PCA, SVD, ICA 정리

[2주차] 2022.01.27 예지혜

논문 1. Principal component analysis: a review and recent developments

Jolliffe IT, Cadima J. 2016. The Royal Society

PCA는 데이터의 차원을 줄이는 방법론으로, 분산을 최대화하며 상관관계가 없는 새로운 변수들을 만드는 방법이다. 주어진 데이터를 이용해 계산하고, 다양한 데이터 타입과 구조에 적용된다는 면에서 활용성이 좋은 방법론이다.

1. Introduction

PCA는 가장 오래되고 많이 사용되는 차원축소 방법론이다. 분산을 최대화하면서 서로 상관성이 없는 변수를 연속적으로 찾아내는 방법이며, 이는 곧 고유값 분해 문제로 직결된다. SVD 방법으로도 대체할 수 있으며 공분산 행렬이나 상관 행렬을 사용할 수 있다.

PCA를 inferential 목적으로 사용한다면 가우시안 분포 가정이 필요하지만, descriptive 목적으로 사용하면 분포 가정이 필요하지 않다. 주로 descriptive 목적으로 사용된다.

2. The basic method

(a) Principal component analysis as an exploratory tool for data analysis

X의 선형결합을 찾는 것이 목표이므로 Var(Xa) = a′Sa (S: sample covariance matrix)를 최대화하는 벡터 a를 찾는 것이 목표이다. 원활한 문제 해결을 위해 unit-norm vectors 조건 즉, a′a=1을 추가한다. 이를 라그랑주 승수법으로 해결하면 다음과 같은 문제로 바뀐다.

$$maximize\ a'Sa - \lambda(a'a - 1)$$

이를 미분하면 다음과 같다.

$$Sa - \lambda a = 0 \Leftrightarrow Sa = \lambda a$$

즉, a는 고유벡터, λ는 고유값인 고유값 분해 문제이다.

이때, 각 주성분이 설명하는 분산은 다음 식에 따라 고유값 λ가 된다.

$$var(Xa) = a'Sa = \lambda a'a = \lambda$$

주성분을 이해할 때, 부호는 급하면 그만이기 때문에 부호 패턴과 계수의 크기가 중요하다.

S와 같은 대칭행렬의 고유벡터들은 정규직교 (i.e. $a_k' a_{k'} = 1\ if\ k = k' and\ 0\ otherwise$) 하다는 특성을 가지고 있기 때문에 다음과 같이 uncorrelatedness를 보일 수 있다.

$$a_{k'}' S a_k = \lambda_k a_{k'}' a_k = 0\ if\ k' \neq k$$

정리하면, 이와 같이 얻은 Xa_k는 원 변수들의 선형결합인 주성분이고 PC scores라고 말하기도 한다. a_k는 PC loadings라 한다.

일반적으로 PCs를 구할 때 centred X (i.e. $x_{ij}^* = x_{ij} - \bar{x}_j$ /각 변수별 centring)를 사용하는데, 이는 결과는 유지하면서 기하학적으로 더 좋은 접근이 가능하다는 장점이 있다.

- SVD 사용 (r은 Y의 rank)

$$Y = ULA'$$

Columns of A (pxr): orthonormal, right singular vectors of Y, Y'Y의 고유벡터
Columns of U (nxr): orthonormal, left singular vectors of Y, YY'의 고유벡터
L (rxr): diagonal matrix, 각 대각원소의 루트는 Y'Y나 YY'의 고유값

이러한 Y에 centred matrix X*를 적용하면 행렬 A의 열이 구하고자 하는 고유벡터 즉, PC loadings가 된다.

