Gaussian Moments for Noisy Independent Component Analysis

Aapo Hyvärinen
Helsinki University of Technology
Laboratory of Computer and Information Science
P.O. Box 5400, FIN-02015 HUT, Finland
aapo.hyvarinen@hut.fi
http://www.cis.hut.fi/~aapo/

February 19, 1999

Abstract

A novel approach for the problem of estimating the data model of independent component analysis (or blind source separation) in the presence of gaussian noise is introduced. We define the gaussian moments of a random variable as the expectations of the gaussian function (and some related functions) with different scale parameters, and show how the gaussian moments of a random variable can be estimated from noisy observations. This enables us to use gaussian moments as one-unit contrast functions that have no asymptotic bias even in the presence of noise, and that are robust against outliers. To implement the maximization of the contrast functions based on gaussian moments, a modification of the fixed-point (FastICA) algorithm is introduced.

EDICS number: SPL.SP.3.2 (higher-order statistical analysis)

1 Introduction

Indendent component analysis [1, 6] is a statistical model where the observed data is expressed as a linear transformation of latent variables ('independent components') that are nongaussian and mutually independent. We may express the model as

$$\mathbf{x} = \mathbf{A}\mathbf{s} + \mathbf{n} \tag{1}$$

where $\mathbf{x} = (x_1, x_2, ..., x_m)$ is the vector of observed random variables, $\mathbf{s} = (s_1, s_2, ..., s_n)$ is the vector of the independent components, and $\mathbf{A}$ is an unknown constant matrix, called the mixing matrix. The vector $\mathbf{n}$ is noise, and is most often omitted. For simplicity, we make in this paper some assumptions that are not strictly necessary: 1) the dimension of $\mathbf{s}$ equals the dimension of $\mathbf{x}$, i.e. $n = m$, 2) the noise $\mathbf{n}$ is gaussian and 3) the noise covariance matrix $\mathbf{\Sigma}$ is known.

A popular approach for estimating the noise-free ICA model is the one-unit (or deflation) method [2, 3, 4, 5]. Denote the noise-free data by $\mathbf{y} = \mathbf{A}\mathbf{s}$. The basic idea in the one-unit approach is to take some measure of nongaussianity and then find projections, say $\mathbf{w}^T\mathbf{y}$, in which this is locally maximized for whitened (sphered) data. Projections in such directions give consistent estimates of the independent components, if the measure of nongaussianity is well chosen. This approach could be used for noisy ICA as well, if only we had measures of nongaussianity which are immune to gaussian noise, or at least, whose values for the original data can be easily estimated from noisy observations. If the measure of nongaussianity is kurtosis [2] (the fourth-order cumulant), this is quite easy, but leads to nonrobust algorithms. The purpose of this letter is to show how to construct corresponding algorithms

for noisy ICA using other one-unit contrast functions than kurtosis. This is based on the concept of gaussian moments. Thus we introduce a new class of algorithms for noisy ICA that are consistent and robust against outliers.

2 Quasi-whitening

To begin with, it must noted that the effect of noise must be taken into account in the preliminary whitening of the data. This is quite simple if the noise covariance matrix is known, as we assume. Denoting by $\mathbf{C} = E\{\mathbf{x}\mathbf{x}^T\}$ the covariance matrix of the observed noisy data, the ordinary whitening should be replaced by the operation $\tilde{\mathbf{x}} = (\mathbf{C} - \mathbf{\Sigma})^{-1/2}\mathbf{x}$. In the following, we call this operation 'quasi-whitening'. The quasi-whitened data $\tilde{\mathbf{x}}$ follows a noisy ICA model as well, with an orthogonal mixing matrix [1, 2], and the following noise covariance matrix:

$$\tilde{\mathbf{\Sigma}} = E\{\tilde{\mathbf{n}}\tilde{\mathbf{n}}^T\} = (\mathbf{C} - \mathbf{\Sigma})^{-1/2}\mathbf{\Sigma}(\mathbf{C} - \mathbf{\Sigma})^{-1/2}. \tag{2}$$

