Proceedings of the International MultiConference of Engineers and Computer Scientists 2010 Vol I,
IMECS 2010, March 17 - 19, 2010, Hong Kong

Statistical Inference for Independent Component Analysis Based on Polynomial Spline Model

Atsushi Kawaguchi *

Abstract— **This paper develops the confidence interval for the independent component analysis. The method is based on the bootstrap method using source density functions estimated by the polynomial splines modeling. A simulation study is conducted to show the numerical example for the proposed method and that the confidence interval has a reasonable coverage probability. Finally, the method is applied to a real fetal electrocardiogram data. One characteristic signal was effectively detected as a favor of the blind source separation by the proposed method.**

Keywords: Blind Source Separation, Bootstrap, Independent Component Analysis, Spline, Statistical Inference

1 Introduction

Independent component analysis (ICA) has been a powerful tool for blind source separation in many applications such as fetal heart monitoring [10] and functional magnetic resonance imaging (fMRI) analysis [2]. In ICA, the objective is to estimate the p by p mixing matrix $\mathbf{A}$ based on a random sample from the random vector given by $\mathbf{X} = \mathbf{AS}$ where $\mathbf{S}$ is a random vector of independently distributed components S_j, $j = 1, 2, \ldots, p$. Many different ICA algorithms have been developed, for example, Infomax [1],FastICA [5],KDICA [3],PSICA [7],In this paper, we assess the variability of the estimator of $\mathbf{A}$ due to a finite number of samples.

Most of studies for the variability of ICA are based on asymptotic variance (for example [9]). We propose to use the bootstrap resampling technique to asses the variability of ICA estimates. We consider the ICA estimator via PSICA algorithm [7].In this method, the logarithmic density functions of the each components of $\mathbf{S}$ are modeled by using polynomial splines. The random resample is drawn based on the estimated density functions. Our method may have the two advantages. First, this resampling method gives us to keep the structure of $\mathbf{S}$ against the method based on the empirical distribution function. Second, the bootstrap method can get us to be free from the assumption for asymptotic property of the ICA estimator.

*Biostatistics Center, Kurume University, 67 Asahi-Machi Kureme 830-0011, Japan. Email: kawaguchi atsushi@med.kurume-u.ac.jp

The remainder of the paper is organized as follows. The proposed method is described in Section 2. In Section 3, simulation studies are presented. A specific application is given in Section 4 where we report an effective separation from fetal electrocardiogram data.

2 Methods

First, we review the PSICA. Suppose each S_j has a density function f_j for $j = 1, 2, \ldots, p$. Then the density function of $\mathbf{X}$ can be expressed as $f_{\mathbf{X}}(\mathbf{x}) = \mathsf{det}(\mathbf{W}) \prod_{j=1}^{p} f_j(\mathbf{w}_j \mathbf{x})$, where $\mathbf{W} = \mathbf{A}^{-1}$ and $\mathbf{w}_j$ is the j-th row of $\mathbf{W}$. Each logarithmic density is modeled using polynomial splines

$$\log(f_j(x; \boldsymbol{\beta}_j)) = C(\boldsymbol{\beta}_j) + \beta_{j01} x + \sum_{i=1}^{m_j} \beta_{j1i} (x - r_{ji})_+^3, \quad (1)$$

where $\boldsymbol{\beta}_j = (\beta_{j01}, \beta_{j11}, \ldots, \beta_{j1m_j})$is a vector of coefficients, $C(\boldsymbol{\beta})$ is a normalized constant, r_{ji} are knots, and $(z)_+ = \max(0, z)$. We obtain the estimate of ($\mathbf{W}$, $\boldsymbol{\beta}_1$, $\ldots$, $\boldsymbol{\beta}_p$) by maximizing the likelihood of $\mathbf{X}$ with respect to ($\mathbf{W}$, $\boldsymbol{\beta}_1$, $\ldots$, $\boldsymbol{\beta}_p$). This is carried out in two steps to be described as follow. In the first step, $\mathbf{W}$ is treated as known and each f_j is estimated with data-dependent knots based on the logspline procedure. In the second step, we estimate $\mathbf{W}$ using the estimated f_j. We alternate these steps until convergence of $\mathbf{W}$. See [7] for the details.

