# Review of Independent Component Analysis

Element of Statistical Learning Book, Chapter 14

Hanchao Zhang

November 9, 2021

## 1 Latent Variables an Factor Analysis

The singular decomposition

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T \tag{1}$$

We can write $\boldsymbol{S} = \sqrt{N}\boldsymbol{U}$ and $\boldsymbol{A}^T = \frac{\boldsymbol{D}\boldsymbol{V}^T}{\sqrt{N}}$
and we have

$$\boldsymbol{X} = \boldsymbol{S}\boldsymbol{A}^T = \sqrt{N}\boldsymbol{U}\frac{\boldsymbol{D}\boldsymbol{V}^T}{\sqrt{N}} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T \tag{2}$$

Where $\boldsymbol{S}$ and $\boldsymbol{X}$ have mean 0, and $\boldsymbol{U}$ is an orthogonal matrix. We can interpret the SVD or the corresponding principal component analysis as an estimate of a latent variable model.

$$\begin{aligned} X_1 &= a_{11}S_1 + a_{12}S_2 + \cdots + a_{1p}S_p \\ X_2 &= a_{21}S_1 + a_{22}S_2 + \cdots + a_{2p}S_p \\ &\vdots \qquad\qquad\qquad \vdots \\ X_p &= a_{p1}S_1 + a_{p2}S_2 + \cdots + a_{pp}S_p \end{aligned} \tag{3}$$

However, for any orthogonal matrix $\boldsymbol{R}$, we can write

$$\begin{aligned} X &= \boldsymbol{A}S \\ &= \boldsymbol{A}\boldsymbol{R}^T\boldsymbol{R}S \\ &= \boldsymbol{A}^*S^* \end{aligned} \tag{4}$$

Hence there are many such decompositions and it is therefore impossible to identify any particular latent variable as unique underlying sources. The classical factor analysis model has the form ($q < p$)

$$\begin{aligned} X_1 &= a_{11}S_1 + a_{12}S_2 + \cdots + a_{1q}S_q + \varepsilon_1 \\ X_2 &= a_{21}S_1 + a_{22}S_2 + \cdots + a_{2q}S_q + \varepsilon_2 \\ &\vdots \qquad\qquad\qquad \vdots \\ X_p &= a_{p1}S_1 + a_{p2}S_2 + \cdots + a_{pq}S_q + \varepsilon_p \end{aligned} \tag{5}$$

or

$$X = \boldsymbol{A}S + \varepsilon \tag{6}$$

Typically the $S_j$ and $\varepsilon_j$ are modeled as Gaussian random variables and the model is fit by maximum likelihood.

## 2 Independent Component Analysis

The ICA model has the form:

$$\begin{aligned} X_1 &= a_{11}S_1 + a_{12}S_2 + \cdots + a_{1p}S_p \\ X_2 &= a_{21}S_1 + a_{22}S_2 + \cdots + a_{2p}S_p \\ &\vdots \qquad\qquad\qquad \vdots \\ X_p &= a_{p1}S_1 + a_{p2}S_2 + \cdots + a_{pp}S_p \end{aligned} \tag{7}$$

or

$$X = \boldsymbol{A}S \tag{8}$$

where the $S_i$ are assumed to be statistically independent rather than uncorrelated. Intuitively, uncorrelated determines the second cross moment Cov(X), and statistically independent determines all cross moments.
We wish to recover the matrix $\boldsymbol{A}$ in $X = \boldsymbol{A}S$. Without loss of generality, we can assume that $X$ has already been whitened to have $\mathrm{Cov}(X) = \mathbf{I}$, which implies that $\boldsymbol{A}$ is orthogonal. Solving the ICA problem amounts to findg and orthogonal $\boldsymbol{A}$ such that the components of the vector random variables $S = \boldsymbol{A}^T X$ are independent and non-Gaussian.
Many of the popular approaches to ICA are based on entropy. The differential entropy $H$ of a random variable $Y$ with density $g(y)$ is given by

$$H(Y) = E[-\log g(Y)] = -\int g(y)\log g(y)dy \tag{9}$$

The quantity $I(Y)$ is called the *Kullback-Leibler* distance or *mutual information*

$$I(Y) = \sum_{j=1}^{p} H(Y_j) - H(Y) \tag{10}$$

This is the measurement of *Kullback-Leibler* distance betweenthe density $g(y)$ of $Y$ and its independence version $\prod_{j=1}^{p} g_j(y_j)$, where $g_j(y_j)$ is the marginal density of $Y_j$. Now, if $X$ has covariance $\mathbf{I}$, and $Y = \boldsymbol{A}^T X$ with $\boldsymbol{A}$ orthogonal, then we will have

$$
\begin{aligned}
I(Y) &= \sum_{j=1}^{p} H(Y_j) - H(Y) && (11)\\
&= \sum_{j=1}^{p} H(Y_j) - H(X) - \log|\det \boldsymbol{A}| && (12)\\
&= \sum_{j=1}^{p} H(Y_j) - H(X) && (13)
\end{aligned}
$$

since we have $Y \sim f_Y(y)$, and $Y = \boldsymbol{A}^T X$, where we can get $X = \boldsymbol{A}Y$, and the $PDF$ of $X$ is $f_X(x) = f_Y(x)||\boldsymbol{A}||$

$$
\begin{aligned}
H(X) = E[-\log f_X(x)] &= -E\Big[\log\big(||\boldsymbol{A}|| \cdot f_Y(x)\big)\Big] && (14)\\
&= -E\Big[\log||A|| + \log f_Y(x)\Big] && (15)\\
&= -\log||A|| - E[-\log f_Y(x)] && (16)\\
&= -\log||A|| + H[Y] && (17)\\
&= H[Y] && (18)
\end{aligned}
$$