PC loadings : A
PCs : X*A = ULA'A = UL
공분산행렬: (n-1)S = X*'X* = (ULA')'(ULA') = ALU'ULA' = AL^2A'

SVD의 특성을 이용하면 PCA의 기하학적 해석이 가능해진다. Original Y와의 차이가 가장 작은 Y_q는 L_q의 대각 원소 중 가장 큰 q개의 원소를 골라 구할 수 있다.

$$Y_q = U_q L_q A_q'$$

이를 산점도 위에서 생각해보면 n개의 데이터를 q차원 subspace에 가장 좋게(original과 거리가 가깝게) 근사한 것으로 볼 수 있다. 원 데이터를 시각화하고자 할 때 2~3개의 PC를 첫 시각화로 종종 사용한다. 또한, q를 늘려갈 때 기존의 best q subspace에 하나의 열을 추가하기만 한 다는 점이 중요하다. (increment nature)

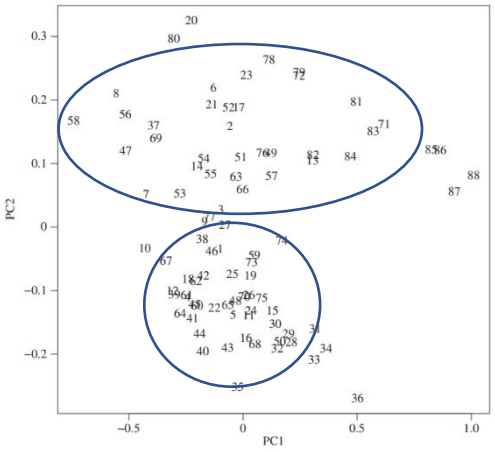
주어진 PC가 총 분산을 설명하는 정도는 다음과 같다. 분산 행렬의 trace가 해당 데이터의 총분산임을 이용한다.

$$\pi_j = \frac{\lambda_j}{\sum_{j=1}^{p} \lambda_j} = \frac{\lambda_j}{tr(S)}$$

또한 increment 성질을 이용하면 q개의 PC로 설명하는 분산은 각 PC가 설명하는 분산을 단순히 더해주면 된다. 일반적으로 70% 정도를 설명하는 q를 찾거나 기하학적으로 2~3개 정도의 PC를 확인해서 사용한다.

이외에도 Optimal subset을 찾기 위한 다양한 기준들이 연구되어왔다.

(b) Example: fossil teeth data (변수 9개)



R의 prcomp 패키지를 사용하여 78.8%와 16.7%를 설명하는 두 개의 PC를 구하였다. PC1은 계수의 부호가 동일하여 변수들의 가중치합으로 이해할 수 있으며 '전반적인 크기'라고 해석하였다. PC2는 길이에 대한 변수만 다른 부호이므로 height와 width 대비 length라고 보았다.

PC1과 PC2를 산점도로 표현하면 크게 두 그룹으로 나뉘며, 아래쪽의 조밀한 군집은 Kuehneotherium의 한 종으로, 위쪽의 넓은 군집은 확인되지 않은 다른 동물로 해석하였다.

(c) Some key issues

(i) Covariance and correlation matrix PCA

공분산행렬로 PCA를 진행하면, 이론적으로는 문제가 없으나 변수간 단위가 다를 때 문제가 생긴다. 이러한 문제를 피하기 위해 각 변수를 표준화시킨 것과 같은 효과를 내는 correlation matrix를 사용할 수 있다. Standardization is useful because most changes of scale are linear transformations of the data, which share the same set of standardized data values.

상관행렬을 이용하면 각 PC가 설명하는 분산도 달라지며, 일반적으로 동일한 정도의 분산을 설명하기 위해 더 많은 PC를 이용해야 하는 것은 상관행렬이다.

상관행렬을 이용했을 때 변수 j와 PC k의 상관계수는 다음과 같다.

$$r_{var_j, PC_k} = \sqrt{\lambda_k} a_{jk}$$

따라서, $a_k' a = 1$ 대신 $\tilde{a}_k' \tilde{a} = \lambda_k$ 정규화를 사용하게 되면 j번째 변수와 k번째 PC의 상관계수가 선형결합의 계수와 같아진다. 이를 위해 일부러 다른 정규화를 사용하기도 한다.

