Fast Parallel Estimation of High Dimensional Information Theoretical Quantities with Low Dimensional Random Projection Ensembles

Zoltán Szabó and András Lőrincz

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

Abstract. The estimation of relevant information theoretical quantities, such as entropy, mutual information, and various divergences is computationally expensive in high dimensions. However, for this task, one may apply pairwise Euclidean distances of sample points, which suits random projection (RP) based low dimensional embeddings. The Johnson-Lindenstrauss (JL) lemma gives theoretical bound on the dimension of the low dimensional embedding. We adapt the RP technique for the estimation of information theoretical quantities. Intriguingly, we find that embeddings into extremely small dimensions, far below the bounds of the JL lemma, provide satisfactory estimates for the original task. We illustrate this in the Independent Subspace Analysis (ISA) task; we combine RP dimension reduction with a simple ensemble method. We gain considerable speed-up with the potential of real-time parallel estimation of high dimensional information theoretical quantities.

Keywords: Independent subspace analysis, random projection, pairwise distances, information theoretical estimations.

1 Introduction

The take-off of information theory goes back to the forties [1]. Tremendous applications have been developed ever since. The computation/estimation of information theoretical quantities (entropy, mutual information, divergence) is still slow. However, consistent estimation of these quantities is possible by nearest neighbor (NN) methods (see, e.g., [2]) that use the pairwise distances of sample points. Although search for nearest neighbors can also be expensive in high dimensions [3], low dimensional approximate isometric embedding of points of high dimensional Euclidean space can be addressed by the Johnson-Lindenstrauss Lemma [4] and the related random projection (RP) methods [5,6]. The RP approach proved to be successful, e.g., in classification, clustering, search for *approximate NN* (ANN), dimension estimation of manifolds, estimation of mixture of Gaussian models, compressions, data stream computation (see, e.g., [7]). We note that the RP approach is also related to compressed sensing [8].

T. Adali et al. (Eds.): ICA 2009, LNCS 5441, pp. 146–153, 2009.

In this paper we show a novel application of the RP technique: we estimate information theoretical quantities using the ANN-preserving properties of the RP technique. We illustrate the method on the estimation of Shannon's multidimensional differential entropy for the Independent Subspace Analysis (ISA) task [9]. The ISA problem extends Independent Component Analysis (ICA) [10] by allowing multidimensional independent components: at a cocktail party, ICA (ISA) is searching for (groups of) people talking independently. Another application area is image registration, where (i) information-theoretical similarity criterion can be advantageous, and (ii) high dimensional features should be handled [11,2] (work in progress). We note that RPs have been applied for ICA, but the underlying considerations differ from ours: [12] picks out random samples using Bernoulli variables and decreases the computational load on ICA, [13] uses RPs for preprocessing before principal component analysis.

The paper is structured as follows: Section 2 formulates the problem domain. In Section 3 the random projection technique is adapted to the estimation of information theoretical quantities and we use it for the estimation of multidimensional differential entropy. Section 4 contains the numerical illustrations. Conclusions are drawn in Section 5.

2 The ISA Model

Let us define the ISA task. Let us assume the observations $\mathbf{x}(t) \in \mathbb{R}^D$, $t = 1, 2, \ldots$ are linear mixtures of multidimensional independent sources, *components* $\mathbf{s}^m(t)$:

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t), \tag{1}$$

where $\mathbf{s}(t)$ concatenates components $\mathbf{s}^m(t) \in \mathbb{R}^{d_m}$; $\mathbf{s}(t) = [\mathbf{s}^1(t); \ldots; \mathbf{s}^M(t)] \in \mathbb{R}^D$ $(D = \sum_{m=1}^M d_m)$. Our assumptions are the following:

1. components are (i) independent: $I(\mathbf{s}^1, \ldots, \mathbf{s}^M) = 0$, where I denotes the mutual information, (ii) i.i.d. (independent identically distributed) in t, and (iii) there is at most one Gaussian component among $\mathbf{s}^m$s.
2. The unknown $\mathbf{A} \in \mathbb{R}^{D \times D}$ *mixing matrix* is invertible.

