3.0708	1/2.8706	6/1.8434	8/2.3543	2/0.0000

图 1 给出了按照非高斯性（或者 L_∞ 范数）排序的第一分量（y_1）和 ATA（或者 TnA）第一分量（y_5）对金融数据 $x_7(t)$ 的预测能力。图 2 给出了四种方法（算法）对应金融数据预测误差的曲线，其中带‘$*$’的实线、带‘□’的虚线、带‘$\times$’的点线和带‘$+$’的点划线分别表示非高斯性、L_∞ 范数、TnA 和 ATA 四种方法（算法）对应的分量个数与预测误差关系曲线。

从表 2 和图 1、图 2 可以看出：(1)基于给定的预测准则，无论是 ATA 还是 TnA 搜索算法的预测效果都要优于非高斯性和 L_∞ 范数方法。如 $m=3$ 时，TnA 和 ATA 算法对应的预测误差分别为 6.9729 和 4.2264，远优于非高斯性和 L_∞ 范数方法对应的 18.3906；(2)即使对于相同的目标函数，ATA 和 TnA 两种搜索算法对应独立分量的排序也不同，对应的预测能力也存在差异。从图 2 中可以看出，以第 5 个独立分量为界，在 $m=4$ 时，ATA 的排序依次为 y_5, y_7, y_3, y_4，后者的排序依次为 y_5, y_1, y_3, y_7，两者包含的元素也不尽相同，前者对应的预测能力优于后者；当 $m<5$ 时，两种排序的预测效果则相反。这也表明了在对应相同的联合优化准则下两种搜索方法的差异，大集合中‘差’在小集合中可能是‘优’的。(3)在 $m<n/2$ 时，ATA 的搜索次数要少于 TnA。如 $m=2$ 时，前者的搜索次数为 15，远小于后者的搜索次数 66，显然搜索效率要高得多。

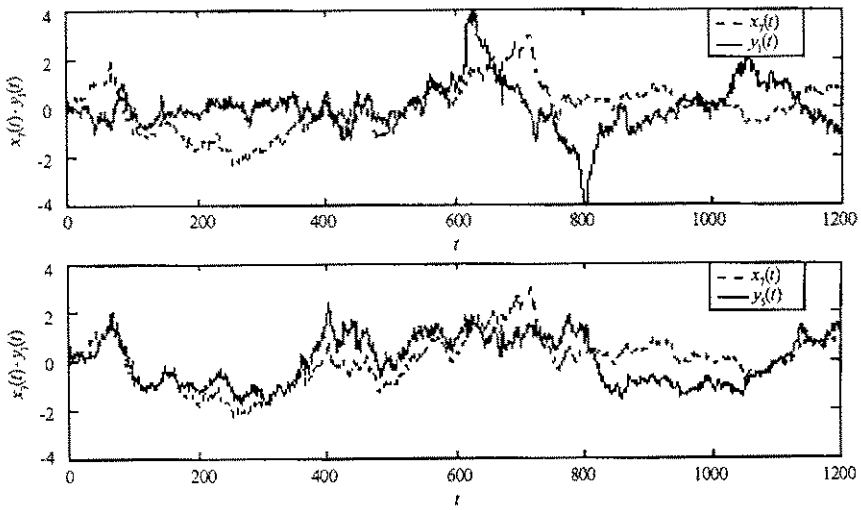

图 1　非高斯性/L_∞ 范数方法(上图),ATA/TnA 方法(下图)的第一分量(实线)对金融数据 $x_7(t)$(虚线)的预测能力

Fig.1　Forecasting ability of the first component (solid line) to economic data $x_7(t)$ (dashed line) NonGaussianity method/L_∞ method correspond upper figure, and ATA/ TnA Upper figure

4　总　结

独立分量的排序是 ICA 估计方法应用的重要环节。结合 ICA 在时间序列预测中的应用,提出了基于时间序列的一阶差分和最小方差误差的多分量联合重构预测排序准则。在目标函数的优化过程中,为了避免海量计算,提出了 ATA 的次优搜索算法。实验结果表明,和传统的排序方法比较,新方法具有优异的预测能力和搜索效率。

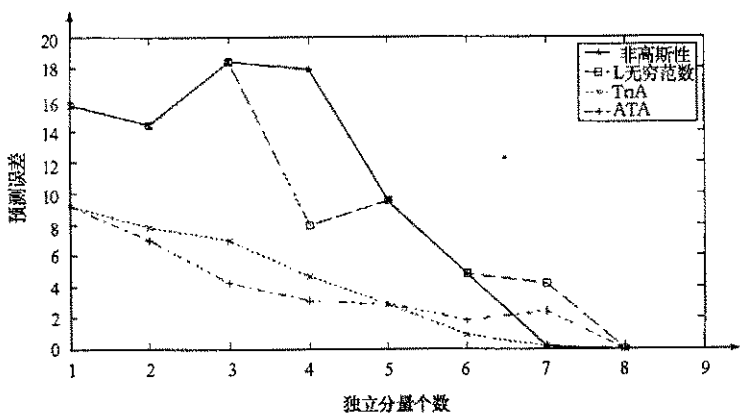

图 2　四种方法(算法)对应金融数据 $x_7(t)$ 预测误差曲线

Fig.2　Forecasting errors of four methods (approaches) to economic data $x_7(t)$

参 考 文 献:

[1] Common P. Independent Component Analysis —a New Concept [J]. Signal Processing, 1994, 36(3): 287 – 314.

[2] Hyvärinen A, Karhunen J, Oja E. Independent Component Analysis [M]. John Wiley, New York, 2001.

[3] Cichocki A, Amari S. Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications [M]. Wiley, 2003.

[4] Hyvärinen A. Survey on Independent Component Analysis [J]. Neural Computing Surveys, 1999, 2: 94 – 128.

[5] Cheung Y, Xu L. Independent Component Ordering in ICA Time Series Analysis [J]. Neural Computing, 2001, 41: 145 – 152.

[6] Back A D, Weigend A S. A First Application of Independent Component Analysis to Extracting Structure from Stock Returns [J]. Neural Systems, 1997, 8(4): 473 – 484.

[7] Wang G, Hu D. Kernel Face Representation Using Independent Component Analysis [A]. The 5th International Symposium on Test and Measurement, 4: 3277 – 3280, Shenzhen, China, June 1 – 5, 2003.

[8] Yuen P C, Lai J H. Face Representation Using Independent Component Analysis [J]. Pattern Recognition, 2002, 35: 1247 – 1257.

[9] http://www – personal.buseco.monash.edu.au/hyndman/TSDL/ [DB], 2003/12/17.