

Neurally Plausible, Non-combinatorial Iterative Independent Process Analysis[★]

András Lőrincz and Zoltán Szabó

*Department of Information Systems, Eötvös Loránd University,
Pázmány P. sétány 1/C, Budapest H-1117, Hungary
E-mail: szzoli@cs.elte.hu, andras.lorincz@elte.hu*

Abstract

It has been shown recently that the identification of mixed hidden independent autoregressive processes (Independent Process Analysis, IPA), under certain conditions, can be free from combinatorial explosion. The key is that IPA can be reduced (i) to Independent Subspace Analysis and then, via a novel decomposition technique called Separation Theorem, (ii) to Independent Component Analysis. Here, we introduce an iterative scheme and its neural network representation that takes advantage of the reduction method and can accomplish the IPA task. Computer simulation illustrates the working of the algorithm.

Key words: independent process analysis, neural network implementation

1 Introduction

Independent Component Analysis (ICA) (see, e.g., [7] and references therein) aims to recover mixed independent hidden sources. There is a broad range of applications for ICA, e.g., blind source separation and deconvolution [1], feature extraction [2], and denoising [6]. ICA algorithms are 1-dimensional: all sources are assumed to be independent real valued random variables. However, applications where not all, but only certain groups of the sources are independent may have high relevance in practice. In this case, independent sources can be multidimensional. For example, consider the generalization of the cocktail-party problem, where independent groups of people are talking

[★] <http://dx.doi.org/10.1016/j.neucom.2006.10.145>

about independent topics, or that more than one group of musicians are playing at the party. The separation task requires an extension of ICA, which can be called Independent Subspace Analysis (ISA), see, e.g., [13].

Most practical problems exhibit considerable temporal correlations. In such cases, one may take advantage of Independent Process Analysis (IPA) [5,3,10,4,11], a generalization of ISA for auto-regressive (AR) processes: 1-dimensional sources are assumed in [5], extensions to AR-mixture of Gaussians are presented in [3,4], and multi-dimensional sources are considered in [10,11]. The IPA task can be rewritten as the identification of Linear Dynamical Systems (LDS) with state space formalization. Further, the identification task can be reduced to the ISA task using linear prediction and the concept of innovations. Typical ISA solutions might need the minimization of mutual information between the estimated processes and the search for (i) the optimal number and (ii) the dimensions of the processes [10]. However, such estimations can be computationally intensive and the dimension search is subject to combinatorial explosion. Recently, a two-phase algorithm that avoided the combinatorial explosion under certain conditions has been introduced for IPA [9]. Here, an iterative scheme that enables neural interpretation is put forth after reviewing the IPA problem. Illustrative numerical simulations are also provided.

2 The IPA Model

Suppose that we have d pieces of m_i -dimensional ($i = 1, \dots, d$) first order auto-regressive (AR) processes

$$\mathbf{s}^i(t+1) = \mathbf{F}^i \mathbf{s}^i(t) + \boldsymbol{\nu}^i(t), \quad (1)$$

where $\mathbf{F}^i \in \mathbb{R}^{m_i \times m_i}$, $\mathbf{s}^i(t) \in \mathbb{R}^{m_i}$, and $\boldsymbol{\nu}^i(t) \in \mathbb{R}^{m_i}$ are non-Gaussian, spatially independent (i.e., $\boldsymbol{\nu}^i$ is independent from $\boldsymbol{\nu}^j$ if $i \neq j$) and temporally independent and identically distributed (i.i.d.) noises. From now on, all the m_i are set equal ($m_i = m \forall i$), but *all results* concern the general case.

Sources \mathbf{s}^i are the *hidden processes* of the external world. Only their mixture,

$$\mathbf{x}(t) = \mathbf{A} \mathbf{s}(t), \quad (2)$$

where $\mathbf{s}(t) = [\mathbf{s}^1(t); \dots; \mathbf{s}^d(t)] \in \mathbb{R}^{dm}$ is the concatenated form of the sources and $\mathbf{A} \in \mathbb{R}^{md \times md}$ is the *mixing matrix*, is available for observation. Equations (1)-(2), together, form an LDS. Estimations of LDS parameters exist for Gaussian noise $\boldsymbol{\nu}(t) = [\boldsymbol{\nu}^1(t); \dots; \boldsymbol{\nu}^d(t)] \in \mathbb{R}^{dm}$ [8], but here noise $\boldsymbol{\nu}$ is non-Gaussian. We assume that the matrix \mathbf{A} is invertible. Then it suffices to solve

the Independent Subspace Analysis (ISA) problem [10], where $\mathbf{F}^i = \mathbf{0}, \forall i$ (see later). Note that ICA is recovered if $\mathbf{F}^i = \mathbf{0}$ for all i and if $m = 1$.