(ii) Biplots

$$X_q^* = GH', where\ G = U_q, H = A_q L_q$$

n rows g_i of matrix G define graphical markers for each individual

p rows h_j of matrix H define markers for each variable

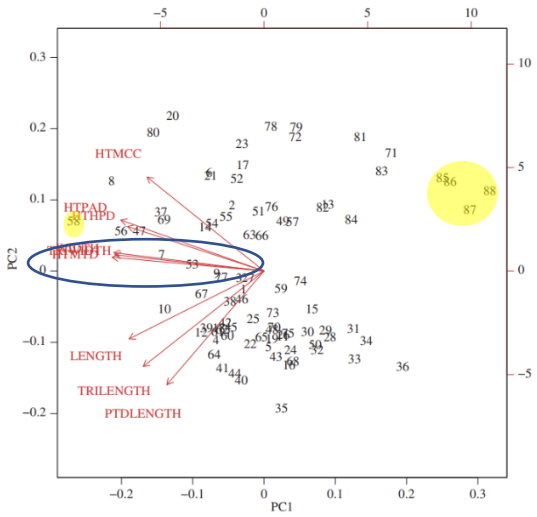
q=p일 때,

어떠한 두 벡터 간의 코사인 값은 그 변수들의 상관계수를 의미한다. (내적 값은 공분산에 비례)

PC 축과 벡터 간의 코사인 값은 PC 변수와 그 변수와의 상관계수를 의미한다.

individual i와 변수 j의 내적 값은 centred value of individual I on variable j

individual i와 변수 j간의 유클리디안 거리는 마할라노비스 거리에 비례한다.



PC1 축에 거의 평행하고 붙어있는 WIDTH, HTMDT, TRIWIDTH 세 변수는 서로 상관관계가 매우 높고, PC1과도 상관성이 매우 높다고 볼 수 있다.

58번 관측치는 모든 변수와 양의 방향으로 projection 되므로 큰 치아라고 볼 수 있고, 85~88 관측치는 그 반대이므로 작은 치아라고 볼 수 있다.

(iii) Centrings

어떤 데이터의 경우에는 centring이 부적절한 경우가 있어 전처리를 피하고 uncentered data를 사용하거나, non-centred second moments T의 행렬을 분해하기도 한다. 이러한 방법을 uncentred PCA라고 부르며, 단순히 centring 되지 않은 데이터를 사용하는 것과 혼동하지 않도록 주의해야 한다. 이는 non-central second moments를 연쇄적으로 최대화하는 방법이다. $\bar{x}$가 0에 근사한 경우를 제외하고는 일반적인 PCA와의 관계를 이해하기 어려운데, 어떤 연구에서는 $\bar{x}$가 매우 큰 경우에 일반적인 PCA와 생각보다 유사함을 보이기도 하였다.

어떠한 경우에는 row centring이나 열과 행을 모두 centring 하는 방법이 유용하기도 하다.

(iv) when n<p

n이 p보다 작을 땐, n이 rank를 결정하며, n개의 변수가 모든 변동을 설명한다. PC를 구하는 것 자체에는 문제가 없지만 일부 프로그램에서는 잘 돌아가지 않는다.

최근엔 모집단과 표본집단을 통한 연구가 이루어졌으나, 모집단의 PC와 표본집단의 PC는 별로 유사하지 않았다. 주로 n이 매우 작은 구조화된 데이터나 노이즈가 많은 데이터로 연구되었다.

3. Adaptations of principal component analysis

(a) Functional principal component analysis
한 관측치에 대해 여러 시점이 기록되는 데이터

(b) Simplified principal components
해석력을 높이고자 분산을 살짝 포기 – rotation, adding a constraint 두 가지 방법이 있다.

(c) Robust principal component analysis
PCA는 이상치에 민감하기 때문에 이를 해결하고자 정의한 방법. 특히 이미지, 머신러닝, 바이오, 웹 데이터와 같이 매우 큰 데이터를 다룬다.

