Asymptotic Analysis of FastICA Algorithm with Finite Sample

Tianwen Wei

Laboratoire Paul Painlevé, USTL

16/4/2012

The generic model for linear *independent component analysis* (ICA) is defined by:

$$\mathbf{x} = \mathbf{A}\mathbf{s}, \tag{1.1}$$

where

- $\mathbf{x} = (x_1, \ldots, x_n)^{\mathsf{T}}$ is the observed signal,
- $\mathbf{s} = (s_1, \ldots, s_m)^{\mathsf{T}}$ is the (unknown) source signal that has mutually independent components,
- $\mathbf{A} \in \mathbb{R}^{n \times m}$ is a (unknown) mixing matrix.

In the sequel, we restrain ourself to the case that

$$\mathrm{Dim}(\mathbf{x}) = \mathrm{Dim}(\mathbf{s}) = d,$$

which is the most common setting.

The generic model for linear *independent component analysis* (ICA) is defined by:

$$\mathbf{x} = \mathbf{A}\mathbf{s}, \tag{1.1}$$

where

- $\mathbf{x} = (x_1, \ldots, x_n)^{\mathsf{T}}$ is the observed signal,
- $\mathbf{s} = (s_1, \ldots, s_m)^{\mathsf{T}}$ is the (unknown) source signal that has mutually independent components,
- $\mathbf{A} \in \mathbb{R}^{n \times m}$ is a (unknown) mixing matrix.

In the sequel, we restrain ourself to the case that

$$\mathrm{Dim}(\mathbf{x}) = \mathrm{Dim}(\mathbf{s}) = d,$$

which is the most common setting.

The aim of ICA is to recover the independent components of $\mathbf{s}$ based on the observed signal $\mathbf{x}$ only.
More precisely, we wish to find vectors $\mathbf{w}_1, \ldots, \mathbf{w}_d$ such that z_i defined by

$$z_i \overset{def}{=} \mathbf{w}_i^{\mathsf{T}} \mathbf{x}, \quad i = 1, \ldots, d$$

are mutually independent.

To achieve the goal of ICA, let us make the following assumptions:

- $s_1, \ldots, s_d$ are mutually independent.
  This is the basic assumption for ICA.
- $\mathbf{s}$ has unit second moment, i.e. $\mathbb{E}[\mathbf{s}\mathbf{s}^T] = \mathbf{I}$.
  This assumption is made to avoid an identifiability issue, which is the fact that both $\mathbf{A}$ and $\mathbf{s}$ being unknown to us, we can never truly determine the magnitude of $\mathbf{s}$, because any scalar multiplier to a component of $\mathbf{s}$ could always be canceled by dividing the corresponding column of $\mathbf{A}$ by the same scalar.
- The components of $\mathbf{s}$ are not Gaussian.
  Gaussian signals contain least information, and hence they are considered as pure noise. Using ICA approach, one cannot estimate recover Gaussian signal.

To achieve the goal of ICA, let us make the following assumptions:

- $s_1, \ldots, s_d$ are mutually independent.
  This is the basic assumption for ICA.
- **s** has unit second moment, i.e. $\mathbb{E}[\mathbf{s}\mathbf{s}^T] = \mathbf{I}$.
  This assumption is made to avoid an identifiability issue, which is the fact that both **A** and **s** being unknown to us, we can never truly determine the magnitude of **s**, because any scalar multiplier to a component of **s** could always be canceled by dividing the corresponding column of **A** by the same scalar.
- The components of **s** are not Gaussian.
  Gaussian signals contain least information, and hence they are considered as pure noise. Using ICA approach, one cannot estimate recover Gaussian signal.