In the ISA problem one estimates hidden source components ($\mathbf{s}^m$) from observations $\mathbf{x}(t)$ alone. (ICA problem: $\forall d_m = 1$). The ISA problem has ambiguities [14,15]: components of equal dimension can be determined up to permutation and up to invertible transformation within the subspaces. Thus, for ISA demixing matrix $\mathbf{W}_{\text{ISA}}$ we have that $\mathbf{W}_{\text{ISA}}\mathbf{A} \in \mathbb{R}^{D \times D}$ is a block-permutation (or block-scaling [16]) matrix. The block-permutation property and the quality of the ISA estimation can be measured by the ISA adapted and normalized Amari-error [17], the *Amari-index* (r) [18], which is 0 for perfect estimation and can not exceed 1.

3 Method

We present our RP based approach through the ISA problem. The ISA task can be viewed as the minimization of the mutual information between the estimated components, or equivalently as the minimization of the sum of Shannon's multidimensional differential entropies of the estimated components on the orthogonal group [19]:

$$J(\mathbf{W}) := \sum_{m=1}^{M} H\left(\mathbf{y}^m\right) \to \min_{\mathbf{W}\in\mathcal{O}^D}, \tag{2}$$

where $\mathbf{y} = \mathbf{W}\mathbf{x}$, $\mathbf{y} = \left[\mathbf{y}^1; \ldots ; \mathbf{y}^M\right]$, $\mathbf{y}^m \in \mathbb{R}^{d_m}$ and d_ms are given. It has been conjectured that the solution of the ISA task can be reduced to ICA followed by grouping of the non-independent ICA elements into ISA subspaces [9]. The conjecture has been rigorously proven by the ISA Separation Theorem for some distribution types [20]. It means that the demixing matrix assumes the form $\mathbf{W}_{\text{ISA}} = \mathbf{P}\mathbf{W}_{\text{ICA}}$ ($\hat{\mathbf{y}}_{\text{ISA}} = [\hat{\mathbf{y}}^1_{\text{ISA}}; \ldots ; \hat{\mathbf{y}}^M_{\text{ISA}}] = \mathbf{P}\hat{\mathbf{y}}_{\text{ICA}}$, $\hat{\mathbf{y}}^m_{\text{ISA}} \in \mathbb{R}^{d_m}$), where the permutation matrix $\mathbf{P} \in \mathbb{R}^{D\times D}$ is to be determined. Estimation of cost function J involves multidimensional entropy estimation, which is computationally expensive in high dimensions, but can be executed by NN methods consistently [21,22]. It has been shown in [11] (in the field of image registration with high dimensional features) that the computational load can be decreased somewhat by (i) dividing the samples into groups and then (ii) computing the averages of the group estimates. We will combine this *parallelizable ensemble approach* with the ANN-preserving properties of RPs and get drastic savings. We suggest the following entropy estimation method[1], for each estimated ISA component $\mathbf{v} := \hat{\mathbf{y}}^m_{\text{ISA}}$: (i) divide the T samples $\{\mathbf{v}(1), \ldots , \mathbf{v}(T)\}$ into N groups indexed by sets $I_1, \ldots , I_N$ so that each group contains K samples, (ii) for all fixed groups take the random projection of $\mathbf{v}$ as $\mathbf{v}_{n,\text{RP}}(t) := \mathbf{R}_n\mathbf{v}(t)$ ($t \in I_n; n = 1, \ldots , N; \mathbf{R}_n \in \mathbb{R}^{d'_m\times d_m}$), (iii) average the estimated entropies of the RP-ed groups to get the estimation $\hat{H}(\mathbf{v}) = \frac{1}{N}\sum_{n=1}^{N} \hat{H}(\mathbf{v}_{n,\text{RP}})$. Our particular choice for $\mathbf{R}_n$ is given in Section 4.2. For the optimization of the estimated cost function $\hat{J}(\mathbf{P})$ one can apply (i) greedy optimization (exchange of 2 coordinates if it decreases $\hat{J}$), or (ii) global methods of higher computational burden, e.g., the cross-entropy (CE) method [23] adapted to permutation searches, because the estimation of $\hat{J}$ is quick.

4 Illustrations

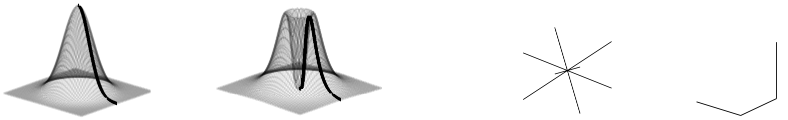
Here, we illustrate the efficiency of the proposed RP based entropy estimation. Section 4.1 is about test cases. Numerical results are presented in Section 4.2.