(d) Symbolic data principal component analysis
- Interval data: 어떠한 관측치나 측정치들의 정확도를 믿지 못할 때, 또한 그 관측치들이 개별적이기보단 그룹으로 묶어질 때, 해당 그룹에 대한 값의 범위를 모델링한다. PCA를 적용하면, PC 또한 interval type으로 찾아진다.
- Histogram data: interval data의 일반화라고 볼 수 있으며, 히스토그램 간의 거리나 히스토그램의 합 또는 평균에 관심이 있다.

4. Conclusion

PCA 자체도 매우 많이 사용되었지만, 다양한 adaptation이 연구되고 사용되었기 때문에 데이터 특성에 맞는 PCA를 보는 것도 좋겠다.

논문 2. Unsupervised Feature Extraction Using Singular Value Decomposition

Kourosh Modarresi. 2015. ICCS

1. Introduction

이 논문에서는 데이터 행렬의 크기를 m x n으로 정의하며, m은 관측치의 개수, n은 변수 개수이다.

2. Feature Extraction and Dimensional Reduction for Modern Data

현대의 데이터는 사이즈가 방대하기도 하지만 고차원, 희소함, 가우시안 분포를 따르지 않는 것, 높은 상관성, 구조화되지 않은 형태, high frequency 등 다양한 특성을 가지고 있다. 이러한 데이터를 다루는 데에 분석하고 모델링하는 어려움도 있고, 계산 비용이 높다는 단점도 있다.

또한 차원의 저주는 다음과 같이 분석을 어렵게 한다.
1. 우연히 패턴을 발견
2. 과적합
3. Random and noisy but false model validation

현대의 데이터는 고차원이고 매우 희소한 경우가 많아 거리 개념을 정의하기 어렵다. 대부분의 거리가 매우 크기 때문에 similarities/dissimilarities를 정의할 때 대부분 매우 유사하지 않다고 결론 내려진다.
또한 정규분포 가정을 하기 어려워 해당 가정을 기반으로 하는 모델을 적용할 수 없다는 단점이 존재한다.

분석에 유리한 조건들도 있는데 다음과 같다. (이 조건이 왜 유리한 조건들일까)
1. Concentration of measure
2. Existence of structure
3. Massive size
4. High correlation
5. Rand deficiency of the data matrix X i.e., rank(X) << min(m,n)

2.1 Feature Selection as a Means for Data Analysis

현대에는 데이터를 얻는 것이 매우 쉬워졌지만 그 크기가 큰 만큼 어떤 부분이 중요한지 파악하는 것이 매우 중요하다. 예를 들어, 온라인 타겟팅이나 캠페인에서 고객의 어떤 특성이 중요한지 추출해낼 수 있다.

머신러닝에는 많은 feature selection 방법이 있지만 크게 지도/비지도 학습으로 나눌 수 있다. 또

다른 관점으로는 원 변수의 선형 결합으로 새로운 변수를 찾는 feature selection과 원 변수 중 몇 개 만을 선택하는 feature extraction으로 나눌 수 있다. Feature selection은 해석이 쉽지 않으나 feature extraction은 원 변수와 직결된다는 면에서 장점이 있다.

일반적으로 차원 축소를 할 때 기준이 되는 것은 데이터의 상관성과 변동을 잘 나타내야 한다는 것이다. 종종 이 둘은 트레이드오프 관계이기도 하다. 어쨌든 데이터의 분산과 상관계수를 최적화 함수로 하는 projection이나 approximation이 주요 목적이다.

3. Description of the Feature Extraction Model Using SVD

3.1 Singular Value Decomposition (SVD)

$$X = UDV^t$$

U, V는 각각 left/right singular vectors 며 m x n, n x n 직교행렬, D = diag(d_1, d_2, … , d_n)

원데이터의 변동을 80~90% 정도 설명하는 임계값을 사용하면 새로운 변수는 원 변수 모두를 사용한 선형결합 형태가 된다. SVD를 feature extraction으로 바로 사용할 수는 없고, 각 PC에 대해 rank 제약조건을 부여하여 0이 되지 않은 변수만을 사용하게 할 수 있다. 알고리즘은 다음과 같다.