Let $\mathbf{F} \in \mathbb{R}^{md \times md}$ denote the block-diagonal matrix constructed from matrices $\mathbf{F}^1, \dots, \mathbf{F}^d$, i.e., $\mathbf{F} = \text{blockdiag}(\mathbf{F}^1, \dots, \mathbf{F}^d)$. Then the goal of IPA is to identify the system, that is, to estimate \mathbf{F}, \mathbf{A} and $\mathbf{s}(t)$ related as

$$\mathbf{s}(t+1) = \mathbf{F}\mathbf{s}(t) + \boldsymbol{\nu}(t), \quad (3)$$

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t). \quad (4)$$

Like ICA, IPA has *undercomplete*, *complete*, and *overcomplete* cases that correspond to $\dim(\mathbf{x}) > \dim(\mathbf{s})$, $\dim(\mathbf{x}) = \dim(\mathbf{s})$ and $\dim(\mathbf{x}) < \dim(\mathbf{s})$, respectively. We treat the complete IPA task. The extension to undercomplete IPA may use principal component analysis. The overcomplete IPA task is challenging and it is not yet solved.

3 Estimation of the IPA Model

Here, we introduce a novel method that (i) can solve the IPA identification problem without combinatorial efforts, and (ii) has a neural implementation. Note that stochastic process $\{\mathbf{x}(t)\}$ is also an AR process (see Eqs. (3)-(4)):

$$\mathbf{x}(t+1) = \mathbf{A}\mathbf{s}(t+1) = \mathbf{A}\mathbf{F}\mathbf{s}(t) + \mathbf{A}\boldsymbol{\nu}(t) = \mathbf{A}\mathbf{F}\mathbf{A}^{-1}\mathbf{x}(t) + \mathbf{A}\boldsymbol{\nu}(t). \quad (5)$$

Let E denote the expectation operator. The innovation $\mathbf{e}(t)$ of the $\{\mathbf{x}(t+1)\}$ process is $\mathbf{e}(t) = \mathbf{x}(t+1) - E[\mathbf{x}(t+1)|\mathbf{x}(t), \mathbf{x}(t-1), \dots]$. It holds that process $\mathbf{e}(t)$ is equal to the noise of the AR process $\mathbf{x}(t)$, i.e., to $\mathbf{A}\boldsymbol{\nu}(t)$ and that $E[\mathbf{x}(t+1)|\mathbf{x}(t), \mathbf{x}(t-1), \dots] = \mathbf{M}\mathbf{x}(t)$ with $\mathbf{M} = \mathbf{A}\mathbf{F}\mathbf{A}^{-1}$, where matrix \mathbf{M} is the ‘predictive matrix’ of the AR process. However, $\mathbf{e}(t) = \mathbf{A}\boldsymbol{\nu}(t)$ is i.i.d, and thus the IPA problem is reduced to an ISA problem, which works on $\mathbf{e}(t)$.

It has been observed that ICA can estimate components of the subspaces of many ISA tasks [10]. The *ISA separation theorem* provides sufficient conditions for this peculiar property. The proof can be found in [12].

When the separation theorem holds, we can take advantage of the hidden AR processes to uncover the unknown permutations of the coordinates of source \mathbf{s} . The dimensions of the subspaces can be revealed by the estimation of matrix \mathbf{F} [9]. In [9], (a) matrix \mathbf{M} was estimated, (b) the innovations were computed, (c) ICA was used to estimate matrix $\mathbf{W} = \mathbf{A}^{-1}$, and (d) matrix \mathbf{F} was approximated by observing that $\mathbf{s}(t+1) = \mathbf{A}^{-1}\mathbf{M}\mathbf{A}\mathbf{s}(t) + \mathbf{A}^{-1}(\mathbf{A}\boldsymbol{\nu}(t))$.

