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, Budapest, Hungary

`szzoli@cs.elte.hu, lorincz@inf.elte.hu`

1. Introduction

Goal: estimation of high dimensional information theoretical quantities (entropy, mutual information, divergence).

- Problem: computation/estimation is quite slow.
- Consistent estimation is possible by nearest neighbor (NN) methods [1] $\rightarrow$ pairwise distances of sample points:
  - expensive in high dimensions [2],
  - approximate isometric embedding into low dimension is possible (Johnson-Lindenstrauss (JL) Lemma [3], random projection (RP) [4]),
  - idea: estimation using the embedded low dimensional samples.

Demo: estimation of multidimensional differential entropy $\rightarrow$ Independent Subspace Analysis (ISA) task [5].

2. The ISA Model

Cocktail party with independent groups of people.

ISA Equations:

- Observations $\mathbf{x}(t) \in \mathbb{R}^D$ are linear mixtures of multidimensional independent sources, *components* $\mathbf{s}^m(t)$:

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

  where $\mathbf{s}(t) = [\mathbf{s}^1(t); \ldots; \mathbf{s}^M(t)] \in \mathbb{R}^D$, $\mathbf{s}^m(t) \in \mathbb{R}^{d_m}$.
- Goal of ISA: estimate hidden source components ($\mathbf{s}^m$) from observations $\mathbf{x}(t)$. ICA problem: $\forall d_m = 1$.

ISA Assumptions:

- Components are
  - independent: $I(\mathbf{s}^1, \ldots, \mathbf{s}^M) = 0$,
  - i.i.d. (independent identically distributed) in t,
  - there is at most one Gaussian among $\mathbf{s}^m$s.
- The unknown $\mathbf{A} \in \mathbb{R}^{D \times D}$ *mixing matrix* is invertible.

ISA Ambiguities [6, 7]:

- permutation (components of equal dimension),
- invertible transformation within the subspaces.

ISA Performance Measure:

- ISA ambiguities $\Rightarrow$ $\mathbf{G} = \mathbf{W}_{\text{ISA}}\mathbf{A}$ is ideally a block-permutation matrix.
- Its measure: *Amari-index* ($r = r(\mathbf{G}) \in [0, 1]$)
  - ICA: Amari-error [8] $\xrightarrow{[7]}$ ISA $\xrightarrow{[9]}$ ISA, $\in [0, 1]$,
  - $r = 0 \leftrightarrow$ perfect estimation, $r = 1 \leftrightarrow$ worst possible.

3. Method

- ISA as entropy optimization on the orthogonal group: ISA task $\Leftrightarrow$ minimization of the mutual information between the estimated components $\Leftrightarrow$ [10]:

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

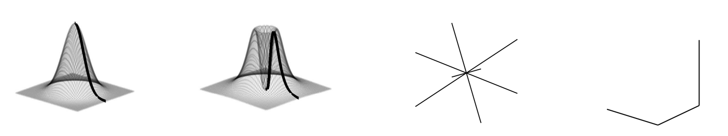
  Here, $\mathbf{y} = [\mathbf{y}^1; \ldots; \mathbf{y}^M] = \mathbf{W}\mathbf{x}$, $\mathbf{y}^m \in \mathbb{R}^{d_m}$; given d_ms.
- ISA Separation Theorem ([5]–conjecture, [11]–proof for certain distribution types):

  ISA = ICA + clustering.
- Cost Estimation [$\hat{H}(\mathbf{v})$, $\mathbf{v} := \hat{\mathbf{y}}^m_{\text{ISA}}$]:
  1. divide the T samples $\{\mathbf{v}(1), \ldots, \mathbf{v}(T)\}$ into N groups (index sets: $I_1, \ldots, I_N$; $|I_n| = K, \forall n$),
  2. for all fixed groups take the random projection of $\mathbf{v}$: $\mathbf{v}_{n,\text{RP}}(t) := \mathbf{R}_n\mathbf{v}(t) \quad (t \in I_n; \forall n; \mathbf{R}_n \in \mathbb{R}^{d'_m \times d_m})$,
  3. average the estimated entropies [12] (*parallelizable ensemble approach*) of the RP-ed groups to get the estimation: $\hat{H}(\mathbf{v}) = \frac{1}{N}\sum_{n=1}^{N} \hat{H}(\mathbf{v}_{n,\text{RP}})$.

4. Illustrations

Databases [11]:

- *d-spherical*: $\mathbf{s}^m \in \mathbb{R}^d$ were spherical random variables (stochastic representation: $\mathbf{v} = \rho\mathbf{u}$, see Fig. 1(a)).
- *d-geom*: $\mathbf{s}^m \in \mathbb{R}^d$s were random variables uniformly distributed on geometric forms (see Fig. 1(b)).
- *all-k-independent*: every k-element subset of $\{s^m_1, \ldots, s^m_{k+1}\}$ is made of independent variables.



(a) (b)

Figure 1: *(a):* d-spherical ($d = 2$)*, (b):* d-geom ($d = 3$)*. (a):* ρ *on the left (right): exp(*$\mu = 1$*), (lognormal(*$\mu = 0, \sigma = 1$*)).*

Questions:

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?

