

文章编号: 2095-4980(2013)01-0146-06

改进白化的 MDL 快速独立分量分析算法

石文斌, 王大鸣, 崔维嘉, 仵国锋

(解放军信息工程大学 信息工程学院, 河南 郑州 450002)

摘要: 采用主分量分析法(PCA)进行的白化处理, 可能会错误估计信号子空间维数, 且未考虑噪声影响。提出了一种基于最小描述长度(MDL)准则信源个数估计改进白化的盲分离算法。通过信源个数估计确定信号子空间的维数, 区分信号与噪声子空间, 并估计噪声平均方差, 对信号特征值进行修正, 进而减小噪声影响, 提高算法分离性能。仿真表明, 在信噪比高于 5 dB 时, MDL 估计正确估计概率趋近于 1, 改进白化的 MDL 快速独立分量分析(FastICA)算法比经典 FastICA 算法分离性能有较为明显的提高。

关键词: 信源个数估计; 最小描述距离; 平均噪声方差; 相似系数

中图分类号: TN911.7

文献标识码: A

MDL-FastICA algorithm based on improved whitening

SHI Wen-bin, WANG Da-ming, CUI Wei-jia, WU Guo-feng

(Institute of Information Engineering, PLA Information Engineering University, Zhengzhou Henan 450002, China)

Abstract: The whitening processing by Principal Component Analysis(PCA) may inaccurately estimate the signal subspace dimension without considering the noise impact. This paper proposes a blind signal separation algorithm based on improved whitening by Minimum Description Length(MDL) for estimating the source signal number. It determines the subspace dimension based on the source signal number estimation, and distinguishes between signal and noise subspace. It can improve the separation performance. The eigenvalues of signal will be corrected by noise average variance estimation, thus it can reduce the impact of noise. Simulation results show that the exact estimate probability by MDL will approach to 1 at hyper-5dB SNR, and the separation performance of improved whitening MDL Fast Independent Component Analysis(FastICA) algorithm can be improved distinctly compared with that of traditional FastICA algorithm.

Key words: source signal number estimation; Minimum Description Length; noise average variance; similarity coefficient

在通信技术不断发展的今天, 无线环境也随之不断恶化, 信号的日益密集, 自然干扰和人为干扰的不断加剧, 使得信号在时域高度密集, 在频域发生混迭, 这使得传统信号处理方法面临挑战。作为通信的非协作方, 在先验信息缺失的情况下, 要在目标信号微弱、存在噪声或人为干扰等背景下, 从中获取有用信息变得越来越难。因此, 实现带噪微弱目标信号的盲分离问题是实现这一目标的先决条件。

盲信号分离就是根据观测到的混合信号确定一变换, 以恢复原始信号或信源^[1-2]。目前多采用独立分量分析(ICA)^[3-5]的方法, 即在源信号不可观测和信号混合方式未知的前提下, 根据输入源信号的统计独立性, 仅由混合后的观测信号检测并分离出源信号中各个独立成分的过程。由于独立分量分析所需源信号的先验知识极少且分离性能良好, 在众多领域已得到有效应用^[6-7]。目前, 较为成熟的盲分离算法有 FastICA^[3-5], JADE^[8], EASI^[9], 自然梯度算法^[10]等。

通常盲分离算法都要对观测信号进行预处理, 即零均值化和白化处理。零均值化处理使得观测信号均值为零, 即去除直流分量, 可简化后续数据处理。白化处理则为了消除观测信号之间的相关性并进行压缩。白化过程通常

收稿日期: 2012-08-20; 修回日期: 2012-09-10

基金项目: 国家科技重大专项资助项目(2011ZX03003-003-02, 2009ZX03003-008-02); 国家高技术研究发展计划(“863”计划)基金资助项目(2009AA011504)

利用 PCA^[11-12]技术实现,其本质是找出数据中的主要元素和结构,去除噪声和冗余,将原有数据降维。传统方法是先将数据协方差矩阵对角化,再根据对角线元素值的大小确定主元个数,然后将观测信号投影到信号子空间,从而实现消除冗余和降维的过程。