According to the ISA separation theorem, matrix $\mathbf{W}\mathbf{M}\mathbf{W}^{-1}$ equals to the block-diagonal matrix \mathbf{F} up to permutation and the sign of the components. Thus, \mathbf{F} couples coordinates of source \mathbf{s} and couplings form blocks in matrix \mathbf{F} . In practice, matrix $\hat{\mathbf{F}}$, the estimation of \mathbf{F} , is only nearly block-diagonal (apart from permutation). The coupled components can be found by the following procedure. We say that two coordinates i and j are $\hat{\mathbf{F}}$ -‘connected’ if $\max(|\hat{F}_{ij}|, |\hat{F}_{ji}|) > \epsilon$, where $\epsilon \geq 0$ and in the ideal case $\epsilon = 0$. Then we group the $\hat{\mathbf{F}}$ -‘connected’ coordinates into separate subspaces as follows: (1) Choose an arbitrary coordinate i and group all $j \neq i$ coordinates to it which are $\hat{\mathbf{F}}$ -‘connected’ with it. (2) Choose an arbitrary and not yet grouped coordinate. Find its connected coordinates. Group them together. (3) Continue until all components are grouped. This is the *gathering procedure* and it is fast. In the worst case, it is quadratic in the number of the coordinates.

Our gathering procedure may fail if one of the matrices $(\mathbf{W}_{\text{ICA}}^d) \mathbf{F}^d (\mathbf{W}_{\text{ICA}}^d)^T$ is itself a block-diagonal matrix made of more than 1 block, or if it becomes block-diagonal during the estimation process (degenerate case). Here, $\mathbf{W}_{\text{ICA}}^d$ denotes the component of separation matrix \mathbf{W}_{ICA} that corresponds to the d^{th} sub-process. In our simulations, degenerate cases did appear, but we found that they became non-degenerate if the sample number was increased.

In the IPA identification problem, the estimation of $\hat{\mathbf{F}}$ and $\hat{\mathbf{W}}$ can be accomplished step-by-step. Now, we provide our novel *two-phase* iterative IPA identification method. Neural implementation is detailed afterwards.

Initialization: random $\hat{\mathbf{F}}$, random orthogonal $\hat{\mathbf{W}}$, $\hat{\mathbf{A}} := \hat{\mathbf{W}}^{-1} (= \hat{\mathbf{W}}^T)$

Repeat

Phase I:

Compute new estimations of $\hat{\mathbf{W}}$ and $\hat{\mathbf{A}}$ by using ICA on estimated innovations of $\mathbf{x}(t)$, that is, on $\hat{\mathbf{e}}(t) = \mathbf{x}(t+1) - \hat{\mathbf{A}}\hat{\mathbf{F}}\hat{\mathbf{W}}\mathbf{x}(t)$.

Phase II:

Improve AR fit on the estimated sources $\hat{\mathbf{s}}(t) := \hat{\mathbf{W}}\hat{\mathbf{x}}(t) \Rightarrow \hat{\mathbf{F}}$

Source estimation ($\hat{\nu}(t)$) at any iteration step:

$\hat{\nu}(t) = \text{permutation of } \hat{\mathbf{W}}(\mathbf{x}(t+1) - \hat{\mathbf{A}}\hat{\mathbf{F}}\hat{\mathbf{W}}\mathbf{x}(t))$

according to $\hat{\mathbf{F}}$ -‘connectedness’,

The individual steps of the two-phase process, i.e., ICA, the estimation of the predictive matrix of the AR process, and matrix inversion become neural in the autoassociative reconstruction network depicted in Fig. 1(a). The neural network has two layers. The lower layer receives input $\mathbf{x}(t)$, the so called *reconstructed input* $\hat{\mathbf{x}}(t) = \hat{\mathbf{A}}\hat{\mathbf{s}}(t)$ and computes the difference, the *reconstruction error* $\hat{\mathbf{e}}(t) = \mathbf{x}(t) - \hat{\mathbf{x}}(t) = \mathbf{x}(t) - \hat{\mathbf{A}}\hat{\mathbf{s}}(t)$. Thus, the reconstruction error estimates the innovation of the $\mathbf{x}(t)$ process. The upper layer holds the hidden representation $\hat{\mathbf{s}}(t)$. There are three matrices: the bottom-up estimation of the

separation matrix $\hat{\mathbf{W}}$, the estimated predictive matrix $\hat{\mathbf{F}}$ of the upper layer, and the top-down estimation of the mixing matrix $\hat{\mathbf{A}}$.

