Deterministic Independent Component Analysis

Ruitong Huang RUITONG@UALBERTA.CA
András György GYORGY@UALBERTA.CA
Csaba Szepesvári SZEPESVA@UALBERTA.CA
Department of Computing Science, University of Alberta, Edmonton, AB T6G2E8 Canada

Abstract

We study independent component analysis with noisy observations. We present, for the first time in the literature, consistent, polynomial-time algorithms to recover non-Gaussian source signals and the mixing matrix with a reconstruction error that vanishes at a $1/\sqrt{T}$ rate using T observations and scales only polynomially with the natural parameters of the problem. Our algorithms and analysis also extend to deterministic source signals whose empirical distributions are approximately independent.

1. Introduction

Independent Component Analysis (ICA) has received much attention in the past decades. In the standard ICA model one can observe a d-dimensional vector X that is a linear mixture of d independent variables $(S_1, \ldots, S_d)$ with Gaussian noise:

$$X = AS + \epsilon, \tag{1}$$

where $\epsilon \sim \mathcal{N}(0, \Sigma)$ is a d-dimensional Gaussian noise with zero mean and covariance matrix Σ, and A is a nonsingular $d \times d$ mixing matrix. The goal of the observer is to recover (separate) the source signals and the mixing matrix given several independent and identically distributed (i.i.d.) observations from the above model. The ICA literature is vast in both practical algorithms and theoretical analyses; we refer to the book of Comon and Jutten (2010) for a comprehensive survey. In this paper we investigate one of he most important problems in ICA: finding consistent, computationally efficient algorithms with finite-sample performance guarantees. In particular, we aim to develop algorithms whose computational and sample complexity are polynomial in the natural parameters of the problem.

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37.

A popular approach to the ICA problem is to find a linear transformation W for X by optimizing a, so-called, *contrast function* that measures dependence or non-gaussianity of the resulting coordinates of WX. The optimal W then can serve as an estimate of A^{-1}, thereby recovering the mixing matrix A. One of the most popular ICA algorithms, FastICA (Hyvarinen, 1999), follows this approach for a specific contrast function. FastICA has been analyzed theoretically from many aspects (Tichavsky et al., 2006; Oja and Yuan, 2006; Ollila, 2010; Dermoune and Wei, 2013; Wei, 2014). In particular, recently Miettinen et al. (2014) showed that in the noise-free case (i.e., when $X = AS$), the error of FastICA (when using a particular forth-moments-based contrast function) vanishes at a rate of $1/\sqrt{T}$ where T is the sample size. In addition, several other methods have been shown to achieve similar error rates in the noise-free setting (e.g., Eriksson and Koivunen, 2003; Samarov et al., 2004; Chen and Bickel, 2005; Chen et al., 2006). However, to our knowledge, no similar finite sample results are available in the noisy case.

On the other hand, several promising algorithms are available in the noisy case that make significant advances towards provably efficient and effective ICA algorithms, albeit fall short of providing a complete solution. Using a quasi-whitening procedure, Arora et al. (2012) reduces the problem to finding all the local optima of a specific function defined using the forth order cumulant, and propose a polynomial-time algorithm to find them with appealing theoretical guarantees. However, the results depend on an unspecified parameter (β in the original paper) whose proper tuning is essential; note that even an exhaustive search over β is problematic, since its valid range is not well understood.

The exploitation of the special algebraic structure of the forth moments induced by the independence leads to several other works related to ICA (Hsu and Kakade, 2013; Anandkumar et al., 2012a;b). A similar idea is also discussed earlier as a intuitive argument to construct a contrast function (Cardoso, 1999). The first rigorous proofs for this idea are developed using matrix perturbation tools in a gen-

eral tensor perspective (Anandkumar et al., 2012a;b; Goyal et al., 2014). A common problem faced by these methods is a minimal gap of the eigenvalues, which may result in an exponential dependence on the number of source signals d. More precisely, these methods all require an eigendecomposition of some flattened tensor where the minimal gap between the eigenvalues plays an essential role. Although the exact size of this gap is not yet understood, a naive analysis introduces an exponential dependence on the dimension d. Such dependence is also observed in the literature (Cardoso, 1999; Goyal et al., 2014). One way to circumvent such dependence is to directly decompose a high-order tensor using the power method, which requires no flattening procedure (Anandkumar et al., 2014). However, when applied to the ICA problem, this introduces a bias term and so the error does not approach 0 as the sample size approaches infinity. Another issue is the well-known fact that the power method is unstable in practice for high-order tensors. Goyal et al. (2014) proposed another method by exploring the characteristic function rather than the forth moments. However, their algorithm requires picking a parameter (σ in the original paper) that is smaller than some unknown quantity, making their algorithm impossible to tune. Recently, Vempala and Xiao (2014) proposed an ICA algorithm based on an elegant, recursive version of the method of Goyal et al. (2014) that avoids dealing with the aforementioned minimal gap; however, they still need an oracle to set the unspecified parameter of Goyal et al. (2014).

In this paper we propose a *provably polynomial-time* algorithm for the noisy ICA model. Our algorithm is a refined version of the ICA method proposed by (Hsu and Kakade, 2013) (HKICA). However, we propose two simpler ways, one inspired by Frieze et al. (1996), Arora et al. (2012), and another based on Vempala and Xiao (2014), to deal with the spacing problem of the eigenvalues under similar conditions to those of Goyal et al. (2014). Unlike the method proposed by Goyal et al. (2014), our first method can force the eigenvalues to be well-separated with a gap that is independent of the mixing matrix A, while our second method, based on the recursive decomposition idea of Vempala and Xiao (2014), avoids dealing with the minimum gap (on the price of introducing other complications). We prove that our methods achieve an $O(1/\sqrt{T})$ error in estimating A and the source signals, with high probability, such that both the convergence rate and the computational complexity scale *polynomially* with the natural parameters of the problem. Our method needs no parameter tuning, which makes it even more appealing.