For convenience, rather than using the entropy $H(Y_j)$, Hyvarinen and Oja (2000) ues the negentropy measure $J(Y_j)$ defined by

**Note**

a

$$J(Y_j) = H(Z_j) - H(Y_j) \tag{19}$$

where $Z_j$ is a Gaussian random variable with the same variance as $Y_j$. They proposed simple approximations to negentropy which can be computed and optimized on the data.

$$J(Y_j) \approx \Big(E\big[G(Y_j)\big] - E\big[G(Z_j)\big]\Big)^2 \tag{20}$$

where $G(u) = \frac{1}{a}\log\cosh(au)$ for $1 < a < 2$.
With pre-whitened data, this amounts to looking for components that are as independent as possible.

# 3 Direct Approach of Independent Component Analysis by a Joint Product Density

Independent component have by definition a joint product density

$$f_S(s) = \prod_{j=1}^{p} f_j(s_j) \tag{21}$$

And in the spirit of representing departures from Gaussianity, we represent each $f_j$ as

$$f_j(s_j) = \phi(s_j)\exp\{g_j(s_j)\} = \frac{1}{\sqrt{2\pi}}\exp\{\frac{1}{2}s_j^2\}\cdot\exp\{g_j(s_j)\} \tag{22}$$

the log-likelihood for the observed data $X = \boldsymbol{A}S$ is

$$\ell(\boldsymbol{A}\{g_j\}_{j=1}^{p}; \boldsymbol{X}) = \sum_{i=1}^{N}\sum_{j=1}^{p}[\log\phi_j(a_j^T x_i) + g_j(a_j^T x_i)] \tag{23}$$

which we want to maximize subjected to $\boldsymbol{A}$ orthogonal and $g_j$ result in density function. So, we instead maximize a regularized version

$$\sum_{j=1}^{p}\Big[\sum_{i=1}^{N}[\log\phi_j(a_j^T x_i) + g_j(a_j^T x_i)] - \underbrace{\int \phi(t)e^{g_j(t)}dt}_{\text{density}} - \underbrace{\lambda_j\int\{g_j'''(t)\}^2(t)dt}_{\text{splines penalty}}\Big] \tag{24}$$

The first integral that controls for density is problematic and requires an approximation. We construct a fine grid of $L$ values $s_\ell^*$ in increments $\Delta$ covering the observed value $s_i$ and count the number of $s_i$ in the resulting bins:

$$y_\ell^* = \frac{\#s_i \in (s_i^* - \frac{\Delta}{2}, s_i^* + \frac{\Delta}{2})}{N} \tag{25}$$

Step 2 (a), we can then approximate the penalized likelihood by

$$\sum_{\ell=1}^{L}\Big\{y_i^*[\log(\phi(s_\ell^*)) + g(s_\ell^*)] - \Delta\phi(s_\ell^*)e^{g(s_\ell^*)}\Big\} - \lambda\int g'''^2(s)ds \tag{26}$$

and in practice, we set all $\lambda_j$ to the same.

Step 2(b), we optimize the $\boldsymbol{A}$ with respect to the penalized likelihood function. Only the first terms in the sum involve $\boldsymbol{A}$, and since $\boldsymbol{A}$ is orthogonal, $\phi$ do not depend on $\boldsymbol{A}$. Hence, we need to maximize

$$C(\boldsymbol{A}) = \frac{1}{N}\sum_{j=1}^{p}\sum_{i=1}^{N}\hat{g}_j(a_j^T x_i) = \sum_{j=1}^{p}C_j(a_j) \tag{27}$$

1. for each $j$ update

$$a_j \longleftarrow E\Big\{X\hat{g}_j'(a_j^T X) - E[g_j''(a_j^T X)]a_j\Big\} \tag{28}$$

2. Orthogonalize $\boldsymbol{A}$ using the symmetric square-root transformation $(\boldsymbol{A}\boldsymbol{A}^T)^{\frac{1}{2}}\boldsymbol{A}$. Let $UDV^T$ be the singular decomposition of $\boldsymbol{A}$, we will have

$$(\boldsymbol{A}\boldsymbol{A}^T)^{\frac{1}{2}}\boldsymbol{A} = (UDV^TVD^TU^T)^{\frac{1}{2}}UDV^T \tag{29}$$

$$= (UD^2U^T)^{\frac{1}{2}}UDV^T \tag{30}$$

$$= D^{-1}UDV^T \tag{31}$$

$$\boldsymbol{A} \leftarrow UV^T \tag{32}$$