Estimating Squared-loss Mutual Information for Independent Component Analysis

Taiji Suzuki[1] and Masashi Sugiyama[2]

[1] Department of Mathematical Informatics, The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
s-taiji@stat.t.u-tokyo.ac.jp
[2] Department of Computer Science, Tokyo Institute of Technology
2-12-1 O-okayama, Meguro-ku, Tokyo 152-8552, Japan
sugi@cs.titech.ac.jp

Abstract. Accurately evaluating statistical independence among random variables is a key component of Independent Component Analysis (ICA). In this paper, we employ a squared-loss variant of mutual information as an independence measure and give its estimation method. Our basic idea is to estimate the *ratio* of probability densities directly without going through density estimation, by which a hard task of density estimation can be avoided. In this density-ratio approach, a natural cross-validation procedure is available for model selection. Thanks to this, all tuning parameters such as the kernel width or the regularization parameter can be objectively optimized. This is a highly useful property in unsupervised learning problems such as ICA. Based on this novel independence measure, we develop a new ICA algorithm named *Least-squares Independent Component Analysis* (LICA).

Key words: Independent component analysis, Mutual information, Squared loss, Density ratio estimation, Cross-validation.

1 Introduction

The purpose of Independent Component Analysis (ICA) [1] is to obtain a transformation matrix that separates mixed signals into statistically-independent sources signals. A direct approach to ICA is to find a transformation matrix such that independence among separated signals is maximized under some independence measure such as *mutual information* (MI).

Various approaches to computing the independence measures from samples have been studied so far. A naive approach is to estimate probability densities based on parametric or non-parametric density estimation. However, finding an appropriate parametric model is not easy without strong prior knowledge and non-parametric estimation is not accurate in high-dimensional problems. Thus this naive approach is not so useful in practice. Another approach is to approximate the *negative entropy* based on the Gram-Charlier expansion [6, 7, 3] or the Edgeworth expansion [5]. An advantage of this approach is that a

Table 1. Summary of existing and proposed ICA methods.

	Model selection	Distribution
Fast ICA (FICA) [2]	**Not Necessary**	Not Free
Natural-gradient ICA (NICA) [3]	**Not Necessary**	Not Free
Kernel ICA (KICA) [4]	Not Available	**Free**
Edgeworth-expansion ICA (EICA) [5]	**Not Necessary**	Nearly normal
Least-squares ICA (LICA) [proposed]	**Available**	**Free**

hard task of density estimation is not directly involved. However, these expansion techniques are based on the assumption that the target density is close to normal and violation of this assumption can cause large approximation error.

The above approaches are based on the probability densities of signals. Another line of research that does not explicitly involve probability densities employs *non-linear correlation*—signals are statistically independent if and only if all non-linear correlations among the signals vanish. Following this line, computationally efficient algorithms have been developed based on the fourth-order statistics [8, 2]. However, these methods ignore higher-order correlation and thus could be inaccurate depending on the target distribution. To cope with this problem, the *kernel trick* has been applied in ICA [4], which allows us to evaluate all non-linear correlations efficiently. However, its practical performance depends on the choice of kernels (more specifically, the Gaussian kernel width) and there seems no theoretically-justified method to determine the kernel width. This is a critical problem in unsupervised learning problem such as ICA.

In this paper, we use a squared-loss variant of MI as an independence measure and give a novel method for estimating it. Our key idea is to estimate the ratio of probability densities contained in squared-loss MI (SMI) directly without going through density estimation. This allows us to avoid a hard task of density estimation. Another practically important advantage of this density-ratio approach is that a natural cross-validation (CV) procedure is available for model selection. Thus all tuning parameters such as the kernel width or the regularization parameter can be objectively optimized through CV.

From an algorithmic point of view, the density-ratio approach *analytically* provides a non-parametric estimate of SMI; furthermore its derivative can also be computed analytically and these useful properties are utilized in deriving a new ICA algorithm—the proposed method is named *Least-squares Independent Component Analysis* (LICA). Characteristics of existing and proposed ICA methods are summarized in Tab. 1, highlighting the advantage of the proposed LICA approach.

2 SMI Estimation for ICA

In this section, we formulate the ICA problem and introduce our independence measure, SMI. Then we give an estimation method of SMI and based on it we derive an ICA algorithm.