但当目标信号微弱,噪声信号较强时,目标信号对应特征值与噪声信号对应特征值之间差别将不再明显,仅从特征值大小上难以区分到底属于哪类特征值。因此会错误估计信号子空间维数及对应的特征向量。因此,本文对盲分离算法中的白化处理进行了改进。首先,利用 MDL^[13-14]算法估计出源信号个数,区分信号与噪声特征值;然后估计平均噪声特征值,并在信号特征值的基础上减去噪声特征值,达到减小噪声影响的目的;最后,将观测信号投影到对应信号子空间,进而实现目标信号盲分离。

1 改进白化的 MDL-FastICA 盲分离算法

1.1 问题模型

设观测信号 $\mathbf{X}=[x_1(t), x_2(t), \dots, x_m(t)]^T$ 为 $m \times N$ 矩阵, m 个观测中任一 X_i 都是由 n 个独立源信号 $\mathbf{S}=[s_1(t), s_2(t), \dots, s_n(t)]^T$ 线性混合后叠加噪声而成,若 $m \times N$ 噪声信号 $\mathbf{N}_0=[n_1(t), n_2(t), \dots, n_m(t)]^T$ 为高斯白噪声,混合矩阵 \mathbf{A} 为 $m \times n$, 则观测信号可表述为:

$$\mathbf{X} = \mathbf{AS} + \mathbf{N}_0 \quad (1)$$

其含义即为 n 个源信号通过混合后,再叠加噪声,得到 m 维的观测信号,其中 N 为采样点数。

如果假设零均值化后信号表示仍然如式(1),则观测信号自相关矩阵 Φ_x' 为:

$$\Phi_x' = \frac{\mathbf{XX}^T}{N-1} = \frac{(\mathbf{AS} + \mathbf{N}_0)(\mathbf{AS} + \mathbf{N}_0)^T}{N-1} = \frac{\mathbf{ASS}^T\mathbf{A}^T + \mathbf{ASN}_0^T + \mathbf{N}_0\mathbf{S}^T\mathbf{A}^T + \mathbf{N}_0\mathbf{N}_0^T}{N-1} = \Phi_x + \Phi_{N_0} + \frac{\mathbf{ASN}_0^T + \mathbf{N}_0\mathbf{S}^T\mathbf{A}^T}{N-1} \quad (2)$$

由于噪声信号为高斯白噪声,若噪声方差为 σ^2 ,且与源信号独立,则有: $\Phi_{N_0} = \sigma^2\mathbf{I}$, $\frac{\mathbf{SN}_0^T}{N-1} = 0$, $\frac{\mathbf{N}_0\mathbf{S}^T}{N-1} = 0$ 。

因此有:

$$\Phi_x' = \Phi_x + \sigma^2\mathbf{I} \quad (3)$$

若 Φ_x 的 m 个特征值 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n > \lambda_{n+1} = \lambda_{n+2} = \dots = \lambda_m = 0$, Φ_x' 的 m 个特征值为 $\mu_1 \geq \mu_2 \geq \dots \geq \mu_m$, 则有 $\mu_1 = \lambda_1 + \sigma^2$, $\mu_2 = \lambda_2 + \sigma^2$, \dots , $\mu_n = \lambda_n + \sigma^2$, $\mu_{n+1} = \sigma^2$, \dots , $\mu_m = \sigma^2$ 。则由 Φ_x 的 m 个特征向量张成的空间即为实际信号子空间。白化处理的关键就是获得该 m 个特征向量及特征值真实值的估计。但在弱信号条件下,由于计算误差、有限采样以及噪声等影响可能使 Φ_x' 的 m 个特征值出现负值,且信号与噪声特征值之间无明显区别,因此准确估计信号特征值及其个数成为白化成功与否的关键。

1.2 增加源信号个数估计的白化方法

白化处理过程中为准确估计信号子空间对应的特征值及其个数,引入了源信号个数估计及信号特征值修正的处理步骤。当自相关矩阵半正定时,出现的负特征值与对应噪声方差的物理意义不符,应当删除。源信号个数估计是为了确定信号空间维数且准确区分信号特征值和噪声特征值,而信号特征值修正则是为了减少噪声影响。此处选用 MDL^[13-14]准则来估计信源个数。