To achieve the goal of ICA, let us make the following assumptions:

- $s_1, \ldots, s_d$ are mutually independent.
  This is the basic assumption for ICA.
- **s** has unit second moment, i.e. $\mathbb{E}[\mathbf{s}\mathbf{s}^T] = \mathbf{I}$.
  This assumption is made to avoid an identifiability issue, which is the fact that both **A** and **s** being unknown to us, we can never truly determine the magnitude of **s**, because any scalar multiplier to a component of **s** could always be canceled by dividing the corresponding column of **A** by the same scalar.
- The components of **s** are not Gaussian.
  Gaussian signals contain least information, and hence they are considered as pure noise. Using ICA approach, one cannot estimate recover Gaussian signal.

Besides, since we can observe $\mathbf{x}$, its mean $\mathbb{E}[\mathbf{x}]$ and covariance matrix $\mathbf{\Sigma_x}$ are available to us. Now define

$$\begin{aligned}\mathbf{y} &\overset{def}{=} \mathbf{\Sigma}^{-1/2}(\mathbf{x} - \mathbb{E}[\mathbf{x}]) \\ \tilde{\mathbf{s}} &\overset{def}{=} \mathbf{s} - \mathbb{E}[\mathbf{s}] \\ \widetilde{\mathbf{A}} &\overset{def}{=} \mathbf{\Sigma}^{-1/2}\mathbf{A}.\end{aligned}$$

Then it is clear that

$$\mathbf{y} = \widetilde{\mathbf{A}}\tilde{\mathbf{s}},$$

where

- The transformed vector $\mathbf{y}$ has zero mean and unit variance.
- The transformed source signal $\tilde{\mathbf{s}}$, in spite of being unknown to us, it has zero mean and unit variance.
- As a result, the matrix $\tilde{\mathbf{A}}$ is orthogonal.

Besides, since we can observe $\mathbf{x}$, its mean $\mathbb{E}[\mathbf{x}]$ and covariance matrix $\mathbf{\Sigma_x}$ are available to us. Now define

$$\begin{aligned}\mathbf{y} &\overset{def}{=} \mathbf{\Sigma}^{-1/2}(\mathbf{x} - \mathbb{E}[\mathbf{x}]) \\ \tilde{\mathbf{s}} &\overset{def}{=} \mathbf{s} - \mathbb{E}[\mathbf{s}] \\ \widetilde{\mathbf{A}} &\overset{def}{=} \mathbf{\Sigma}^{-1/2}\mathbf{A}.\end{aligned}$$

Then it is clear that

$$\mathbf{y} = \widetilde{\mathbf{A}}\tilde{\mathbf{s}},$$

where

- The transformed vector $\mathbf{y}$ has zero mean and unit variance.
- The transformed source signal $\tilde{\mathbf{s}}$, in spite of being unknown to us, it has zero mean and unit variance.
- As a result, the matrix $\tilde{\mathbf{A}}$ is orthogonal.

In view of the discussion above, it is reasonable to make the following two additional assumptions :

- The source signal $\mathbf{s}$ has zero mean, i.e. $\mathbb{E}[\mathbf{s}] = 0$.
- The mixing matrix $\mathbf{A}$ is orthogonal, i.e. $\mathbf{A}\mathbf{A}^{\mathsf{T}} = \mathbf{I}$.

Clearly enough, these two assumptions allow us to work with the observed signal $\mathbf{x}$ that is already perfectly centered and whitened, i.e.

$$\mathbb{E}[\mathbf{x}] = 0,$$
$$\mathbb{E}[\mathbf{x}\mathbf{x}^{\mathsf{T}}] = \mathbf{I}.$$

In view of the discussion above, it is reasonable to make the following two additional assumptions :

- The source signal $\mathbf{s}$ has zero mean, i.e. $\mathbb{E}[\mathbf{s}] = 0$.
- The mixing matrix $\mathbf{A}$ is orthogonal, i.e. $\mathbf{A}\mathbf{A}^{\mathsf{T}} = \mathbf{I}$.

