

Independent components analyze

Conf.dr. Cătălina-Lucia COCIANU
Catedra de Informatică Economică, ASE București

The work reported in this paper focused on the independent component analysis. Independent component analysis (ICA) is a statistical technique for revealing hidden factors that underlie sets of random variables, measurements or signals. ICA defines a generative model for the observed multivariate data, which is typically given as a large database of samples. In the model, the data variables are assumed to be mixtures of some unknown latent variables and the mixing system is also unknown. Some methods for estimating the model of independent component analysis based on maximization of nongaussianity are reported in the final part of the paper.

Keywords: principal components, independent components, nongaussianity, kurtosis, negentropy.

Analiza în componente independente (ICA) definește un model generativ pentru date multimodale observate, care constau în general în eșantioane grupate în baze de date de dimensiuni mari. În cadrul modelării ICA, variabilele observate sunt definite prin combinații liniare sau neliniare de variabile latente necunoscute; de asemenea, sistemul de combinare a variabilelor latente este necunoscut. Variabilele latente care definesc datele observate sunt repartizate non-gaussian, sunt presupuse mutual independente și sunt numite componentele independente ale datelor observate. Tehnicile ICA permit calcularea componentelor independente din cadrul datelor observate.

Tehnicile ICA sunt aplicate în analiza imaginilor digitale, a bazelor de date document, dar și la măsurători psihometrice sau în cadrul calculului unor indicatori economici. În general, datele observate (măsurătorile) sunt definite prin intermediul unei mulțimi de semnale captate în paralel sau prin utilizarea seriilor temporale. Câteva exemple de date asupra cărora sunt aplicate tehnici de analiză de tip ICA sunt: combinații de semnale sonore captate simultan prin intermediul microfoanelor, undele cerebrale înregistrate prin aplicarea unor senzori multipli, seriile temporale paralele obținute în cadrul unor procese industriale ș.a.m.d.

1. Definirea modelului de analiză în componente independente

Modelul statistic ICA este definit prin intermediul variabilelor latente astfel. Fie x_1, x_2, \dots, x_n variabile aleatoare, definite prin combinația liniară a variabilelor aleatoare s_1, s_2, \dots, s_n : pentru orice $1 \leq i \leq n$,

$$(1) \quad x_i = a_{i1}s_1 + a_{i2}s_2 + \dots + a_{in}s_n,$$

unde, pentru orice $1 \leq i, j \leq n$, $a_{ij} \in \mathbf{R}$. Prin definiție, variabilele s_1, s_2, \dots, s_n sunt mutual independente. Modelul ICA este generativ, în sensul că, datele observate sunt obținute printr-un proces de mixare a componentelor $s_i, 1 \leq i \leq n$. Componentele independente $s_i, 1 \leq i \leq n$ sunt variabile latente (nu sunt direct observabile). De asemenea, coeficienții $a_{ij}, 1 \leq i, j \leq n$ sunt necunoscuți. Problema este de a estima coeficienții $a_{ij}, 1 \leq i, j \leq n$ și variabilele latente $s_i, 1 \leq i \leq n$ pe baza observațiilor obținute asupra variabilelor aleatoare x_1, x_2, \dots, x_n într-un cadru cât mai puțin restrictiv.

Fie $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$, $\mathbf{s} = (s_1, s_2, \dots, s_n)^T$ și $\mathbf{A} = \|a_{i,j}\|_{1 \leq i, j \leq n}$. Modelul ICA este definit de relația matriceală:

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

Fie $\mathbf{a}_i, 1 \leq i \leq n$ coloanele matricei \mathbf{A} . Modelul ICA poate fi definit prin,

$$(3) \mathbf{x} = \sum_{i=1}^n \mathbf{a}_i s_i$$

Modelul ICA cu zgomot este definit prin,

$$(4) \mathbf{x} = \mathbf{A}\mathbf{s} + \boldsymbol{\eta},$$

unde $\boldsymbol{\eta} = (\eta_1, \eta_2, \dots, \eta_n)^T$ este vectorul zgomot. În general, tehnicile ICA în cadrul modelelor cu zgomot sunt dezvoltate în ipotezele în care zgomotul este gaussian și independent de componentele principale $s_i, 1 \leq i \leq n$. În continuare este tratat cazul modelelor ICA fără zgomot.

Estimarea coeficienților și a componentelor independente în modelul definit de (2) este realizată în cadrul următorului sistem de restricții.