Another contribution of the present paper is that our analysis is conducted in a deterministic manner. In practice, ICA is also known to work well for unmixing the mixture of various deterministic signals. One of the classical

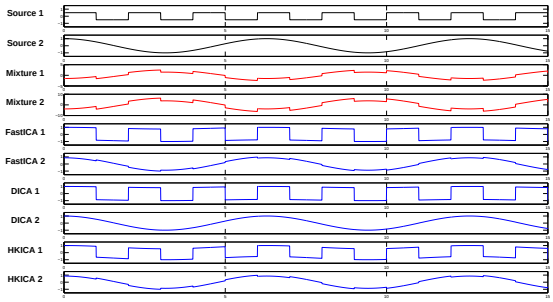

Figure 1. Example of ICA for deterministic sources: The first two rows show the source signals $s_1(t) = 0.5 - \lfloor t - 2\lfloor t/2 \rfloor \rfloor$, $s_2 = \cos(t)$, the next two rows present the observations with mixing matrix $A = \begin{pmatrix} 1 & -2 \\ 2.6 & -5.1 \end{pmatrix}$. The reconstructed (and rescaled) signals are shown for FastICA, HKICA, and DICA after sampling $As(t)$ at 10000 uniformly spaced points in the interval $[0, 15]$.

demonstrations of ICA is showing that two periodic signals can be well recovered from their mixtures (Hyvärinen and Oja, 2000). Such an example is shown in Figure 1. It can be seen that our algorithm, DICA, in this particular example, can solve the problem better than other algorithms, FastICA (Hyvarinen, 1999) and HKICA (Hsu and Kakade, 2013). Such phenomenon suggests that the usual probabilistic notion is unsatisfactory if one wishes to have deeper understanding of ICA. Our deterministic analysis helps investigate this curious phenomenon without losing any generality to the traditional stochastic setting. Formally, instead of observing T i.i.d. samples from (1), the source signals are defined by the function $s : \mathbb{N} \to \mathbb{R}^d$ be a d-dimensional deterministic "signal", and the observations are $x(t) = As(t) + \epsilon_t$, where $(\epsilon_t)_{t=1}^{\infty}$ is an i.i.d. sequence of d-dimensional $\mathcal{N}(0, \Sigma)$ random variables.

The rest of this paper is organized as follows: The ICA problem is introduced in detail in Section 2 and our main results are highlighted in Section 3. The polynomial-time algorithms underlying these results are developed through the next two sections: Section 4.1 is devoted to the analysis of the HKICA algorithm, also showing its disadvantages, while our new algorithms are presented in Section 5. Experimental results are reported in Section 6. Proofs are presented in the full version of the paper (Huang et al., 2015).

1.1. Notation

We denote the set of real and natural numbers by $\mathbb{R}$ and $\mathbb{N}$, respectively. A vector $v \in K^d$ for a field K is assumed to be a column vector. Let $\|v\|_2$ denote its L_2-norm, and for any matrix Z let $\|Z\|_2 = \max_{v:\|v\|_2=1} \|Zv\|_2$ denote the corresponding induced norm. Denote the maximal and minimal singular value of Z by $\sigma_{\max}(Z)$ and $\sigma_{\min}(Z)$, respectively. Also, let Z_i and $Z_{i:}$ denote the ith column and, resp., row of Z, and let $Z_{(2,\min)} = \min_i \|Z_i\|_2$, $Z_{(2,\max)} = \max_i \|Z_i\|_2$ and $Z_{\max} = \max_{i,j} |Z_{i,j}|$.

Clearly, $\sigma_{\max}(Z) = \|Z\|_2 \geq Z_{(2,\max)} \geq Z_{\max}$, and $\sigma_{\min}(Z) \leq Z_{(2,\min)}$. For a tensor (including vectors and matrices) T, its Frobenious norm (or L_2 norm) $\|T\|_F$ is defined as the square root of the sum of the square of all the entries. For a vector $v = (v_1, \ldots, v_d) \in K^d$, $|v|$ is defined coordinatewise, that is $|v| = (|v_1|, \ldots, |v_d|)$. The transpose of a vector/matrix Z is denoted by $Z^\top$, while the inverse of the transpose is denoted by $Z^{-\top}$. The outer product of two vectors $v, u \in K^d$ is denoted by $u \otimes v = uv^\top$. $v^{\otimes k}$ denotes the k-fold outer product of v with itself, that is, $v \otimes v \otimes v \ldots \otimes v$, which is a k-dimensional tensor. Given a 4-dimensional tensor T, we denote the matrix Z by $T(\eta, \eta, \cdot, \cdot)$ that is generated by marginalizing the first two coordinates of T on the direction η, that is, $Z_{i,j} = \sum_{k_1,k_2=1}^{d} \eta_{k_1}\eta_{k_2}T_{k_1,k_2,i,j}$. (Similar definitions for marginalizing different coordinates of the tensor.) For a real vector v and some real number C, $v \leq C$ means that all the entries of v are at most C. The bold symbol $\mathbf{1}$ denotes a vector with all entries being 1 (the dimension of this vector will always be clear from the context). Finally, $\mathrm{Poly}\,(\cdot, \cdots, \cdot)$ denotes a polynomial function of its argument.

2. The ICA Problem

In this paper we consider the following non-stochastic version of the ICA problem. Assume that we can observe the d dimensional mixed signal $x(t) \in \mathbb{R}^d, t \in [T] := \{1, 2, \ldots, T\}$ generated by

$$x(t) = As(t) + \epsilon(t), \tag{2}$$

where A is a $d \times d$ nonsingular mixing matrix, $s : [T] \to [-C, C]^d$ is a bounded, d-dimensional source function for some constant $C \geq 1$. , and $\epsilon : [T] \to \mathbb{R}^d$ is the noise function. We will denote the ith component of s by s_i. Furthermore, we will use the notation $\sigma_{\min} = \sigma_{\min}(A)$ and $\sigma_{\max} = \sigma_{\max}(A)$

For any $t, k \geq 1$ and signal $u : [t] \to \mathbb{R}^k$, we introduce the empirical distribution $\nu_t^{(u)}$ defined by $\nu_t^{(u)}(B) = \frac{1}{t}|\{\tau \in [t] : u(t) \in B\}|$ for all Borel sets $B \subset \mathbb{R}^k$. Next we will impose assumptions on the empirical measure that guarantee that on the average we do not deviate too much from the stochastic model. The next assumption implies that the empirical distributions of the source signals are approximately zero mean, and that the noise is approximately zero-mean Gaussian.

Assumption 2.1. Assume there exists a constant L and a function $g : \mathbb{N} \to \mathbb{R}$ such that $g(t) \to 0$ as $t \to \infty$ and

(i) $\|\mathbb{E}_{S_i \sim \nu_t^{(s_i)}}[S_i]\|_F$, $\|\mathbb{E}_{Y \sim \nu_t^{(\epsilon)}}[Y]\|_F \leq g(t)$;

(ii) $\|\mathbb{E}_{Y \sim \nu_t^{(\epsilon)}}[Y^{\otimes 2}]\|_F$, $\|\mathbb{E}_{Y \sim \nu_t^{(\epsilon)}}[Y^{\otimes 3}]\|_F \leq L$;

(iii) $\left\| \left(\mathbb{E}_{Y \sim \nu_t^{(\epsilon)}}[Y^{\otimes 4}] - (\mathbb{E}_{Y \sim \nu_t^{(\epsilon)}}[Y^{\otimes 2}])^{\otimes 2} \right) (\eta, \eta, \cdot, \cdot) \right.$
$\left. - 2(\mathbb{E}_{Y \sim \nu_t^{(\epsilon)}}[Y^{\otimes 2}])^{\otimes 2}(\eta, \cdot, \eta, \cdot) \right\|_F \leq g(t)\|\eta\|_2^2.$

Here L and the function g may depend on $\{A, \Sigma, C, d\}$.