For a centered X ;

- Step (1.1) compute $\underbrace{min}_{U_q,V_q,D_q} \|X - U_q D_q V_q\|$ to obtain U_q , D_q and V_q

For q= numerical rank of the matrix.

- Step(1.2) compute the rank –q of X;

$$X_q = U_q D_q V_q$$

using newly computed X_q, we have new values for the missing entries.

Step (1.3) Iterate steps (1.1) and (1.2) till convergence ;

$$\|X_{q(i+1)} - X_{q(i)}\| / \|X_{q(i)}\| \leq \delta$$

for small δ.

Step 1. Centred X에 대해 q차원의 SVD 분해를 진행한다.
(왜 iterative한 과정이 되어야 하는지..)

Step (2) Computing rank constrained SVD;

- Using Rank-1 approximation to our data matrix X;

$\text{argmin}_{(u,v,\sigma)} \|X - \sigma uv^t\|_2^2 \;\; s.t. \quad \|v\|_2^2=\|u\|_2^2=1$

With the rank constraints;

$$min \; \|v\|_0 \text{ and } min \; \|u\|_0$$

Step 2. Rank 제약 조건을 반영한다. 이때, norm-zero computation은 NP hard problem 이므로 second norm 제약 조건을 이용한다. 그러면 다음과 같은 최적화 문제가 된다.

$$\text{argmin}_{(u,v,\sigma)} \|X - Xvu^t\|_2^2 \;\; s.t. \;\; \|v\|_2^2=\|u\|_2^2=1 \text{ also } \|v\|_0 \leq \delta \text{ and } \|u\|_0 \leq \eta$$

4. Results

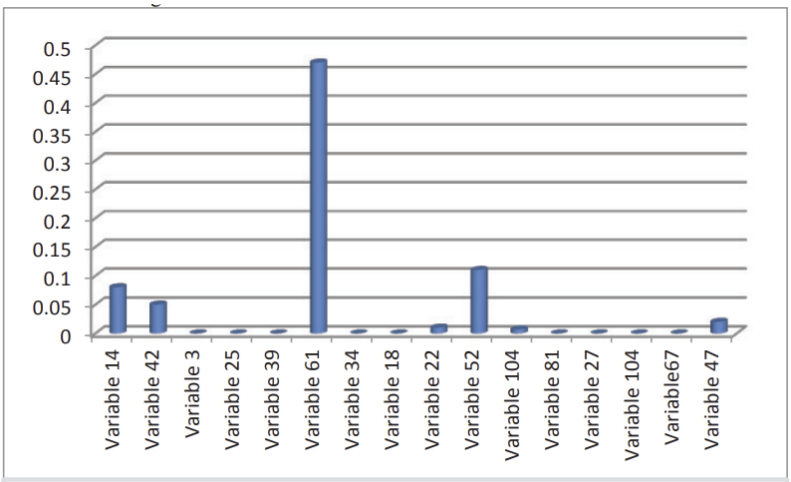


Figure 1. the most significant features with their significance for the data matrix in Example1.

실제 데이터에 적용해본 것으로 12개의 각 변수의 데이터에 대한 significance를 보여주는 plot이다. (어떻게 계산했는지 잘 모르겠음)

We see the difference between k-rank svd for both cases in terms of the relative error based of Euclidean difference between the new coordinates in both cases.

3. Independent Component Analysis

14.7 Independent Component Analysis and Exploratory Projection Pursuit

다변량 데이터에는 지능, 정서적 능력, 뇌 활동 등 직접 측정할 수 없는 것들을 간접적으로 측정한 데이터들이 많다. 이러한 잠재된 특성을 분석하기 위해 전통적으로 연구되어 온 것이 인자분석이다. 인자분석은 정규분포 가정을 필요로 하는데, 비정규성을 기반으로 하는 ICA가 새로운 강자로 떠올랐다.

