Hiányosan Megfigyelt Független Altér Analízis

Szabó Zoltán

Programozáselmélet és Szoftvertechnológiai Tanszék, Informatikai Kar,
Eötvös Loránd Tudományegyetem,
Pázmány Péter sétány 1/C, Budapest, 1117
WWW: `http://nipg.inf.elte.hu`
`szzoli@cs.elte.hu`

Kivonat

Jelen munka célja független többdimenziós folyamatok keresése hiányosan megfigyelt keverés mellett. A feladatnak megfelelő koktél parti probléma több sikeres alkalmazással bír, azonban a hiányosan megfigyelt eset csak a legegyszerűbb ICA (Independent Component Analysis) megfogalmazásra kidolgozott, ahol a rejtett független folyamatok (i) 1-dimenziósak, és (ii) időben i.i.d. eloszlásúak. Munkánkban a független folyamat keresést a hiányosan megfigyelt esetben kiterjesztjük (i) dinamikával rendelkező (AR, autoregresszív), (ii) többdimenziós független folyamatok esetére. A megoldásra szeparációs elvet származtatunk, miszerint a megoldás szétbontható: hiányosan megfigyelt AR becslésre, és független altér analízisre (ISA, Independent Subspace Analysis), amelyeket már meg tudunk oldani. Megközelítésünk hatékonyságát numerikus példákkal illusztráljuk.

1. Bevezetés–Koktél Parti Feladat

- Koktél parti probléma:
  - D beszélő, D mikrofon,
  - feladat: a kevert jelekből az eredetiek helyreállítása.

- Feltevések: független források (ICA) $\xrightarrow{\text{enyhítés}}$ független forráscsoportok (ISA).
- Alkalmazások:
  - ICA: jellemző kivonatolás, zajtalanítás, pénzügyi és neurobiológiai adatok elemzése, arcfelismerés.
  - ISA: EEG, fMRI, ECG, gén adatok elemzése, mintázat és arcirány felismerés.
- Hiányos megfigyelések esete:
  - Irodalomban csak a legegyszerűbb i.i.d., 1D források esete kidolgozott (ICA) [1, 2].
  - Most: többdimenziós források esete (ISA) és AR dinamika.

2. Hiányosan Megfigyelt AR-ISA Modell

2.1 Egyenletek

Többdimenziós független AR források keverékének ($\mathbf{x}$) hiányosan mért változata a megfigyelésünk ($\mathbf{y}$):

$$\mathbf{s}_{t+1} = \sum_{j=1}^{L} \mathbf{F}_j \mathbf{s}_{t+1-j} + \mathbf{e}_{t+1}, \tag{1}$$

$$\mathbf{x}_t = \mathbf{A}\mathbf{s}_t, \tag{2}$$

$$\mathbf{y} = \mathcal{M}(\mathbf{x}). \tag{3}$$

Itt:

- $\mathbf{s}^m \in \mathbb{R}^{d_m}$ $(m = 1, \ldots, M)$ a rejtett források, $\mathbf{s}(t) := [\mathbf{s}^1(t); \ldots; \mathbf{s}^M(t)] \in \mathbb{R}^D$, $D = \sum_{m=1}^{M} d_m$,
- $\mathbf{F}_j$ mátrixok és az $\mathbf{e}$ meghajtó zaj írja le $\mathbf{s}$ dinamikáját,
- $\mathbf{A} \in \mathbb{R}^{D\times D}$ az ismeretlen *keverő mátrix*,
- a nem-hiányzó mérések időpontjait, és koordinátáit adja az $\mathcal{M}$ maszk leképezés..

Feladat: az $\mathbf{y}$ megfigyelésből a rejtett $\mathbf{s}$ forrás és az $\mathbf{A}$ keverőmátrix (avagy $\mathbf{W}$ inverzének) becslése.