1) MDL 信源个数估计原理

用 Schwartzand 和 Rissanen 提出的基于信息论的模型 MDL 准则进行信源个数估计。给定观测信号数据 $\mathbf{X}=[x_1, x_2, \dots, x_N]^T$ 和概率密度函数含参形式 $f(\mathbf{X}|\hat{\Theta})$, 以 MDL 为准则选择与观测数据拟合最好的模型。对应模型为:

$$MDL = -\log f(\mathbf{X}|\hat{\Theta}) + \frac{1}{2}\varphi \log N \quad (4)$$

式中: $f(\mathbf{X}|\hat{\theta})$ 为 x_1, x_2, \dots, x_N 的联合概率密度函数; $\hat{\Theta}$ 为参数 Θ 的极大似然估计; φ 为 Θ 中独立参数个数。第 1 部分为 $f(\mathbf{X}|\Theta)$ 与 $f(\mathbf{X}|\hat{\Theta})$ 间基于 $K-L$ 距离最大似然参数估计模型;第 2 部分是为了保证估计的无偏性而引入的惩罚因子。

对于独立同分布高斯随机变量,其联合概率密度函数为:

$$f(x_1, x_2, \dots, x_N | \Theta) = \prod_{i=1}^N \frac{1}{\pi^k \det \Phi_x} \exp(-x_i^T \Phi_x^{-1} x_i) \quad (5)$$

去掉与 Θ 无关的项, 则得到似然函数如下:

$$L(\Theta) = -N \log \det \Phi_x - \text{tr}[\Phi_x^{-1}] \hat{\Phi}_x \quad (6)$$

其中 $\hat{\Phi}_x$ 为观测数据自相关函数的估计, 形式如下:

$$\hat{\Phi}_x = \frac{1}{N} \sum_{i=1}^N x_i x_i^T \quad (7)$$

若 Φ_x 的 m 个特征值为 $\mu_1 \geq \mu_2 \geq \dots \geq \mu_m$, 则噪声方差为:

$$\hat{\sigma}^2 = \frac{1}{m-k} \sum_{i=k+1}^m \hat{\mu}_i \quad (8)$$

对式(6)用最大似然估计, 可得:

$$L(\hat{\Theta}) = \log \left(\frac{\prod_{i=k+1}^m (\hat{\mu}_i)^{1/(m-k)}}{\frac{1}{m-k} \sum_{i=k+1}^m \hat{\mu}_i} \right)^{(m-k)N} \quad (9)$$

由 $\Theta^{(k)}$ 张成的空间自由度为 $k+1+2mk$, 但所有参数并不独立, 在正交化过程中, 自由度会减少 $k(k-1)$, 因此, $\Theta^{(k)}$ 中独立参数个数为:

$$\varphi = k+1+2mk - k(k+1) \approx k(2m-k) \quad (10)$$

从对应表达式为:

$$\text{MDL}(k) = -\log \left(\frac{\prod_{i=k+1}^m (\hat{\mu}_i)^{1/(m-k)}}{\frac{1}{m-k} \sum_{i=k+1}^m \hat{\mu}_i} \right)^{(m-k)N} + \frac{1}{2} k(2m-k) \log N \quad (11)$$

然后求 $\text{MDL}(k)_m = \min \{\text{MDL}(i), i=1, 2, \dots, m-1\}$, 则此时得到的 k 即为源个数估计 $\hat{n} = k$ 。