14.7.1 Latent Variables and Factor Analysis

SVD에서 X=UDV'일 때 $S=\sqrt{N}U$(무엇을 의미?), $A^T=DV^T/\sqrt{N}$ 이라 두면 $X=SA^T$이므로 X는 S의 컬럼들의 선형결합으로 볼 수 있다. 따라서 SVD와 PCA는 잠재 변수 모델이라고 볼 수 있다.

$$\begin{aligned} X &= \mathbf{A}S \\ &= \mathbf{A}\mathbf{R}^T\mathbf{R}S \\ &= \mathbf{A}^*S^*, \qquad \mathrm{Cov}(S^*) = \mathbf{R}\,\mathrm{Cov}(S)\,\mathbf{R}^T = \mathbf{I} \end{aligned}$$

이러한 조건을 통해 unique한 A를 찾는다.

인자분석은 심리학자들에 의해 주로 연구되었으며, q<p에 대해 다음과 같이 표현할 수 있다.

$$\begin{aligned} X_1 &= a_{11}S_1 + \cdots + a_{1q}S_q + \varepsilon_1 \\ X_2 &= a_{21}S_1 + \cdots + a_{2q}S_q + \varepsilon_2 \\ &\vdots \qquad\qquad \vdots \\ X_p &= a_{p1}S_1 + \cdots + a_{pq}S_q + \varepsilon_p, \end{aligned}$$

이때 S는 잠재 변수 또는 요인이라고 할 수 있으며, A는 factor loadings, 입실론은 uncorrelated zero-mean disturbances으로, 설명되지 않은 남은 변동이다. 주로 S와 입실론은 정규분포로 가정되며 모델 적합은 최대우도 방식으로 이루어진다.

Identifiability issue와 rotated version이 더 좋은 해석력을 가질 때가 있다는 한계점이 있어 많이 사용되지는 않았다.

14.7.2 Independent Component Analysis

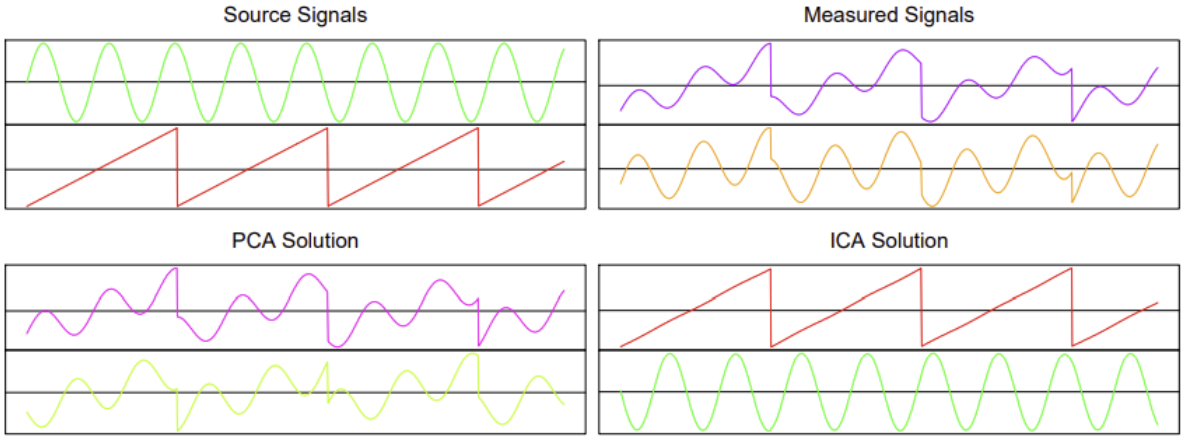
ICA는 식 자체는 PCA의 X=AS와 동일한데, S가 무상관이 아닌 독립인 경우이다.