1. Variabilele latente $s_i, 1 \leq i \leq n$ sunt independente.
2. Componentele independente $s_i, 1 \leq i \leq n$ au distribuție non-gaussiană.
3. Matricea \mathbf{A} este pătratică și inversabilă. Fără a pierde din generalitate, pentru simplificarea algoritmilor ICA, vom presupune în continuare că variabilele mixtură și componentele independente au medie 0. În cazul în care presupunerea este falsă, este aplicată centrarea variabilelor utilizate. Dacă vectorul variabilelor mixtură este inițial \mathbf{x}' , cu $E(\mathbf{x}') \neq (0, 0, \dots, 0)^T$, atunci algoritmi ICA sunt aplicați vectorului centrat $\mathbf{x} = \mathbf{x}' - E(\mathbf{x}')$. Deoarece $E(\mathbf{s}) = \mathbf{A}^{-1} E(\mathbf{x})$, componentele independente rezultă cu medie 0 și matricea de mixare \mathbf{A} rămâne nemodificată în urma procesului de centrare [6].

2. Estimarea modelelor ICA prin maximizarea non-gaussianității

Non-gaussianitatea este o ipoteză de bază în tehnicile ICA (vezi observațiile de la secțiunea 1). Măsurarea și maximizarea non-gaussianității unui vector aleator determină obținerea unor modalități de estimare a modelelor ICA.

Fie \mathbf{x} vector aleator observat distribuit conform modelului ICA, $\mathbf{x} = \mathbf{A}\mathbf{s}$. Vom presupune pentru început că toate componentele independente sunt identic distribuite. Estimarea variabilelor latente este realizată prin

$\mathbf{s} = \mathbf{A}^{-1}\mathbf{x}$, deci o componentă independentă poate fi estimată printr-o combinație liniară a componentelor lui \mathbf{x} ,

$$y = \mathbf{b}^T \mathbf{x} = \sum_{i=1}^n b_i x_i, \text{ unde } \mathbf{b} \text{ este vectorul care}$$

trebuie determinat. Obținem,

$$(5) y = \mathbf{b}^T \mathbf{x} = \mathbf{b}^T \mathbf{A}\mathbf{s}.$$

Pe baza relației (5) rezultă că y este reprezentat prin combinația liniară a componentelor independente $s_i, 1 \leq i \leq n$ cu coeficienții

$$\mathbf{q}^T = \mathbf{b}^T \mathbf{A}, \quad y = \mathbf{b}^T \mathbf{x} = \mathbf{q}^T \mathbf{s} = \sum_{i=1}^n q_i s_i.$$

În practică, vectorul \mathbf{b} poate fi estimat prin utilizarea teoremei limită centrală. Ideea care stă la baza estimării lui \mathbf{b} este următoarea. Deoarece combinația liniară a unor variabile aleatoare este mai aproape de o variabilă distribuită normal, comparativ cu oricare dintre variabilele utilizate, $y = \mathbf{q}^T \mathbf{s}$ este mai "gaussiană" decât oricare din componentele independente și este cu atât mai aproape de una din variabilele s_i cu cât apropierea de repartiția normală este mai mică. Pentru $1 \leq i \leq n$, dacă $y = s_i$, atunci q_i este singura componentă nenulă a vectorului \mathbf{q} . Rezultă că vectorul \mathbf{b} este aproximat prin maximizarea non-gaussianității variabilei $y = \mathbf{b}^T \mathbf{x}$. Dacă \mathbf{b} este astfel încât non-gaussianitatea lui $\mathbf{b}^T \mathbf{x}$ este maximă, atunci $\mathbf{q} = \mathbf{A}^T \mathbf{b}$ are o singură componentă nenulă și $y = \mathbf{b}^T \mathbf{x} = \mathbf{q}^T \mathbf{s}$ este egală cu una din componentele independente. În spațiul n -dimensional al vectorilor \mathbf{b} , prin optimizarea non-gaussianității componenteii $y = \mathbf{b}^T \mathbf{x}$ sunt obținute $2n$ puncte de maxim local, corespunzând variabilelor $s_i, -s_i, 1 \leq i \leq n$. Pentru optimizarea spațiului de căutare a vectorului \mathbf{b} este parcursă o etapă de preprocesare, constând în transformarea vectorului aleator \mathbf{x} într-un vector alb (vezi secțiunea 1), $\mathbf{z} = \mathbf{V}\mathbf{x} = \mathbf{V}\mathbf{A}\mathbf{s}$, cu $E(\mathbf{z}\mathbf{z}^T) = \mathbf{I}_n$. Pentru vectorul \mathbf{z} este determinat \mathbf{w} care maximizează non-gaussianitatea combinației liniare $\mathbf{w}^T \mathbf{z}$. Deoarece $\mathbf{q} = \mathbf{A}^T \mathbf{b} = (\mathbf{V}\mathbf{A})^T \mathbf{w}$, rezultă

$$(6) \|\mathbf{q}\|^2 = (\mathbf{w}^T \mathbf{V}\mathbf{A})(\mathbf{A}^T \mathbf{V}^T \mathbf{w}) = \|\mathbf{w}\|^2.$$

