엔트로피 최대화를 이용한 새로운 밀도추정자의 설계

김웅명° 이현수
경희대학교 컴퓨터공학과 컴퓨터구조 및 뉴럴네트워크 연구실
{wmkim°, leehs}@khu.ac.kr

Design of New Density Estimator with Entropy Maximization

Woong Myung Kim° Hyon Soo Lee
Dept. of Computer Engineering, Kyung Hee University

요 약

본 연구에서는 엔트로피 이론을 사용하여 ICA(Independent Component Analysis) 점수함수를 생성하는 새로운 밀도추정자(Density Estimator)를 제안한다. 원 신호에 대한 밀도함수의 추정은 적당한 점수함수를 생성하기 위해 필요하고, 미분 가능한 밀도함수인 커널을 이용한 밀도추정법(Kernel Density Estimation)을 이용하여 점수함수를 생성하였다. 보다 빠른 점수함수의 생성을 위해서 식의 형태를 convolution 형태로 표현하였으며, ICA 학습을 위해서 결합엔트로피를 최대화(Joint Entropy Maximization)하는 방향으로 커널의 폭을 학습하였다. 이를 위해서 기울기 강하법(Gradient descent method)를 사용하였으며, 이러한 제약 사항은 새로운 밀도 추정자를 설계하기 위한 기본적인 개념을 나타낸다. 실험결과, 커널의 폭을 담당하는 smoothing parameters들이 일정한 값으로 학습함을 알 수 있었다.

1. 독립성분분석

ICA는 BSS(Blind Source Separation)문제를 풀기 위한 방법 중 하나이다. ICA는 수식형태에 대한 것을 살펴보면, 단지 통계적인 추정문제를 다루는 문제이다. ICA는 PCA를 이용한 decorrelation 조건 보다 좀 더 강력한 제약사항을 가진다[1]. 첫째, 독립요소(Independent Component)는 통계적으로 독립이라고 가정한다. 둘째, 독립요소는 반드시 nongaussian 분포를 가진다. 셋째, 혼합행렬(unknown mixing matrix)는 반드시 정방행렬(square matrix)이다. 이러한 제약 사항은 ICA를 특정한 조건에서 소스 신호에 대한 추정을 가능하게 한다[2][6].

여기서, s를 $m\times n$크기의 통계적인 독립성분(소스)이라고 가정하고, x는 식 1과 같이 원 신호 s들에 대한 선형 변환이 된다. 그리고 A는 혼합행렬(mixing matrix)이라고 가정한다. 이것은 $m\times m$ 사이즈의 비정칙(nonsingular) 행렬이 된다. 선형적인 혼합 신호들을 만들기 위한 식은 다음과 같이 표현된다.

$$x = As \tag{1}$$

x는 혼합되어진 신호이며, 다음과 같은 식 2에 의해 신호를 분리하는 모델로 나타낼 수 있다.

$$u = Wx \tag{2}$$

신호를 분리하는 행렬은 W로 표현하며, 분리행렬 $W = A^{-1}$와 같은 조건을 가질 때 분리된 신호 u를 추정할 수 있다. 여기에서 $f(u)$를 u의 결합 분포(joint density)에 대한 모델이라고 가정하는 경우 각 출력 신호에 대한 주변분포(marginal densities)를 계산하면 아래의 식 3과 같다[3][4].

$$f(u) = \prod_{j=1}^{m} f_j(u^{(j)}) \tag{3}$$

여기에서 $u^{(j)}$는 벡터 u의 j번째 신호이고 $f_j(u^{(j)})$는 신호 u에 대한 확률밀도함수이다. 측정된 신호 x에 대한 확률밀도함수는 식 4와 같이 표현 할 수 있다[2][3][4]. 여기에서 x에 대한 분포는 W행렬과 분리 신호 u의 확률밀도함수의 구성으로 나누어진다.

$$f(x) = |detW| f(Wx) \tag{4}$$

ICA에서 분리행렬 W값을 최적화 하는 학습 과정이기 때문에, 다음과 같이 로그우도함수(log-likelihood function)을 사용하면 식 5와 같은 비용함수를 만들 수 있다[4].