Remark 2.2. The first assumption forces the average of s and ϵ decay to 0 at a rate of $g(t)$. The next one requires that both the second and third moments of the noise be bounded. The last assumption basically says that the induced measure of the noise function ϵ has 0 kurtosis in the limit.

We will also need to guarantee that the source signals and the noise be approximately independent:

Assumption 2.3. Assume the source signal function and the noise function are 'independent' up to the 4th moment in the sense that for any $i_1, i_2, j_1, j_2 \geq 0$ such that $i_1 + i_2 + j_1 + j_2 \leq 4$,

$$\|\mathbb{E}_{S \sim \nu_t^{(s)}}[(AS)^{\otimes i_1} \otimes \mathbb{E}_{Y \sim \nu_t^{(\epsilon)}}[Y^{\otimes j_1}] \otimes (AS)^{\otimes i_2}] - \mathbb{E}_{(S,Y) \sim \nu_t^{(s,\epsilon)}}[(AS)^{\otimes i_1} \otimes Y^{\otimes j_1} \otimes (AS)^{\otimes i_2}]\|_F \leq g(t),$$

$$\|\mathbb{E}_{Y \sim \nu_t^{(\epsilon)}}[Y^{\otimes j_1} \otimes \mathbb{E}_{S \sim \nu_t^{(s)}}[(AS)^{\otimes i_1}] \otimes Y^{\otimes j_2}] - \mathbb{E}_{(S,Y) \sim \nu_t^{(s,\epsilon)}}[Y^{\otimes j_1} \otimes (AS)^{\otimes i_1} \otimes Y^{\otimes j_2}]\|_F \leq g(t),$$

for the same function g in Assumption 2.1, where (s, ϵ) is the function obtained by concatenating s and ϵ together.

The sufficiency of such weaker assumptions is also discussed in the paper of Frieze et al. (1996). The next proposition shows that these assumptions are all satisfied, with high probability, for the traditional stochastic setting of the ICA model with Gaussian noise independent to the source signals.

Proposition 2.4. *In the traditional stochastic setting of ICA, that is, when $(s(t))_{t \in [T}$ is an i.i.d. sequence, independent of the i.i.d. Gaussian noise sequence $(\epsilon(t))_{t \in [T]}$,there exists $L = \mathrm{Poly}\left(A_{\max}, \|\Sigma\|_2, C, d, \frac{1}{\delta}\right)$ and $g(t) = L/\sqrt{t}$, such that Assumptions 2.1 and 2.3 hold with probability at least $1 - \delta$.*

On the other hand, our setting can also cover some other examples excluded by the traditional setting, such as the example of Figure 1 in Section 1.

Example 2.5. Assume that the unknown sources s_i ($1 \leq i \leq d$) are deterministic and periodic. Our observation $x = As + \epsilon$ is a linear mixture of s contaminated by i.i.d. Gaussian noise for each time step, where A is a nonsingular matrix and $\epsilon \sim \mathcal{N}(0, \Sigma)$ is Gaussian. Even though ϵ is i.i.d. for every time step, the observations cannot satisfy the traditional i.i.d. assumption, since the source s is deterministic. However, it can be proved that if the ratio of the periods of each pair of (s_i, s_j) is irrational, this example satisfies all the assumptions above for T large enough.

Our setting also extends the traditional one to a practically important case, Markov sources.

Example 2.6. Assume that s_i is a stationary and ergodic Markov source, and the sources are independent of each other for $1 \leq i \leq d$. Our observations are similar to the setting in Example 2.5. Because of the Markov property, the observations do not satisfy the i.i.d. assumptions. On the other hand, it can be verified that this example satisfies the above assumptions.

3. Main Results

The ICA approach requires that the components s_i of the source signal s be statistically independent. In our setup, we require that the empirical distribution $\nu_T^{(s)}$ be close to a product distribution.

Fix some product distribution $\mu = \mu_1 \otimes \ldots \otimes \mu_d$ over $\mathbb{R}^d$ such that $\mathbb{E}_{S_i \sim \mu_i}[S_i] = 0$ and $\kappa_i := |\mathbb{E}_{S_i \sim \mu_i}[S_i^4] - 3\left(\mathbb{E}_{S_i \sim \mu_i}[S_i^2]\right)^2| \neq 0$. Let K denote the diagonal matrix $\mathrm{diag}(\kappa_1, \cdots, \kappa_d)$, and define $\kappa_{\max} = \max_i \kappa_i$ and $\kappa_{\min} = \min_i \kappa_i$.

To measure the distance of $\nu_T^{(s)}$ from μ, define the following family of "distances" to measure the closeness of two distributions: Given two distributions ν_1 and ν_2 over $\mathbb{R}^d$, let $D_k(\nu_1, \nu_2) = \sup_{f \in \mathcal{F}} |\int f(s) d\nu_1(s) - \int f(s) d\nu_2(s)|$, where $\mathcal{F} = \{f : \mathbb{R}^d \to \mathbb{R} : f(s) = \prod_{j=1}^k s_{i_j}, 1 \leq i_1, \ldots, i_k \leq d\}$ is the set of all monomials up to degree k. Finally let

$$\xi = \left(6C^2 D_2(\mu, \nu_T^{(s)}) + D_4(\mu, \nu_T^{(s)})\right). \tag{3}$$

In general, we will need a condition that ξ is small enough, so that the components of s are "independent" enough. To this end, one should choose μ to minimize ξ; however, such a minimizer does not always exists. Generally, μ could be selected as the product of the limit distributions, if applicable, of the individual sources. On the other hand, in the traditional stochastic setting where the observations are i.i.d. samples, the empirical distribution will converge to the population distribution, which, based on the independence assumption, is a product probability measure. Therefore, in this case, ξ will be small for large enough sample sizes.

Example 3.1. Let μ_1 be a Bernoulli distribution $\mu_1(\{0.5\}) = 1/2$ and $\mu_1(\{-0.5\}) = 1/2$, and μ_2 to be a distribution with density function $p(x) = \frac{1}{\pi\sqrt{1-x^2}}$ for $-1 \leq x \leq 1$. For the demonstration example in Figure 1, pick $\mu = \mu_1 \otimes \mu_2$. It is easy to see that μ_1 (μ_2) is the limit distribution of source 1 (respectively, source 2). Let $T = 2 * u + b$ as the division with remainder, where u is integer and $0 \leq b < 2$. Moreover, assume $b \leq 1$ (similar analysis will go through for the case of $b > 1$). The induced distribution $\nu_T^{s_1}$ of source 1 is $\nu_T^{s_1}(\{0.5\}) = \frac{u+b}{T}$ and $\nu_T^{s_1}(\{-0.5\}) = \frac{u}{T}$. Thus the total variation distance of μ_1 and $\nu_T^{s_1}$ is at most $1/(2T)$. Similarly, it can be verified that the total variation distance of ν_T and μ also decays as $1/T$. Thus, D_4 is $O(1/T)$, since the monomials $f(s)$ in the definition of D_4 are bounded from above by 1. Lastly, note that D_2 is upper bounded by D_4 by definition, so ξ decays at a $1/T$ rate.

