Chapter 3: Independent Component Analysis[a]

As in the chapter on factor analysis (Chapter 2) we assume that the observations $\{x_i^0\}_{i=1}^n$ are n realizations of an $\mathbb{R}^p$-valued random variable X^0, and we let $X = X^0 - \mathbb{E}[X]$.

Then, the independent component analysis (ICA) model assumes that

$$X = \boldsymbol{B}S, \quad \boldsymbol{B} \in \mathbb{R}^{p\times p}, \quad \mathbb{E}[S] = 0, \quad \mathrm{Var}(S) = \boldsymbol{I}_p \tag{3.1}$$

where the p components of the $\mathbb{R}^p$-valued random variable S are independent random variables and where $\boldsymbol{B}$ is invertible.

Remark: The ICA model holds true if X is Gaussian and if $\mathrm{Var}(X)$ is full rank, in which case $\boldsymbol{B}$ and S can be obtained using population PCA[b].

Unlike the factor analysis model, where the factor F has typically no physical existence, in ICA the variable S is a real signal that we want to recover from X. Remark that this is the reason why the matrix $\boldsymbol{B}$ is assumed to be invertible.

A typical (toy) application of ICA is to recover the $p = 2$ signals S_1 and S_2 emitted by two persons speaking simultaneously in a room from the signals X_1 and X_2 recorded by two microphones.

[a] The main reference for this chapter is [**?**].

[b] To see this recall that $X = \boldsymbol{\Gamma} Y$ where $\boldsymbol{\Gamma} \in O(p)$ and where $Y \sim \mathcal{N}_p(0, \boldsymbol{L})$, with $\boldsymbol{L}$ a diagonal matrix having the eigenvalues of $\mathrm{Var}(X)$ as non-zero entries (see Chapter 1, page 24). If $\mathrm{Var}(X)$ is full rank the matrix $\boldsymbol{L}$ is invertible, and thus $X = (\boldsymbol{\Gamma}\boldsymbol{L}^{1/2})(\boldsymbol{L}^{-1/2}Y)$. This shows that the ICA model holds with $\boldsymbol{B} = \boldsymbol{\Gamma}\boldsymbol{L}^{1/2}$ and with $S = \boldsymbol{L}^{-1/2}Y \sim \mathcal{N}_p(0, \boldsymbol{I}_p)$ (recall that $S \sim \mathcal{N}_p(0, \boldsymbol{I}_p)$ implies that $S_1, \dots, S_p \overset{\text{iid}}{\sim} \mathcal{N}_1(0, 1)$).

ICA model: Discussion of its assumptions and problem formulation

- In (3.1) there is no loss of generality to assume that $\mathrm{Var}(S) = \boldsymbol{I}_p$. Indeed, if $X = \tilde{\boldsymbol{B}}\tilde{S}$ with $\boldsymbol{\Psi} := \mathrm{Var}(\tilde{S}) \neq \boldsymbol{I}_p$ then, letting $\boldsymbol{B} = \tilde{\boldsymbol{B}}\boldsymbol{\Psi}^{1/2}$ and $S = \boldsymbol{\Psi}^{-1/2}\tilde{S}$, we have

$$X = \tilde{\boldsymbol{B}}\tilde{S} = \big(\tilde{\boldsymbol{B}}\boldsymbol{\Psi}^{1/2}\big)\big(\boldsymbol{\Psi}^{-1/2}\tilde{S}\big) = \boldsymbol{B}S, \quad \mathrm{Var}(S) = \boldsymbol{I}_p \tag{3.2}$$

  implying that (3.1) holds.

  Remark: (3.2) shows that $\mathrm{Var}(\tilde{S})$ cannot be recovered from X if we only know that $X = \tilde{\boldsymbol{B}}\tilde{S}$ for some matrix $\tilde{\boldsymbol{B}} \in \mathbb{R}^{p\times p}$.

- Recalling that S is a true signal that we want to recover, the assumption that the components of S are independent is necessary to make $\boldsymbol{B}$ identifiable (the identifiability of $\boldsymbol{B}$ will be discussed more precisely later). In particular, if we only assume that the components of S are uncorrelated then S can only be recovered up to an orthogonal transformation since, for all $\boldsymbol{G} \in O(p)$, $X = \boldsymbol{B}S = (\boldsymbol{B}\boldsymbol{G}^\top)(\boldsymbol{G}S)$ where $\mathrm{Var}(\boldsymbol{G}S) = \mathrm{Var}(S)$.

