独立分量分析方法及其在地理与环境科学中的应用*

粟泽毅，陈阿林
(重庆师范大学 地理科学学院，重庆 400047)

摘 要：综述了独立分量分析(ICA)问题的模型、原理，以及定义在基本 ICA 问题上的各种算法，包括信息最大化 infomax 算法、负熵最大化法、最大似然法等，并对各算法性能作了比较。并介绍了可获得的 ICA 算法系统 fastica 和 oolabss 及其运行平台，说明了 ICA 在地理与环境问题中的应用。

关键词：盲信号分离；ICA；负熵；盲反褶积；污染源
中图分类号：K909；O213 **文献标识码**：A **文章编号**：1672-6693(2005)04-0016-05

The Independent Component Analysis and Its Application in the Geographical and Environmental Fields

SU Ze-yi, *CHEN A-lin*
(College of Geography Science, Chongqing Normal University, Chongqing 400047, China)

Abstract: This paper summarizes the principle and the model of Independent Component Analysis, and various algorithms based on the basic ICA problem, including the HJ neural network, the infomax and maximum negentropy algorithm. The performances of the these methods are compared, the gainable system of ICA algorithm on the web is introduced, the applications of ICA in the Geographical and Environmental fields are discribed and explained.

Key words: blind signal separate; ICA; negentropy; blind deconvolution; polluting source

独立分量分析(Independent Component Analysis，ICA)是近年来发展起来的一种新的信号处理技术。它是伴随着盲信源问题而发展起来的，故又称盲分离(Blind Source Separate)。盲信号处理(Blind Signal Processing，BSP)是 20 世纪最后 10 年迅速发展起来的一个研究领域。它又可以分成若干个互相关联而目标有所区别的子领域，如盲信号分离(Blind Signal Separation，BSS)以及盲解卷(Blind Deconvolution)、盲均衡(Blind Equalization)等。按照所取的假设条件和研究途径不同，可以包含独立分量分析(ICA)、因子分析(Factor Analysis，FA)和独立因子分析(Independent Factor Analysis，IFA)等若干课题。

ICA 是由 C. Jutten 和 J. Herault 于 1991 年提出[1]，目的是寻求统计上独立的非高斯分布数据的线性变换。在 ICA 问题中，观测信号是几个相互独立源信号的线性组合，寻找一个线性变换，使得变换后得到的估计信号在统计上尽可能独立，从而仅从线性混合的观测信号就可以分离出原始的源信号。ICA 就是从已知的观测数据中捕捉(提取)数据的基本结构，具有稀疏性和减少冗余性。

独立分量分析在阵列信号处理、生理医学信号处理、语音信号处理、信号分析及过程控制的信号去噪和特征提取、模式识别等领域有着广泛的应用。此外，独立分量分析也应用在数据挖掘(data mining)[2~4]中。

1 独立分量分析

1.1 盲信号分离问题

设有 N 个未知的源信号 $S_i(t)$，$i=1,2,\cdots,N$，构

* 收稿日期：2005-05-09
作者简介：粟泽毅(1976-)，女，重庆人，硕士研究生，研究方向为地理信息系统。

成一个 n 维列向量 $S=(s_1(t),s_2(t),\cdots,s_n(t))^{\mathrm{T}}$,其中 t 是离散时刻,取值为 0,1,2,…。设 A 是一个 $M\times N$ 维矩阵,一般称为混合矩阵(mixing matrix)。设 $X(t)=(x_1(t),x_2(t),\cdots,x_m(t))^{\mathrm{T}}$ 是由 M 个可观察信号 $x_i(t)$,$i=1,2,\cdots,M$ 构成的列向量,且满足方程

$$X(t)=A\cdot S(t),\quad M\geqslant N \tag{1}$$

BSS 的命题是:对任何 t,根据已知的 $X(t)$ 在 A 未知的条件下求未知的 $S(t)$。这构成一个无噪声的盲分离问题。设 $N(t)=(N_1(t),N_2(t),\cdots,N_m(t))^{\mathrm{T}}$ 是由 M 个白色、高斯、统计独立噪声信号 $n_i(t)$ 构成的列向量,且 $X(t)$ 满足方程

$$X(t)=A\cdot S(t)+N(t),\quad M\geqslant N \tag{2}$$

则由已知的 $X(t)$ 在 A 未知时求 $S(t)$ 的问题是一个有噪声盲分离问题。

1.2 独立分量分析[5]