3 Gaussian moments

It has been argued e.g. in [3, 5] that kurtosis may be a rather poor measure of nongaussianity (contrast function) in many applications. This is because it gives estimators that are very sensitive to outliers, and may have large mean-square errors. Therefore, in [3, 5] an approach was developed in which the higher-order statistics of the projection $\mathbf{w}^T\mathbf{y}$ are taken into account through general contrast functions of the form

$$J_G(\mathbf{w}^T\mathbf{y}) = |E\{G(\mathbf{w}^T\mathbf{y})\} - E\{G(\nu)\}| \tag{3}$$

where the function G is a sufficiently regular nonquadratic function, and ν is a standardized gaussian variable.

The main point of this paper is to show that for certain choices of G, it is simple to estimate the values of J_G consistently from noisy observations, generalizing this approach for noisy ICA. The basic idea is to choose G to be the density function of a zero-mean gaussian random variable, or a related function.

Denote by

$$\varphi_c(x) = \frac{1}{c}\varphi(\frac{x}{c}) = \frac{1}{\sqrt{2\pi}c}\exp(-\frac{x^2}{2c^2}) \tag{4}$$

the gaussian density function of variance c^2, and by $\varphi_c^{(k)}(x)$ the k-th ($k > 0$) derivative of $\varphi_c(x)$. Denote further by $\varphi_c^{(-k)}$ the k-th integral function of $\varphi_c(x)$, obtained by $\varphi_c^{(-k)}(x) = \int_0^x \varphi_c^{(-k+1)}(\xi)d\xi$, where we define $\varphi_c^{(0)}(x) = \varphi_c(x)$. (The lower integration limit 0 is here quite arbitrary, but has to be fixed.) Then we have the following theorem (proven in Appendix A):

Theorem 1 *Let z be any nongaussian random variable, and denote by n an independent gaussian noise variable of variance σ^2. Define the gaussian function φ as in (4). Then for any constant $c > \sigma^2$ we have*

$$E\{\varphi_c(z)\} = E\{\varphi_d(z+n)\} \tag{5}$$

with $d = \sqrt{c^2 - \sigma^2}$. Moreover, (5) still holds when φ is replaced by $\varphi^{(k)}$ for any integer index k.

We call the statistics of the form $E\{\varphi_c^{(k)}(\mathbf{w}^T\mathbf{y})\}$ the *gaussian moments* of the data. Thus we can estimate the noisy ICA model by maximizing, for quasi-whitened data $\tilde{\mathbf{x}}$, the following contrast function:

$$\max_{\|\mathbf{w}\|=1} |E\{\varphi_{d(\mathbf{w})}^{(k)}(\mathbf{w}^T\tilde{\mathbf{x}})\} - E\{\varphi_c^{(k)}(\nu)\}| \tag{6}$$

with $d(\mathbf{w}) = \sqrt{c^2 - \mathbf{w}^T \tilde{\mathbf{\Sigma}} \mathbf{w}}$. This gives a consistent (i.e. asymptotically unbiased) method of estimating the noisy ICA model due to the theorem in [5].

4 Fixed-point algorithm for gaussian moments

To perform the optimization in (6), we can derive a modification of the (general form of the) fixed-point, or FastICA, algorithm [3, 4]. A detailed derivation will be presented elsewhere. An important point in the derivation is that the algorithm can be considerably simplified by adapting the value of c at every iteration, e.g. so that $d(\mathbf{w}) = 1$ always. At the same time, this solves the problem of choosing values for the parameter c. Such an adaptation of c is justified by the fact that the function G needs only to be of a given shape, so that the signs of certain non-polynomial cumulants do not change [5].