2.1 Problem Formulation

Suppose there is a d-dimensional random signal

$$\boldsymbol{x} = (x^{(1)}, \ldots, x^{(d)})^\top$$

drawn from a distribution with density $p(\boldsymbol{x})$, where $\{x^{(m)}\}_{m=1}^d$ are statistically independent of each other. Thus, $p(\boldsymbol{x})$ can be factorized as

$$p(\boldsymbol{x}) = \prod_{m=1}^{d} p_m(x^{(m)}).$$

We cannot directly observe the *source* signal $\boldsymbol{x}$, but a linearly mixed signal $\boldsymbol{y}$:

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x},$$

where $\boldsymbol{A}$ is a $d \times d$ invertible matrix called the *mixing matrix*. The goal of ICA is, given mixed signal samples $\{\boldsymbol{y}_i\}_{i=1}^n$, to obtain a *demixing matrix* $\boldsymbol{W}$ that recovers the original source signal $\boldsymbol{x}$—we denote the demixed signal by $\boldsymbol{z}$:

$$\boldsymbol{z} = \boldsymbol{W}\boldsymbol{y}.$$

The ideal solution is given by $\boldsymbol{W} = \boldsymbol{A}^{-1}$, but we can only recover it up to permutation and scaling of components of $\boldsymbol{x}$ due to non-identifiability of the ICA setup [1].

A direct approach to ICA is to determine $\boldsymbol{W}$ so that components of $\boldsymbol{z}$ are as independent as possible. Here, we adopt SMI as the independence measure:

$$I_s(Z^{(1)}, \ldots, Z^{(d)}) := \frac{1}{2} \int \left(\frac{q(\boldsymbol{z})}{r(\boldsymbol{z})} - 1 \right)^2 r(\boldsymbol{z}) d\boldsymbol{z}, \tag{1}$$

where $q(\boldsymbol{z})$ denotes the joint density of $\boldsymbol{z}$ and $r(\boldsymbol{z})$ denotes the product of marginal densities $\{q_m(z^{(m)})\}_{m=1}^d$:

$$r(\boldsymbol{z}) = \prod_{m=1}^{d} q_m(z^{(m)}).$$

Since I_s vanishes if and only if $q(\boldsymbol{z}) = r(\boldsymbol{z})$, the degree of independence among $\{z^{(m)}\}_{m=1}^d$ may be measured by SMI. Note that Eq.(1) corresponds to the *f-divergence* from $q(\boldsymbol{x})$ to $r(\boldsymbol{z})$ with the squared loss, while ordinary MI corresponds to the f-divergence with the log loss. Thus SMI could be regarded as a natural generalization of ordinary MI.

Based on the independence detection property of SMI, we try to find the demixing matrix $\boldsymbol{W}$ that minimizes SMI estimated from the demixed samples:

$$\{\boldsymbol{z}_i \mid \boldsymbol{z}_i = (z_i^{(1)}, \ldots, z_i^{(d)})^\top := \boldsymbol{W}\boldsymbol{y}_i\}_{i=1}^n.$$

Our key constraint when estimating SMI is that we want to avoid density estimation. Below, we show how this could be accomplished.

2.2 SMI Inference via Density Ratio Estimation

Using *convex duality* [9], we can express SMI as

$$I_s(Z^{(1)},\dots,Z^{(d)}) = \sup_g \left[\int \Big(g(\boldsymbol{z})q(\boldsymbol{z}) - \frac{1}{2}g(\boldsymbol{z})^2 r(\boldsymbol{z})\Big)d\boldsymbol{z} - \frac{1}{2}\right], \tag{2}$$

where $\sup_g$ is taken over all measurable functions. Thus computing I_s is reduced to solving the following optimization problem:

$$\inf_g \left[\int \Big(\frac{1}{2}g(\boldsymbol{z})^2 r(\boldsymbol{z}) - g(\boldsymbol{z})q(\boldsymbol{z})\Big)d\boldsymbol{z}\right]. \tag{3}$$

We can confirm that the optimal solution g^* of the problem (3) is given as

$$g^*(\boldsymbol{z}) = \frac{q(\boldsymbol{z})}{r(\boldsymbol{z})}. \tag{4}$$

Thus, solving the problem (3) amounts to inferring the density ratio (4).