Pe baza relației (6) rezultă că, dacă \mathbf{q} aparține

sferii unitate, atunci și \mathbf{w} este situat pe sfera unitate. În consecință, optimizarea non-gaussianității combinației liniare $\mathbf{w}^T \mathbf{z}$ este realizată în ipoteza $\|\mathbf{w}\|=1$.

2.1. Măsurarea non-gaussianității prin kurtosis

Fie y o variabilă aleatoare de medie 0. Funcția kurtosis a variabilei y este definită pe baza funcției caracteristice de gradul II,

$$\phi(\omega) = \ln(\varphi(\omega)) = \ln(E(\exp(j\omega y))),$$

unde $j = \sqrt{-1}$ și φ este funcția caracteristică de ordinul I a variabilei y . Cumulanții de ordin k , $k \geq 1$, ai variabilei aleatoare y sunt coeficienții dezvoltării în serie Taylor a funcției

$$\phi, \kappa_k = (-j)^k \left. \frac{d^k \phi(\omega)}{d\omega^k} \right|_{\omega=0}.$$

Funcția kurtosis a variabilei aleatoare y este cumulantul de ordin 4, $\text{kurt}(y) = E(y^4) - 3(E(y^2))^2$.

În continuare vom presupune că y este normalizat, $E(y^2) = 1$. Obținem, $\text{kurt}(y) = E(y^4) - 3$, deci, în această situație, kurtosis-ul este o versiune normalizată a momentului de ordin 4 al variabilei aleatoare y . Dacă y este distribuită normal, atunci $E(y^4) = 3(E(y^2))^2$ și

$$\text{kurt}(y) = 3(E(y^2))^2 - 3 = 0.$$

În cazul repartiției normale de medie 0 și varianță 1, kurtosis-ul este nul; pentru majoritatea repartițiilor non-gaussiene, măsura kurtosis este, însă, nenulă. Variabila aleatoare y se numește subgaussiană dacă $\text{kurt}(y) < 0$; dacă $\text{kurt}(y) > 0$, atunci y este supragaussiană. O variabilă aleatoare supragaussiană, y , are densitatea de repartiție cu valori mari în 0 și în pentru y mare, respectiv cu valori mici pentru valori intermediare ale lui y . Dacă y este subgaussiană, atunci densitatea de repartiție este cu variații mici într-o vecinătate a lui 0 și cu valori mici în rest. Datorită modului simplu de calcul, funcția $|\text{kurt}|$ este des utilizată ca măsură a non-gaussianității în estimarea modelelor ICA [6]. Măsura kurtosis este estimată uzual prin momentul de ordin 4 al datelor observate, analiza teoretică fiind simplificată de proprie-

tatea de liniaritate: dacă x_1, x_2 sunt variabile aleatoare independente și α este constantă, atunci,

$$\text{kurt}(x_1 + x_2) = \text{kurt}(x_1) + \text{kurt}(x_2) \text{ și}$$

$$\text{kurt}(\alpha x_1) = \alpha^4 \text{kurt}(x_1).$$

Algoritmul gradient bazat pe funcția kurtosis

Fie \mathbf{x} vector aleator observat distribuit conform modelului ICA, $\mathbf{x} = \mathbf{A}\mathbf{s}$ și $\mathbf{z} = \mathbf{V}\mathbf{x} = \mathbf{V}\mathbf{A}\mathbf{s}$, cu $E(\mathbf{z}\mathbf{z}^T) = \mathbf{I}_n$ vectorul obținut după preprocesare. Maximizarea funcției $|\text{kurt}(\mathbf{w}^T \mathbf{z})|$ este

realizată pe baza algoritmului gradient astfel. Fie \mathbf{w} inițial selectat aleator. Vectorul \mathbf{w} este modificat conform direcției în care