2) 考虑噪声影响的白化处理

通过信源个数估计可将观测数据自相关矩阵 Φ_x 的 m 个特征值 $\mu_1 \geq \mu_2 \geq \dots \geq \mu_m$ 进行划分, 即前 k 个特征值为 $\mu_1 = \lambda_1 + \sigma_1^2, \mu_2 = \lambda_2 + \sigma_2^2, \dots, \mu_k = \lambda_k + \sigma_k^2$, 而后 $(m-k)$ 个特征值, $\mu_{k+1} = \sigma_{k+1}^2, \dots, \mu_m = \sigma_m^2$ 对应为噪声特征值。因此, 对平均噪声方差做如下估计:

$$\bar{\sigma}^2 = \frac{1}{m-k} \sum_{i=k+1}^m \mu_i \quad (12)$$

并以此平均噪声方差对前 k 个信号特征值进行修正, 即:

$$\hat{\lambda}_i = \mu_i - \bar{\sigma}^2, \quad \hat{\lambda}_i = \mu_i - \bar{\sigma}^2, i=1, 2, \dots, k \quad (13)$$

各特征值对应特征向量所组成的矩阵为: $\mathbf{U}^H = \begin{bmatrix} \mathbf{U}_K \\ \mathbf{U}_{M-K} \end{bmatrix}$, 其中 \mathbf{U}_K 对应信号子空间, \mathbf{U}_{M-K} 对应噪声子空间。

经典白化处理是通过 PCA 实现的, 白化矩阵取 $\sum_K^{-1/2} \mathbf{U}_K^H$, 其中 $\sum_K = \text{diag}(\mu_1, \mu_2, \dots, \mu_k)$, 观测信号经过白化处理后则具有单位协方差矩阵, 即 $E[\sum_K^{-1/2} \mathbf{U}_K^H x(t) x^H(t) \mathbf{U}_K \sum_K^{-1/2}] = \mathbf{I}_K$ 。

而此处选用的白化矩阵为:

$$\mathbf{Q} = \sum_K^{-1/2} \mathbf{U}_K^H \quad (14)$$

式中 $\hat{\sum}_K = \text{diag}(\hat{\lambda}_1, \hat{\lambda}_2, \dots, \hat{\lambda}_k) = \sum_K - \bar{\sigma}^2 \mathbf{I}$, $\bar{\sigma}^2$ 为平均噪声方差, 以此来代替传统的白化矩阵。因而白化信号为:

$$\mathbf{z} = \mathbf{Q}\mathbf{x} \quad (15)$$

信号空间的特征值减去了噪声方差的估计均值, 从而有助于提高微弱目标信号的盲分离性能。

1.3 基于信源个数估计改进白化的 MDL-FastICA 算法

FastICA^[3-5]算法是由芬兰赫尔辛基大学 Hyvarinent 等人提出的一种基于定点迭代的盲分离算法,此处选用负熵作为高斯性度量。处理过程针对白化后的数据 z 进行。负熵定义如下:

$$J(z) = H(z_{\text{Gauss}}) - H(z) \quad (16)$$

其中,

$$H(z) = \int f(z) \log f(z) dz \quad (17)$$

由于负熵具有非负性,当且仅当 z 服从高斯分布时,负熵值为零,故可以此作为高斯性度量。为简化计算,采用多项式函数逼近概率密度函数,且将 $y = \mathbf{w}^T z$ 代入,则负熵可转化为:

$$J(z) \propto \{E[G(\mathbf{w}^T z)] - E[G(v)]\}^2 \quad (18)$$

在 $\|\mathbf{w}\|=1$ 的约束条件下,根据 Kuhn-Tucker 条件,最优点能满足下式:

$$F(\mathbf{w}) = E[zg(\mathbf{w}^T z)] - \beta \mathbf{w} = 0 \quad (19)$$

此处, β 为一定值, $g(\bullet)$ 为 $G(\bullet)$ 的导数。对 $F(\mathbf{w})$ 求取关于 \mathbf{w} 的梯度,并由牛顿法则可得如下迭代公式:

$$\mathbf{w}_{n+1} = \mathbf{w}_n - \frac{E[zg(\mathbf{w}_n^T z)] - \beta \mathbf{w}}{E[g'(\mathbf{w}_n^T z)] - \beta} \quad (20)$$