However, directly solving the problem (3) is not possible due to the following two reasons. The first reason is that finding the minimizer over all measurable functions is not tractable in practice since the search space is too vast. To overcome this problem, we restrict the search space to some linear subspace $\mathcal{G}$:

$$\mathcal{G} = \{\boldsymbol{\alpha}^\top \boldsymbol{\varphi}(\boldsymbol{z}) \mid \boldsymbol{\alpha} = (\alpha_1,\dots,\alpha_b)^\top \in \mathbb{R}^b\}, \tag{5}$$

where $\boldsymbol{\alpha}$ is a parameter to be learned from samples, $^\top$ denotes the transpose of a matrix or a vector, and $\boldsymbol{\varphi}(\boldsymbol{z})$ is basis function such that

$$\boldsymbol{\varphi}(\boldsymbol{z}) = (\varphi_1(\boldsymbol{z}),\dots,\varphi_b(\boldsymbol{z}))^\top \geq \boldsymbol{0}_b \quad \text{for all } \boldsymbol{x}.$$

$\boldsymbol{0}_b$ denotes the b-dimensional vector with all zeros. Note that $\boldsymbol{\varphi}(\boldsymbol{z})$ could be dependent on the samples $\{\boldsymbol{z}_i\}_{i=1}^n$, i.e., *kernel* models are also allowed. We explain how the basis functions $\boldsymbol{\varphi}(\boldsymbol{z})$ are chosen in Section 2.3.

The second reason why directly solving the problem (3) is not possible is that the true probability densities $q(\boldsymbol{z})$ and $r(\boldsymbol{z})$ contained in the density ratio (4) are unavailable. To cope with this problem, we approximate them by their empirical distributions—then the optimization problem is reduced to

$$\widehat{\boldsymbol{\alpha}} := \operatorname*{argmin}_{\boldsymbol{\alpha}\in\mathbb{R}^b} \left[\frac{1}{2}\boldsymbol{\alpha}^\top \widehat{\boldsymbol{H}}\boldsymbol{\alpha} - \widehat{\boldsymbol{h}}^\top\boldsymbol{\alpha} + \frac{1}{2}\lambda\boldsymbol{\alpha}^\top\boldsymbol{\alpha}\right], \tag{6}$$

where we included $\lambda\boldsymbol{\alpha}^\top\boldsymbol{\alpha}$ ($\lambda > 0$) for regularization purposes and

$$\widehat{\boldsymbol{H}} := \frac{1}{n^d}\sum_{i_1,\dots,i_d=1}^{n} \boldsymbol{\varphi}(z_{i_1}^{(1)},\dots,z_{i_d}^{(d)})\boldsymbol{\varphi}(z_{i_1}^{(1)},\dots,z_{i_d}^{(d)})^\top,\ \ \widehat{\boldsymbol{h}} := \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\varphi}(z_i^{(1)},\dots,z_i^{(d)}).$$

Differentiating the objective function (6) with respect to $\boldsymbol{\alpha}$ and equating it to zero, we can obtain an analytic-form solution as

$$\widehat{\boldsymbol{\alpha}} = (\widehat{\boldsymbol{H}} + \lambda\boldsymbol{I}_b)^{-1}\widehat{\boldsymbol{h}},$$

where $\boldsymbol{I}_b$ is the b-dimensional identity matrix. Thus, the solution can be computed very efficiently just by solving a system of linear equations. Using the solution $\widehat{\boldsymbol{\alpha}}$, we can approximate SMI as

$$\widehat{I}_s = -\frac{1}{2} - \frac{1}{2}\widehat{\boldsymbol{\alpha}}^\top \widehat{\boldsymbol{H}}\widehat{\boldsymbol{\alpha}} + \widehat{\boldsymbol{h}}^\top \widehat{\boldsymbol{\alpha}}. \tag{7}$$

Ordinary MI can also be estimated similarly using the density ratio [10]. However, the use of SMI is more advantageous due to the analytic-form solution.

2.3 Design of Basis Functions and Model Selection

As basis functions, we propose using a Gaussian kernel model:

$$\varphi_\ell(\boldsymbol{z}) = \exp\left(-\frac{\|\boldsymbol{z} - \boldsymbol{v}_\ell\|^2}{2\sigma^2}\right) = \prod_{m=1}^{d} \exp\left(-\frac{(z^{(m)} - v_\ell^{(m)})^2}{2\sigma^2}\right), \tag{8}$$

where $\{\boldsymbol{v}_\ell \mid \boldsymbol{v}_\ell = (v_\ell^{(1)}, \ldots, v_\ell^{(d)})^\top\}_{\ell=1}^{b}$ are Gaussian centers randomly chosen from $\{\boldsymbol{z}_i\}_{i=1}^{n}$—more precisely, we set $\boldsymbol{v}_\ell = \boldsymbol{z}_{c(\ell)}$, where $\{c(\ell)\}_{\ell=1}^{b}$ are randomly chosen from $\{1, \ldots, n\}$ without replacement. An advantage of the Gaussian kernel lies in the factorizability in Eq.(8), contributing to reducing the computation of the matrix $\widehat{\boldsymbol{H}}$ significantly:

$$\widehat{H}_{\ell,\ell'} = \frac{1}{n^d} \prod_{m=1}^{d} \left[\sum_{i=1}^{n} \exp\left(-\frac{(z_i^{(m)} - v_\ell^{(m)})^2 + (z_i^{(m)} - v_{\ell'}^{(m)})^2}{2\sigma^2}\right)\right].$$

In the experiments, we fix the number of basis functions at

$$b = \min(100, n),$$

and choose the Gaussian width σ and the regularization parameter λ by CV with grid search as follows. First, the samples $\{\boldsymbol{z}_i\}_{i=1}^{n}$ are divided into K disjoint subsets $\{\mathcal{Z}_k\}_{k=1}^{K}$ of (approximately) the same size (we use $K = 5$ in the experiments). Then an estimator $\widehat{\boldsymbol{\alpha}}_{\mathcal{Z}_k}$ is obtained using $\{\mathcal{Z}_j\}_{j\neq k}$ (i.e., without $\mathcal{Z}_k$) and the approximation error for the hold-out samples $\mathcal{Z}_k$ is computed:

$$J_{\mathcal{Z}_k}^{(K\text{-CV})} = \frac{1}{2}\widehat{\boldsymbol{\alpha}}_{\mathcal{Z}_k}^\top \widehat{\boldsymbol{H}}\widehat{\boldsymbol{\alpha}}_{\mathcal{Z}_k} - \widehat{\boldsymbol{h}}^\top \widehat{\boldsymbol{\alpha}}_{\mathcal{Z}_k},$$

where $|\mathcal{Z}_k|$ denotes the number of sample pairs in the set $\mathcal{Z}_k$. This procedure is repeated for $k = 1, \ldots, K$ and its average $J^{(K\text{-CV})}$ is outputted:

$$J^{(K\text{-CV})} = \frac{1}{K}\sum_{k=1}^{K} J_{\mathcal{Z}_k}^{(K\text{-CV})}.$$

For model selection, we compute $J^{(K\text{-CV})}$ for all model candidates (the Gaussian width σ and the regularization parameter λ in the current setting) and choose the model that minimizes $J^{(K\text{-CV})}$. We can show that $J^{(K\text{-CV})}$ is an almost unbiased estimate of the objective function in Eq.(3), where the 'almost'-ness comes from the fact that the number of samples is reduced in the CV procedure due to data splitting.

2.4 The LICA Algorithm

Finally, we show how the above SMI estimation idea could be employed in the context of ICA.

Here, we use a simple gradient technique for obtaining a minimizer of the estimated SMI. The update rule of the demixing matrix $\boldsymbol{W}$ is given by

$$\boldsymbol{W} \longleftarrow \boldsymbol{W} - \varepsilon \frac{\partial \widehat{I}_s}{\partial \boldsymbol{W}}, \tag{9}$$

where ε (> 0) is the step size. We can show that the gradient is given by

$$\frac{\partial \widehat{I}_s}{\partial W_{\ell,\ell'}} = \frac{\partial \widehat{\boldsymbol{h}}^\top}{\partial W_{\ell,\ell'}}(-\widehat{\boldsymbol{\beta}} + 2\widehat{\boldsymbol{\alpha}}) + \widehat{\boldsymbol{\alpha}}^\top \frac{\partial \widehat{\boldsymbol{H}}}{\partial W_{\ell,\ell'}}(\widehat{\boldsymbol{\beta}} - \frac{3}{2}\widehat{\boldsymbol{\alpha}}), \tag{10}$$