如果按照以下的几个基本假设条件来解决 BSS 问题,则称之为 ICA。这些条件如下。

1)各源信号 $S_i(t)$ 均为 0 均值、实随机变量,各源信号之间统计独立。如果每个源信号 $S_i(t)$ 的概率密度函数(简写为 pdf)为 $P_i(S_i)$,则 $S(t)$ 的 pdf 为 $P_s(S)$,可以用下式计算

$$P_s(S)=\prod_{i=1}^{n}P_i(S_i)\ ; \tag{3}$$

2)源信号数 M 与观察信号数 N 相同,即 $M=N$,这时混合阵 A 是一个确定且未知的 $N\times N$ 维方阵。假设 A 是满秩的,逆矩阵 A^{-1} 存在;

3)各个 $S_i(t)$ 的 pdf 中只允许有一个具有高斯分布,如果具有高斯分布的源信号个数超过一个,则各个源信号是不可分的。Darmois 定理严格证明了这一结论[6];

4)各观察器引入的噪声很小,可以忽略不计。这时可以用(1)式描述源信号与观察信号之间的关系且 $M=N$;

5)关于各源信号的 pdf $P_i(S_i)$,略有一些先验知识。

ICA 的思路是设置一个 $N\times N$ 维反混合矩阵 $W=(W_{ij})$,$X(t)$ 经过 W 变换后得到 N 维输出列向量 $Y(t)=(y_1(t),y_2(t),\cdots,y_n(t))^{\mathrm{T}}$,即有

$$Y(t)=W\cdot X(t)=W\cdot A\cdot X(t) \tag{4}$$

图1 所示为 ICA 问题简单网络模型——最简单的盲源分离问题。独立信号 $s_1,s_2,\cdots,s_n$,经过一未知线性混合,产生观察信号 $x_1,x_2,\cdots,x_m$。观察信号 x 通过该盲源分离系统后得到近似于 s 的输出 y。

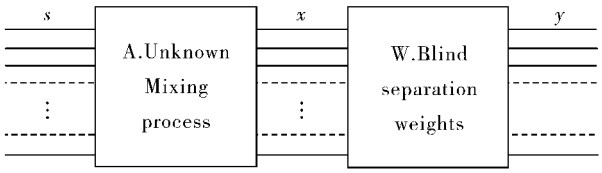

图1 ICA 的简单网络模型

2 ICA 的几种主要算法

如果通过学习得以实现 $WA=I$(I 是 $N\times N$ 维单位阵),则 $Y(t)=S(t)$,从而达到了源信号分离目标。由于没有任何参照目标,这一学习只能是自组织的。学习过程的第一步是建立一个以 W 为变元的目标函数 $L(w)$,如果某个 $\hat{w}$ 能使 $L(w)$ 达到极大(小)值,该 $\hat{w}$ 即为所需的解。第二步即是用一种有效的算法求 $\hat{w}$。按照 $L(w)$ 定义的不同和求 $\hat{w}$ 的方法不同可以构成各种 ICA 算法。令人吃惊的是这些算法都特别简单,而且,即使关于源信号 pdf 的先验知识很不充分,分离效果也特别好。

2.1 求矩阵 w 的两个估计原理

ICA 估计原理 1 非线性不相关。找到矩阵 w 以使分量 y_i 和 y_j(对任意 $i\neq j$)是不相关的,变换分量 $g(y)$ 和 $h(y)$ 是不相关的,式中 g 与 h 是某个合适的非线性函数。