Illustrations: In the test examples

- number of components: minimal ($M = 2$).
- performance measure: Amari-index over 50 random $(\mathbf{A}, \mathbf{s})$ runs.
- dimension of the components: $d = d_1 = d_2$–used only in the Amari-index.
- compared ISA cost functions:
  - H: RADICAL [13], NN method [10],
  - I: Kernel Canonical Correlation Analysis (KCCA) [14].
- optimization method of $\hat{J}(\mathbf{P})$: greedy, global (CE) [15], NCut [16],
- ICA step: fastICA.
- RP group sizes: $|I_n| = 2{,}000$ (and $5{,}000$ for $d = 50$).
- RP $(\mathbf{R}_n)$: *database-friendly projection* $P(r_{n,ij} = \pm 1) = 1/2$; possible more general constructions [4].

Answers: quartiles (Q_1, Q_2, Q_3),

1. *d-spherical*, *d-geom* databases: $d = 2, 10, 20, 50$; extreme RP case ($d' = 1$). Fig. 2(a)-(b):
   - power law estimation error decrease:

$$r(T) \propto T^{-c} \quad (c > 0). \tag{3}$$

   - estimation works up to about $d = 50$, Fig. 2(c): for sample number $T = 100{,}000$ 5 and 9 outliers (outside of interval $[Q_1 - 1.5(Q_3 - Q_1), Q_3 + 1.5(Q_3 - Q_1)]$) from 50 random runs $\leftrightarrow$ 90% and 82% accuracy.

   Demo: in Fig. 3 ($d = 2$) and Fig. 4 ($d = 50$).
2. Comparison with NN; $d = 20$, RP dimensions $d' = 1, 5$, Fig. 2(e)-(f):
   - similar performances,
   - 8 to 30 times speed-up at $T = 100{,}000$ *for serial implementations*.
3. When MI-graph clustering fails, e.g., for the *all-4-independent* database: RP with CE provides accurate estimations, see Fig. 2(d).