The two-phase learning process can be implemented in neural networks:

Phase I: Matrix $\hat{\mathbf{W}}$ receives the reconstruction error and is tuned by ICA to separate. Rule $\Delta\hat{\mathbf{W}}(t) = \alpha(t) [\mathbf{I} - 2f(\hat{\mathbf{W}}(t)\hat{\mathbf{e}}(t))(\hat{\mathbf{W}}(t)\hat{\mathbf{e}}(t))^T] \hat{\mathbf{W}}(t)$, where $\alpha(t)$ denotes the time dependent learning rate, is neural [7] and could be used.

Phase II: Matrix $\hat{\mathbf{F}}$ in the hidden layer learns to predict by minimizing objective $J_1(\hat{\mathbf{F}}) = \frac{1}{2}\|\hat{\mathbf{s}}(t+1) - \hat{\mathbf{F}}\hat{\mathbf{s}}(t)\|^2$. Top-down matrix $\hat{\mathbf{A}}$ learns to invert matrix $\hat{\mathbf{W}}$ by minimizing objective $J_2(\hat{\mathbf{A}}) = \frac{1}{2}\|\mathbf{x}(t+1) - \hat{\mathbf{A}}\hat{\mathbf{s}}(t+1)\|^2$. Taking the gradients of J_1 and J_2 , local Hebbian learning rules emerge:

$$\Delta\hat{\mathbf{A}}(t) = \beta(t)\nabla_{\hat{\mathbf{A}}}J_1[\hat{\mathbf{A}}(t)] = \beta(t)\hat{\mathbf{e}}(t)\hat{\mathbf{s}}^T(t+1), \quad (6)$$

$$\Delta\hat{\mathbf{F}}(t) = \gamma(t)\nabla_{\hat{\mathbf{F}}}J_2[\hat{\mathbf{F}}(t)] = \gamma(t)\hat{\boldsymbol{\nu}}(t)\hat{\mathbf{s}}^T(t). \quad (7)$$

However, we note that the *gathering procedure* is not in a neural form, yet.

4 Results

We shall demonstrate the working of the algorithm. The task was to identify the hidden processes from the mixed AR signals without using our knowledge about the number and the dimensions of the hidden processes. The hidden AR components were defined by their driving noises $\boldsymbol{\nu}^i$ (Eq. (1)), and two test were carried out. In the first case, the conditions of the separation theorem were satisfied: 3 pieces of 4-dimensional noise sources with spherical distributions were chosen; this is the *4-spherical* database, total dimension is 12. In the other case, noise sources were constructed from images of tale figures. These images were considered as density functions and $\boldsymbol{\nu}^i$ s were generated by sampling 2-dimensional coordinates proportional to the corresponding pixel intensities. This is the *tale* database. The \mathbf{F}^i predictive matrices of the \mathbf{s}^i AR processes were chosen randomly under the constraint that the AR processes were stable. Mixing matrix \mathbf{A} was chosen randomly from the orthogonal group. The two-phase algorithm was applied for the mixed data in batch learning that avoids parameter dependencies for $\alpha(t)$, $\beta(t)$, and $\gamma(t)$. For speed, FastICA was applied [7]. Coordinates of matrix $\hat{\mathbf{F}}$ were initialized with independent, uniformly distributed random variables over $[0, 1]$. The normalized Amari-error was used to measure the precision of the estimation. This error takes values in interval $[0, 1]$; 0 (1) means perfect (worst possible) estimation. 50 independent runs were averaged, sample number T changed between 1,000 and 20,000.

The quality of the gathering procedure depends on parameter ε . We have studied (i) the Amari-error for a suitable ε as a function of the sample and iteration numbers, and (ii) the dynamic range, i.e., the ratio of the highest and lowest ‘good ε values’: We divided interval $[0, F_m]$ ($F_m = \max_{i,j} |\hat{\mathbf{F}}_{ij}|$) into 200 equal parts. For different sample and iteration numbers in both databases and at each division point we used the gathering procedure to group the ICA elements. For each of the 50 random trials we have computed the Amari-errors separately. For the smallest Amari-error, we collected the corresponding set of ε ’s, these are the ‘good ε values’. Then we took the ratio of the largest and smallest ε values in this set and averaged the ratios over the 50 runs. The average is called the dynamic range.