Now we are ready to state our main result, which shows the existence of polynomial-time algorithms for ICA that reconstructs the mixing matrix A with error that vanishes at an $O(1/\sqrt{T})$ rate for T samples and is also polynomial in the natural parameters of the problem:

Theorem 3.2. *Consider the ICA problem (2). There exists an algorithm that estimates the mixing matrix A from T samples of x such that (i) the computational complexity of the algorithm is $O(d^3 T)$; and (ii) if Assumptions 2.1 and 2.3 are satisfied,*

$$T \geq \mathrm{Poly}\left(d, \frac{1}{\kappa_{\min}}, \frac{1}{\delta}, L, C, \sigma_{\max}, \frac{1}{\sigma_{\min}}\right),$$

and there exists a product distribution μ such that

$$D_4(\mu, \nu_T) \leq \mathrm{Poly}\left(\frac{1}{C}, \sigma_{\min}, \frac{1}{\sigma_{\max}}, \frac{1}{d}, \delta, \kappa_{\min}\right),$$

then, with probability at least $1 - \delta$, there exists a permutation π and constants $\{c_1, \ldots, c_d\}$, such that for all $1 \leq k \leq d$,

$$\|c_k \hat{A}_{\pi(k)} - A_k\|_2 \leq \mathcal{C}\left(D_4(\mu, \nu_T) + g^2(T) + g(T)\right),$$

where $\mathcal{C} = \mathrm{Poly}\left(\sigma_{\max}, 1/\sigma_{\min}, 1/\kappa_{\min}, 1/\delta, d, C, L\right)$, and $\hat{A}$ is the output of the algorithm.

In particular, in the traditional stochastic setting, if S has distribution μ and

$$T \geq \mathrm{Poly}\left(d, \frac{1}{\kappa_{\min}}, \frac{1}{\delta}, C, \sigma_{\max}, \frac{1}{\sigma_{\min}}, \|\Sigma\|_2\right),$$

then, with probability at least $1 - \delta$, there exists a permutation π and constants $\{c_1, \ldots, c_d\}$, such that for all $1 \leq k \leq d$,

$$\|c_k \hat{A}_{\pi(k)} - A_k\|_2 \leq \frac{\mathrm{Poly}\left(C, \sigma_{\max}, \frac{1}{\sigma_{\min}}, \frac{1}{\kappa_{\min}}, \frac{1}{\delta}, d\right)}{\sqrt{2T}}.$$

Remark 3.3. Note that the result is polynomial in $1/\delta$ which is weaker than being polynomial in $\log(1/\delta)$.

In the next sections, we will present two algorithms, DICA (Algorithm 2 and HKICA.R (Algorithm 3) in Section 5 that satisfy the theorem.

4. Estimating Moments: the HKICA Algorithm

In this section we introduce the ICA method of Hsu and Kakade (2013) which is based on the well-known excess-kurtosis-like quantity defined as follows:

For any $p \geq 1$, $\eta \in \mathbb{R}^d$, and distribution ν over $\mathbb{R}^d$, let

$$m_p^{(\nu)}(\eta) = \mathbb{E}_{X\sim\nu}[(\eta^\top X)^p], \quad (4)$$

$$f_\nu(\eta) = \tfrac{1}{12}\left(m_4^{(\nu)}(\eta) - 3m_2^{(\nu)}(\eta)^2\right). \quad (5)$$

Hsu and Kakade (2013) showed that $\nabla^2 f_{\nu_T^{(x)}}(\eta)$, the second derivative of the function $f_{\nu_T^{(x)}}$, is extremely useful for the ICA problem: They showed that if $\mu^{(X)}$ is the distribution of the observations X in the stochastic setting where S comes from the product distribution μ, then $f_{\mu^{(X)}}(\eta) = f_{A\mu}(\eta)$ for all η (where $A\mu$ denotes the distribution of AS) and, consequently, the eigenvectors[1] of the matrix $M = \nabla^2 f_{\mu^{(X)}}(\phi)(\nabla^2 f_{\mu^{(X)}}(\psi))^{-1}$ are the rescaled columns of A if $\frac{\phi^\top A_i}{\psi^\top A_i}$ are distinct for all i. Thus, to obtain an algorithm, one needs to estimate $\nabla^2 f_{\mu^{(X)}}$ in such a way that the noise ϵ could still be neglected.

An estimate $\nabla^2\hat{f}$ of $\nabla^2 f_{\mu^{(X)}}$ is not hard, since for any ν, $\nabla^2 f_\nu(\eta)$ can be computed as

$$\nabla^2 f_\nu(\eta) = G_\nu(\eta) := G_1^{(\nu)}(\eta) - G_2^{(\nu)}(\eta) - 2G_3^{(\nu)}(\eta), \quad (6)$$

where

$$G_1^{(\nu)}(\eta) = \mathbb{E}_{X\sim\nu}[(\eta^\top X)^2 XX^\top];$$
$$G_2^{(\nu)}(\eta) = \mathbb{E}_{X\sim\nu}[(\eta^\top X)^2]\mathbb{E}_{X\sim\nu}[XX^\top];$$
$$G_3^{(\nu)}(\eta) = \mathbb{E}_{X\sim\nu}[(\eta^\top X)X]\mathbb{E}_{X\sim\nu}[(\eta^\top X)X^\top],$$

and these quantities can be estimated using the observed samples. In what follows, we will use the estimate $\nabla^2\hat{f} := \nabla^2 f_{\nu_T^{(x)}}$ and, in general, we will add a "hat" to quantities which are derived from the empirical distribution $\nu_T^{(x)}$. It is important to note that, under our assumptions, the noise ϵ has limited effect in the estimation procedure, as shown in the full version of the paper (Huang et al., 2015). In particular, the difference in the estimation of the Hessian matrix caused by the noise is $\mathrm{Poly}\left(L_\eta, L, d, \sigma_{\max}, C\right)\left(g(T)+1\right)g(T)$. Denote this quantity by $P(L_\eta)$. Note that this error caused by the noise decays at a rate of $\sqrt{T}$. Putting everything together, we obtain the algorithm HKICA, named after Hsu and Kakade (2013), which is shown in Algorithm 1,

Algorithm 1 The HKICA algorithm.

input $x(t)$ for $1 \leq t \leq T$.
output An estimation of the mixing matrix A.