ICA 估计原理 2 极大非高斯性。在 y 的方差是常量的约束下,找到线性混合 $y=\sum b_ix_i$ 的非高斯局部最大值,每个局部极大值给出了一个独立分量。

2.2 ICA 目标函数

2.2.1 信息最大化(infomax)判据[7]

Bell 和 Sejnowsk 于 1995 年提出了 infomax 判据。迄今为止,许多判据和算法都与这个判据有或多或少的联系。该方法是利用独立分量分析网络中非线性单元的最大信息,得到的一种盲源分离方法。

此方法选择 Y 的熵作为目标函数,用 $L_H(W)$ 表示,即有

$$L_H(W)=H(Y)=-E[\ln P_Y(Y)] \tag{7}$$

如图2 所示,在网络的输出端引入非线性激励函数 $g(\cdot)$,使其输出信号 y_i 的熵 $H(Y)$ 最大(其中 $y=[y_1,y_2,\cdots,y_n]$),网络权值 W 通过梯度下降法得到,可以证明,单调有界非线性函数 $g(\cdot)$ 应取对应源的累积分布函数。对超高斯分布的源,W 学习规则可表示为

$$\Delta W\propto\frac{\partial \mathrm{H}(Y)}{\partial W}W^{\mathrm{T}}W=[\mathrm{I}-\varphi(u)u^{\mathrm{T}}]W \tag{8}$$

这里得到 W 后,便可求出 $u=Wx$ 时,u 是 s 的估

计。其中 $u=[u_1,u_2,\cdots,u_n]$,$x=[x_1,x_2,\cdots,x_n]$,

$$\varphi(u)=-\frac{\partial p(u)}{\partial u}\Big/p(u)=\left[-\frac{\partial p(u_1)}{\partial u_1}\Big/p(u_1),\cdots,-\frac{\partial p(u_n)}{\partial u_n}\Big/p(u_n)\right]$$

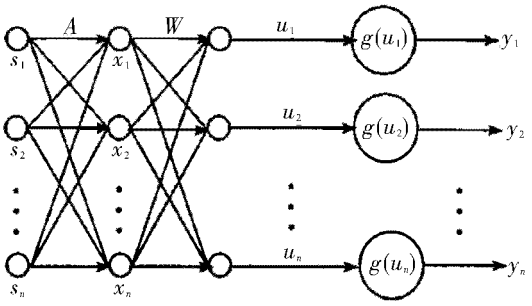

图 2　infomax 方法的原理

从上面的论述可以看出,infomax 方法的关键是激励函数 $g(\cdot)$ 的选取。在更多情况下,源的分布函数是未知的。上面两种算法都有一个不足,就是当源的分布为亚高斯时,没有相应的求 W 的算法。原因是一旦 $g(\cdot)$ 确定,就意味着源的概率分布已确定。Lee 等人提出求 W 较为韧性的算法。

$$\Delta W\propto\begin{cases}[I+f(u)u^{\mathrm{T}}-\alpha uu^{\mathrm{T}}]W\\ [I-f(u)u^{T}-\alpha uu^{\mathrm{T}}]W\end{cases}\qquad(9)$$

在上式的右边分别适用于源是超高斯分布和亚高斯分布。$f(u)$ 是单调非线性奇函数,其导数是源的分布函数,$f(u)$ 可以取 $\mathrm{sign}(u)$,$\mathrm{banh}(u)$,$\mathrm{abs}(u^{0.9})\times\sin(u)$ 等。α 是调整系数,对超高斯分布 α 没有严格限制,对亚高斯分布,$\alpha>0$。

经过验证,解决 ICA 问题主要归结于对被分离信号源 pdf $P_i(S_i)$(或等效的 $\hat{P}_i(y_i)$)的确定或在学习过程中对它的确定。很多实验表明,关键在于确定这些 pdf 是超高斯的或亚高斯的,而其具体形式细节对分离效果的影响不大。