Our bootstrap procedure to construct the confidence interval is as follows. First, the PSICA algorithm is applied to the original data $\mathbf{X}_1$, $\mathbf{X}_2$, $\ldots$, $\mathbf{X}_n$ and obtain the estimate $\widehat{\mathbf{A}}$. We obtain bootstrap samples S_j^* by generating random numbers from estimated f_j (j = 1, 2, $\ldots$, p). Using $\mathbf{S}^*$ consisting of S_j^* and $\mathbf{A}^* = \widehat{\mathbf{A}}$, we generate $\mathbf{X}^*$ where $\mathbf{X}^* = \mathbf{A}^*\mathbf{S}^*$. We apply the PSICA algorithm to $\mathbf{X}^*$ to obtain the estimate $\widehat{\mathbf{A}}^*$ of $\mathbf{A}^*$. The columns of $\widehat{\mathbf{A}}^*$ are permutated to agree with sign and order of columns of $\mathbf{A}^*$. This process is iterated by B times and the corresponding estimates $\widehat{\mathbf{A}}^{*1}$, $\widehat{\mathbf{A}}^{*2}$, $\ldots$, $\widehat{\mathbf{A}}^{*B}$ are obtained. Let denote (i, j)–elements of $\mathbf{A}$ and $\widehat{\mathbf{A}}^{*b}$ by a_{ij} and a_{ij}^{*b}, respectively (b = 1, 2, $\ldots$, B), we define the $100(1-\alpha)$ percent confidence limit of a_{ij} as $[\widehat{a}_{ij}^{(L_{ij})}, \widehat{a}_{ij}^{(U_{ij})}]$ where $\widehat{a}_{ij}^{(\alpha)}$ is the α percentile of a_{ij}^{*b} (b = 1, 2, $\ldots$, B), $L_{ij} = \Phi(\overline{r}_{ij} - z_\alpha \sigma_{ij})$, $U_{ij} = \Phi(\overline{r}_{ij} + z_\alpha \sigma_{ij})$,

ISBN: 978-988-17012-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

$\Phi(\cdot)$ and z_α are the cumulative distribution function and the $100(1-\alpha)$ percentile of the standard normal distribution, respectively, and $\overline{r}_{ij}$ and σ_{ij} are a mean and a standard deviation of normal scores for a_{ij}^{*b} ($b = 1, 2, \ldots, B$), respectively. The point estimate of $\mathbf{A}$ is defined as the matrix $\widehat{\mathbf{A}}^{(0.5)}$ whose elements are $a_{ij}^{(0.5)}$.

3 Simulation Studies

3.1 Numerical Example

The methods in Section 2 have their application illustrated for a simulated data. The true three independent source components S_1, S_2 and S_3 (p=3) with $n = 1000$ as the sample size are randomly generated from distributions represented as solid lines in Figure 1.

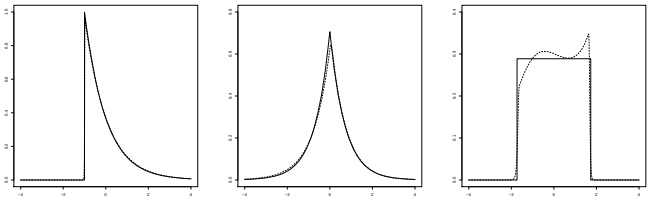

Figure 1: True density function for independent components (solid) and estimated density (dot) (S_1, S_2 and S_3 from the left)

Each elements of true 3 by 3 mixing $\mathbf{A}$ is randomly generated from an uniform distribution with the range [–1, 1], while the (3, 1)–element is set to be 0. Data $\mathbf{X} = (X_1, X_2, X_3)'$ to be applied are generated using the equation (2) with independent components S_1, S_2 and S_3 and the true mixing matrix $\mathbf{A}$.