$|\text{kurt}(\mathbf{w}^T \mathbf{z})|$ are o creștere mai "rapidă". Deoarece vectorul \mathbf{z} este cu proprietatea că

$$E(\mathbf{z}\mathbf{z}^T) = \mathbf{I}_n, \text{ rezultă că } E((\mathbf{w}^T \mathbf{z})^2) = \|\mathbf{w}\|^2 \text{ și}$$

gradientul funcției $|\text{kurt}(\mathbf{w}^T \mathbf{z})|$ este calculat prin,

$$(7) \frac{\partial |\text{kurt}(\mathbf{w}^T \mathbf{z})|}{\partial \mathbf{w}} = 4 \text{sign}(\text{kurt}(\mathbf{w}^T \mathbf{z})) [E(\mathbf{z}(\mathbf{w}^T \mathbf{z})^3) - 3\mathbf{w}\|\mathbf{w}\|^2].$$

Funcția $|\text{kurt}(\mathbf{w}^T \mathbf{z})|$ este optimizată pe sfera unitate, deci vectorul \mathbf{w} este normalizat după fiecare pas. Pentru a simplifica algoritmul putem omite termenul $-3\mathbf{w}\|\mathbf{w}\|^2$, deoarece acesta schimbă doar norma lui \mathbf{w} , nu și direcția. Algoritmul gradient este descris, pe baza relației (7), astfel [6].

$$\Delta \mathbf{w} = \mathbf{w}(t) - \mathbf{w}(t-1)$$

$$\Delta \mathbf{w} \propto \text{sign}(\text{kurt}(\mathbf{w}^T \mathbf{z})) E(\mathbf{z}(\mathbf{w}^T \mathbf{z})^3)$$

$$\mathbf{w} \leftarrow \frac{\mathbf{w}}{\|\mathbf{w}\|}, \text{ unde } \propto = \text{"este proporțional cu"}.$$

O variantă adaptivă a algoritmului gradient pentru maximizarea non-gaussianității poate fi obținută prin omiterea operației de medie; rezultă relațiile,

$$\Delta \mathbf{w} \propto \text{sign}(\text{kurt}(\mathbf{w}^T \mathbf{z})) \mathbf{z}(\mathbf{w}^T \mathbf{z})^3, \mathbf{w} \leftarrow \frac{\mathbf{w}}{\|\mathbf{w}\|}.$$

În varianta adaptivă, fiecare observație $\mathbf{z}(t)$ poate fi utilizată în algoritm o singură dată.

Observații

1. Pentru calculul valorii $\text{sign}(\text{kurt}(\mathbf{w}^T \mathbf{z}))$, nu poate fi omisă operația de mediere din defini-

ția funcției $|\text{kurt}(\mathbf{w}^T \mathbf{z})|$. În acest scop, funcția kurtosis poate fi estimată prin,

$$(8) \Delta\gamma \propto \left(\left(\mathbf{w}^T \mathbf{z} \right)^4 - 3 \right) - \gamma$$

Relația (8) definește estimarea funcției kurtosis de tip medie mobilă.

2. În multe situații practice, este cunoscută proprietatea de subgaussianitate, respectiv supra gaussianitate a distribuțiilor de probabilitate a componentelor independente, deci $\text{sign}(\text{kurt}(\mathbf{w}^T \mathbf{z}))$ este cunoscut.

Algoritmul FastICA bazat pe utilizarea funcției kurtosis

Algoritmul FastICA pe baza funcției kurtosis este mai eficient comparative cu algoritmul gradient, datorită următoarelor considerente. La momentul stabilizării algoritmului gradient, gradientul funcției $|\text{kurt}(\mathbf{w}^T \mathbf{z})|$ indică direcția vectorului \mathbf{w} , adică este egal cu \mathbf{w} multiplicat cu o constantă scalară. Doar în această situație, adăugarea gradientului la \mathbf{w} nu modifică direcția vectorului și algoritmul converge [6].

Algoritmul FastICA este derivat astfel. Din (7) obținem,

$$(9) \mathbf{w} \propto \left[E\left(\mathbf{z}(\mathbf{w}^T \mathbf{z})^3 \right) - 3\mathbf{w}\|\mathbf{w}\|^2 \right].$$

Relația (9) permite utilizarea următoarei reguli de actualizare, urmată de scalare,

$$(10) \mathbf{w} \leftarrow \left[E\left(\mathbf{z}(\mathbf{w}^T \mathbf{z})^3 \right) - 3\mathbf{w} \right], \quad \mathbf{w} \leftarrow \frac{\mathbf{w}}{\|\mathbf{w}\|}.$$

Vectorul obținut pe baza relației (10) permite calculul uneia din componentele independente prin $\mathbf{w}^T \mathbf{z}$.