We assume for now that $\mathrm{Var}(X) = \boldsymbol{I}_p$. In this case, the ICA model (3.1) implies that

$$\mathrm{Var}(X) = \mathrm{Var}(\boldsymbol{B}S) = \boldsymbol{B}\boldsymbol{B}^\top = \boldsymbol{I}_p$$

showing that if (3.1) holds then $\boldsymbol{B}$ must be an orthogonal matrix[a]

Therefore, under the assumption $\mathrm{Var}(X) = \boldsymbol{I}_p$ we have $X = \boldsymbol{B}S$ if and only if $S = \boldsymbol{B}^\top X$, and thus under (3.1) and the assumption $\mathrm{Var}(X) = \boldsymbol{I}_p$ the matrix $\boldsymbol{B} \in O(p)$ is such that the components of the random variable $\boldsymbol{B}^\top X$ are independent.

[a]Recall that the ICA model assumes that $\boldsymbol{B}$ is invertible

Defining $\boldsymbol{B}$ through an optimization problem

For a given random variable Z we denote by $g_Z(z)\mathrm{d}z$ its probability distribution (w.r.t. some reference measure $\mathrm{d}z$) and we let

$$H(Z) = -\int g_Z(z)\log\big(g_Z(z)\big)\,\mathrm{d}z.$$

Remark: The quantity $H(Z)$ is called the entropy of Z.

We also recall that the Kullback-Leibler (KL) divergence between the distributions $p(z)\mathrm{dz}$ and $q(z)\mathrm{dz}$ is given by

$$\mathrm{KL}(p||q) = \int \log\Big(\frac{p(z)}{q(z)}\Big)p(z)\mathrm{dz}$$

and we also recall the following result:

Lemma 3.1 *For any distributions $p(z)\mathrm{dz}$ and $q(z)\mathrm{dz}$ we have $KL(p||q)\geq 0$, where the equality holds if and only if $p=q$.*

Easy computations show that if $Z=(Z_1,\dots,Z_p)$ is and $\mathbb{R}^p$-valued random variable then the KL divergence between the distributions $g_Z(z)\mathrm{dz}$ and $\prod_{j=1}^p g_{Z_j}(z_j)\mathrm{dz}$ can be written as follows:

$$\mathrm{KL}\Big(g_Z||\prod_{j=1}^p g_{Z_j}\Big) = \sum_{j=1}^p H(Z_j) - H(Z) =: I(Z).$$

By Lemma 3.1, $I(Z)\geq 0$ and $I(Z)=0$ if and only if all the components of Z are independent. Therefore, $I(Z)$ can be interpreted as a measure of independence between the components of Z. (The quantity $I(Z)$ is called the mutual information of Z.)

Under the assumption $\mathrm{Var}(X)=\boldsymbol{I}_p$, it follows that the matrix $\boldsymbol{B}$ in (3.2) verifies

$$\boldsymbol{B}\in \operatorname*{argmin}_{\boldsymbol{G}\in O(p)} I(\boldsymbol{G}^\top X). \tag{3.3}$$

Two key lemmas for finding an approximate solution to (3.3)

Lemma 3.2 *Let Y be a continuous and real-valued random variable with $\mathbb{E}[Y]=0$ and $\mathrm{Var}(Y)=1$, and let $Z\sim\mathcal{N}_1(0,1)$. Then, $H(Z)\geq H(Y)$, where the equality holds if and only if $Y\sim\mathcal{N}_1(0,1)$.*

Proof: We have

$$\begin{aligned}
\mathrm{KL}(p_Y\|p_Z)=\int\log\Big(\frac{p_Y(y)}{p_Z(y)}\Big)p_Y(y)\mathrm{d}y&=-H(Y)-\int\log(p_Z(y))p_Y(y)\mathrm{d}y\\
&=-H(Y)-\int\Big(-\frac{1}{2}\log(2\pi)-\frac{y^2}{2}\Big)p_Y(y)\mathrm{d}y\\
&=-H(Y)-\int\Big(-\frac{1}{2}\log(2\pi)-\frac{z^2}{2}\Big)p_Z(z)\mathrm{d}z\\
&=-H(Y)-\int\log(p_Z(z))p_Z(z)\mathrm{d}z\\
&=-H(Y)+H(Z)
\end{aligned}$$

where the third equality uses the fact that $\mathbb{E}[Y^2]=\mathbb{E}[Z^2]$. Then, the result follows from Lemma 3.1. □

To proceed further for any $\mathbb{R}$-valued random variable Y such that $\mathbb{E}[Y]=0$ and $\mathrm{Var}(Y)=1$ we let $J(Y)=H(Z)-H(Y)$, with $Z\sim\mathcal{N}_1(0,1)$.