1: Sample ϕ and ψ independently from a standard Gaussian distribution of dimension d;
2: Evaluate $\nabla^2\hat{f}(\phi)$ and $\nabla^2\hat{f}(\psi)$,
3: Compute $\hat{M} = (\nabla^2\hat{f}(\phi))(\nabla^2\hat{f}(\psi))^{-1}$;
4: Compute all the eigenvectors of $\hat{M}$, $\{\mu_1, \ldots, \mu_d\}$;
5: Return $\hat{A} = (\mu_1, \ldots, \mu_d)$.

[1] Throughout the paper eigenvectors always mean right eigenvectors, unless specified otherwise.

4.1. Analysis of HKICA

Hsu and Kakade (2013) claimed that HKICA is easy to analyze using matrix perturbation techniques. In this section we provide a rigorous analysis of the algorithm, which reveals some unexpected complications.

Definition 4.1. Let $\mathcal{E}_\psi$ denote the following event: For some fixed $C_1 = \frac{\sqrt{\pi}A_{(2,\min)}}{\sqrt{2}d}\ell$ for $0 \leq \ell \leq 1$, and $L_u \geq \sqrt{2d}$, $\min_i |\psi^\top A_i| \geq C_1$ and $\|\psi\|_2 \leq L_u$ hold simultaneously.

The performance of the HKICA algorithm will essentially depend on the parameter , as shown in the following theorem, where

$$\gamma_A = \min_{i,j:i\neq j}\left|\left(\frac{\phi^\top A_i}{\psi^\top A_i}\right)^2 - \left(\frac{\phi^\top A_j}{\psi^\top A_j}\right)^2\right|. \quad (7)$$

Theorem 4.2. *Suppose Assumptions 2.1 and 2.3 hold. Furthermore, assume that*

$$T \geq \mathrm{Poly}\left(d, L_u, C, \sigma_{\max}, \kappa_{\max}, L, \frac{1}{\ell}, \frac{1}{\kappa_{\min}}, \frac{1}{\sigma_{\min}}, \frac{1}{\gamma_A}\right),$$

and that there exist a product measure μ such that

$$\xi \leq \mathrm{Poly}\left(\gamma_A, \frac{1}{d}, \frac{1}{L_u}, \frac{1}{\sigma_{\max}}, \frac{1}{\kappa_{\max}}, \kappa_{\min}, \sigma_{\min}, \ell\right).$$

Then, on the event $\mathcal{E}_\psi$, there exists a permutation π and constants $\{c_1, \ldots, c_d\}$, such that for any k,

$$\max_{1\leq k\leq d}\|c_1\hat{A}_{\pi(k)} - A_k\|_2 \leq \frac{1}{\gamma_A}(\xi + P(L_u))Q \quad (8)$$

where $\hat{A}$ is the output of the HKICA algorithm, and

$$Q = \mathrm{Poly}\left(d, L_u, \sigma_{\max}, \kappa_{\max}, \frac{1}{\kappa_{\min}}, \frac{1}{\sigma_{\min}}, \frac{1}{\ell}\right).$$

Remark 4.3. (i) Note that the bound in (8) goes to zero at an $O(1/\sqrt{T})$ rate whenever $D_4(\mu, \nu_T^{(s)}) = O(1/\sqrt{T})$ and $g(T) = O(1/\sqrt{T})$, as, e.g., in the stochastic setting. (ii) The parameter $1/\gamma_A$ is essential in the above theorem, in the sense that not only the reconstruction error bound is linear in $1/\gamma_A$, but the condition also requires a small $1/\gamma_A$ so that the above error bound is valid. Also, since γ_A is the minimum spacing of the eigenvalues of $M = \nabla^2 f_{A\mu}(\phi)(\nabla^2 f_{A\mu}(\psi))^{-1}$, the eigenvalue perturbations imposed by the noise cannot be too large compared to γ_A without potentially ruining the eigenvectors of M; thus, the dependence on γ_A seems to necessary.

Despite the important role that γ_A plays in the efficiency of the HKICA algorithm, it is not clear how it depends on different properties of A. To the best of our knowledge, even a polynomial (in the dimension d) lower bound of γ_A is not yet available in the literature. Similar problems have been discussed by Hüsler (1987) and Goyal et al. (2014), but there solutions are not applicable to our case.

5. A Refined HKICA Algorithm

The problems with γ_A motivate us to refine the HKICA algorithm. The idea is inspired by Arora et al. (2012) and Frieze et al. (1996) using a quasi-whitening procedure:

One can show that $\nabla^2 f_\mu(\psi) = AKD_\psi A^\top$ where $D_\psi = \text{diag}\left((\psi^\top A_1)^2, \cdots, (\psi^\top A_d)^2\right)$, and so $B = AK^{1/2}D_\psi^{1/2}R^\top$ for some orthonormal matrix R. Defining $T_i = \nabla^2 f_\mu(B^{-\top}\phi_i)$, one can calculate that $T_i = AK^{1/2}D_\psi^{-1/2}\Lambda_i A^\top$ where $\Lambda_i = \text{diag}\left((\phi_i^\top R_1)^2, \ldots, (\phi_i^\top R_d)^2\right)$ and R_i denote the ith column of R. Then $M = T_1 T_2^{-1} = A\Lambda A^{-1}$ with $\Lambda = \Lambda_1\Lambda_2^{-1} = \text{diag}\left(\left(\frac{\phi_1^\top R_1}{\phi_2^\top R_1}\right)^2, \ldots, \left(\frac{\phi_1^\top R_d}{\phi_2^\top R_d}\right)^2\right)$. Thus, A_i are again the eigenvectors of M, but now the eigenvalues of M are defined in terms of the orthogonal matrix R instead of A, and so the resulting minimum spacing

$$\gamma_R = \min_{i,j:i\neq j}\left|\left(\frac{\phi_1^\top R_i}{\phi_2^\top R_i}\right)^2 - \left(\frac{\phi_1^\top R_j}{\phi_2^\top R_j}\right)^2\right| \tag{9}$$

is much easier to handle.

The resulting algorithm, called Deterministic ICA (DICA), is shown in Algorithm 2. Note that on the event $\mathcal{E}_\phi$, $\|\phi_j^\top R\|_2 \leq L_u$, $j \in \{1,2\}$. We will show later that this event $\mathcal{E}_\phi$, as well as other events defined later, will hold simultaneously with high probability.

Algorithm 2 Deterministic ICA (DICA)

input $x(t)$ for $1 \leq t \leq T$.
output An estimation of the mixing matrix A.