Intuitively, lack of correlation determines the second-degree cross-moments (covariances) of a multivariate distribution, while in general statistical independence determines all of the cross-

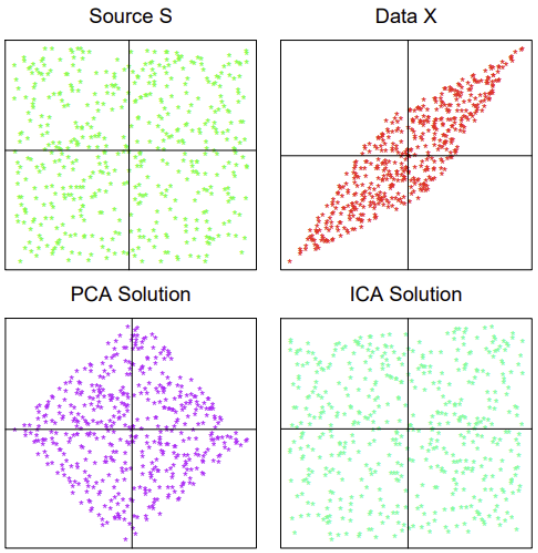
moments. These extra moment conditions allow us to identify the elements of A uniquely. Since the multivariate Gaussian distribution is determined by its second moments alone, it is the exception, and any Gaussian independent components can be determined only up to a rotation, as before.

S가 독립이고 비정규성을 만족하면 앞서 직면했던 identifiability 문제는 피할 수 있다.

결국 ICA 문제는 S가 독립이고 비정규성을 만족하도록 하는 직교행렬 A를 찾는 SVD 문제이다.



이 예시를 보면 ICA가 두 개의 신호를 잘 분리해낸 것을 확인할 수 있다.



Source: 500 realizations from the two independent uniform sources

Data X: their mixed versions

=> PCA, ICA 해석

PCA는 분산이 가장 큰 쪽을 투영하여 데이터를 추출해냈지만, ICA는 서로 다른 두 데이터를 잘 분리하여 원래 Source와 유사한 형태로 분리해냈다.

- entropy를 이용한 접근

Differential entropy H: $H(Y) = -\int g(y) \log g(y) dy$
모든 랜덤변수 중 가우시안 분포가 가장 큰 엔트로피를 가지고 있다.

Mutual information I: $I(Y) = \sum_{j=1}^{p} H(Y_j) - H(Y)$

이러한 I(Y)는 Y의 분포 g(y)와 그의 독립 버전 $\prod_{j=1}^{p} g_j(y_j)$ 간의 Kullback-Leibler 거리이다.

여기서 X가 공분산행렬로 I를 가지고, Y=A'X, A가 직교행렬이면 I(Y)를 최소화하는 A를 찾는 것이 가장 독립인 경우이다.

- negentropy를 이용한 접근

편의성을 위해 negentropy $J(Y_j) = H(Z_j) - H(Y_j)$ 를 이용한 연구도 있다. Negentropy는 음수가 아니며, Y_j가 가우시안 분포로부터 얼마나 먼 지, 즉 얼마나 비정규성을 따르는지 측정하는 척도이다.