This gives the following *fixed-point algorithm with bias removal* for quasi-whitened data:

$$\mathbf{w}^* = E\{\tilde{\mathbf{x}} g(\mathbf{w}^T \tilde{\mathbf{x}})\} - (\mathbf{I} + \tilde{\mathbf{\Sigma}}) \mathbf{w} E\{g'(\mathbf{w}^T \tilde{\mathbf{x}})\} \tag{7}$$

where $\mathbf{w}^*$, the new value of $\mathbf{w}$, is normalized to unit norm after every iteration, and $\tilde{\mathbf{\Sigma}}$ is given by (2). Surprisingly, (7) is of the same form as the corresponding algorithm that maximizes the modulus of kurtosis for quasi-whitened data (obtained by modifying slightly the derivation in [4]). The function g is here the derivative of G, and can thus be choosen, for example, among the following:

$$g_1(u) = \tanh(u), \quad g_2(u) = u \exp(-u^2/2), \quad g_3(u) = u^3, \tag{8}$$

where g_1 is an approximation of $\varphi^{(-1)}$, which is the gaussian cdf (these relations hold up to some irrelevant constants), g_2 equals $\varphi^{(1)}$, and g_3 is obtained by using kurtosis. These functions cover essentially the nonlinearities ordinarily used in the fixed-point or FastICA algorithm [3]. It can be seen that the addition of $\tilde{\mathbf{\Sigma}}$ in (7) is the key to removing bias. Several independent components can be found using different orthogonalization schemes, exactly as in the noise-free case [4].

5 Simulations

To test our algorithm in (7), we made the simulations depicted in Fig. 1. The dimension of the data was 4, the independent components had i.i.d. Laplace distributions, and noise covariance was $.25\,\mathbf{I}$, At each trial, a 4×4 mixing matrix was randomly generated, and normalized so that the total energy of the signals was equal to 1, corresponding to a signal-to-noise ratio of 4. Only one independent component was estimated at each trial, and the resulting error was measured as: error $= \min_i |1 - |\mathbf{w}^T \mathbf{a}_i| / \|\mathbf{w}^T \mathbf{A}\||$, where $\mathbf{a}_i$ is the i-th row of the mixing matrix after quasi-whitening. Sample size N was varied from 1000 to 64000, and the error was estimated as the median of the errors of 200 trials. The results are depicted in Fig. 1. One can see clearly that for the modified estimators, the errors tend to zero, showing lack of bias. This is not the case for the original estimators, which show considerable bias.

6 Conclusion

We introduced a new approach to estimation of the noisy ICA model, using the concept of gaussian moments. The useful property of gaussian moments

is that the gaussian moments of underlying random variables can be simply estimated from noisy observations. Higher-order cumulants have the same property, but they lead to estimators that are sensitive to outliers. Thus we derived a fixed-point (FastICA) algorithm [4, 3] for noisy ICA that is statistically consistent, i.e. without asymptotic bias, and robust against outliers (for suitable choice of g, and for robustly whitened data). Moreover, it inherits from the noise-free fixed-point algorithm the advantages of being computationally simple and very fast [4, 3].

References

[1] P. Comon. Independent component analysis – a new concept? *Signal Processing*, 36:287–314, 1994.

[2] N. Delfosse and P. Loubaton. Adaptive blind separation of independent sources: a deflation approach. *Signal Processing*, 45:59–83, 1995.

[3] A. Hyvärinen. A family of fixed-point algorithms for independent component analysis. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'97)*, pages 3917–3920, Munich, Germany, 1997.

[4] A. Hyvärinen and E. Oja. A fast fixed-point algorithm for independent component analysis. *Neural Computation*, 9(7):1483–1492, 1997.

[5] A. Hyvärinen and E. Oja. Independent component analysis by general nonlinear Hebbian-like learning rules. *Signal Processing*, 64(3):301–313, 1998.

[6] C. Jutten and J. Herault. Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24:1–10, 1991.