1: Sample ψ from a d-dimensional standard Gaussian distribution;
2: Evaluate $\nabla^2\hat{f}(\psi)$,
3: Compute $\hat{B}$ such that $\nabla^2\hat{f}(\psi) = \hat{B}\hat{B}^\top$;
4: Sample ϕ_1 and ϕ_2 independently from the standard Gaussian distribution;
5: Compute $\hat{T}_1 = \nabla^2\hat{f}(\hat{B}^{-\top}\phi_1)$ and $\hat{T}_2 = \nabla^2\hat{f}(\hat{B}^{-\top}\phi_2)$;
6: Compute all the eigenvectors of $\hat{M} = \hat{T}_1\left(\hat{T}_2\right)^{-1}$, $\{\mu_1, \ldots, \mu_d\}$;
7: Return $\hat{A} = \{\mu_1, \ldots, \mu_d\}$.

Definition 5.1. Let $\mathcal{E}_\phi$ denote the following event: For some fixed constant $L_u \geq \sqrt{2d}$ and ℓ_l such that $\ell_l = \frac{\sqrt{\pi}}{\sqrt{2d}}\ell$ for $0 \leq \ell \leq 1$, $\|\phi_1\|_2 \leq L_u$, $\|\phi_2\|_2 \leq L_u$, and $\min_i\{|\phi_2^\top R_i|\} \geq \ell_l$ hold simultaneously.

Similarly to Theorem 4.2, one can show that under some technical assumptions, which hold with probability 1 if ξ, $P(L_u)$, and $P\left(\frac{\sqrt{3}L_u}{\sqrt{2}\sigma_{\min}\kappa_{\min}^{1/2}C_1}\right)$ are small enough, on the event $\mathcal{E}_\psi \cap \mathcal{E}_\phi$, there exists a permutation π and constants $\{c_1, \ldots, c_d\}$, such that for $1 \leq k \leq d$,

$$\|c_k\hat{A}_{\pi(k)} - A_k\|_2 \leq \frac{4\sigma_{\max}^2}{\gamma_R\sigma_{\min}}\tilde{Q},$$

where $\hat{A}$ is the output of the DICA algorithm and $\tilde{Q}$ is polinomial in the usual problem parameters and decays roughly as $(\xi + P(L_u))$. Details are given in the full version of the paper (Huang et al., 2015). It is very similar to the result of Theorem 4.2, with γ_R in place of γ_A, as required.

To analyze γ_R analytically, note that ϕ_1 and ϕ_2 are independently sampled from standard Gaussian distribution. Thus, $\{\phi_1^\top R_1, \cdots, \phi_1^\top R_d, \phi_2^\top R_1, \cdots, \phi_2^\top R_d\}$ are $2d$ independent standard Gaussian random variables. Let $Z_i = \frac{\phi_1^\top R_i}{\phi_2^\top R_i}$. Therefore, Z_i, $1 \leq i \leq d$ are d independent Cauchy$(0,1)$ random variables. Using this observation, we show in the full version (Huang et al., 2015) that, among others, $\gamma_R \geq \frac{\delta}{2d^2}$ with probability at least $1-\delta$.

Based on the above, one can show that Theorem 3.2 holds for DICA (Huang et al., 2015). Furthermore, a heuristic modification of DICA can also be derived that performs better in the experiments, but proving performance guarantees for that algorithm has defied our efforts so far (details are given in the full version of the paper, Huang et al. 2015).

5.1. Recursive Versions

Recently, Vempala and Xiao (2014) proposed a recursion idea to improve the sample complexity of the Fourier PCA algorithm of Goyal et al. (2014). Instead of recovering all the columns of A in a single eigen-decomposition, the recursive algorithm only decomposes the whole space into two subspaces according to the maximal spacing of the eigenvalues, then recursively decomposes each subspaces until they are all 1-dimensional. The insight of this recursive procedure is the following: when the maximal spacing of the eigenvalues are much larger than the minimal one, the algorithm may win over a single decomposition even with the accumulating errors through the recursion. However, this algorithm is based on the assumption that the mixing matrix is orthonormal, so that the projection to its subspaces can always eliminate some component of the source signal.

We adapt the above idea to our algorithms. Due to space limitations, we will only consider the simplest recursive algorithm, the recursive version of HKICA, as an example.

To force an orthonormal mixing matrix, we will first compute the square root matrix B of $\nabla^2 f(\psi) = AD_\psi KA^\top$. Thus $B = AD_\psi^{1/2}K^{1/2}R^\top$ for some orthonormal matrix R. Transforming our observations by B^{-1}, we have the new observations $y(t) = B^{-1}x(t) + B^{-1}\epsilon(t) = RD_\psi^{1/2}K^{1/2}s(t) + B^{-1}\epsilon(t)$. Note that transformed noise

vector $B^{-1}\epsilon(t)$ is still Gaussian. Also, $D_\psi^{1/2}K^{1/2}$ is diagonal, thus $RD_\psi^{1/2}K^{1/2}s(t)$ is an orthonormal mixture of independent sources. We then apply the recursive algorithm to recover the mixing matrix R. Finally, BR gives an estimate of A up to scaling.

To recover R using a recursive algorithm, we follow the idea of HKICA (and DICA) to compute two Hessian matrices $T_1 = RD_\psi^{-1}\Lambda_1 R^\top$ and $T_2 = RD_\psi^{-1}\Lambda_2 R^\top$. Then, instead of computing the eigen-decomposition of $T_0 = T_1 T_2^{-1}$ (as in HKICA), we only decompose its eigenspace into two subspaces, according to the maximal spacing of the eigenvalues of T_0. The *Decompose* helper function takes a projection matrix P of a subspace spanned by some columns of R (WLOG we assume it is the first k columns of R). Then we compute the projection of T_0 as $M = P^\top T_0 P$. Thus the eigenspace of $PMP^\top$ is in the span of P. Lastly, by separating the eigenvectors of M according to its eigenvalues into PP_1 and PP_2, the Decompose function repeatedly decomposes the subspaces into two smaller subspaces.

Algorithm 3 Recursive version of HKICA (HKICA.R)

input $x(t)$ for $1 \le t \le T$.
output An estimation of the mixing matrix A.

1: Sample ψ from a d-dimensional standard Gaussian distribution;
2: Evaluate $\nabla^2 \hat{f}(\psi) = \hat{G}(\psi)$;
3: Compute $\hat{B}$ such that $\nabla^2 \hat{f}(\psi) = \hat{B}\hat{B}^\top$;
4: Compute $\hat{y}(t) = \hat{B}^{-1}x(t)$ for $1 \le t \le T$;
5: Let $P = I_d$;
6: Compute $\hat{R} = \text{Decompose}(\hat{y}, P)$;
7: Return $\hat{B}\hat{R}$;

Algorithm 4 The Decompose helper function

input $x(t)$ for $1 \le t \le T$, a projection matrix $P \in \mathbb{R}^{d\times k}$ ($d \ge k$).
output An estimation of the mixing matrix $A \in \mathbb{R}^{d\times k}$.