4.1 Databases

We define three databases [20] to study our RP based ISA identification algorithm. In the *d-spherical* test hidden sources $\mathbf{s}^m$ were spherical random variables. Since spherical variables assume the form $\mathbf{v} = \rho\mathbf{u}$, where $\mathbf{u}$ is uniformly distributed on the d-dimensional unit sphere, and ρ is a non-negative scalar random variable independent of $\mathbf{u}$, they can be given by means of ρ (see Fig. 1(a)).

[1] The idea can be used for a number of information theoretical quantities, provided that they can be estimated by means of pairwise Euclidean distances of the samples.

In the *d-geom* dataset $\mathbf{s}^m$s were random variables uniformly distributed on d-dimensional geometric forms (see Fig. 1(b)). In the *all-k-independent* database, the d-dimensional hidden components $\mathbf{v} := \mathbf{s}^m$ were created as follows: coordinates v_i $(i = 1, \ldots, k)$ were independent uniform random variables on the set {0,...,k-1}, whereas v_{k+1} was set to $mod(v_1 + \ldots + v_k, k)$. In this construction, every k-element subset of $\{v_1, \ldots, v_{k+1}\}$ is made of independent variables.



(a) (b)

Fig. 1. Illustration of the (a): *d-spherical* ($d = 2$), and (b): *d-geom* ($d = 3$) databases. ρ of the stochastic representation on the left (right): exponential with parameter $\mu = 1$ (lognormal with parameters $\mu = 0$, $\sigma = 1$).

4.2 Simulations

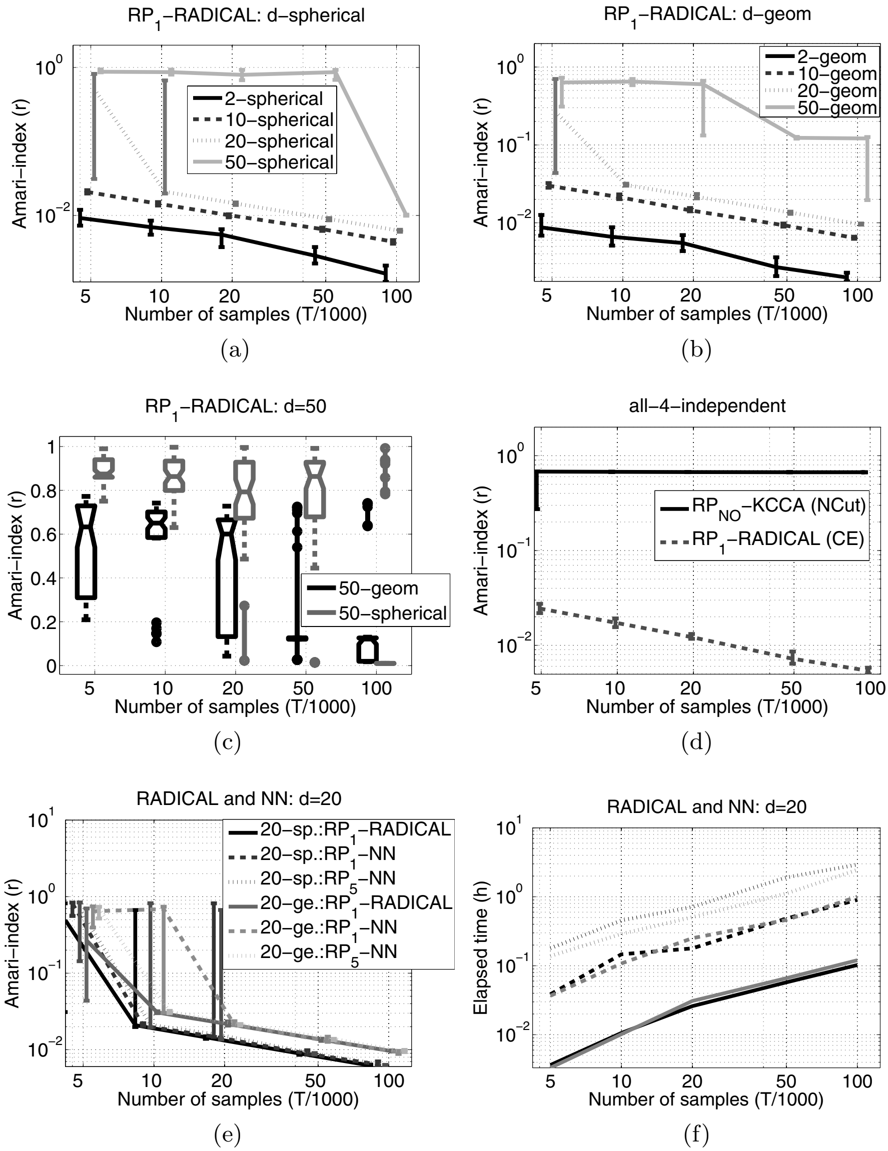
Results on databases *d-spherical*, *d-geom*, and *all-k-independent* are provided here. These experimental studies focused on the following issues:

1. What dimensional reduction can be achieved in the entropy estimation of the ISA problem by means of random projections?
2. What speed-up can be gained with the RP dimension reduction?
3. What are the advantages of our RP based approach in global optimization?

In our experiments the number of components was minimal ($M = 2$). We used the Amari-index to measure and compare the performance of the different methods. For each individual parameter, 50 random runs were averaged. Our parameters included T, the sample number of observations $\mathbf{x}(t)$ and d, the dimension of the components ($d = d_1 = d_2$[2]). We also studied different estimations of the ISA cost function: we used the RADICAL procedure[3] [24] and the NN method [19] for entropy estimation and the Kernel Canonical Correlation Analysis (KCCA) [25] for mutual information estimation. The reduced dimension d' in RP and the optimization method (greedy, global (CE), NCut [26]) of the ISA cost were also varied in different tests. Random run means random choice of quantities $\mathbf{A}$ and $\mathbf{s}$. The ICA step was performed by the well-known fastICA method. The size of the randomly projected groups was set to $|I_n| = 2,000$, except for the case $d = 50$, when it was $5,000$. RP was realized by the *database-friendly projection* technique, i.e., the $r_{n,ij}$ coordinates of $\mathbf{R}_n$ were drawn independently from distribution $P(r_{n,ij} = \pm 1) = 1/2$, but more general constructions could also be used [5,6].

[2] This constraint was used only for the evaluation of the performance (Amari-index) of the algorithm.

[3] We chose RADICAL, because it is consistent, asymptotically efficient, converges rapidly, and it is computationally efficient.



(a) (b) (c) (d) (e) (f)

Fig. 2. Performance of the RP method. Notations: 'RP$_{d'}$ - method of cost estimation (method of optimization if not greedy)'. (a), (b): accuracy of the estimation versus the number of samples for the *d-spherical* and the *d-geom* databases on log-log scale. (c): notched boxed plots for $d = 50$, (d): Performance comparison on the *all-4-independent* database between the RP method using global optimization and the NCut based grouping of coordinates using the pairwise mutual information graph (on log-log scale). (e)-(f): Accuracy and computation time comparisons with the NN based method for the *20-spherical* and the *20-geom* databases (on log-log scale).