$$L(u, W) = -\sum_{j=}^{m} \log f(u_j)$$

$$= -n\log|detW| - \sum_{j=1}^{m}\sum_{i=1}^{n} \log f_j(y_i) \tag{5}$$

일반적으로 식 5는 $W = A^{-1}$의 조건을 가질 경우 전역해(global minimum)를 가진다. 또한 natural gradient 방법을 사용한 ICA 학습을 이용하면 전역해에 근사할 수 있다. 이러한 전역해에 근사하기 위해, 위 식을 W에 대해서 로그우도값을 최대로 하게 되면 식 6와 같이 나타낼 수 있다(최대 로그우도값을 구하기 위하여 W에 대해서 미분을 수행한다)[2][3][5][7].

$$W = W - \eta(E[\varphi(u)u^T] - I)W \tag{6}$$

여기서 η는 학습상수이며, E는 기대치이다. 대부분의 경우에 신호에 대한 샘플 평균으로 계산을 한다. I는 $m\times m$크기의 항등행렬을 나타낸다. 일반적으로 식 6에서 사용된 φ는 비선형 점수함수(nonlinear score function)라고 불리어지고 식 7과 같이 표현된다[2][4][7].

$$\varphi(u) = -[\log f(u)]' = \frac{f'(u)}{f(u)} \tag{7}$$

여기서 $f(u)$는 u에 대한 확률밀도함수이고 $f'(u)$는 밀도함수를 u 대해 미분하여 유도된 식이다. 본 연구는 ICA 학습시, 신호에 대한 추정오차를 줄이기 위해서, 원 신호를 추정하고 이를 바탕으로 결합엔트로피(Joint Entropy)를 최대화 하는 제약조건을 가진 점수함수를 생성한다. 즉, 결합엔트로피를 최대화하는 새로운 밀도추정자(Density Estimator)를 구성하는 것을 목적으로 한다.

2. SFG(Score Function Generation) 알고리즘

통계학적으로 확률밀도를 추정하는 방법 중 히스토그램이나 커널을 이용하여 밀도를 추정하는 방법이 있다. 이를 커널을 이용한 추정방법이라고 하고 일반적으로 KDE(Kernel Density Estimate)방법으로 부른다[8]. KDE를 이용한 추정방법은 미분이 가능하므로[9], 본 연구에서는 커널을 이용한 확률밀도 추정방법을 사용한다. 커널기반의 추정은 아래와 같이 정의되어진다.

$$\hat{f}(u)\tilde{=}f(c)=\frac{1}{nh}\sum_{i=1}^{n}K(\frac{c-u_i}{h}) \quad (8)$$

$$K(z)=\frac{1}{\sqrt{2\pi}}\exp^{-z^2/2} \quad (9)$$

식 7와 식8를 이용하여 c에 대해서 미분을 하게 되면 식 10과 같은 형태로 나타나게 된다. 여기에서 c는 가우시안 커널들의 중심점이고 일차원 벡터로 표현된다. l을 커널의 개수라고 할 때, $c=[c_1,c_2,....,c_l]$로 되어진다. 여기에서 u 대신에 c로 미분을 하는 이유는 KDE 출력결과가 c의 차원과 동일하기 때문이다. 또한 식 9에서 n과 π부분은 상수이기 때문에 생략할 수 있다. 따라서 미분을 하여 유도된 식은 아래와 같다.

$$\hat{\varphi}(u)\tilde{=}\varphi(c)=\frac{\frac{1}{h^2}\sum_{i=1}^{n}(c-u_i)\exp(-\frac{(c-u_i)}{2h^2})}{\sum_{i=1}^{n}(-\frac{(c-u)^2}{2h^2})} \quad (10)$$

위 식에서 n은 ICA서 신호 샘플의 수이다. 즉 u는 ICA 신호 출력으로부터 측정된 값이다. h는 일반적으로 양수이고, 각 커널의 넓이에 해당되는 부분이다. 일반적으로 bandwidth 혹은 smoothing parameter라고 한다. 커널을 이용한 확률추정 방법은 시간이 많이 소요되기 때문에, 위의 식을 컨볼류션 형태로 바꾼다.