Remark: By Lemma 3.2, for any $\mathbb{R}$-valued random variable Y such that $\mathbb{E}[Y]=0$ and $\mathrm{Var}(Y)=1$ we have $J(Y)\geq 0$, where the equality holds if and only if $Y\sim\mathcal{N}_1(0,1)$. Hence, the quantity $J(Y)$, called the negentropy of Y, is a measure of distance to normality.

Lemma 3.3 *Assume $\mathrm{Var}(X)=\boldsymbol{I}_p$. Then, the matrix $\boldsymbol{B}$ in (3.2) is such that $\boldsymbol{B}\in\mathrm{argmax}_{\boldsymbol{G}\in O(p)}\sum_{j=1}^p J(g_{(j)}^\top X)$.*

Proof: Let $p_X(\cdot)$ be the density of X and let $Y=\boldsymbol{G}^\top X$. Then, using the change of variables formula and the fact that $|\det(\boldsymbol{G})|=1$, the density $p_Y(\cdot)$ of Y is defined by $p_Y(y)=p_X(\boldsymbol{G}y)$ and thus

$$\begin{aligned}
H(\boldsymbol{G}^\top X)=H(Y)=-\int p_Y(y)\log\big(p_Y(y)\big)\,\mathrm{d}y=-\int p_X(\boldsymbol{G}y)\log\big(p_X(\boldsymbol{G}y)\big)\,\mathrm{d}y&=-\int|\det(\boldsymbol{G}^{-1})|p_X(x)\log\big(p_X(x)\big)\,\mathrm{d}x\\
&=-\int p_X(x)\log\big(p_X(x)\big)\,\mathrm{d}x\\
&=H(X)
\end{aligned}$$

where the third equality uses the change of variables formula for integrals. Then the result follows from (3.3) and from the definition of $I(\boldsymbol{G}^\top X)$. □

Approximating the solution to (3.3)

Using Lemma 3.2 and Lemma 3.3, it follows if $\mathrm{Var}(X) = \boldsymbol{I}_p$ then the matrix $\boldsymbol{B}$ in (3.2) verifies

$$\begin{aligned}\boldsymbol{B} &\in \mathrm{argmax}_{\boldsymbol{G}\in O(p)} \sum_{j=1}^{p} J(g_{(j)}^\top X) \qquad (3.4)\\ &= \mathrm{argmax}_{\boldsymbol{G}\in O(p)} \Big\{ \sum_{j=1}^{p} \text{"departure from Gaussianitiy of } g_{(j)}^\top X\text{"} \Big\}.\end{aligned}$$

The quantity $J(g_{(j)}^\top X)$ is usually intractable but its interpretation in term of distance to normality provides a simple way to compute an approximate solution to (3.4): replace $J(g_{(j)}^\top X)$ by another measure of distance to normality!

In ICA, the departure from Gaussianitiy of $g_{(j)}^\top X$ is often measured by

$$\Big(\mathbb{E}[\varphi(g_{(j)}^\top X)] - \mathbb{E}[\varphi(Z)]\Big)^2, \quad Z \sim \mathcal{N}_1(0,1) \qquad (3.5)$$

for some function $\varphi : \mathbb{R} \to \mathbb{R}$.

Remark: Using (3.5) with the function $\varphi(u) = \frac{1}{a}\log\cosh(au)$ for some $a \in [1,2]$, or with the function $\varphi(u) = -e^{-u^2/2}$, often works well in practice and provide a reasonable approximation of the negentropy $J(g_{(j)}^\top X)$ [**?**].

Then, for a given choice of $\varphi : \mathbb{R} \to \mathbb{R}$ and assuming $\mathrm{Var}(X) = \boldsymbol{I}_p$, the matrix $\boldsymbol{B}$ in (3.2) can be approximated by the matrix $\tilde{\boldsymbol{B}}$ verifying

$$\tilde{\boldsymbol{B}} \in \mathrm{argmax}_{\boldsymbol{G}\in O(p)} \sum_{j=1}^{p} \Big(\mathbb{E}[\varphi(g_{(j)}^\top X)] - \mathbb{E}[\varphi(Z)]\Big)^2 \qquad (3.6)$$

where $Z \sim \mathcal{N}_1(0,1)$.