$$\mathbf{X} = \mathbf{AS} \tag{2}$$
$$= \begin{pmatrix} -0.469 \\ 0.816 \\ 0.000 \end{pmatrix} S_1 + \begin{pmatrix} -0.256 \\ -0.597 \\ 0.322 \end{pmatrix} S_2 + \begin{pmatrix} 0.146 \\ 0.797 \\ 0.258 \end{pmatrix} S_3$$

The plots for independent components and observed data are shown in the left and middle panel of Figure 2, respectively. The proposed method with $B = 100$ bootstrap samples is applied to the data. The estimated densities which are used to generate the bootstrap samples are represented with dotted lines in Figure 1.

The 95% confidence intervals constructed by the proposed method were represented in Table 1 whose each elements of $\mathbf{A} = \{a_{ij}\}$ was (lower, upper) limits. All intervals contained the true values.

The reproduced sources via $\widehat{\mathbf{S}} = \mathbf{X}\widehat{\mathbf{A}}^{(0.5)}$ are shown in the right panel of Figure 2, where a (3,1)–element of $\widehat{\mathbf{A}}^{(0.5)}$ was replaced with 0 according to the resulting confidence interval. We may say that the reproducing was quite good comparing with the true sources in the left panel of Figure 2. All Pearson's correlation coefficients between the estimates and the true sources were 0.999.

Table 1: 95% confidence intervals for each element a_{ij}

	95%CI		
(i, j)	(Lower,	Upper)	True
(1, 1)	(−0.513,	−0.430)	−0.469
(1, 2)	(−0.279,	−0.247)	−0.256
(1, 3)	(0.122,	0.155)	0.146
(2, 1)	(0.706,	0.912)	0.816
(2, 2)	(−0.676,	−0.585)	−0.597
(2, 3)	(0.780,	0.860)	0.797
(3, 1)	(−0.004,	0.040)	0.000
(3, 2)	(0.322,	0.361)	0.322
(3, 3)	(0.220,	0.258)	0.258

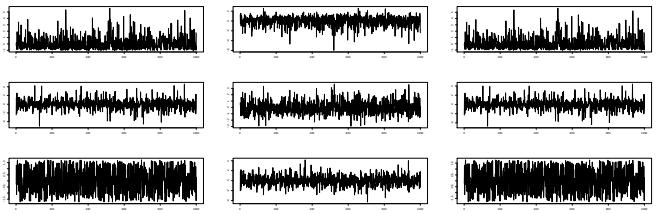

Figure 2: True sources (left: S_1, S_2, S_3 from the top), observed data (middle: X_1, X_2, X_3 from the top), estimated sources (right: $\widehat{S}_1$, $\widehat{S}_2$, $\widehat{S}_3$ from the top)

3.2 Coverage Probability

We demonstrate evaluate the proposed confidence interval by estimating the coverage probability. We generated randomly independent components from same densities as previous section and created observed data following the equation (2) with the sample size $n = 100$, 250, 500, 1000, 2000. The proposed method is applied and the 95 % confidence intervals for the elements from the simulations have transformations to indicator variables with the value of 1 if containing the true and the value of 0 otherwise. The means of these indicator variables across 200 of the simulations are the resulting estimates of the coverage probability. The results from the simulations pertaining to the coverage probability are shown in Figure 3. We may say that there is somewhat departure from the nominal level of 0.95 in small sample sizes, but the proposed confidence intervals have a reasonable coverage probability with increase in sample size.