식 10을 FFT 식을 이용한 형태로 나타내면 식 11과 12로 나타난다. 여기서 g를 가우시안 커널이라고 정의하고, g'은 g를 미분한 값이다. h는 측정되어진 히스토그램이라고 한다. F는 $f(u)$이고, H는 $f'(u)$이다.

$$F \equiv ifft(fft(h)*fft(g)) \quad (11)$$

$$F \equiv ifft(fft(h)*fft(g')) \quad (12)$$

FFT-convolution 기법을 사용한 점수함수는 계산 복잡도가 $O(n^2)$ 에서 $O(nlogn)$으로 줄어들게 된다. 따라서 점수함수를 빠른 시간 내에 생성하여 계산할 수 있다[9][10].

3. 엔트로피 최대화를 가진 SFG 알고리즘

Bell과 Sejnowski는 선형의 혼합행렬로부터 엔트로피 최대화를 단순한 전방향 신경망을 에 대한 ICA 학습을 제안했다. 신경망 출력에서 결합엔트로피는 다음과 같이 표현가능하다.

$$H(y_1,...y_m)=H(y_1)+.....+H(y_m)-I(y_1,.....,y_m) \quad (13)$$

여기서 $H(y_i)$는 ICA 출력에 대한 주변엔트로피이고, $I(y_1,...,y_m)$는 엔트로피에 대한 상호정보(mutual information)이다. 일반적으로 주변엔트로피의 합인 $H(y)$는 결합엔트로피이다. 주로 각각의 주변엔트로피는 $H(y_i)=-E[\log p(y_i)]$와 같이 정의된다. 출력 밀도 $p(y_i)$와 원 신호를 추정한 출력 신호의 밀도 $p(u_i)$ 사이의 관계는 다음과 같다.

$$p(y_i)=p(u_i)/|\frac{\partial y_i}{\partial u_i}| \quad (14)$$

결합엔트로피와 W행렬을 나타내면 아래와 같다.

$$\partial H(y)/\partial W=\partial(-I(y))/\partial W-\partial\sum_{i=1}^{m}E[\log p(y_i)]/\partial W \quad (15)$$

따라서 상호정보량 $I(y)$가 최소화 될 때, 결합 엔트로피 $H(y)$는 최대화 한다. 전방향 신경망의 가중치 W에 의해서 미분을 하게 되면 아래와 같은 관계를 가지게 된다.

$$\partial H(y)/\partial W=\partial(-E[\log|J|])/\partial W \quad (16)$$

여기서 J는 2차 미분을 나타내는 Jacobian 행렬이다.

$$p(y)=p(u)/J(u), J=|\partial\varphi(u)/\partial u| \quad (17)$$

본 연구에서는 점수함수를 추정하기 위해서 기울기 강하법을 사용하여 결합 엔트로피를 최대화하는 방법을 사용하고, 엔트로피 최대화를 위하여 smoothing parameters를 알맞게 조정한다. 따라서 h_k는 $[h_1,...,h_k]$와 같은 벡터로 정의한다. 즉, 기울기 강하법을 채용하여 h_k에 대해서 적당한 갱신양을 계산할 수 있다. 식 18과 같이 비선형 점수함수를 먼저 유도한 후

$$\tilde{\varphi}(u)\cong\frac{f(c)}{f(c)}=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_k}\exp(-\frac{1}{2h^2}(c-u_i)^2)$$

$$f(c)=(\partial f(c)/\partial c) \quad (18)$$

ICA 학습에서, 분리행렬 W와 smoothing parameter h는 엔트로피를 최대화하기 위해, 다음과 같이 표현된다.

$$\Delta(W,h)=\eta\frac{\partial H(y)}{\partial(W,h)}=\frac{\partial H(u)}{\partial(W,h)}+\frac{\partial E(\ln J)}{\partial(W,h)} \quad (19)$$

식 20은 h와 엔트로피 최대화의 관계를 나타낸 것이다.

$$\Delta h=\alpha\frac{\partial H(y)}{\partial h}=\frac{\partial H(u)}{\partial h}+\frac{\partial E(\ln J)}{\partial h} \quad (20)$$