2.2. Măsurarea non-gaussianității prin negentropie

Negentropia este definită pe baza principiului entropiei maxime, ca o măsură a non-gaussianității unui vector aleator. Fie \mathbf{y} vector aleator cu densitatea de repartiție p_y . Entropia vectorului \mathbf{y} este definită prin,

$$H(\mathbf{y}) = - \int p_y(\boldsymbol{\eta}) \log p_y(\boldsymbol{\eta}) d\boldsymbol{\eta}.$$

$$H(y) \approx - \int \varphi(\xi) \left(1 + \kappa_3 \frac{H_3(\xi)}{3!} + \kappa_4 \frac{H_4(\xi)}{4!} \right) \left[\log(\varphi(\xi)) + \kappa_3 \frac{H_3(\xi)}{3!} + \kappa_4 \frac{H_4(\xi)}{4!} - \frac{\left(\kappa_3 \frac{H_3(\xi)}{3!} + \kappa_4 \frac{H_4(\xi)}{4!} \right)^2}{2} \right] d\xi$$

Conform principiului entropiei maxime, în clasa distribuțiilor cu medie și matrice de covarianță date, entropia maximă este obținută pentru distribuția normală. Rezultă că entropia poate fi utilizată ca măsură a non-gaussianității. Negentropia unui vector aleator \mathbf{y} cu medie \mathbf{m}_y și matrice de covarianță Σ_y este definită prin,

$$(11) J(\mathbf{y}) = H(\mathbf{y}_{Gauss}) - H(\mathbf{y}), \text{ unde } \mathbf{y}_{Gauss} \text{ este distribuit } N(\mathbf{m}_y, \Sigma_y).$$

Negentropia este întotdeauna pozitivă, valoarea 0 este atinsă dacă și numai dacă \mathbf{y} este distribuit normal. Utilizarea funcției negentropie ca măsură a non-gaussianității unui vector aleatoru este justificată de teoria informației (principiul entropiei maxime). Principalul inconvenient în măsurarea non-gaussianității prin negentropie este de ordin practic; estimarea negentropiei pe baza definiției (11) necesită estimarea (posibil non-parametrică) a densității de repartiție.

O variantă alternativă de aproximare a negentropiei unei variabile aleatoare y , cu medie 0 și dispersie 1 poate fi obținută prin dezvoltarea Gram-Charlier a densității de repartiție p_y , [6]

$$(12) p_y(\xi) \approx \hat{p}_y(\xi) = \varphi(\xi) \left(1 + \kappa_3 \frac{H_3(\xi)}{3!} + \kappa_4 \frac{H_4(\xi)}{4!} \right),$$

unde φ este funcția caracteristică asociată lui y și H_i sunt polinoame de gradul i , cu proprietatea,

$$\int \varphi(\xi) H_i(\xi) H_j(\xi) d\xi = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}.$$
 Utilizând

relația (12), obținem, $H(y) = - \int \hat{p}_y(\xi) \log \hat{p}_y(\xi) d\xi$.

Entropia este aproximată în ipoteza că densitatea de repartiție este apropiată de repartiția normală standard, pentru care cumulanții din (12) sunt foarte mici, deci $\log(1 + \varepsilon) \approx \varepsilon - \frac{\varepsilon^2}{2}$.

Obținem,

și, în continuare,

$$H(y) \approx - \int \varphi(\xi) \log(\varphi(\xi)) d\xi - \frac{(\kappa_3)^2}{2 \cdot 3!} - \frac{(\kappa_4)^2}{2 \cdot 4!}.$$

Rezultă că negentropia poate fi estimată prin,

$$(13) J(y) \approx \frac{1}{12} (E(y^3))^2 + \frac{1}{48} \text{kurt}(y)^2.$$

O abordare alternativă a calculului negentropiei unei variabile aleatoare y presupune înlocuirea în (13) a funcțiilor polinomiale y^3, y^4 cu funcțiile G^i , nu neapărat polinomiale și aproximarea negentropiei prin mediile $E(G^i(y))$ [6]. De exemplu, pentru G^1 funcție nepolinomială impară și G^2 funcție nepolinomială pară, obținem aproximarea,

$$(14) J(y) \approx k_1 (E(G^1(y)))^2 + k_2 (E(G^2(y)) - E(G^2(v)))^2$$

unde k_1, k_2 sunt constante pozitive și v este variabilă normală standard. În cazul în care este folosită o singură funcție nepolinomială G , aproximarea negentropiei este,