对该式简化可得 FastICA 计算式为:

$$\begin{aligned} \mathbf{w}^* &= E[zg(\mathbf{w}^T z)] - E[g'(\mathbf{w}^T z)]\mathbf{w} \\ \mathbf{w} &= \frac{\mathbf{w}^*}{\mathbf{w}} \end{aligned} \quad (21)$$

其中非线性函数 $g(\bullet)$ 可选取, $g_1(y) = \tanh(a_1 y)$, $g_2(y) = y(\exp(-y^2/2))$, $g_3(y) = y^3$ 其中之一,此处选择 $g_3(y) = y^3$ 为非线性函数。

因此,基于 MDL 改进白化的 FastICA 算法基本步骤如下:

- 1) 零均值化数据;
- 2) 依据 MDL 准则进行信源估计,得信源个数 m ;
- 3) 按照式(12)估计平均噪声方差,并依据式(13)~式(14)实现白化处理;
- 4) 设置迭代次数 $p=1$,并随机初始化权向量 \mathbf{w}_p ;
- 5) 令 $\mathbf{w}_p = E[zg(\mathbf{w}^T z)] - E[g'(\mathbf{w}^T z)]\mathbf{w}$;
- 6) $\mathbf{w}_p = \mathbf{w}_p - \sum_{i=1}^{p-1} (\mathbf{w}_p^T \mathbf{w}_i) \mathbf{w}_i$;
- 7) $\mathbf{w}_p = \mathbf{w}_p / \|\mathbf{w}_p\|$,若不收敛则返回步骤 5);
- 8) 令 $p = p+1$,若 $p \leq m$,则返回步骤 4),否则结束。

2 仿真分析

为验证以上算法的分离性能,设置以下仿真实验。此处选取源信号分别为超高斯型语音、亚高斯型正弦、chirp,2ASK,4PSK 及两者的混合源对算法进行仿真验证,采样点数为 1 000 点。由随机生成的 10×4 混合矩阵混合后,用于仿真实验。其中,信噪比在 10 dB~30 dB 范围内变化,在各信噪比下分别作 1 000 次蒙特卡罗实验,计算平均相似系数和平均输出信噪比,以此来衡量算法的分离性能。

为衡量算法的分离性能,采用分离信号输出信噪比以及与源信号之间的相似系数 ρ_y 作为评价指标^[15],其定义分别为:

$$R_{\text{SN},i} = 10 \log \left(\frac{\sum_{k=1}^K s_i^2(k)}{\sum_{k=1}^K [y_i(k) - s_i^2(k)]^2} \right) \quad (22)$$

$$\rho_{ij} = \frac{\left| \sum_{i=1}^N s_i(t)y_j(t) \right|}{\sqrt{\sum_{i=1}^N s_i^2(t) \sum_{i=1}^N y_j^2(t)}} \quad (23)$$

同时为衡量算法整体的分离性能，定义平均输出信噪比和评价相似系数为：

$$\bar{R}_{SN} = \frac{1}{n} \sum_{i=1}^n R_{SN,i} \quad (24)$$

$$\bar{\rho} = \frac{1}{n} \sum_{i=1}^n \max \{ \rho_{ij}, j = 1, 2, \dots, m \} \quad (25)$$

当平均相似系数 $\bar{\rho}$ 较大时，每一路分离信号与源信号的相似程度都较高，当 \bar{R}_{SN} 较大时，说明分离信号输出信噪比性能较好。通过以上指标计算即可说明算法整体的分离性能的好坏。仿真结果见图 1~图 3。