Figure 1 summarizes our findings for the *tale* and the *4-spherical* tasks. According to the figure, there are good ε parameters for the $\hat{\mathbf{F}}$ -‘connectedness’ already for 1,000-2,000 samples: our neural method can find the hidden components with high precision. Increasing the sample number the Amari-error decreases, it is 1.54% for the *tale* and 0.85% for the 4-spherical databases, at 20,000 samples on the average. Two iterations were satisfactory for the optimization, further iterations did not improve the results. The dynamic range is better by a factor about 2.5 for the 4-spherical database that satisfies the conditions of the separation theorem. Using 20,000 samples, the dynamic range was about 10 for the *tale* database. These results demonstrate the robustness of the neural approach.

5 Summary

Recently, it has been shown that for hidden and mixed AR processes, which are driven by non-Gaussian noises, ICA is satisfactory for the estimation of the independent directions, provided that innovations satisfy certain technical conditions. It then followed that combinatorial searches for the hidden processes can be lessened or fully avoided except for degenerate cases. Here, we have introduced an iterative approach and its neural network architecture that iterates in two-phases and enables the finding and the grouping of the hidden components. In one phase, the network makes use of non-linear Hebbian learning rule that estimates the ICA components. In the other phase, Delta-rules tune the predictive matrix of the hidden process and the inverse of the ICA matrix that improves the ICA estimation of the iterative procedure. Numerical studies indicate the robustness of the neural algorithm.

References

- [1] A. J. Bell and T. J. Sejnowski, *An information maximisation approach to blind separation and blind deconvolution*, *Neural Computation* **7** (1995), 1129–1159.
- [2] ———, *The ‘independent components’ of natural scenes are edge filters*, *Vision Research* **37** (1997), 3327–3338.
- [3] Y. Cheung and L. Xu, *Dual multivariate auto-regressive modeling in state space for temporal signal separation*, *IEEE Transactions on Systems, Man, Cybernetics—Part B* **33** (2003), 386–398.
- [4] K. E. Hild, H. T. Attias, and S. S. Nagarajan, *An EM method for spatio-temporal blind source separation using an AR-MOG source model*, *ICA 2006*, LNCS, vol. 3889, Springer, 2006, pp. 98–105.
- [5] A. Hyvärinen, *Independent component analysis for time-dependent stochastic processes*, *ICANN 1998*, 1998, pp. 541–546.
- [6] A. Hyvärinen, *Sparse code shrinkage: Denoising of nongaussian data by maximum likelihood estimation*, *Neural Computation* **11** (1999), 1739–1768.
- [7] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent component analysis*, John Wiley, New York, NY, 2001.
- [8] L. Ljung and T. Söderström, *Theory and practice of recursive identification*, MIT Press, Cambridge, MA, 1983.
- [9] B. Póczos and A. Lőrincz, *Non-combinatorial estimation of independent autoregressive sources*, *Neurocomputing Letters* (2006).
- [10] B. Póczos, B. Takács, and A. Lőrincz, *Independent subspace analysis on innovations*, *ECML 2005*, LNAI, vol. 3720, Springer Verlag, 2005, pp. 698–706.
- [11] Z. Szabó, B. Póczos, and A. Lőrincz, *Cross-entropy optimization for independent process analysis*, *ICA 2006*, LNCS, vol. 3889, Springer, 2006, pp. 909–916.
- [12] Z. Szabó, B. Póczos, and A. Lőrincz, *Separation theorem for independent subspace analysis with sufficient conditions*, Tech. report, Eötvös Loránd University, Budapest, Hungary, 2006, <http://arxiv.org/abs/math.ST/0603535>.
- [13] F. J. Theis, *Blind signal separation into groups of dependent signals using joint block diagonalization*, *ISCAS 2005*, 2005, pp. 5878–5881.