where

$$\begin{aligned}
\frac{\partial \widehat{h}_k}{\partial W_{\ell,\ell'}} =& \frac{1}{n\sigma^2}\sum_{i=1}^{n}(z_i^{(\ell)} - v_k^{(\ell)})(u_k^{(\ell')} - y_i^{(\ell')}) \exp\left(-\frac{\|\boldsymbol{z}_i - \boldsymbol{v}_k\|^2}{2\sigma^2}\right), \\
\frac{\partial \widehat{H}_{k,k'}}{\partial W_{\ell,\ell'}} =& \frac{1}{n^{d-1}} \prod_{m=1,m\neq\ell}^{d} \left[\sum_{i=1}^{n} \exp\left(-\frac{(z_i^{(m)} - v_k^{(m)})^2 + (z_i^{(m)} - v_{k'}^{(m)})^2}{2\sigma^2}\right)\right] \\
& \times \left[\frac{1}{n\sigma^2}\sum_{i=1}^{n}\left((z_i^{(\ell)} - v_k^{(\ell)})(u_k^{(\ell')} - y_i^{(\ell')}) + (z_i^{(\ell)} - v_{k'}^{(\ell)})(u_{k'}^{(\ell')} - y_i^{(\ell')})\right)\right. \\
& \left. \times \exp\left(-\frac{(z_i^{(\ell)} - v_k^{(\ell)})^2 + (z_i^{(\ell)} - v_{k'}^{(\ell)})^2}{2\sigma^2}\right)\right], \\
\boldsymbol{u}_\ell =& \boldsymbol{y}_{c(\ell)}, \quad \boldsymbol{y}_i = (y_i^{(1)}, \ldots, y_i^{(d)})^\top, \quad \text{and} \quad \widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{H}} + \lambda \boldsymbol{I}_b)^{-1}\widehat{\boldsymbol{H}}\widehat{\boldsymbol{\alpha}}.
\end{aligned}$$

In ICA, scaling of components of $\boldsymbol{z}$ can be arbitrary. This implies that the above gradient updating rule can lead to a solution with bad scaling, which is not preferable from a numerical point of view. To avoid numerical instability, we normalize $\boldsymbol{W}$ at each gradient iteration as

$$W_{\ell,\ell'} \longleftarrow \frac{W_{\ell,\ell'}}{\sqrt{\sum_{m=1}^{d} W_{\ell,m}^2}}. \tag{11}$$

The proposed ICA algorithm, which we call *Least-squares Independent Component Analysis* (LICA), is summarised below.

1. Initialize demixing matrix $\boldsymbol{W}$ and normalize it by Eq.(11).
2. Optimize Gaussian width σ and regularization parameter λ by CV.
3. Compute gradient $\frac{\partial \widehat{I}_s}{\partial \boldsymbol{W}}$ by Eq.(10).
4. Choose step-size ε such that $\widehat{I}_s$ (see Eq.(7)) is minimized (*line-search*).
5. Update $\boldsymbol{W}$ by Eq.(9).
6. Normalize $\boldsymbol{W}$ by Eq.(11).
7. Repeat 2.–6. until $\boldsymbol{W}$ converges.

3 Numerical Examples

In this section, we illustrate how our algorithm behaves using the following three 2-dimensional datasets:

(a) Sub·Sub-Gaussians: $p(\boldsymbol{x}) = U(x^{(1)}; -0.5, 0.5)U(x^{(2)}; -0.5, 0.5)$,
(b) Super·Super-Gaussians: $p(\boldsymbol{x}) = L(x^{(1)}; 0, 1)L(x^{(2)}; 0, 1)$,
(c) Sub·Super-Gaussians: $p(\boldsymbol{x}) = U(x^{(1)}; -0.5, 0.5)L(x^{(2)}; 0, 1)$,

where $U(x; a, b)$ $(a, b \in \mathbb{R}, a < b)$ denotes the uniform density on $[a, b]$ and $L(x; \mu, v)$ $(\mu \in \mathbb{R}, v > 0)$ denotes the Laplacian density with mean μ and variance v. Let the number of samples be $n = 300$ and we observe mixed samples $\{\boldsymbol{y}_i\}_{i=1}^n$ through the following mixing matrix:

$$\boldsymbol{A} = \begin{pmatrix} \cos(\pi/4) & \sin(\pi/4) \\ -\sin(\pi/4) & \cos(\pi/4) \end{pmatrix}.$$

The observed samples are plotted in Figure 1.

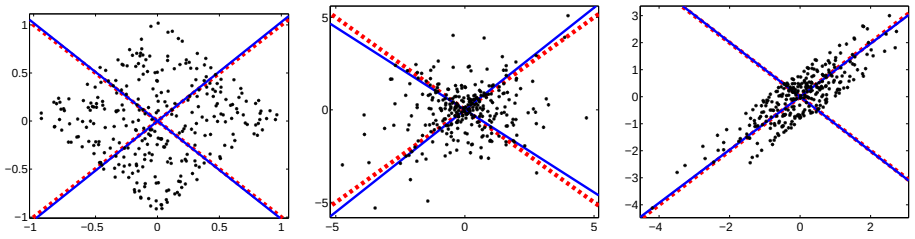
Figure 2 depicts the value of estimated SMI (7) over iterations and Figure 3 depicts the elements of the demixing matrix $\boldsymbol{W}$ over iterations. The true independent directions as well as the estimated independent directions are plotted in Figure 1. The results show that estimated SMI decreases rapidly and good solutions are obtained for all the datasets.