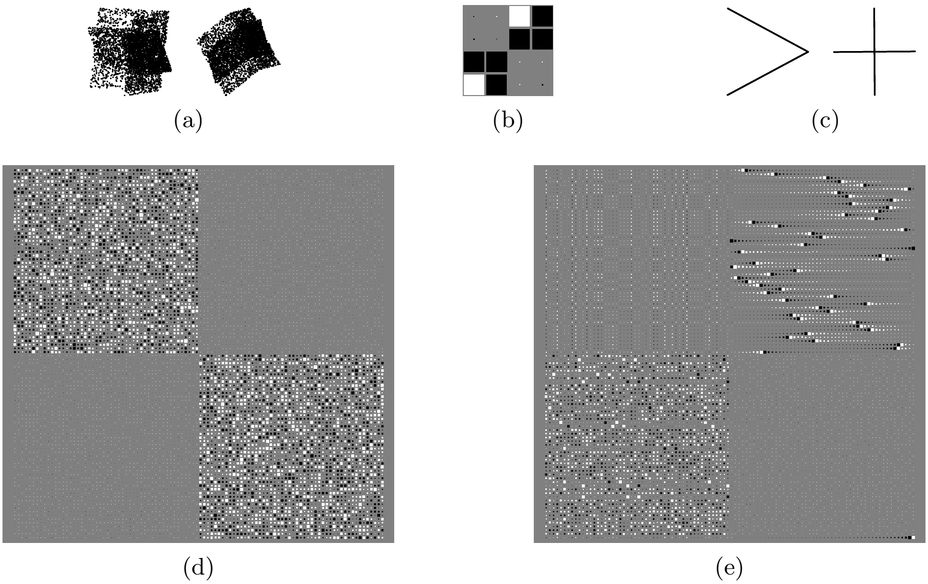


Fig. 3. Estimated components and Hinton-diagrams. Number of samples: $T = 100,000$. Databases *2-geom*: (a)-(c), *50-spherical*: (d), *50-geom*: (e). (a): observed signals $\mathbf{x}(t)$, (b): Hinton-diagram: the product of the mixing matrix of the ISA task and the estimated demixing matrix is approximately a block-permutation matrix with 2×2 blocks, (c): estimated components $\hat{\mathbf{s}}^m$, recovered up to the ISA ambiguities, (d)-(e): Hinton-diagrams of the *50-spherical* and the *50-geom* tests, respectively. Hinton-diagrams have average Amari-indices: for (b) 0.2%, for (d) 1%, for (e) 12%.

In the first study we were interested in the limits of the RP dimension reduction. We increased dimension d of the subspaces for the *d-spherical* and the *d-geom* databases ($d = 2, 10, 20, 50$) and studied the extreme case, the RP dimension d' was set to 1. Results are summarized in Fig. 2(a)-(b) with quartiles (Q_1, Q_2, Q_3). We found that the estimation error decreases with sample number according to a power law $[r(T) \propto T^{-c}\ (c > 0)]$ and the estimation works up to about $d = 50$. For the $d = 50$ case we present notched boxed plots (Fig. 2(c)). We show the quartiles and depict the outliers, i.e., those that fall outside of interval $[Q_1 - 1.5(Q_3 - Q_1), Q_3 + 1.5(Q_3 - Q_1)]$ by circles. According to the figure, the error of estimation drops for sample number $T = 100,000$ for both types of datasets: for databases *50-geom* and *50-spherical*, respectively, we have 5 and 9 outliers from 50 random runs and thus with probability 90% and 82%, the estimation is accurate. As for question two, we compared the efficiency (Q_1, Q_2, Q_3) of our method for $d = 20$ with the NN methods by RP-ing into $d' = 1$ and $d' = 5$ dimensions. Results are shown in Fig. 2(e)-(f).[4] The figure demonstrates that for database *20-geom* performances are similar, but for database *20-spherical*

[4] We note that for $d = 20$ and without dimension reduction the NN methods are very slow for the ISA tasks.

our method has smaller standard deviation for $T = 20,000$. At the same time our method offers 8 to 30 times speed-up at $T = 100,000$ *for serial implementations*. Figure 3 presents the components estimated by our method for dimensions $d = 2$ and $d = 50$, respectively. With regard to our third question, the ISA problem can often be solved by grouping the estimated ICA coordinates based on their mutual information. However, this method, as illustrated by (Q_1, Q_2, Q_3) in Fig. 2(d), does not work for our *all-4-independent* database. Inserting the RP based technique into global optimization procedure, we get accurate estimation for this case, too. CE optimization was used here. Results are presented in Fig. 2(d).

5 Conclusions

In this paper we have shown that random projections (RP) can be used for the estimation of information theoretical quantities. The underlying thought of our approach is that RP approximately preserves the Euclidean distances between sample points and that a number of information theoretical quantities can be estimated from the pairwise distances of sample points. The proposed technique has been demonstrated on the estimation of Shannon's multidimensional differential entropy for the solution of the Independent Subspace Analysis task. The promise of this work is a considerable speed-up that results from the *RP technique* and the *parallel nature of the ensemble method* that we applied. Promising applications emerge – among many others – in the field of image registration.

Acknowledgments. This work has been supported by the National Office for Research and Technology and by the EC NEST 'Perceptual Consciousness: Explication and Testing' grant under contract 043261. Opinions and errors in this manuscript are the author's responsibility, they do not necessarily reflect the opinions of the EC or other project members.