1: if $k == 1$, Return P;
2: Sample ϕ_1 and ϕ_2 independently from a standard Gaussian distribution of dimension d;
3: Evaluate $\nabla^2 \hat{f}(\phi_1)$ and $\nabla^2 \hat{f}(\phi_2)$,
4: Compute $\hat{T} = (\nabla^2 \hat{f}(\phi_1))(\nabla^2 \hat{f}(\phi_2))^{-1}$;
5: Compute $\hat{M} = P^\top \hat{T} P$;
6: Compute all the eigen-decomposition of $\hat{M}$, its eigenvalues$\{\sigma_1, \ldots, \sigma_d\}$ where $\sigma_1 \ge \ldots \ge \sigma_k$ and their corresponding eigenvectors $\{\mu_1, \ldots, \mu_k\}$;
7: Find the index $m = \arg\max \sigma_m - \sigma_{m+1}$;
8: Let $P_1 = (\mu_1, \ldots, \mu_m)$, and $P_2 = (\mu_{m+1}, \ldots, \mu_k)$;
9: Compute $W_1 = \text{Decompose}(x, PP_1)$, and $W_2 = \text{Decompose}(x, PP_2)$;
10: Return $[W_1, W_2]$;

Remark 5.2. Other algorithms can be modified into a recursive version in a similar way.

Theorem 5.3. *Under the conditions of Theorem 3.2, with probability at least $1-\delta$, the recursive version of HKICA returns a mixing matrix $\hat{A}$ with an error $\|\hat{A} - ADP\|_2$ bounded by*

$$\text{Poly}\left(d, \frac{1}{\kappa_{\min}}, \frac{1}{\sigma_{\min}}, \frac{1}{\ell}, L_u, L, C, \sigma_{\max}\right)(\tilde{Q}^2 + \bar{\xi})$$

for some diagonal matrix D and permutation matrix P.

Remark 5.4. Note that when T is large enough, the term $\tilde{Q}^2$ will be dominated by $\bar{\xi}$, which is the error carried over from quasi-whitening. The recursion idea improves the sample complexity of the eigen-decomposition (to recover the orthonormal mixing matrix R).

6. Experimental Results

In this section we compare the performance of different ICA algorithms in some synthetic examples, with mixing matrices of different coherences.

We test 9 algorithms: HKICA (HKICA), and its recursive version (HKICA.R); DICA (DICA), and its recursive version (DICA.R); the modified version of DICA (MDICA), and its recursive version (MDICA.R); the default FastICA algorithm from the 'ITE' toolbox (Szabó et al., 2012) (FICA); the recursive Fourier PCA algorithm of Xiao (2014) (FPCA); and random guessing (Random). FPCA is modified so that it can be applied to the case of non-orthogonal mixing matrix.

In the simulation, a common mixing matrix A of dimension 6 is generated in the following ways: We construct four kinds of matrices: $A_1 = P$; $A_2 = v_b \times \mathbf{1}' + 0.3 \times P$; $A_3 = v_b \times \mathbf{1}' + 0.05 \times P$; and $A_4 = v_b \times \mathbf{1}' + 0.005 \times P$. Here the vector v_b and the matrix P are both generated from standard normal distribution (with different dimensions). Then all the mixing matrices are rescaled to a same magnitude. We also generate an orthonormal mixing matrix R, obtained by computing the left column space of a non-singular random matrix (from standard normal distribution). Then we generate a 6-dimensional BPSK signal s as follows. Let $p = (\sqrt{2}, \sqrt{5}, \sqrt{7}, \sqrt{11}, \sqrt{13}, \sqrt{19})$. We generate a $\{+1, -1\}$ valued sequence $q(t)$ uniformly at random for $1 \le t \le T$, and set $s_i(t) = q(t)i \times \sin(p_i t)$. Note that in order to have the components of s close to independent, we need the ratio of their frequencies are irrational.

Lastly, the observed signal is generated as $x = As + c\epsilon$ where ϵ is the noise generated from a d-dimensional normal distribution with randomly generated covariance. We take $T = 20000$ instances of the observed signal on time steps $t = 1, \ldots, 20000$. We test the noise ratio c from 0 (noise-

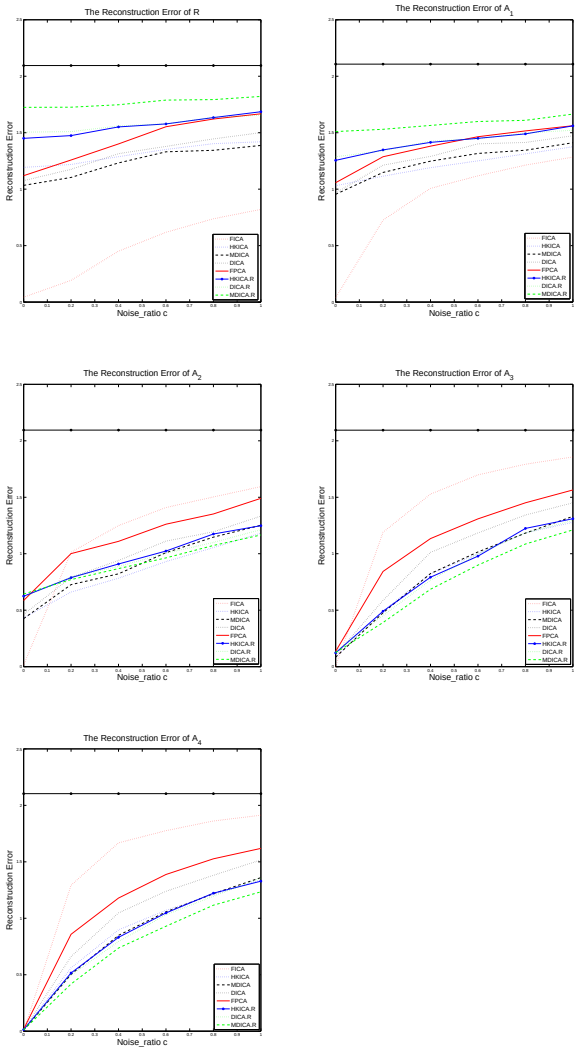

Figure 2. Reconstruction Error

free) to 1 (very noisy). All the algorithms are evaluated on a 150 repetitions. For each repetition, we try 3 times and report the best.

We measure the performances of the algorithms by its actual reconstruction error. In particular, we evaluate the following quantity between the true mixing matrix A and the estimate $\hat{A}$ returned by the algorithms: $\min_{\Pi,S} \|\hat{A}\Pi S - A\|_{\text{Frob}}$, where Π is a permutation matrix, and S is a column scaling matrix (diagonal). The calculation of this measure would require a exhaust search for the optimal permutation.

6.1. Results

We report the reconstruction errors for different kinds of mixing matrices and noise ratios.

The experimental results suggest that moment methods are more robust to high-coherence mixing matrices and Gaussian noise than FastICA. FastICA achieves the best performance in case of low coherence. As the coherence of the mixing matrix A increases, its performance decreases quickly and becomes sensitive to noise.