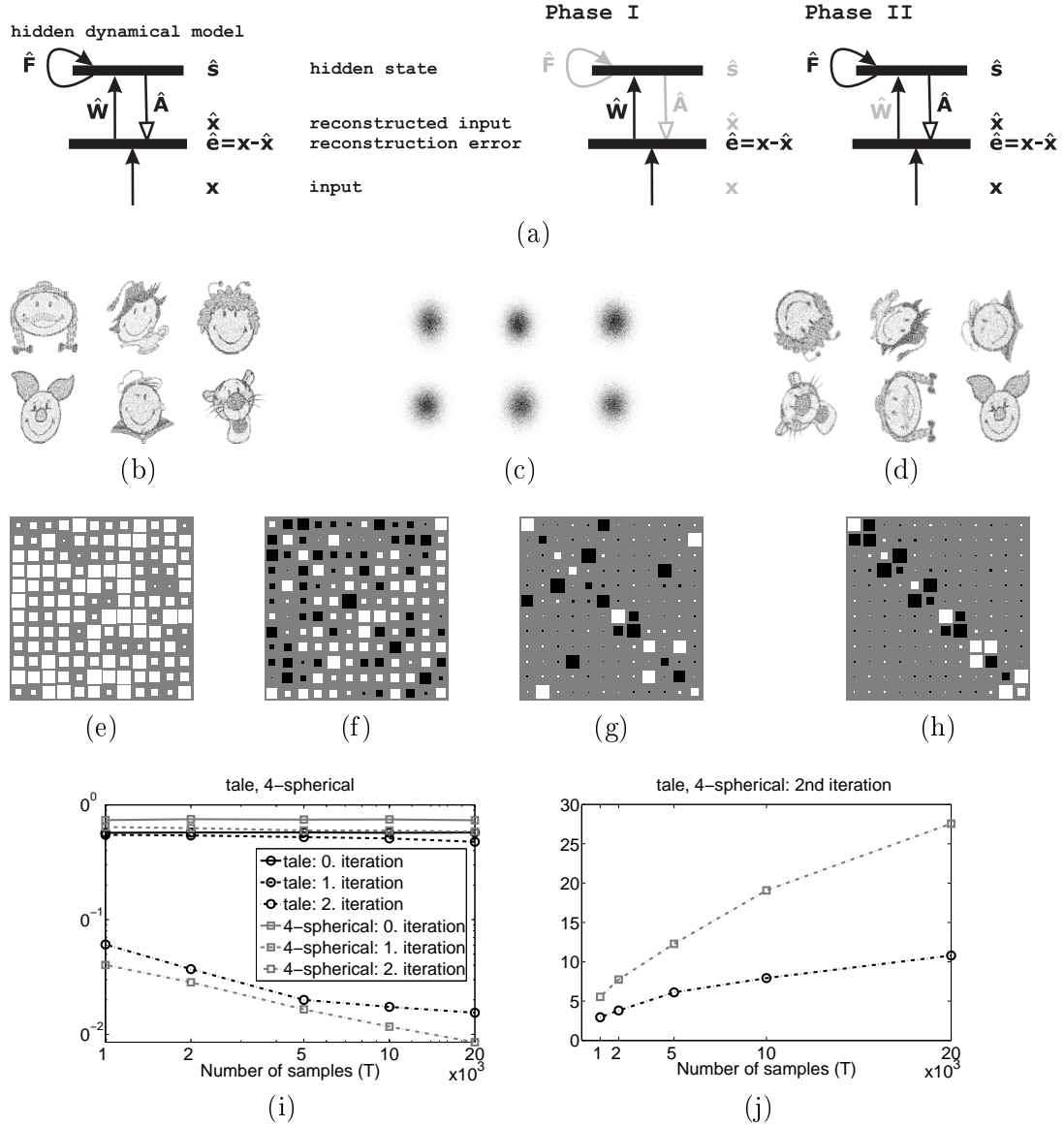


Fig. 1. Neural network implementation and computer study. (a): neural network and its two-phase operation. Gray matrix: not trained, gray vector: not used in the respective phases. (b): noise sources, sample number 20000. (c): 2-dimensional projections of the observed mixed signal, (d): estimated noise sources ($\hat{\mathbf{e}}^i$). (e)-(f)-(g): $\hat{\mathbf{F}}$ in 0^{th} , 1^{st} , and 2^{nd} iterations, (h): like (g), but after the gathering procedure. (i): Amari-error as a function of iterations and sample number for the *tale* (black lines with circles) and the *4-spherical* (gray lines with squares) databases, (j): dynamic range vs. number of samples for the two databases after the second iteration.