2.2 Feltételek

- $\mathbf{e}^m$ meghajtó zajok függetlenek $[I(\mathbf{e}^1, \ldots, \mathbf{e}^M) = 0]$, nem-Gaussok, és időben i.i.d.-k,
- $\mathbf{A}$: invertálható,
- $\mathbf{F}[z] = \mathbf{I} - \sum_{j=1}^{L} \mathbf{F}_j z^j$ polinommátrix stabil, azaz

$$\det(\mathbf{F}[z]) \neq 0, (\forall z \in \mathbb{C}, |z| \leq 1). \tag{4}$$

2.3 Speciális esetek

$\mathcal{M}$ = identitás és $L = 0$: (i.i.d.-)ISA feladat adódik. Ha plusszban még $\forall d_m = 1$, akkor az ICA-t kapjuk.

2.4 Megoldási Stratégia

- $\mathbf{s}$ AR, $\mathbf{x}$ az ő invertálható lineáris transzformáltja $\Rightarrow$ $\mathbf{x}$ AR, $\mathbf{A}\mathbf{e}$ innovációval:

$$\mathbf{x}_{t+1} = \sum_{j=1}^{L} \mathbf{A}\mathbf{F}_j\mathbf{A}^{-1}\mathbf{x}_{t+1-j} + \mathbf{A}\mathbf{e}_{t+1}. \tag{5}$$

- $\mathbf{A}\mathbf{e}$: közelítőleg Gauss ($\Leftarrow$ d-függő CHT [3]),
- $\mathbf{y}$ hiányosan megfigyelt AR-re: fit, majd a becsült innováción ISA.

3. Illusztráció

3.1 Jóságmérce

ISA egyértelműség [4] miatt:

- a rejtett forráskomponensek: (i) permutáció, és (ii) altéren belüli lineáris transzformáció erejéig állíthatóak helyre $\Rightarrow$
- Ideális esetben: $\mathbf{G} = \hat{\mathbf{W}}_{\text{ISA}}\mathbf{A}$ egy blokk-permutációs mátrix. Ez a tulajdonság lemérhető az Amari-index-szel [5]:
  - az általánosság rovása nélkül feltehető, hogy a forrás dimenziók monoton növők: $d_1 \leq d_2 \leq \ldots \leq d_M$,
  - definíció:

$$r(\mathbf{G}) := \frac{1}{2M(M-1)} \left[\sum_{i=1}^{M} \left(\frac{\sum_{j=1}^{M} g^{ij}}{\max_j g^{ij}} - 1 \right) + \sum_{j=1}^{M} \left(\frac{\sum_{i=1}^{M} g^{ij}}{\max_i g^{ij}} - 1 \right) \right], \tag{6}$$

ahol $\mathbf{G} = \left[\mathbf{G}^{ij} \in \mathbb{R}^{d_i \times d_j}\right]_{i,j=1,\ldots,M}$, $g^{ij} = \sum_k \sum_l |\mathbf{G}^{ij}_{kl}|$.

  - Tulajdonsága: $r(\mathbf{G}) \in [0,1]$, $r(\mathbf{G}) = 0$–a tökéletes.