Clearly enough, these two assumptions allow us to work with the observed signal $\mathbf{x}$ that is already perfectly centered and whitened, i.e.

$$\mathbb{E}[\mathbf{x}] = 0,$$
$$\mathbb{E}[\mathbf{x}\mathbf{x}^{\mathsf{T}}] = \mathbf{I}.$$

The primitive idea of ICA is stated as follows:

Sum (or linear mixture) of independent random variables
tends to be gaussian, by Central Limit Theorem

$$\Updownarrow$$

independent components are less gaussian
than their mixture

Inspired by the idea above, to recover one of the source signal s_i, we look for the vector $\mathbf{w}_*$ such that the random variable $\mathbf{w}_*^T\mathbf{x}$ is "the least gaussian". Here we restrict our searching of $\mathbf{w}_*$ on the unit sphere $\mathcal{S}$, since by assumption both $\mathbf{x}$ and $\mathbf{s}$ have unit variance.

The primitive idea of ICA is stated as follows:

Sum (or linear mixture) of independent random variables
tends to be gaussian, by Central Limit Theorem

$$\Updownarrow$$

independent components are less gaussian
than their mixture

Inspired by the idea above, to recover one of the source signal s_i, we look for the vector $\mathbf{w}_*$ such that the random variable $\mathbf{w}_*^T\mathbf{x}$ is "the least gaussian". Here we restrict our searching of $\mathbf{w}_*$ on the unit sphere $\mathcal{S}$, since by assumption both $\mathbf{x}$ and $\mathbf{s}$ have unit variance.

Example: kurtosis as measure of non-gaussianity

The classical measure of non-gaussianity is kurtosis or the fourth-order cumulant. The kurtosis of a random variable Y is classically defined by

$$k_4(Y) = \mathbb{E}[Y^4] - 3\mathbb{E}[Y^2].$$

Since kurtosis is zero for a gaussian random variable and non zero for most non gaussian random variables, we can use its square or absolute value to measure the non-gaussianity, i.e. we propose

$$\mathbf{w}_* = \underset{\mathbf{w}\in\mathcal{S}}{\operatorname{argmax}} \left(k_4(\mathbf{w}^\mathsf{T}\mathbf{x})\right)^2. \tag{1.2}$$

There exists many other measures of non-gaussianity, e.g. neg-entropy, etc.

Problems concerning the maximization of non-gaussianity, as shown in the example of kurtosis, always involve searching the extremum on the unit sphere of a function of the type

$$\Phi(w) = \mathbb{E}[G(\mathbf{w}^{\mathsf{T}}\mathbf{x})],$$

where G is some smooth function. In the sequel, we shall refer to $\Phi(\cdot)$ as the contrast function.

The following result confirms that the extrema of Φ on the unit sphere are the solution to our ICA problem:

Proposition

Suppose that $\mathbf{x}$ *follows model*

$$\mathbf{x} = \mathbf{A}\mathbf{s}$$

and the assumptions stated above hold. Denote $g = G'$. *If*

$$\mathbb{E}[g'(s_i) - s_i g(s_i)] \neq 0,$$

then the i*th column vectors* $\mathbf{a}_i$ *of the mixing matrix* $\mathbf{A}$ *is a local extrema of* $\Phi(\cdot)$ *on the unit sphere* $\mathcal{S}$.

By assumption $\mathbf{A}\mathbf{A}^{\mathsf{T}} = \mathbf{I}$, we have for $i = 1, \ldots, d$,

$$\mathbf{a}_i^{\mathsf{T}}\mathbf{x} = \mathbf{a}_i^{\mathsf{T}}\mathbf{A}\mathbf{s} = \mathbf{e}_i^{\mathsf{T}}\mathbf{s} = s_i.$$

The fixed point algorithm, also known as FastICA, is one of the most successful algorithms for ICA in terms of accuracy and low computational complexity. It can be implemented as follows:

- To begin with, choose an arbitrary initial vector $\mathbf{w}$ on the unit sphere $\mathcal{S}$.
- Run the following iteration until convergence:

$$\mathbf{w}^+ \leftarrow \mathbb{E}[g(\mathbf{w}^\mathsf{T}\mathbf{x})\mathbf{x} - g'(\mathbf{w}^\mathsf{T}\mathbf{x})\mathbf{w}] \tag{1.3}$$

$$\mathbf{w} \leftarrow \frac{\mathbf{w}^+}{\|\mathbf{w}^+\|} \tag{1.4}$$

The following theoretical result concerning FastICA is well known:

Theorem

If the initial choice $\mathbf{w}$ *is close enough to* $\mathbf{a}_i$ *for some* i*, then iteration (1.3) and (1.4) converges to* $\mathbf{a}_i$*, provided that*

$$\mathbb{E}[g'(s_i) - s_i g(s_i)] \neq 0.$$

In practice, people have independent observations $\mathbf{x}(1), \ldots, \mathbf{x}(N)$ that follow model (1.1), i.e.

$$\mathbf{x}(k) = \mathbf{A}\mathbf{s}(k), \quad k = 1, \ldots, N,$$

where $\mathbf{s}(1), \ldots, \mathbf{s}(k)$ are accordingly N independent realizations of $\mathbf{s}$. To implement FastICA, one has to compute the mathematical expectations appeared in

$$\mathbf{w}^+ \leftarrow \mathbb{E}[g(\mathbf{w}^{\mathsf{T}}\mathbf{x})\mathbf{x} - g'(\mathbf{w}^{\mathsf{T}}\mathbf{x})\mathbf{w}].$$

This is usually achieved by using the empirical approximation:

$$\begin{aligned}\mathbf{w}^+ &\leftarrow \widehat{\mathbb{E}}_N[g(\mathbf{w}^{\mathsf{T}}\mathbf{x})\mathbf{x} - g'(\mathbf{w}^{\mathsf{T}}\mathbf{x})\mathbf{w}] \\ &\stackrel{def}{=} \frac{1}{N}\sum_{k=1}^{N}\Big[g\big(\mathbf{w}^{\mathsf{T}}\mathbf{x}(k)\big)\mathbf{x}(k) - g'\big(\mathbf{w}^{\mathsf{T}}\mathbf{x}(k)\big)\mathbf{w}\Big].\end{aligned}$$

Now question arises: how does the empirical approximation alter the behavior of FastICA algorithm? This question can be decomposed into several smaller questions:

- Does the algorithm that uses empirical approximation converge?

  Due to the probabilistic nature of the empirical approximation, we cannot expect to give a deterministic answer.
- If it does converge, what is its limit?

  Since the limit will be dependent on the specific realizations $\mathbf{x}(1), \ldots, \mathbf{x}(N)$ which are random, we're pretty sure that the limit will not coincide with $\mathbf{a}_i$ for any i.
- If the limit exists, what do we know about its asymptotic behavior?

  No idea for the first glance.

Now question arises: how does the empirical approximation alter the behavior of FastICA algorithm? This question can be decomposed into several smaller questions:

- Does the algorithm that uses empirical approximation converge?

  Due to the probabilistic nature of the empirical approximation, we cannot expect to give a deterministic answer.
- If it does converge, what is its limit?

  Since the limit will be dependent on the specific realizations $\mathbf{x}(1), \ldots, \mathbf{x}(N)$ which are random, we're pretty sure that the limit will not coincide with $\mathbf{a}_i$ for any i.
- If the limit exists, what do we know about its asymptotic behavior?

  No idea for the first glance.

Now question arises: how does the empirical approximation alter the behavior of FastICA algorithm? This question can be decomposed into several smaller questions:

- Does the algorithm that uses empirical approximation converge?

  Due to the probabilistic nature of the empirical approximation, we cannot expect to give a deterministic answer.

- If it does converge, what is its limit?