Estimation of the approximate solution $\tilde{B}$

A possible approach for estimating the matrix $\tilde{\boldsymbol{B}}$ defined in (3.6) from the observations $\{x_i\}_{i=1}^n$ is as follows. For all $\boldsymbol{G} \in O(p)$ let

$$L(\boldsymbol{G}) = \sum_{j=1}^{p} \Big(\mathbb{E}[\varphi(g_{(j)}^\top X)] - \mathbb{E}[\varphi(Z)] \Big)^2$$

and

$$\hat{L}_n(\boldsymbol{G}, \{x_i\}_{i=1}^n) = \sum_{j=1}^{p} \Big(\frac{1}{n} \sum_{i=1}^{n} \varphi(g_{(j)}^\top x_i) - \mathbb{E}[\varphi(Z)] \Big)^2$$

so that $\hat{L}_n(\boldsymbol{G}, \{x_i\}_{i=1}^n) \approx L(\boldsymbol{G})$[a].

Then, noting that $\tilde{\boldsymbol{B}} \in \mathrm{argmax}_{\boldsymbol{G} \in O(p)} L(\boldsymbol{G})$, we can estimate $\tilde{\boldsymbol{B}}$ using the matrix $\tilde{\boldsymbol{B}}_n$ defined by

$$\tilde{\boldsymbol{B}}_n \in \mathrm{argmax}_{\boldsymbol{G} \in O(p)} \hat{L}_n(\boldsymbol{G}, \{x_i\}_{i=1}^n). \tag{3.7}$$

The definition of $\hat{L}_n(\boldsymbol{G}, \{x_i\}_{i=1}^n)$ requires to compute $\mathbb{E}[\varphi(Z)]$, a quantity which, depending on the choice of φ, may be intractable. However, being a one dimensional integral, we can easily (and efficiently) estimate $\mathbb{E}[\varphi(Z)]$ using numerical integration methods.

Letting $\hat{\varphi}$ be an estimate of $\mathbb{E}[\varphi(Z)]$, a computable estimate $\tilde{\boldsymbol{B}}_n'$ of $\tilde{\boldsymbol{B}}$ is therefore defined by

$$\tilde{\boldsymbol{B}}_n' \in \mathrm{argmax}_{\boldsymbol{G} \in O(p)} \sum_{j=1}^{p} \Big(\frac{1}{n} \sum_{i=1}^{n} \varphi(g_{(j)}^\top x_i) - \hat{\varphi} \Big)^2. \tag{3.8}$$

[a] The set $O(p)$ being compact if follows that if $\{X_i\}_{i=1}^n$ are i.i.d. copies of X then, under some conditions on φ, we have $\sup_{\boldsymbol{G} \in O(p)} |\hat{L}_n(\boldsymbol{G}, \{X_i\}_{i=1}^n) - L(\boldsymbol{G})| \to 0$ in probability. Note that, whatever the distribution of X is, this uniform convergence result holds if φ is continuous an bounded, as this is the case for the choice $\varphi(u) = -e^{-u^2/2}$ mentioned earlier.

Estimation of the approximate solution $\tilde{B}$ (end)

A simple way to solve the optimization problem (3.8) is to use a projected gradient descend algorithm:

A projected gradient descend algorithm for solving (3.8)

Input: Matrix $\tilde{\boldsymbol{B}}^{(0)} \in O(p)$ and step-size $\gamma > 0$

for $s \geq 1$ **do**

(i) Let $\boldsymbol{C}_s = \tilde{\boldsymbol{B}}^{(s-1)} + \gamma \nabla_{\boldsymbol{B}} \sum_{j=1}^{p} \left(\frac{1}{n} \sum_{i=1}^{n} \varphi(x_i^\top \tilde{b}_{(j)}^{(s-1)}) - \hat{\varphi} \right)^2$

(ii) Let $\tilde{\boldsymbol{B}}^{(s)} = (\boldsymbol{C}_s \boldsymbol{C}_s^\top)^{-1/2} \boldsymbol{C}_s \in O(p)$.

if Convergence=TRUE **then**

(iii) **return** $\tilde{\boldsymbol{B}}^{(s)}$.

(iv) **break**

end if

end for

Remark: In practice, more sophisticated (and more efficient) methods are used to estimate the matrix $\tilde{\boldsymbol{B}}$ defined in (3.6) [see **?** , Section 6].

Remark: Estimating $\boldsymbol{B}$ through the estimation of the matrix $\tilde{\boldsymbol{B}}$ is called the FastICA method.

Estimation of $\boldsymbol{B}$ when $\mathrm{Var}(X) \neq \boldsymbol{I}_d$

We now consider the general case where $\mathrm{Var}(X) = \boldsymbol{\Sigma}$ for some $\boldsymbol{\Sigma} \neq \boldsymbol{I}_p$. We assume below that $\boldsymbol{\Sigma}$ is full rank and let $\boldsymbol{L}$ be a diagonal matrix containing the eigenvalues of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Gamma} \in O(p)$ be the corresponding matrix of orthonormal eigenvectors.