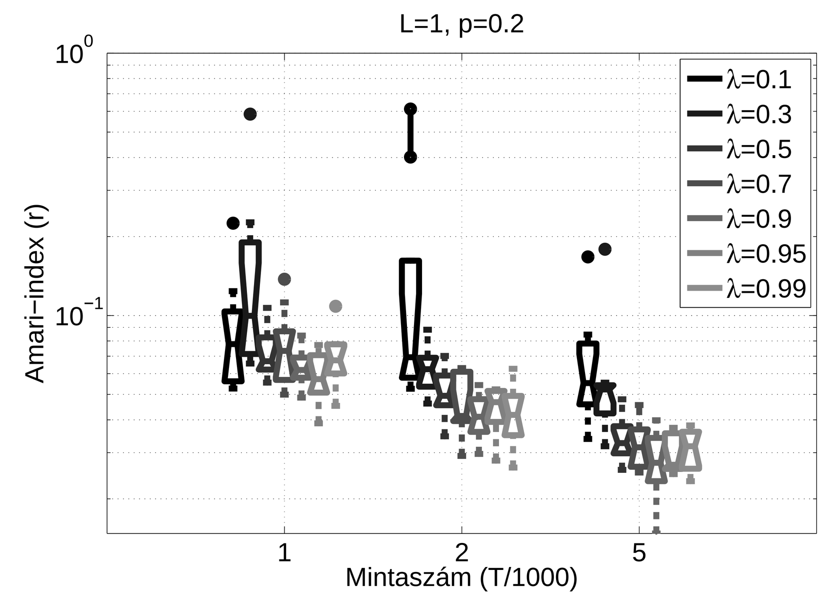
3.2 Demo

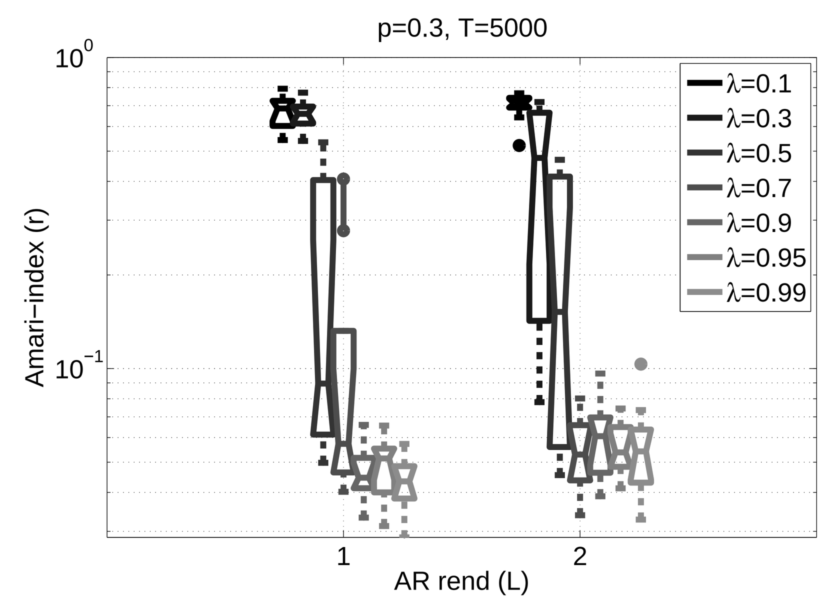
- Részfeladatok: (i) hiányos AR identifikáció maximum-likelihood elven [6, 7], (ii) ISA részfeladat a becsült ICA elemek csoportosításával (ISA szeparációs tétel [8]), (iii) ICA lépés fastICA-val [9], (iv) klaszterezésben a kölcsönös információ becslése KCCA-val [10].
- Jóságmérce: Amari-index (r), 10 véletlen ($\mathbf{e}, \mathbf{A}, \mathbf{F}[z]$) futtatásra átlagolva, fix paraméterek (T, L, λ, p) mellett:
  - rejtett forráskomponensek ($\mathbf{e}^m$): 3 db 2-dimenziós ($d_m = 2$, $M = 3$), ABC betűin egyenletesek.
  - $\mathbf{A}$ keverőmátrix: véletlen ortogonális,
  - $\mathbf{F}[z]$ dinamika stabil:

$$\mathbf{F}[z] = \prod_{j=1}^{L} (\mathbf{I} - \lambda \mathbf{O}_i z) \quad (|\lambda| < 1, \lambda \in \mathbb{R}) \tag{7}$$

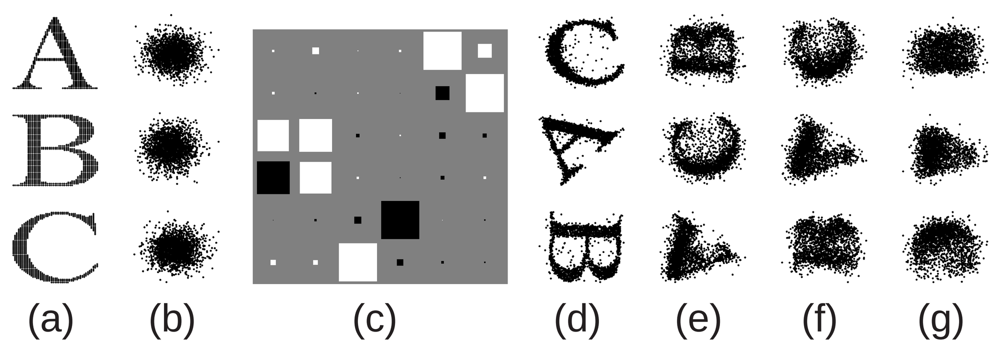
ahol $\lambda \to 1$ ($\lambda \in \{0.1, 0.3, 0.5, 0.7, 0.9, 0.95, 0.99\}$), $\mathbf{O}_i$-k véletlen ortogonális mátrixok, AR rend $L \in \{1, 2\}$.