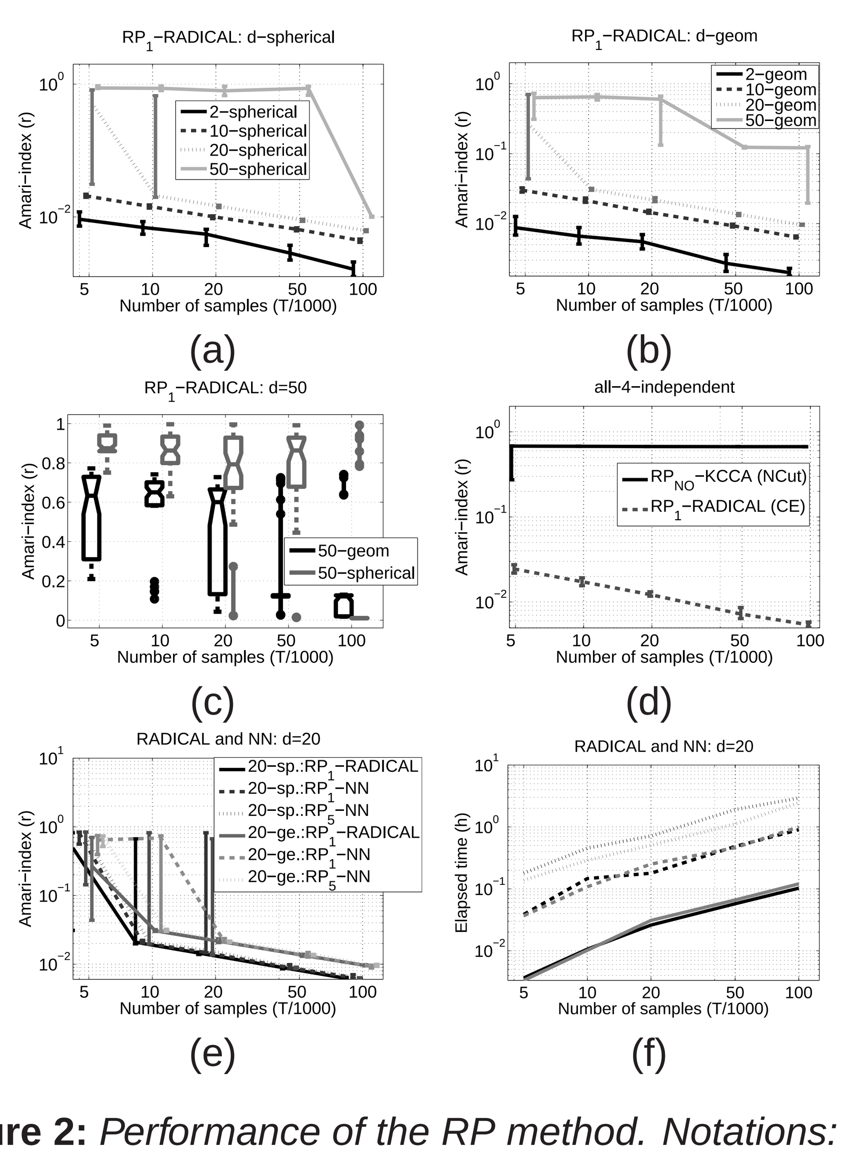


Figure 2: *Performance of the RP method. Notations: 'RP*$_{d'}$ *- method of cost estimation (method of optimization if not greedy)'. (a), (b): accuracy for the* d-spherical *and the* d-geom *databases, log-log scale. (c): notched boxed plots for* $d = 50$*. (d): RP+global optimization vs. pairwise MI+NCut on the* all-4-independent *dataset, log-log scale. (e)-(f): accuracy, computation time comparisons with NN for the* 20-spherical *and the* 20-geom *databases (on log-log scale).*

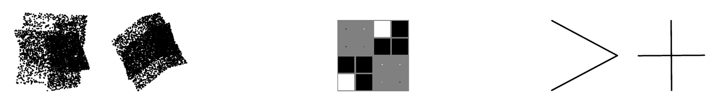


Figure 3: *Illustration on the* 2-geom *test (*$T = 100{,}000$*). Left: observed signals* $\mathbf{x}(t)$*; center: Hinton-diagram of* $\mathbf{G}$*, ideally block-permutation matrix with* 2×2 *blocks; right: estimated components* $\hat{\mathbf{s}}^m$*, recovered up to the ISA ambiguities.*

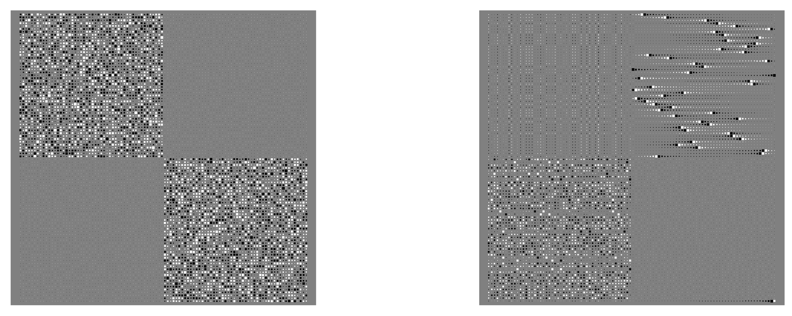


Figure 4: *Hinton-diagrams with average Amari-indices on the* 50-spherical *(left) and the* 50-geom *(right) datasets.*

Acknowledgments.

This work has been supported by the National Office for Research and Technology.

References

[1] Neemuchwala, H., Hero, A., Zabuawala, S., Carson, P.: Image registration methods in high-dimensional space. Int. J. Imaging Syst. and Technol. **16** (2007) 130–145

[2] 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** (1998) 891 – 923

[3] Johnson, W., Lindenstrauss, J.: Extensions of Lipschitz maps into a Hilbert space. Contemporary Mathematics **26** (1984) 189–206

[4] Matoušek, J.: On variants of the Johnson-Lindenstrauss lemma. Random Structures and Algorithms **33** (2008) 142–156

[5] Cardoso, J.: Multidimensional independent component analysis. In: ICASSP'98. Volume 4. (1998) 1941–1944

[6] Theis, F.J.: Uniqueness of complex and multidimensional independent component analysis. Signal Processing **84** (2004) 951–956

[7] Theis, F.J.: Multidimensional independent component analysis using characteristic functions. In: EUSIPCO'05. (2005)

[8] Amari, S., Cichocki, A., Yang, H.H.: A new learning algorithm for blind signal separation. NIPS 1996 **8** (1996) 757–763

[9] Szabó, Z., Póczos, B., Lőrincz, A.: Cross-entropy optimization for independent process analysis. In: ICA'06. Volume 3889 of LNCS., Springer (2006) 909–916

[10] Póczos, B., Lőrincz, A.: Independent subspace analysis using k-nearest neighborhood distances. LNCS **3697** (2005) 163–168

[11] Szabó, Z., Póczos, B., Lőrincz, A.: Undercomplete blind subspace deconvolution. J. of Machine Learning Res. **8** (2007) 1063–1095

[12] Kybic, J.: High-dimensional mutual information estimation for image registration. In: ICIP'04, IEEE Computer Society (2004) 1779–1782

[13] Learned-Miller, E. G., Fisher, J. W. III: ICA using spacings estimates of entropy. J. of Machine Learning Res. **4** (2003) 1271–1295

[14] Bach, F.R., Jordan, M.I.: Beyond independent components: Trees and clusters. J. of Machine Learning Res. **4** (2003) 1205–1233

[15] Rubinstein, R.Y., Kroese, D.P.: The Cross-Entropy Method. Springer (2004)

[16] Póczos, B., Szabó, Z., Kiszlinger, M., Lőrincz, A.: Independent process analysis without a priori dimensional information. In: ICA'07. Volume 4666 of LNCS., Berlin Heidelberg, Springer-Verlag (2007) 252–259