  Since the limit will be dependent on the specific realizations $\mathbf{x}(1), \ldots, \mathbf{x}(N)$ which are random, we're pretty sure that the limit will not coincide with $\mathbf{a}_i$ for any i.

- If the limit exists, what do we know about its asymptotic behavior?

  No idea for the first glance.

Now question arises: how does the empirical approximation alter the behavior of FastICA algorithm? This question can be decomposed into several smaller questions:

- Does the algorithm that uses empirical approximation converge?

  Due to the probabilistic nature of the empirical approximation, we cannot expect to give a deterministic answer.

- If it does converge, what is its limit?

  Since the limit will be dependent on the specific realizations $\mathbf{x}(1), \ldots, \mathbf{x}(N)$ which are random, we're pretty sure that the limit will not coincide with $\mathbf{a}_i$ for any i.

- If the limit exists, what do we know about its asymptotic behavior?

  No idea for the first glance.

The purpose of my work is to answer these questions. The main results are:

- For any $\varepsilon > 0$, there exists $N(\varepsilon)$ such that for sample size $N > N(\varepsilon)$, the empirical FastICA converges to some local extremum of the empirical contrast function

$$\widehat{\Phi}(\mathbf{w}) \stackrel{def}{=} \frac{1}{N} \sum_{k=1}^{N} G\left(\mathbf{w}^{\mathsf{T}} \mathbf{x}(k)\right)$$

  on the unit sphere with probability larger than $1 - \varepsilon$.

- The limit of empirical FastICA is asymptotically normal:

$$N^{1/2}(\hat{\mathbf{v}}_N - \mathbf{v}) \xrightarrow[N\to\infty]{\mathscr{D}} \mathcal{N}(0, \mathbf{\Sigma}),$$

where

- $\hat{\mathbf{v}}_N$ is the limit of the empirical FastICA,
- $\mathbf{v}$ is the corresponding limit of the theoretical FastICA,
- $\mathbf{\Sigma}$ is the asymptotic covariance matrix

$$\mathbf{\Sigma} = \frac{\mathbb{E}\left[\left(g(\mathbf{v}^{\mathsf{T}}\mathbf{x})\right)^2\right](\mathbf{I} - \mathbf{v}\mathbf{v}^{\mathsf{T}})}{\left(\mathbb{E}[g'(\mathbf{v}^{\mathsf{T}}\mathbf{x}) - g(\mathbf{v}^{\mathsf{T}}\mathbf{x})\mathbf{v}^{\mathsf{T}}\mathbf{x}]\right)^2}.$$

Denote

$$\mathbf{T}(\mathbf{w}) = \frac{\mathbb{E}[g(\mathbf{w}^{\mathsf{T}}\mathbf{x})\mathbf{x} - g'(\mathbf{w}^{\mathsf{T}}\mathbf{x})\mathbf{w}]}{\|\mathbb{E}[g(\mathbf{w}^{\mathsf{T}}\mathbf{x})\mathbf{x} - g'(\mathbf{w}^{\mathsf{T}}\mathbf{x})\mathbf{w}]\|}$$

$$\mathbf{T}_N(\mathbf{w}) = \frac{\widehat{\mathbb{E}}_N[g(\mathbf{w}^{\mathsf{T}}\mathbf{x})\mathbf{x} - g'(\mathbf{w}^{\mathsf{T}}\mathbf{x})\mathbf{w}]}{\|\widehat{\mathbb{E}}_N[g(\mathbf{w}^{\mathsf{T}}\mathbf{x})\mathbf{x} - g'(\mathbf{w}^{\mathsf{T}}\mathbf{x})\mathbf{w}]\|}.$$

Then theoretical and empirical FastICA can respectively be represented by the function iteration $\mathbf{T}^{(n)}(\mathbf{w})$ and $\mathbf{T}_N^{(n)}(\mathbf{w})$.
It can be shown that

$$\partial\mathbf{T}(\mathbf{a}_i) = 0,$$

where ∂ denotes the derivative. That's why theoretical FastICA converges by the theorem of fixed point.