$$(15) J(y) \propto (E(G(y)) - E(G(v)))^2.$$

Pentru estimarea parametrilor modelelor ICA, este utilizată una din variantele,

$$G(y) = \frac{1}{a_1} \log \cosh(a_1 y), \quad 1 \leq a_1 \leq 2,$$

$$\text{respectiv, } G(y) = - \exp\left(-\frac{y^2}{2}\right).$$

Algoritmul gradient pe baza negentropiei

Aproximarea unei componente independente a modelului ICA poate fi realizată pe baza algoritmului gradient pentru maximizarea negentropiei, astfel, [6]

$$\Delta \mathbf{w} = \mathbf{w}(t) - \mathbf{w}(t-1)$$

$$\Delta \mathbf{w} \propto \gamma E(\mathbf{z}g(\mathbf{w}^T \mathbf{z}))$$

$$\mathbf{w} \leftarrow \frac{\mathbf{w}}{\|\mathbf{w}\|}, \quad \text{unde } \gamma = E(G(y)) - E(G(v)), \quad v$$

este variabilă aleatoare normală standardizată și g este derivata funcției G , alegerile cele mai frecvente fiind,

$$(16) g(y) = \tanh(a_1 y),$$

$$(17) g(y) = y \exp\left(-\frac{y^2}{2}\right), \text{ respectiv}$$

$$(18) g(y) = y^3.$$

Similar algoritmului gradient pentru maximizarea non-gaussianității prin kurtosis, poate fi

obținută o variantă adaptivă prin eliminarea medierii în calculul $\Delta \mathbf{w}$. Constanta γ este estimată prin,

$$\Delta \gamma \propto (G(\mathbf{w}^T \mathbf{z}) - E(G(v))) - \gamma.$$

Obținem următorul algoritm [6] pentru calculul vectorului \mathbf{w} și derivarea componentei independente $\mathbf{w}^T \mathbf{z}$.

Date de intrare: $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$, vectorul datelor observate

Preprocesare:

1. Centrează datele observate

2. Calculează vectorul alb $\mathbf{z} = \mathbf{V}\mathbf{x}$,

$\mathbf{V} = \Lambda^{-\frac{1}{2}} \Phi^T$, unde $\Phi = (\varphi_1, \varphi_2, \dots, \varphi_n)$ este matricea cu coloane vectorii proprii normați ai matricei de covarianță $\Sigma = E(\mathbf{x}\mathbf{x}^T)$ și $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$.

Pas 1. Selectează aleator vectorul unitar \mathbf{w} și valoarea inițială a lui γ

Pas 2. Calculează noul vector \mathbf{w} prin $\Delta \mathbf{w} \propto \gamma \mathbf{z}g(\mathbf{w}^T \mathbf{z})$, cu g definită prin (16), (17) sau (18).

Pas 3. $\mathbf{w} \leftarrow \frac{\mathbf{w}}{\|\mathbf{w}\|}$

Pas 4. Dacă semnul valorii γ nu este a priori cunoscut, calculează noua valoare γ prin

$$\Delta \gamma \propto (G(\mathbf{w}^T \mathbf{z}) - E(G(v))) - \gamma.$$

Pas 5. Dacă algoritmul nu s-a stabilizat, reia **Pas 2 – Pas 4**.

Algoritmul FastICA pentru maximizarea negentropiei

Similar algoritmului FastICA pentru maximizarea non-gaussianității prin funcția kurtosis, metoda gradient sugerează regula de actualizare,

$$(17) \mathbf{w} \leftarrow E(\mathbf{z}g(\mathbf{w}^T \mathbf{z})),$$

urmată de normalizarea vectorului \mathbf{w} . Coeficientul γ poate fi omis, datorită faptului că γ este eliminat prin normalizare. Derivarea unui algoritm FastICA pentru maximizarea negentropiei nu poate fi realizată însă prin (17), datorită convergenței slabe a algoritmului rezultat. Regula de actualizare (17) este modificată astfel [6].

$\mathbf{w} = E(\mathbf{z}g(\mathbf{w}^T \mathbf{z}))$ este echivalentă cu

$$(18) (1 + \alpha)\mathbf{w} = E(\mathbf{z}g(\mathbf{w}^T \mathbf{z})) + \alpha\mathbf{w}.$$