2.2.2　**负熵判决准则**[8]　由中心极限定理可知,一随机量如果由许多相互独立的随机量的和组成,只要各独立的随机量具有有限的均值和方差,则不论各独立随机量为何种分布,则该随机量必接近高斯分布。因此可以在分离过程中,通过对分离结果非高斯性的度量来监测分离结果间的相互独立性。当非高斯性度量达到最大时,表明已完成对各独立分量的分离,对于一概率密度函数为 $p(y)$ 随机量 y,负熵定义为

$$Ng(y)=H(y_{\mathrm{Gauss}})-H(y)\qquad(10)$$

当该变量为高斯分配时,负熵的值将会成为 0,在独立成分分析法中所要追求的基础是负熵最大。在独立成分分析的过程中,也往往在追求各独立成分间彼此重复信息(mutual information)的最小,可以用(11)式来表示这样的概念。

$$I(y_1,y_2,\cdots,y_m)=\sum_{i=1}^{m}H(y_i)-H(y)\qquad(11)$$

同时,经推演,可得到(12)式,负熵和相互信息最小这两个概念只相差一个参数值及正负值。

$$I(y_1,y_2,\cdots,y_n)=C-\sum_i J(y_i)\qquad(12)$$

因此,也可以得知,独立成分分析法的两个基础是各独立变量间的相互共同信息最小,以及负熵最大。这也是独立成分分析法与其它统计法的最大的不同之处。

实际应用中,由于(10)式的计算需要知道概率密度分布函数,这显然是不切实际的。文献[9]给出了一种近似公式可进行非高斯性度量,即

$$Ng(y)\propto[E\{G(y)\}-E\{G(y_{\mathrm{Gauss}})\}]^2\qquad(13)$$

其中 $E(\cdot)$ 为均值运算,$G(\cdot)$ 可取 $G_1(u)=a_1^{-1}\cdot\lg\cos(a_1u)$ 或 $G_2(u)=-\exp(-u^2/2)$ 等非线性函数。不难理解,上式同样可以对分离结果的非高斯性即独立性进行度量,且可用于实际计算。

此外,ICA 的目标函数还有最大似然(maximum likelihood)目标函数[10],统计独立性目标函数等。

3　技术产品

3.1　FASTICA 算法系统

该系统是 One-unit algorithm for ICA 算法发展的结果,其算法是由 Aapo Hyvarinen 提出,后来由 Helsinki University of Technology(HUT)的 Laboratory of Computer and Information Science 发展成 fixed-point algorithm。目前 FastICA 系统有几个版本,最新的是 Fastica5。此系统属公开使用的性质,可通过 ICA Central 这个网站获得。

Fastica 系统具有图形使用者接口,其算法是采用 fixed-point algorithm。这是一个高计算效率的方法,用来执行独立分量分析法的运算。它使用一个 fixed-point iteration scheme,它的运算速度约为传统梯度下降法的 10 ~ 100 倍,适用于 Matlab 系统语言环境。

3.2　OOLABSS 系统

另外一个已受各界肯定的独立分量分析法系统就是 OOLABSS,是由日本的 Brain-Style Information Processing Group,Brain Science Institute-Riken 所发展起来的,这个研究群是由 S. Amari 所领导,称为 OOLABSS(Object Oriented Laboratory for Blind Source

Separation),是一个研究独立成分分析法及盲目来源分离的研究室。OOLABSS是由B. Orsier在Windows95/NT上以C++程序语言所设计及执行的学习算法为基础,在B. Cichocki及B. Orsier合作下所发展而成的。使用这个程序,用户可以自定神经元的激发函数,也可以加上一些干扰到传感器讯号中。

4 ICA在地理与环境科学中的应用

ICA不仅解除了信号的相关(二阶统计量),而且降低了高阶统计意义上的依赖性,使信号成分尽可能独立。这样它就产生了两个方面的突出应用,一是用于盲源分离,二是用于特征提取。

随着发展,它越来越多地被应用在地理与环境科学中,产生了许多相关的应用和研究。经过尝试与验证,独立分量分析在地震波信号处理和水污染的污染源分离应用中取得了较好的分离结果。