A Proof of Theorem

Denote by $p(.)$ the pdf of z. For $k = 0$, we have

$$E\{\varphi_d(z+n)\} = \int \varphi_d(y)[\int \varphi_\sigma(y-t)p(t)dt]dy$$
$$= \int p(t)[\int \varphi_\sigma(y-t)\varphi_d(y)dy]dt = E\{\varphi_c(z)\} \quad (9)$$

which proves the theorem for $k = 0$. For other values of k, introduce a hypothetical location parameter θ. Taking the k-th derivative (resp. integral) under the expectation of the both sides of $E\{\varphi_c(z+\theta)\} = E\{\varphi_d(z+n+\theta)\}$, and setting $\theta = 0$, we obtain the theorem for $k > 0$ (resp. $k < 0$).

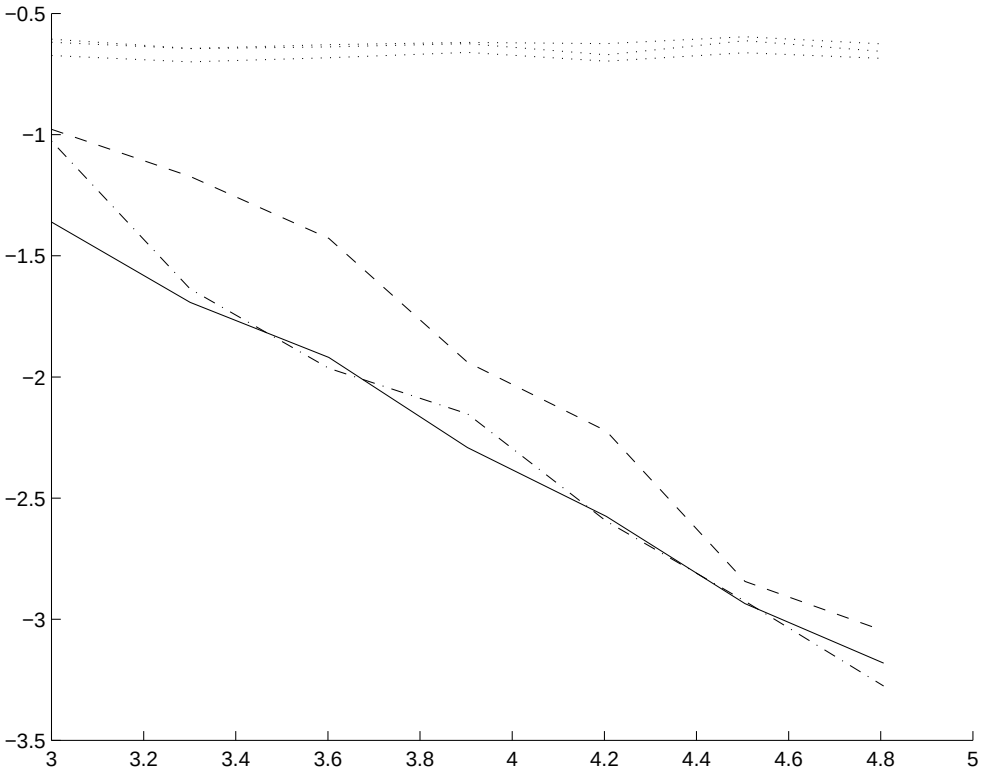

Figure 1: Consistency of the estimators for fixed noise level (SNR=4) and sample size varying from 1000 to 64000. Horizontal axis: log10 of sample size. Vertical axis: log10 of error measure as given in text. Dotted lines: estimators without bias correction, for the three nonlinearities in (8). Other lines: estimators with bias correction (solid: g_3, dashed: g_2, dot-dashed: g_1). Only the estimators with bias correction have errors that tend to zero.

Figure caption:

Consistency of the estimators for fixed noise level (SNR=4) and sample size varying from 1000 to 64000. Horizontal axis: log10 of sample size. Vertical axis: log10 of error measure as given in text. Dotted lines: estimators without bias correction, for the three nonlinearities in (8). Other lines: estimators with bias correction (solid: g_3, dashed: g_2, dot-dashed: g_1). Only the estimators with bias correction have errors that tend to zero.