  - Maszk leképezés ($\mathcal{M}$): minden (koordináta,időpont) pár egymástól függetlenül p valószínűséggel nem volt megfigyelhető, $p \in \{0.01, 0.1, 0.2, 0.3\}$.
  - Mintaszám: $1,000 \leq T \leq 5,000$.
- Statisztikák megjelenítése: boxplot-okkal. Benne:
  - kvartilisek (Q_1, Q_2, Q_3),
  - legnagyobb/-kisebb nem kiugrók (kiugró$\notin$ $[Q_1 - 1.5IQR, Q_3 + 1.5IQR]$, $IQR = Q_3 - Q_1$) kinyúló tüskékkel,
  - kiugrók körrel.
- Numerikus tapasztalatok: a bemutatott technika
  - 20–30%-os megfigyelés hiányig ($p = 0.2 - 0.3$),
  - az AR folyamat invertálhatósági tartományának peremén ($\lambda \to 1$) is stabilan becsli a forrásokat. Statisztikák összesítéséért lásd 2. ábra. Becslés illusztrációjáért lásd 3. ábra.





ábra 2: *Becslési hiba illusztrációja. Fent: Amari-index mintaszám függvényében kül. λ kontrakciós paraméterekre, $L = 1$, $p = 0.2$. Lent: Amari-index $L = 1$ illetve $L = 2$-re, mintaszám $T = 5,000$, megfigyelés hiánya $p = 0.3$.*



ábra 3: *Becslés illusztrációja ($T = 5,000$, $\lambda = 0.9$): (a): rejtett $\mathbf{e}^m$ komponensek sűrűségfüggvényei. (b): megfigyelés az $\mathcal{M}$ „ritkítás” előtt ($\mathbf{x}$). (d): Amari-index szerint átlagos jóságú becsült komponensek ($\hat{\mathbf{e}}^m$) 1%-osan ($p = 0.01$) hiányos megfigyelés mellett. (c): (d)-nek megfelelő $\mathbf{G}$ mátrix Hinton-diagramja, közelítőleg 2×2-es blokkokból álló blokk-permutációs mátrix. (e)-(g): mint (d), csak a megfigyelési hiány mértéke $p = 0.1$, $p = 0.2$ és $p = 0.3$.*

Hivatkozások

[1] Chan, K., Lee, T.W., Sejnowski, T.J.: Variational Bayesian learning of ICA with missing data. Neural Computation **15** (2003) 1991 – 2011

[2] Cemgil, A.T., Févotte, C., Godsill, S.J.: Variational and stochastic inference for Bayesian source separation. Digital Signal Processing **17** (2007) 891–913

[3] Petrov, V.: Central limit theorem for m-dependent variables. In: Proceedings of the All-Union Conference on Probability Theory and Mathematical Statistics. (1958) 38–44

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

[5] Amari, S., Cichocki, A., Yang, H.H.: A new learning algorithm for blind signal separation. Advances in Neural Information Processing Systems **8** (1996) 757–763

[6] Lomba, J.T.: Estimation of Dynamic Econometric Models with Errors in Variables. Volume 339 of Lecture notes in economics and mathematical systems. Berlin; New York:Springer-Verlag (1990)

[7] Casals, J., Sotoca, S.: Exact initial conditions for maximum likelihood estimation of state space models with stochastic inputs. Economics Letters **57** (1997) 261–267

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

[9] Hyvärinen, A., Oja, E.: A fast fixed-point algorithm for independent component analysis. Neural Computation **9** (1997) 1483–1492

[10] Bach, F.R., Jordan, M.I.: Kernel independent component analysis. Journal of Machine Learning Research **3** (2002) 1–48