References

1. Shannon, C.: The mathematical theory of communication. Bell System Technical Journal 27, 623–656 (1948)
2. Neemuchwala, H., Hero, A., Zabuawala, S., Carson, P.: Image registration methods in high-dimensional space. Int. J. Imaging Syst. and Technol. 16(5), 130–145 (2007)
3. Arya, S., Mount, D.M., Netanyahu, N.S., Silverman, R., Wu, A.Y.: An optimal algorithm for approximate nearest neighbor searching in fixed dimensions. J. of the ACM 45(6), 891 (1998)
4. Johnson, W., Lindenstrauss, J.: Extensions of Lipschitz maps into a Hilbert space. Contemporary Mathematics 26, 189–206 (1984)
5. Arriga, R.I., Vempala, S.: An algorithmic theory of learning: Robust concepts and random projections. Machine Learning 63, 161–182 (2006)
6. Matoušek, J.: On variants of the Johnson-Lindenstrauss lemma. Random Structures and Algorithms 33(2), 142–156 (2008)

7. Vempala, S.S.: The Random Projection Method. DIMACS Series in Discrete Math., vol. 65 (2005), http://dimacs.rutgers.edu/Volumes/Vol65.html
8. Baraniuk, R., Davenport, M., DeVore, R., Wakin, M.: A simple proof of the restricted isometry property for random matrices (to appear, 2008)
9. Cardoso, J.: Multidimensional independent component analysis. In: ICASSP 1998, vol. 4, pp. 1941–1944 (1998)
10. Jutten, C., Hérault, J.: Blind separation of sources: An adaptive algorithm based on neuromimetic architecture. Signal Processing 24, 1–10 (1991)
11. Kybic, J.: High-dimensional mutual information estimation for image registration. In: ICIP 2004, pp. 1779–1782. IEEE Computer Society, Los Alamitos (2004)
12. Gaito, S., Greppi, A., Grossi, G.: Random projections for dimensionality reduction in ICA. Int. J. of Appl. Math. and Comp. Sci. 3(4), 154–158 (2006)
13. Bingham, E.: Advances in Independent Component Analysis with Applications to Data Mining. PhD thesis, Helsinki University of Technology (2003), http://www.cis.hut.fi/ella/thesis/
14. Theis, F.J.: Uniqueness of complex and multidimensional independent component analysis. Signal Processing 84(5), 951–956 (2004)
15. Theis, F.J.: Multidimensional independent component analysis using characteristic functions. In: EUSIPCO 2005 (2005)
16. Theis, F.J.: Blind signal separation into groups of dependent signals using joint block diagonalization. In: ISCAS 2005, Kobe, Japan, pp. 5878–5881 (2005)
17. Amari, S., Cichocki, A., Yang, H.H.: A new learning algorithm for blind signal separation. In: NIPS 1996, vol. 8, pp. 757–763 (1996)
18. Szabó, Z., Póczos, B., Lőrincz, A.: Cross-entropy optimization for independent process analysis. In: Rosca, J.P., Erdogmus, D., Príncipe, J.C., Haykin, S. (eds.) ICA 2006. LNCS, vol. 3889, pp. 909–916. Springer, Heidelberg (2006)
19. Póczos, B., Lőrincz, A.: Independent subspace analysis using k-nearest neighborhood distances. In: Duch, W., Kacprzyk, J., Oja, E., Zadrożny, S. (eds.) ICANN 2005. LNCS, vol. 3697, pp. 163–168. Springer, Heidelberg (2005)
20. Szabó, Z., Póczos, B., Lőrincz, A.: Undercomplete blind subspace deconvolution. J. of Machine Learning Res. 8, 1063–1095 (2007)
21. Kozachenko, L.F., Leonenko, N.N.: On statistical estimation of entropy of random vector. Problems Infor. Transmiss. 23(2), 95–101 (1987)
22. Hero, A., Ma, B., Michel, O., Gorman, J.: Applications of entropic spanning graphs. Signal Processing 19(5), 85–95 (2002)
23. Rubinstein, R.Y., Kroese, D.P.: The Cross-Entropy Method. Springer, Heidelberg (2004)
24. Learned-Miller, E.G., Fisher III, J.W.: ICA using spacings estimates of entropy. J. of Machine Learning Res. 4, 1271–1295 (2003)
25. Bach, F.R., Jordan, M.I.: Beyond independent components: Trees and clusters. J. of Machine Learning Res. 4, 1205–1233 (2003)
26. Póczos, B., Szabó, Z., Kiszlinger, M., Lőrincz, A.: Independent process analysis without a priori dimensional information. In: Davies, M.E., James, C.J., Abdallah, S.A., Plumbley, M.D. (eds.) ICA 2007. LNCS, vol. 4666, pp. 252–259. Springer, Heidelberg (2007)