ISBN: 978-988-17012-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

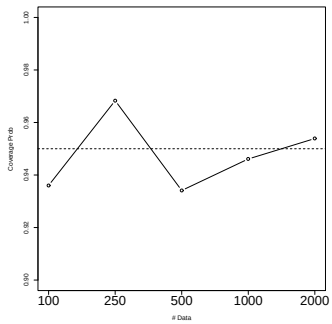

Figure 3: Coverage probabilities

4 Application

The method described in the Section 2 is applied to analyze cutaneous potential recordings data of a pregnant woman measured at 8-chanels with 2500 time points. The data is available on the ICA CENTRAL website. The objective of the analysis is to extract the fetal electrocardiogram which may have a fractal structure (for example, [8]).We estimate the correlation dimension which is one of measurements for a fractal dimension by the method proposed by [6]from both the raw signals and the estimated sources. The signals were embedded into the vectors $\mathbf{Y}_t = (Y_t, Y_{t-1}, \ldots, Y_{t-d+1})'$ where t is a time point and d is an embedding dimension. Figure 4 shows the result for estimation of correlation dimension for each d for the raw data (left) and estimated sources (right). The saturation of lines indicates that the signal has a fractal structure [4].

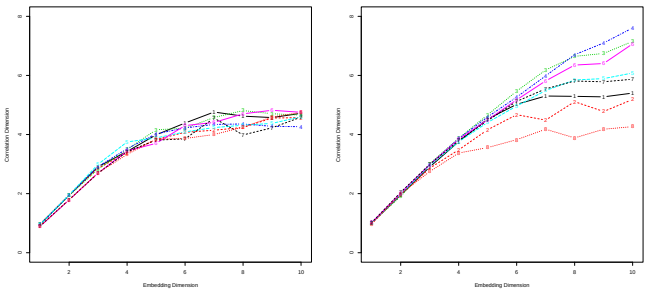

Figure 4: Estimation of correlation dimension (left: observed data, right: estimated sources)

The eighth component of estimated sources may have more remarkable saturation than others. This indicates that this estimated source is most close to the fetal heart rate variability. On the other hand, the result for all components of raw signals showed the saturation and such clear difference among components has not noted. This might come from a domination of fractal structure of the fetal heart rate variabilities which contain in all components of raw signals. However we can expect that these potentially also contain mother's heart rate variability and other artifacts.

Acknowledgement

The author appreciates the basic idea and the constructive comments of Dr. Young K. Truong. This work was supported by Grant-in-Aid for Young Scientists (B) (21700312).

References

[1] Bell, A. J., Sejnowski, T. J., "An information maximisation approach to blind separation and blind deconvolution," *Neural Computation*, V7, N6, pp. 1129-1159, 11/95

[2] Calhoun, V.D., Adali, T., "Unmixing fMRI with independent component analysis," *Engineering in Medicine and Biology Magazine, IEEE*, V25, N2, pp. 79-90, 3-4/06

[3] Chen, A., Bickel, P.J., "Efficient independent component analysis," *Annals of Statistics*, V34, N6, pp. 2825-2855, 12/06

[4] Grassberger, P., Procaccia, I., "Characterization for Strange Attractors," *Physical Review Letters*, V50, N5, pp. 346-349, 1/83

[5] Hyvärinen, A., Oja, E., "A fast fixed point algorithm for independent component analysis," *Neural Computation*, V9, N7, 1483-1492, 10/97

[6] Kawaguchi, A., "Estimating the correlation dimension from chaotic dynamical systems by U-statistics, " *Bulletin of Informatics and Cybernetics*, V34, N2, pp. 143-150, 12/02.

[7] Kawaguchi, A., Truong, Y. "Spline Independent Component Analysis, " Manuscript. [http://www.bios.unc.edu/˜truong/MRI/sica.pdf]

[8] Okamura, K, "Know the Fetus, See the Fetus (in Japanese)," *Acta Obstetrica et Gynaecologica Japonica*, V60, N8, pp. 1567-1576, 8/08.

[9] Shimizu, S., Hyvärinen, A., Kano, Y., Hoyer, P. O. and Kerminen, A. J., "Testing significance of mixing and demixing coefficients in ICA," *Proc. International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2006)*, Charleston, SC, USA. 3/06

[10] Stone, J. V., *Independent component analysis: a tutorial introduction*, MIT Press, 2004.

ISBN: 978-988-17012-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)