It is easily seen that, by the SLLN

$$\partial \mathbf{T}_N(\mathbf{w}) \xrightarrow[N\to\infty]{a.s.} \partial \mathbf{T}(\mathbf{w}), \quad \forall \mathbf{w} \in \mathcal{S}.$$

Therefore, for large N we have

$$\|\partial \mathbf{T}_N(\mathbf{a}_i)\| \leq \epsilon < 1,$$

and hence the convergence of $\mathbf{T}_N^{(n)}(\mathbf{w})$.

As for the empirical contrast function

$$\widehat{\Phi}_N(\mathbf{w}) = \widehat{\mathbb{E}}_N[G(\mathbf{w}^\mathsf{T}\mathbf{x})],$$

we have the following approximation by Taylor's Theorem

$$\widehat{\Phi}_N(\mathbf{w}) \approx \widehat{\Phi}_N(\mathbf{u}) + (\mathbf{w}-\mathbf{u})^T\gamma_N(\mathbf{u}) + \frac{1}{2}(\mathbf{w}-\mathbf{u})^T[\alpha_N(\mathbf{u})\mathbf{I} + \mathbf{H}_N(\mathbf{u})](\mathbf{w}-\mathbf{u}),$$

where

$$\gamma_N(\mathbf{u}) \in \mathbb{R}^d, \quad \alpha_N(\mathbf{u}) \in \mathbb{R}, \quad H_N(\mathbf{u}) \in \mathbb{R}^{d\times d}.$$

Then it is not difficult to see that $\widehat{\Phi}_N(\mathbf{w})$ has extremum at $\mathbf{u}$ if and only if:

- $\gamma_N(\mathbf{u}) = 0$,
- The matrix $\alpha_N(\mathbf{u})\mathbf{I} + \mathbf{H}_N(\mathbf{u})$ is positive or negative definite.

Suppose that $\mathbf{T}_N^{(n)} \to \hat{\mathbf{v}}_N$. We can show that

- $\gamma_N(\hat{\mathbf{v}}_N) = 0$,
- $\mathbf{H}_N(\hat{\mathbf{v}}_N) \xrightarrow[N\to\infty]{\mathbb{P}} 0$,
- $\alpha_N(\hat{\mathbf{v}}_N) \xrightarrow[N\to\infty]{\mathbb{P}} \alpha(\mathbf{v}) \neq 0$.

It follows that $\hat{\mathbf{v}}_N$ is a local extremum of $\widehat{\Phi}$.

Finally, since

$$\mathbf{v} = \underset{\mathbf{w}\in\mathcal{S}\cap V}{\operatorname{argmin}}\ \mathbb{E}[G(\mathbf{w}^{\mathsf{T}}\mathbf{x})]$$
$$\hat{\mathbf{v}}_N = \underset{\mathbf{w}\in\mathcal{S}\cap V}{\operatorname{argmin}}\ \widehat{\mathbb{E}}_N[G(\mathbf{w}^{\mathsf{T}}\mathbf{x})],$$

we can use the technique of M-estimator to obtain

$$N^{1/2}(\hat{\mathbf{v}}_N - \mathbf{v}) \xrightarrow[N\to\infty]{\mathscr{D}} \mathcal{N}(0, \mathbf{\Sigma}),$$

where

$$\mathbf{\Sigma} = \frac{\mathbb{E}\left[\left(g(\mathbf{v}^{\mathsf{T}}\mathbf{x})\right)^2\right](\mathbf{I} - \mathbf{v}\mathbf{v}^{\mathsf{T}})}{\left(\mathbb{E}[g'(\mathbf{v}^{\mathsf{T}}\mathbf{x}) - g(\mathbf{v}^{\mathsf{T}}\mathbf{x})\mathbf{v}^{\mathsf{T}}\mathbf{x}]\right)^2}.$$