We expected that DICA will achieve smaller error for an extremely coherent A, since $1/\gamma_A$ will be much larger than $1/\gamma_R$. However, the experimental results indicate the opposite. Note that high coherence implies small minimal singular value. In this case, the estimation error of M in DICA could be much larger than that in HKICA, because of the fourth degree of A^{-1}. This error overwhelms the improvement brought by larger eigenvalue spacings, if the sample size is not large enough. The investigation of this phenomenon is left for future work.

On the other hand, MDICA tries to achieve a small estimation error, meanwhile we expect it to keep the eigenvalue spacing large (intuitively, it is approximately the spacing of the square of d Gaussian random variables), leading to good performance. This is confirmed by the experimental results, in both the non-recursive and recursive versions.

The recursive idea is not always helpful for the moment methods. For a highly coherent A, the recursive versions outperform their non-recursive counterparts. Note that in this case, A is close to singular (small minimal singular value), and thus it requires more samples. On the other hand, when A has relatively low coherence, the estimation error of the fourth moments contributes more to the reconstruction error. Recursive algorithms suffers from making several such estimations.

In summary, the results suggest that these moment methods are comparable to each other in practice, while FastICA is better for mixing matrices with low coherence or mild coherence with low noise. If the mixing matrix is orthonormal, then FPCA performs better than the other algorithms. If the observations have large noise and the mixing matrix is not extremely coherent, then HKICA may be the best choice. In the case of an extremely coherent mixing matrix, MDICA performs the best. Also, the recursive idea is very helpful for small sample sizes.

7. Conclusions

We considered the problem of independent component analysis with noisy observation. For the first time in the literature, we presented ICA algorithms that can recover non-Gaussian source signals with polynomial computational complexity and provable performance guarantees on the reconstruction error that guarantee that for T samples the reconstruction error vanishes at a $1/\sqrt{T}$ rate and depends only polynomially on the natural parameters of the problem. The algorithms do not depend on unknown problem parameters, and also extend to deterministic sources with approximately independent empirical distributions.

Acknowledgements

This work was supported by the Alberta Innovates Technology Futures and NSERC.

References

A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky. Tensor decompositions for learning latent variable models. *CoRR*, abs/1210.7559, 2012a.

A. Anandkumar, D. Hsu, and S. M. Kakade. A method of moments for mixture models and hidden markov models. *arXiv preprint arXiv:1203.0683*, 2012b.

A. Anandkumar, R. Ge, and M. Janzamin. Guaranteed non-orthogonal tensor decomposition via alternating rank-1 updates. *arXiv preprint arXiv:1402.5180*, 2014.

S. Arora, R. Ge, A. Moitra, and S. Sachdeva. Provable ica with unknown gaussian noise, with implications for gaussian mixtures and autoencoders. In *Advances in Neural Information Processing Systems*, pages 2375–2383, 2012.

J. Cardoso. High-order contrasts for independent component analysis. *Neural computation*, 11(1):157–192, 1999.

A. Chen and P. J Bickel. Consistent independent component analysis and prewhitening. *Signal Processing, IEEE Transactions on*, 53(10):3625–3632, 2005.

A. Chen, P. J Bickel, et al. Efficient independent component analysis. *The Annals of Statistics*, 34(6):2825–2855, 2006.

P. Comon and C. Jutten. *Handbook of Blind Source Separation: Independent component analysis and applications*. Academic press, 2010.

A. DasGupta. Finite sample theory of order statistics and extremes. In *Probability for Statistics and Machine Learning*, pages 221–248. Springer, 2011.

A. Dermoune and T. Wei. FastICA algorithm: Five criteria for the optimal choice of the nonlinearity function. *IEEE transactions on signal processing*, 61(5-8):2078–2087, 2013.

J. Eriksson and V. Koivunen. Characteristic-function-based independent component analysis. *Signal Processing*, 83 (10):2195–2208, 2003.

A. Frieze, M. Jerrum, and R. Kannan. Learning linear transformations. In *37th IEEE Annual Symposium on Foundations of Computer Science*, pages 359–359. IEEE Computer Society, 1996.

N. Goyal, S. Vempala, and Y. Xiao. Fourier PCA and robust tensor decomposition. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, pages 584–593. ACM, 2014.

D. Hsu and S. M. Kakade. Learning mixtures of spherical Gaussians: moment methods and spectral decompositions. In *Proceedings of the 4th conference on Innovations in Theoretical Computer Science*, pages 11–20. ACM, 2013.

R. Huang, A. György, and Cs. Szepesvári. Deterministic independent component analysis. in preparation, 2015.

J. Hüsler. Minimal spacings of non-uniform densities. *Stochastic processes and their applications*, 25:73–81, 1987.

A. Hyvarinen. Fast and robust fixed-point algorithms for independent component analysis. *Neural Networks, IEEE Transactions on*, 10(3):626–634, 1999.

A. Hyvärinen and E. Oja. Independent component analysis: algorithms and applications. *Neural Networks*, 13(4–5): 411–430, 2000.

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338, 2000.

S. Miettinen, J.and Taskinen, K. Nordhausen, and H. Oja. Fourth moments and independent component analysis. *arXiv preprint arXiv:1406.4765*, 2014.

E. Oja and Z. Yuan. The FastICA algorithm revisited: Convergence analysis. *Neural Networks, IEEE Transactions on*, 17(6):1370–1381, 2006.

E. Ollila. The deflation-based FastICA estimator: statistical analysis revisited. *Signal Processing, IEEE Transactions on*, 58(3):1527–1541, 2010.

A. Samarov, A. Tsybakov, et al. Nonparametric independent component analysis. *Bernoulli*, 10(4):565–582, 2004.

G.W. Stewart and J.-g. Sun. *Matrix perturbation theory*. Computer science and scientific computing. Academic Press, 1990. ISBN 9780126702309.

Z. Szabó, B. Póczos, and A. Lőrincz. Separation theorem for independent subspace analysis and its consequences. *Pattern Recognition*, 45:1782–1791, 2012.

P. Tichavsky, Z. Koldovsky, and E. Oja. Performance analysis of the FastICA algorithm and Cramér-Rao bounds for linear independent component analysis. *Signal Processing, IEEE Transactions on*, 54(4):1189–1203, 2006.

S. Vempala and Y. Xiao. Max vs min: Independent component analysis with nearly linear sample complexity. *CoRR*, abs/1412.2954, 2014. URL http://arxiv.

org/abs/1412.2954.

T. Wei. The convergence and asymptotic analysis of the generalized symmetric FastICA algorithm. *arXiv preprint arXiv:1408.0145*, 2014.

Y. Xiao. Fourier pca package. *GitHub*, 2014. URL https://github.com/yingusxiaous/libFPCA.