Coeficientul α este determinat prin metoda de aproximare Newton. Pentru aplicarea metodei Newton, este utilizat faptul că maximum negentropyiei variabilei latente $\mathbf{w}^T \mathbf{z}$ este obținut pentru un optim al mediei $E(G(\mathbf{w}^T \mathbf{z}))$. Conform condițiilor Lagrange, optimul lui $E(G(\mathbf{w}^T \mathbf{z}))$ în ipoteza $E((\mathbf{w}^T \mathbf{z})^2) = \|\mathbf{w}\|^2 = 1$ este obținut în punctul care anulează Lagrangian-ul, adică

$$(19) E(\mathbf{z}g(\mathbf{w}^T \mathbf{z})) + \beta\mathbf{w} = 0.$$

Ecuția (19) este rezolvată prin metoda Newton astfel. Fie $F = E(\mathbf{z}g(\mathbf{w}^T \mathbf{z})) + \beta\mathbf{w}$. Gradientul funcției F este,

$$(20) \frac{\partial F}{\partial \mathbf{w}} = E(\mathbf{z}\mathbf{z}^T g'(\mathbf{w}^T \mathbf{z})) + \beta\mathbf{I}.$$

Pentru a simplifica determinarea inversei în (20), primul termen este aproximat prin

$$E(\mathbf{z}\mathbf{z}^T g'(\mathbf{w}^T \mathbf{z})) \approx E(\mathbf{z}\mathbf{z}^T)E(g'(\mathbf{w}^T \mathbf{z})) = E(g'(\mathbf{w}^T \mathbf{z}))\mathbf{I},$$

deci gradientul devine matrice diagonală și inversa poate fi calculată direct. Rezultă următoarea iterație Newton,

$$(21) \mathbf{w} \leftarrow \mathbf{w} - \frac{E(\mathbf{z}g(\mathbf{w}^T \mathbf{z})) + \beta\mathbf{w}}{E(g'(\mathbf{w}^T \mathbf{z})) + \beta}.$$

Regula de actualizare (21) poate fi simplificată în continuare prin înmulțirea cu $E(g'(\mathbf{w}^T \mathbf{z})) + \beta$ și rezultă,

$$(22) \mathbf{w} \leftarrow E(\mathbf{z}g(\mathbf{w}^T \mathbf{z})) - E(g'(\mathbf{w}^T \mathbf{z}))\mathbf{w}.$$

Algoritmul FastICA rezultă pe baza relației (22) și este descris astfel [6].

Date de intrare: $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$, vectorul datelor observate

Preprocesare:

1. Centrează datele observate
2. Calculează vectorul alb $\mathbf{z} = \mathbf{V}\mathbf{x}$,

$\mathbf{V} = \Lambda^{-\frac{1}{2}}\Phi^T$, unde $\Phi = (\phi_1, \phi_2, \dots, \phi_n)$ este matricea cu coloane vectorii proprii normați ai matricei de covarianță $\Sigma = E(\mathbf{x}\mathbf{x}^T)$ și $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$.

Pas 1. Selectează aleator vectorul unitar \mathbf{w}

Pas 2. Calculează

$$\mathbf{w} \leftarrow E(\mathbf{z}g(\mathbf{w}^T \mathbf{z})) - E(g'(\mathbf{w}^T \mathbf{z}))\mathbf{w}, \text{ cu } g \text{ defini-$$

tă prin (16), (17) sau (18).

$$\text{Pas 3. } \mathbf{w} \leftarrow \frac{\mathbf{w}}{\|\mathbf{w}\|}$$

Pas 4. Dacă algoritmul nu s-a stabilizat, reia **Pas 2 – Pas 3.**

2.3. Estimarea mai multor componente independente ale modelului ICA

În cadrul secțiunilor 2.1 și 2.2. au fost prezentați algoritmi pentru calculul unei singure componente independente $\mathbf{w}^T \mathbf{z}$. În continuare sunt descriși doi algoritmi pentru estimarea mai multor componente independente ale modelului ICA definit de (2). Posibilitatea calculului mai multor componente independente este asigurată de faptul că orice doi vectori $\mathbf{w}_i, \mathbf{w}_j, i \neq j$ corespunzători componentelor independente s_i, s_j sunt ortogonali și $E((\mathbf{w}_i^T \mathbf{z})(\mathbf{w}_j^T \mathbf{z})) = \mathbf{w}_i^T \mathbf{w}_j$.

Algoritmul de ortogonalizare prin deflatare

Ortogonalizarea prin deflatare utilizează metoda Gram-Schmidt, deci componentele independente sunt estimate secvențial. Pentru $p \geq 1$, dacă au fost calculați vectorii $\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_p$, atunci \mathbf{w}_{p+1} este calculat astfel:

1. aplică regula de actualizare a unui algoritm pentru estimarea unei singure componente independente;
2. extrage din \mathbf{w}_{p+1} proiecțiile vectorilor

$\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_p$ (vectorii proiecție sunt $(\mathbf{w}_{p+1}^T \mathbf{w}_j)\mathbf{w}_j, 1 \leq j \leq p$)

Rezultă următorul algoritm pentru estimarea secvențială a m componente independente ale modelului ICA, [6]

Date de intrare: $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$, vectorul datelor observate

Preprocesare:

1. Centrează datele observate
2. Calculează vectorul alb $\mathbf{z} = \mathbf{V}\mathbf{x}$,

$\mathbf{V} = \Lambda^{-\frac{1}{2}}\Phi^T$, unde $\Phi = (\phi_1, \phi_2, \dots, \phi_n)$ este matricea cu coloane vectorii proprii normați

ai matricei de covarianță $\Sigma = E(\mathbf{x}\mathbf{x}^T)$ și $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$.