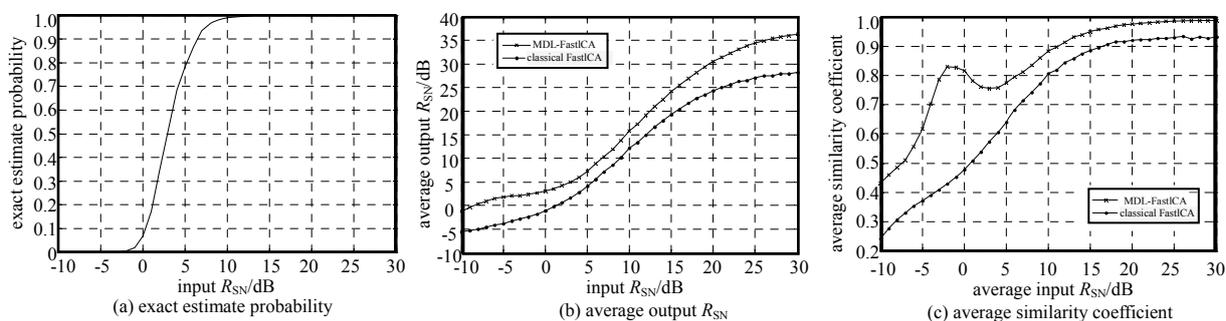


Fig.1 Super-Gaussian source simulation results

图 1 超高斯源仿真结果

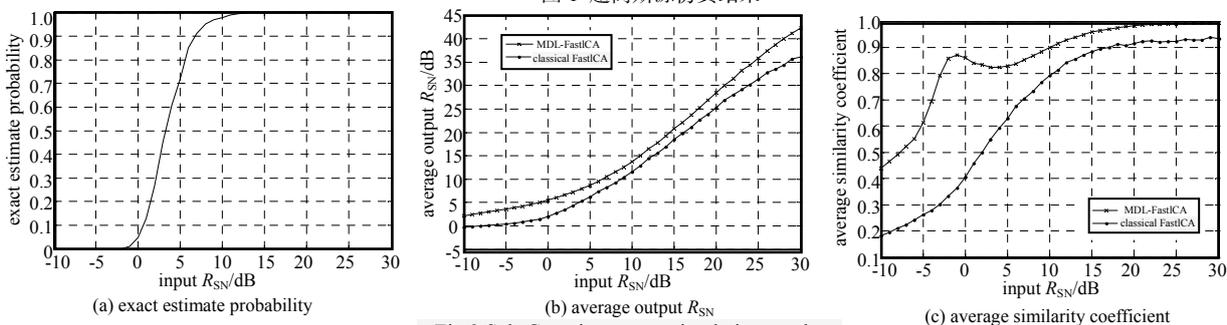


Fig.2 Sub-Gaussian source simulation results

图 2 亚高斯源仿真结果

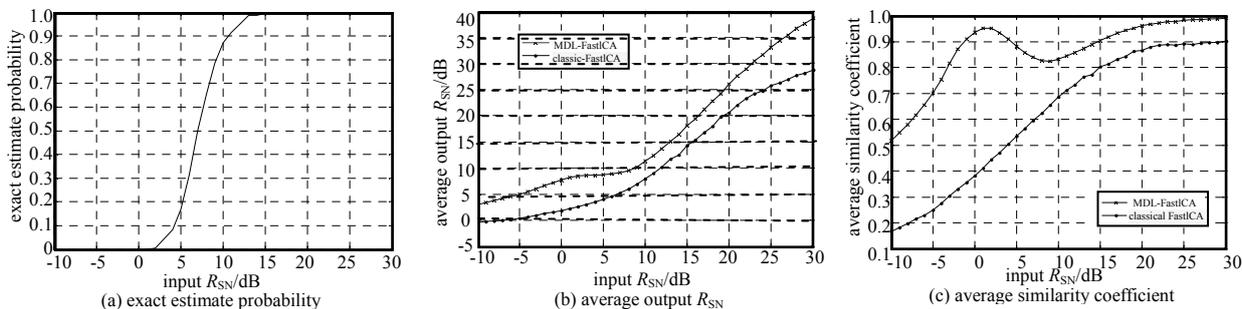


Fig.3 Gaussian mixture source simulation results

图 3 混合高斯源仿真结果

由 MDL 正确估计概率曲线可以看出，该估计方法在信噪比较低的情况下，实现正确估计的概率很低，但由于该估计为信源个数的无偏、一致估计，所以在信噪比高于 5 dB 后正确估计概率逐渐趋近于 1。