예시 – 숫자 손글씨 데이터

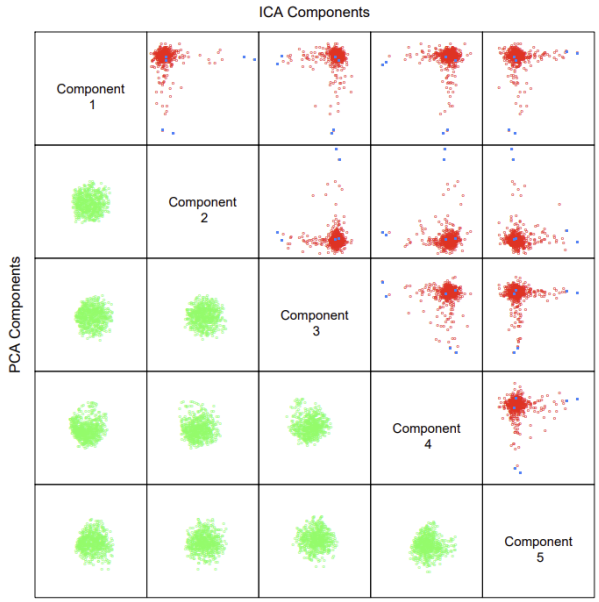
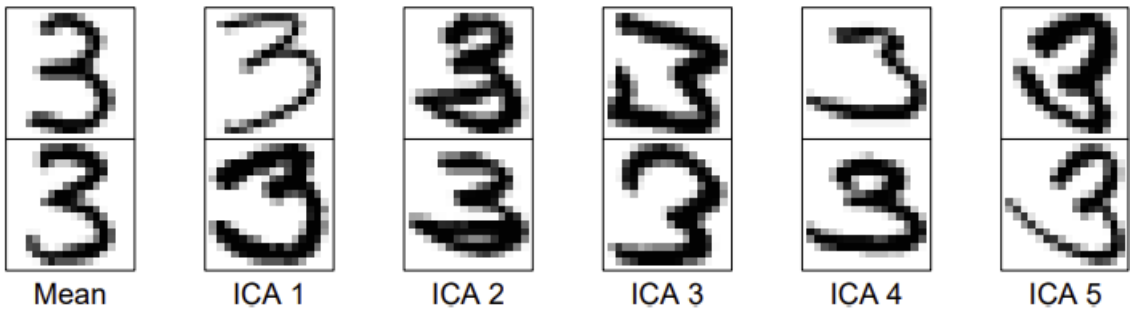


FIGURE 14.39. *A comparison of the first five ICA components computed using* `FastICA` *(above diagonal) with the first five PCA components(below diagonal). Each component is standardized to have unit variance.*

이 데이터의 첫번째 5개 PCA, ICA component를 시각화한 결과이다.

PCA는 결합정규분포 형태를 보이고, ICA는 긴 꼬리를 가진 분포이다.

PCA는 분산에, ICA는 비정규성에 초점을 두었기 때문이라고 이해할 수 있다.



이는 ICA component를 시각화 한 것으로 각 요소들이 설명하는 부분을 확인할 수 있다.

ICA는 뇌 과학 연구에도 많이 쓰이는데, 두피의 각 전극으로부터 기록되는 데이터들이 독립적으로 결합하여 잠재 정보를 가지고 있다고 간주한다.

14.7.3 Exploratory Projection Pursuit

이 방법은 고차원의 데이터를 저차원으로 투영하면 정규분포처럼 보일 것이라는 관점에서 시작되었다. 시작은 다르지만, 이 방법과 ICA는 비슷한 점이 많다.

14.7.4 A Direct Approach to ICA

IC는 다음과 같은 joint product density를 가지고 있다.

$$f_S(s) = \prod_{j=1}^{p} f_j(s_j)$$

In the spirit of representing departures from Gaussianity,

$$f_j(s_j) = \phi(s_j)e^{g_j(s_j)} : a\ tilted\ Gaussian\ density$$

이때 X의 log-likelihood는 다음과 같으며, 이를 최대화하는 것이 목적이다. (A 직교행렬 제약조건)

$$\ell(\mathbf{A}, \{g_j\}_1^p; \mathbf{X}) = \sum_{i=1}^{N}\sum_{j=1}^{p} \left[\log \phi_j(a_j^T x_i) + g_j(a_j^T x_i)\right],$$

이 모델은 over-parametrized이므로 다른 버전의 식을 사용한다.

$$\sum_{j=1}^{p}\left[\frac{1}{N}\sum_{i=1}^{N}\left[\log \phi(a_j^T x_i) + g_j(a_j^T x_i)\right] - \int \phi(t)e^{g_j(t)}dt - \lambda_j \int \{g_j'''(t)\}^2(t)dt\right]$$

이를 구하려는 Product Density ICA 알고리즘을 설명하고 있으나..