2차 미분을 수행하면 반복적인 식으로 표현이 되어지는데, 이를 간략화 된 식으로 표현하면 다음과 같다.

$$K1=\exp(-\frac{1}{2h_k^2}(c-u_j)^2),\qquad K2=(c-u_j)\exp(-\frac{1}{2h_k^2}(c-u_j)^2)$$

$$K3=(c-u_j)^2\exp(-\frac{1}{2h_k^2}(c-u_j)^2),\quad K4=(c-u_j)^3\exp(-\frac{1}{2h_k^2}(c-u_j)^2)$$

$$K5=(c-u_j)^4\exp(-\frac{1}{2h_k^2}(c-u_j)^2) \quad (21)$$

비선형 점수함수를 생성하기 위해서, 식 20의 조건을 사용하여 c에 대해서 1차 미분을 하였다.

$$\partial(\varphi(u))/\partial u\cong\partial(\tilde{\varphi}(c))/\partial c$$

$$=\left(\left[\sum_{i=1}^{n}K1-\frac{1}{h_k^2}K3\right]\left[\sum_{i=1}^{n}K1\right]-\left[\sum_{i=1}^{n}K2\right]\left[\sum_{i=1}^{n}-\frac{1}{h_k^2}K2\right]\right)/\left[\sum_{i=1}^{n}K1\right]^2 \quad (22)$$

Jacobian 행렬에 대해서 h로 미분을 한 결과는 식 23과 같다.

$$\Delta h\cong\partial J/\partial h=(\frac{\partial(\tilde{\varphi}(c))}{\partial c})/\partial h$$

$$=\left((P1)/\left[\sum_{i=1}^{n}K1\right]^2\right)-\left(P2/\left[\sum_{i=1}^{n}K1\right]^4\right)-\left(P3/\left[\sum_{i=1}^{n}K1\right]^4\right) \quad (23)$$

$$P1=\left[\sum_{i=1}^{n}\frac{1}{h_k^3}K3-\frac{1}{h_k^3}K5\right]\left[\sum_{i=1}^{n}K1\right]-\left[\sum_{i=1}^{n}K1-\frac{1}{h_k^3}K3\right]\left[\sum_{i=1}^{n}\frac{1}{h_k^3}K3\right]$$

$$P2=\left(\left[\sum_{i=1}^{n}\frac{1}{h_k^3}K4\right]\left[\sum_{i=1}^{n}-\frac{1}{h_k^2}K2\right]+\left[\sum_{i=1}^{n}K2\right]\left[\sum_{i=1}^{n}\frac{-2}{h_k^3}K2-\frac{1}{h_k^5}K4\right]\right)\left[\sum_{i=1}^{n}K1\right]^2$$

$$B3=\left(\left[\sum_{i=1}^{n} K2\right]\left[\sum_{i=1}^{n}-\frac{1}{h_k^2}K2\right]\right)2[K1]\left[\sum_{i=1}^{n}\frac{1}{h_k^3}K3\right]$$

4. 시뮬레이션

SFG 알고리즘의 성능을 평가하기 위해서, 3개의 신호에 대해서 BSS 실험을 수행하였으며, super-gaussian 및 sub-gaussian 신호 각 1개와 왜도가 기울어진 분포를 가지는 신호 1개, 총 3개의 신호를 가지고 실험을 수행하였다. 전체 학습회수는 400회를 수행하였다. 그림 1은 100개의 커널수를 가지고, 엔트로피 최대화를 가진 SFG 알고리즘을 200번 학습하여 나타난 결과이다. 식 23을 이용하여 smoothing parameter에 대한 적당한 갱신량을 구할 수 있다. 학습회수 130번 이후에는 smoothing parameter의 값이 안정화되어 진다는 것을 알 수 있다. 따라서, 본 연구에서 제안된 SFG알고리즘이 성공적으로 학습을 함을 알 수 있었다.