由平均输出信噪比及平均相似系数对比曲线可知，在不同高斯型源信号条件下，改进白化的 MDL-FastICA 算法分离性能提高较为明显。尤其当信噪比高于 5 dB 时，估计概率高，性能的提升幅度更为可观。而信噪比较低时，MDL 会出现少估现象，造成分离信号个数减少，这也造成了平均相似系数突然增大的现象，原因是仅分

离出性能最好的一个或少数几个源信号。但总体来说 MDL-FastICA 算法较传统采用 PCA 白化的 FastICA 算法而言,由平均输出信噪比和平均相似系数刻画的算法分离性能均获得了明显的提升。而分离性能的提升是以增加算法运算复杂度为代价,其中 MDL 估计运算复杂度较高,而特征值修正部分仅设计多次连加运算。

3 结论

算法在白化处理之前先利用 MDL 算法对信源个数进行估计,从而区分观测信号自相关矩阵特征值分解后的目标信号特征值和噪声特征值。然后对平均噪声方差进行估计,并修正信号特征值,接着进行白化处理,从而减小信号子空间维数估计错误及噪声所带来的影响。通过仿真实验表明,改进白化的 MDL-FastICA 算法在正确估计概率较高的情况下,目标信号分离性能比采用经典 PCA 白化的 FastICA 算法性能有较为明显的提高。

参考文献:

- [1] TITEN C,HERAULT J. Blind separation of sources,part I:an adaptive algorithm based on neuromimetic architecture[J]. Signal Processing, 1991,24(1):1-10
- [2] 张贤达,保铮. 盲信号分离[J]. 电子学报, 2001,29(12A):1766-1771.
- [3] Hyvarinen A,Oja E A. Independent component analysis: algorithm and applications[J]. Neural Networks, 2000,13(4):411-430.
- [4] Hyvarinen A. Fast and Robust Fixed-point Algorithms for Independent Component Analysis[J]. IEEE Trans. Neural Networks, 1999, 10(3):626-634.
- [5] Hyvarinen A,Oja E A. Fast Fixed-Point Algorithm for Independent Component Analysis[J]. Neural Computation, 1997,9(7): 1483-1492.
- [6] 付卫红,杨小牛,刘乃安,等. 独立分量分析(ICA)的通信信号盲侦察技术[J]. 四川大学学报, 2007,44(6):1245-1249.
- [7] 黄振川,杨小牛,张旭东. 基于独立分量分析的通信侦察复信号盲分离[J]. 清华大学学报, 2010,50(1):86-91.
- [8] Cardoso J F, Souloumiac A. Blind Beam Forming for Non-Gaussian Signals[J]. IEE Proceedings-F, 1993,140(6):362-370.
- [9] 牛奕龙,马建仓,王毅. 一种新的基于峰度的盲源分离算法[J]. 系统仿真学报, 2005,17(1):185-188.
- [10] Amari S. Natural gradient works efficiently in learning[J]. Neural Computation, 1998,10(2):251-276.
- [11] JOLLIFE I T. Principal Component Analysis[M]. New York:Springer, 1986.
- [12] Lindsay I Smith. A tutorial on Principal Components Analysis[EB/OL]. (2002-02-26)[2012-08-20]. http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.
- [13] Akaike H. Information theory and all extension of the maximum likelihood principle[C]// Problems of control and Inform Theory. Budapest:[s.n.], 1973:267-281
- [14] Schwartz Ct. Estimating the dimension of a model[J]. Ann. Stat, 1978,6(2):243-258.
- [15] 谢胜利,何昭水,高鹰. 信号处理的自适应理论[M]. 北京:科学出版社, 2006.

作者简介:



石文斌(1979-),男,甘肃省陇南市人,硕士,助理工程师,主要研究方向为无线与移动通信信号盲处理。email:shiwenbinwangxiao@163.com.

王大鸣(1971-),男,辽宁省大连市人,博士,副教授,主要研究方向为无线与移动通信。

崔维嘉(1976-),男,辽宁省丹东市人,博士,讲师,主要研究方向为自组织网、蜂窝网络融合技术。

仵国锋(1974-),男,河南省灵宝市人,讲师,主要研究方向为无线与移动通信。