4 Conclusions

In this paper, we have proposed a new estimation method of a squared-loss variant of mutual information, and based on this, we developed an ICA algorithm. The proposed ICA method, named least-squares ICA (LICA), has several preferable properties, e.g., it is distribution-free and model selection by cross-validation is available. Our future work includes development of efficient optimization algorithm beyond gradient techniques.

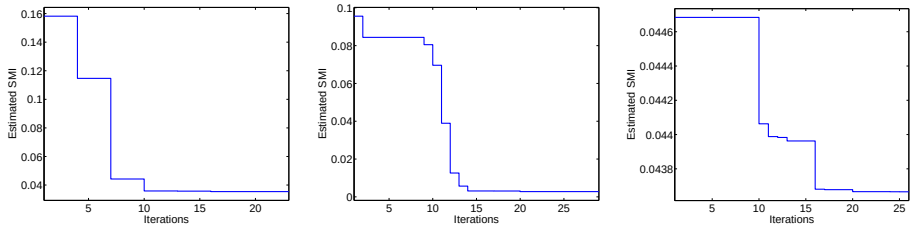
References

1. Hyvärinen, A., Karhunen, J., Oja, E.: Independent Component Analysis. Wiley, New York (2001)
2. Hyvärinen, A.: Fast and robust fixed-point algorithms for independent component analysis. IEEE Transactions on Neural Networks **10**(3) (1999) 626–634
3. Amari, S., Cichocki, A., Yang, H.H.: A new learning algorithm for blind signal separation. In: Advances in Neural Information Processing Systems, MIT Press (1996) 757–763
4. Bach, F.R., Jordan, M.I.: Kernel independent component analysis. Journal of Machine Learning Research **3** (2002) 1–48
5. Hulle, M.M.V.: Sequential fixed-point ICA based on mutual information minimization. Neural Computation **20**(5) (2008) 1344–1365



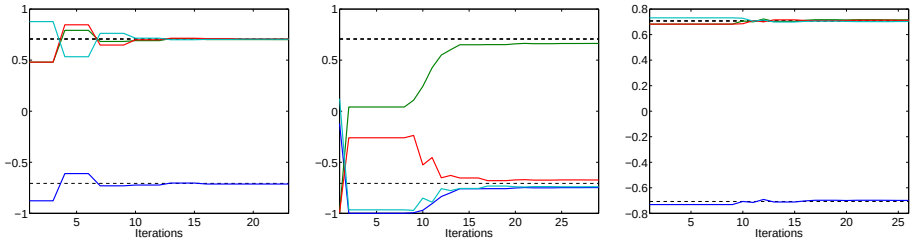
(a) (b) (c)

Fig. 1. Observed samples (black asterisks), true independent directions (red dotted lines) and estimated independent directions (blue solid lines).



(a) (b) (c)

Fig. 2. Estimated SMI $\widehat{I}_s$ over iterations.



(a) (b) (c)

Fig. 3. The elements of the demixing matrix $\boldsymbol{W}$ over iterations. The blue, green, red, cyan lines correspond to $W_{1,1}$, $W_{1,2}$, $W_{2,1}$, and $W_{2,2}$, respectively. The black dotted lines denote the true values.

6. Cardoso, J.F., Souloumiac, A.: Blind beamforming for non-Gaussian signals. Radar and Signal Processing, IEE Proceedings-F **140**(6) (1993) 362–370
7. Comon, P.: Independent component analysis, a new concept? Signal Processing **36**(3) (1994) 287–314
8. Jutten, C., Herault, J.: Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. Signal Processing **24**(1) (1991) 1–10
9. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press, Cambridge (2004)
10. Suzuki, T., Sugiyama, M., Sese, J., Kanamori, T.: Approximating mutual information by maximum likelihood density ratio estimation. In: New Challenges for Feature Selection in Data Mining and Knowledge Discovery. Volume 4 of JMLR Workshop and Conference Proceedings. (2008) 5–20