4.1 ICA在地震信号处理中的应用[11]

地壳、上地幔内部具有强速度间断面或过渡带,如莫霍面410km间断面和670km间断面,经过这些间断面的远震转换波(P—S)具有较强信号。利用密集布设的地震台站记录大量地震波,可以得到多次经过同一转换‘点”(面元)的记录,再利用远震体波的径向分量反褶积垂向分量,扣除震源和仪器因素,可以得到接收函数。接收函数保留了地下介质信息,通过动校正,将经过同一转换面元的接收函数叠加,增强有效信息,抑制干扰。通过追踪这些叠加的转换波震相,可以获得相应间断面的图像。然而,当地震台站位于较厚沉积层地区时,由于沉积层内形成的多次地震反射波振幅较强,且又与间断面的转换波在同一时段出现,就会将转换波信息淹没掉。目前,现有的技术手段尚无法将混合的转换波和多次反射波进行有效地分离,从而导致在沉积区无法从接收函数中提取间断面信息。通过ICA方法可以实现较有效的沉积层较厚区远震转换波与多次反射波的分离。

地震勘探中,反射序列和地震记录在一定条件下具备ICA模型特点,在信号处理中探索其应用很有意义。尤其在深层弱信号特征提取上可能发挥作用。另外,地震记录中常常包含干扰信号,若设地震记录中的有效信号和随机干扰信号在统计上独立,且服从非高斯分布,那么根据ICA思想方法,只需该道的两次观测现实,或者近似取邻近的两道,就可以采用ICA直接从地震道中分解出有效信号和随机干扰,从而达到去噪目的。因此,可将ICA方法用于叠前去噪。

4.2 ICA在解决一般化地震盲反褶积问题上的重要应用[12]

地震反褶积基本上是一个盲过程。通常地表爆炸激发的地震子波是未知的,而在地震记录上相邻反射的地震波又是重叠的,无法从地震记录中分离出地震子波,这是由于地下地层的厚度通常要小于地震子波的波长。一般地,在地震子波和反射函数都未知时,常常要做统计性假设,反褶积称为统计性反褶积。这些假设和相应的方法在实际中一般效果较好,但是不能保证假设条件总是正确的,因此无任何假设条件下完成地震反褶积—盲道识别或盲反褶积,十分有意义。

文献[13]提出一种用在通信系统上盲均衡和信道参数估计方法。该系统与地震单道线性褶积极其相似,极有可能用于地震盲反褶积。该方法利用过采样技术和一种新的独立分量分析(ICA)神经网络,仅通过接收信号完成盲道均衡,然后基于接收信号的高阶累积量和线形系统的特性,利用进化规划算法和已估计的均衡序列,估计信道参数。与已有的方法相比,提出的方法利用估计序列辨识系统参数,不必另外产生训练序列,网络结构简单,收敛性好,可以在线得到均衡输出。这种算法在单道地震盲反褶积和子波估计上,具有研究意义和应用前景。

4.3 ICA应用于水污染中的污染源研究[14]

水环境中,属于自然环境作用力以外,从人类土地利用的过程中所产生的各种污染源,各具有不同类型的污染物质及能量,这也就是污染源的特征,可以用水样水质检测记录中各种水质参数的异常变动加以表示。因此,如果将现有的水样水质检测记录加以分析,则水质参数所表现的特征,也就是代表了被排放到水体中的污染物质及能量,更同时显示出造成该污染源的土地利用类型及强度。ICA抽取水质测站资料内涵特征,然后,以数据的内涵特征为污染源的特征,可以推论从污染源排入水体中的物质及能量,更可进而推论可能的土地利用类型。这个简单的逻辑就是ICA进行污染源分离的思考架构。