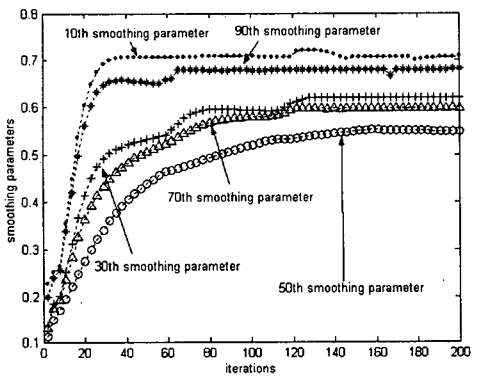

그림 1. 학습에 따른 smoothing parameters의 변화

그림 2는 SFG알고리즘과 기존의 Extended Infomax알고리즘, Fixed Point 알고리즘의 에러를 비교한 것이다. 식 24는 SNR(Signal to Noise Ratio)를 나타낸다.

$$SNR\,(dB) = 10\log_{10}\left(\frac{\sum_{j=1}^{m} s_j^2}{\sum_{j=1}^{m}(\hat{s}_j - s_j)^2}\right) \qquad (24)$$

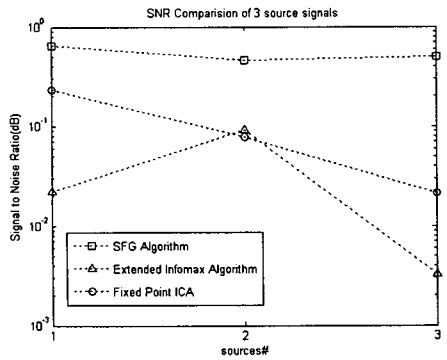

그림 2. SFG 알고리즘의 SNR

실험결과 측정된 SNR 값을 살펴보면, 기존의 고정된 점수함수로 학습했던 알고리즘 보다 SFG알고리즘이 보다 원 신호를 완벽하게 분리했음을 알 수 있었다.

5. 결론

본 연구에서는 최대우도방법을 이용하여, 고정된 점수함수를 사용하는 기존의 ICA의 문제점을 해결하기 위한 접근방법을 제시하였다. 이와 같은 문제점은 원 신호의 분포와 출력신호의 분포에서 에러를 포함한다는 것이다. 따라서 이를 줄이기 위하여 SFG 알고리즘을 제안하였으며, 학습시간을 줄이기 위하여 식의 형태를 convolution으로 변경하였다. 또한 엔트로피 최대화 방법을 이용하여 점수함수를 학습하였으며, smoothing parameters를 학습하기 위하여 기울기 강하법을 이용하였다. 실험결과, 제안되어진 SFG 알고리즘이 기존의 ICA 학습보다 분리 성능이 향상되어짐을 알 수 있었다.

참고문헌

[1] P. Comon, "Independent component analysis, a new concept?," *Signal Processing*, vol.36, no.3, pp.287-314, 1994.

[2] A. Hyvärinen, J. Karhunen, E. Oja, "Independent Component Analysis", *Wiley Interscience*, 2001

[3] A.J. Bell and T. Sejnowski, "An information maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol.7, no.6, pp.1129-1159, 1995.

[4] J.-F. Cardoso, "Blind signal separation: Statistical principles," *Proc. IEEE. Special Issue on Blind Identification and Estimation*, vol.9, pp.2009-2025, Oct. 1998.

[5] T.-W. Lee, M. Girolami, and T.J. Sejnowski, "Independent component analysis using an extended infomax algorithm for mixed sub-gaussian and super-gaussian sources," *Neural Computation.*, vol.11, no.2, pp.417-441, 1999.

[6] A. Hyvärinen, "Survey on independent component analysis," *Neural Computing Surveys*, vol.2, pp.94-128, 1999.

[7] S.-i. Amari, A. Cichocki, and H. H. Yang, "A new learning algorithm for blind source separation," *in Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press, vol.8, pp 757-763, 1996

[8] B. W. Silverman, "Density Estimation for Statistics and Data Analysis", New York: Chapman and Hall, 1985.

[9] N. Vlassis and Y. Motomura, "Efficient source adaptivity in independent component analysis," *IEEE Transaction on Neural Networks*, vol.12, pp.559-566, May. 2001.

[10] Woong-Myung Kim, Hyon Soo Lee, "An Efficient Score Function Generation Algorithm with Information Maximization", *Advanced in Natural Computation*, Lecture note on Computer Science, LNCS3610, part1, Aug. 2005.