Pas 1. Alege m , numărul componentelor independente de estimat; $p \leftarrow 1$

Pas 2. Inițializează \mathbf{w}_p (aleator)

Pas 3. Aplică regula de actualizare a unui algoritm pentru estimarea unei singure componente independente pentru calculul lui \mathbf{w}_p

Pas 4. Aplică regula de ortogonalizare

$$\mathbf{w}_p \leftarrow \mathbf{w}_p - \sum_{j=1}^{p-1} (\mathbf{w}_p^T \mathbf{w}_j) \mathbf{w}_j$$

Pas 5.
$$\mathbf{w}_p \leftarrow \frac{\mathbf{w}_p}{\|\mathbf{w}_p\|}$$

Pas 6. Dacă algoritmul pentru calculul vectorului \mathbf{w}_p nu s-a stabilizat, reia **Pas3** – **Pas 6**.

Pas 7. $p \leftarrow p + 1$; dacă $p \leq m$, reia **Pas2** – **Pas 7**.

Algoritmul de ortogonalizare simetrică

Fie m , numărul componentelor independente de estimat. Ortogonalizarea simetrică permite obținerea vectorilor $\mathbf{w}_p, 1 \leq p \leq m$, în paralel, astfel.

1. aplică regula de actualizare a unui algoritm pentru estimarea unei singure componente independente în paralel pentru fiecare $\mathbf{w}_p, 1 \leq p \leq m$;

2. ortogonalizează vectorii $\mathbf{w}_p, 1 \leq p \leq m$, prin aplicarea unei metode de ortogonalizare simetrică a matricei $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_m)^T$.

Ortogonalizarea simetrică a matricei $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_m)^T$ este realizată prin,

$$(23) \mathbf{w} \leftarrow (\mathbf{W}\mathbf{W}^T)^{-\frac{1}{2}} \mathbf{W}. \text{ În relația (23), matricea } (\mathbf{W}\mathbf{W}^T)^{-\frac{1}{2}} \text{ este obținută prin descompunerea } \mathbf{W}\mathbf{W}^T = \Phi' \Lambda' (\Phi')^T,$$

unde $\Lambda' = \text{diag}(\lambda'_1, \lambda'_2, \dots, \lambda'_n)$ este matricea diagonală a valorilor proprii ale matricei $\mathbf{W}\mathbf{W}^T$ și Φ' este matricea cu coloane vectorii proprii normați corespunzători $(\lambda'_1, \lambda'_2, \dots, \lambda'_n)$. Obținem,

$$(24) (\mathbf{W}\mathbf{W}^T)^{-\frac{1}{2}} = \Phi' \Lambda'^{-\frac{1}{2}} (\Phi')^T.$$

Bibliografie

- [1] Diamantaras, K.I., Kung S.Y., 1996, *Principal Component Neural Networks: theory and applications*, John Wiley & Sons
- [2] Devroye, L., Györfi, L., Lugosi, G., 1996, *A Probabilistic Theory of Pattern Recognition*, Springer Verlag
- [3] Haykin, S., 1999, *Neural Networks A Comprehensive Foundation*, Prentice Hall, Inc.
- [4] Haykin, S., Kosko, B., 2001, *Intelligent Signal Processing*, IEEE Press
- [5] Hastie, T., Tibshirani, R., Friedman, J., 2001, *The Elements of Statistical Learning*, Springer
- [6] Hyvarinen, A., Karhunen, J., Oja, E., 2001, *Independent Component Analysis*, John Wiley & Sons
- [7] Kung, S.Y., Diamantaras, K.I., 1990, A Neural Network Learning Algorithm for Adaptive Principal Component Extraction (APEX), In *IEEE Transaction on Acoustics, Speech and Signal Processing*
- [8] Oja, E., 1992, Principal Components, Minor Components and Linear Neural Networks, In *Neural Networks, vol. 5*
- [9] Schervish, M., 1995, *Theory of Statistics*, Springer