5 ICA前沿的研究

本文中讨论的几种方法是针对源的线性组合,且满足前文中的5个假设条件的解法。应当说明的是,这是较理想的情况,实际中往往不能同时满足上

述这些假设条件。最近,许多学者都涉及了减弱这几个假设条件的 ICA 方法的研究,例如,非线性 ICA[15],信号有时间延迟的混合,卷积和的情况,带噪声的 ICA[16],源的不稳定问题等等。

带噪声的 ICA 问题是 ICA 的另一热点之一。这种噪声是指传感器噪声,关于这方面的研究引起了许多学者的关注。此外,卷积和及源的不稳定问题等等,都是 ICA 的热点问题。

6　结论

综上所述,独立分量分析法这个研究领域,在现今的学术环境中呈现出不断上升的态势。以上所提及的各种算法及系统,与基于特征分析如奇异值分解(SVD)、主分量分析(PCA)等传统信号分离方法相比,独立分量分析(ICA)是基于高阶统计特性的分析方法。在很多应用中,对高阶统计特性的分析更符合实际。另外,ICA 方法与传统的滤波方法和累加平均的方法[17]相比,ICA 在消除噪声的同时,对其它信号的细节几乎没有破坏,其去噪性能也往往要比传统的滤波方法好得多。

参考文献:

[1] HYVARINEN A,OJA E. Independent Component Analysis: Algoritbms and Application. [J]. Neural Networks,2000 (13):411-430.

[2] 彭煊,刘金福,王炳锡. 基于独立分量分析的语音增强[J]. 信号处理,2002,18(5):477-479.

[3] 曾生根,夏德深. 独立分量分析在多光谱遥感图像分类中的应用[J]. 计算机工程与应用,2004,21:108-110,145.

[4] 曾生根,朱宁波,包晔,等. 一种改进的快速独立分量分析算法及其在图象分离中的应用[J]. 中国图象图形学报,2003,8(10):1159-1165.

[5] 杨行峻,郑君里. 人工神经网络与盲信号处理[M]. 北京:清华大学出版社,2003.

[6] CAO X R,LIU R W. General Approach to Blind Source Separation. IEEE Trans[J]. Signal Processing, 1996,78: 753-766.

[7] 张旭秀,邱天爽. 独立分量分析原理及其应用[J]. 大连铁道学院学报,2003,24(2):64-68.

[8] 吴小培,冯焕清. 基于独立分量分析的图象分离技术及应用[J]. 中国图象图形学报,2001,6(2):133-137.

[9] HYVARINEN A. Independent Component Analysis:A Tutorial[EB/OL]. http://www. cis. hut. fi/aapo/papers/IJCNN99 _ tutorialweb/node15. html,2005-05-02.

[10] CARDPSP J F. Informax and Maximum Likelihood for Blind Source Separation[J]. IEEE Signal Proces- sing letters,1997,4:112-114.

[11] 刘喜武,刘洪,李幼铭. 独立分量分析及其在地震信息处理中应用初探[J]. 地球物理学进展, 2003,18(1): 90-96.

[12] 刘喜武,刘洪. 地震盲反褶积综述[J]. 地球物理学进展,2003,18(2):203-209.

[13] 何振亚,刘琚,杨绿溪,等. 盲均衡和信道参数估计的一种 ICA 和进化计算方法[J]. 中国科学(E 辑),2000,30(2):142-149.

[14] 杨士兴,应用独立成分分析方法分离集水区河川污染源之研究[EB/OL]. http://www. geog. ntu. edu. tw/main/paper/d81208004,2005-05-02.

[15] HYVARINEN A. PAJUNER P. Nonlinear Independent Component Analysis:Existence and Uniqueness Results [J]. Signal Processing,1998,64:301-313.

[16] HYVARINEN A. Independent Component Analysis in the Presence of Gaussian Noise by Maximizing Joint Likelihood [J]. Neuro Computing, 1998,22:49-67.

[17] 王开发,余小平. 具有周期和时滞的 Hopfield 型连续神经网络的周期解[J]. 重庆师范学院学报(自然科学版),1998,15(4):24-26.

(责任编辑　游中胜)