Let $\boldsymbol{K} = \boldsymbol{L}^{-1/2}\boldsymbol{\Gamma}^\top$, $\boldsymbol{B}_W = \boldsymbol{K}\boldsymbol{B}$ and $W = \boldsymbol{K}X$, and note that

(i) Under the ICA model the random variable W is such that $W = (\boldsymbol{K}\boldsymbol{B})S = \boldsymbol{B}_W S$ and such that

$$\mathrm{Var}(W) = \boldsymbol{L}^{-1/2}\boldsymbol{\Gamma}^\top\boldsymbol{\Sigma}\boldsymbol{\Gamma}\boldsymbol{L}^{-1/2} = \boldsymbol{L}^{-1/2}\boldsymbol{\Gamma}^\top(\boldsymbol{\Gamma}\boldsymbol{L}\boldsymbol{\Gamma}^\top)\boldsymbol{\Gamma}\boldsymbol{L}^{-1/2} = \boldsymbol{I}_p.$$

(ii) Since $\boldsymbol{B}_W = \boldsymbol{K}\boldsymbol{B}$ we have $\boldsymbol{B} = \boldsymbol{K}^{-1}\boldsymbol{B}_W = \boldsymbol{\Gamma}\boldsymbol{L}^{1/2}\boldsymbol{B}_W$.

Let $\boldsymbol{\Lambda}$ be the diagonal matrix containing the eigenvalues of the matrix $\boldsymbol{S} := \frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}$, $\boldsymbol{A} \in O(p)$ be the corresponding matrix of orthonormal eigenvectors and $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{A}$ be the matrix of principal components.

Then, (i)-(ii) suggest the following three steps for estimating $\boldsymbol{B}$:

1. Compute $\boldsymbol{W} = \boldsymbol{Y}\boldsymbol{\Lambda}^{-1/2}$. Remark that $\frac{1}{n}\boldsymbol{W}^\top\boldsymbol{W} = \boldsymbol{I}_p$ and recall that (under some conditions) we can interpret $\{w_i\}_{i=1}^n$ as "approximate" realizations of W (see Chapter 1, pages 24–25). The transformation $\boldsymbol{X} \mapsto \boldsymbol{W}$ of the data is called whitening.
2. Use $\{w_i\}_{i=1}^n$ and the approach introduced in this chapter for estimating $\boldsymbol{B}$ when $\mathrm{Var}(X) = \boldsymbol{I}_p$ to compute an estimate $\tilde{\boldsymbol{B}}'_{n,W}$ of $\boldsymbol{B}_W$.
3. Estimate $\boldsymbol{B}$ by

$$\hat{\boldsymbol{B}}_n := \boldsymbol{A}\boldsymbol{\Lambda}^{1/2}\tilde{\boldsymbol{B}}'_{n,W}. \tag{3.9}$$

Recovering the signals from the data

Recall that under the ICA model the observations $\{x_i\}_{i=1}^n$ are realizations of a random variable X such that $X = \boldsymbol{B}S$ with $\boldsymbol{B}$ and S as in (3.1).

Under the ICA model we therefore have $x_i = \boldsymbol{B}s_i$ for all i and thus, denoting by $\boldsymbol{S}_{\text{sig}}$ the $n \times p$ matrix with rows $\{s_i\}_{i=1}^n$,

$$\boldsymbol{X} = \boldsymbol{S}_{\text{sig}}\boldsymbol{B}^\top.$$

Given the estimate $\hat{\boldsymbol{B}}_n$ of $\boldsymbol{B}$ defined in (3.9) we can estimate the matrix signals $\boldsymbol{S}_{\text{sig}}$ using

$$\hat{\boldsymbol{S}}_{\text{sig}} := \boldsymbol{X}\big(\hat{\boldsymbol{B}}_n^\top\big)^{-1}.$$

Recalling that $\tilde{\boldsymbol{B}}'_{n,W} \in O(p)$, we have

$$\begin{aligned}\hat{\boldsymbol{S}}_{\text{sig}} = \boldsymbol{X}\big(\hat{\boldsymbol{B}}_n^\top\big)^{-1} &= \boldsymbol{X}\Big(\big(\tilde{\boldsymbol{B}}'_{n,W}\big)^\top \boldsymbol{\Lambda}^{1/2}\boldsymbol{A}^\top\big)\Big)^{-1} \\ &= \boldsymbol{X}\boldsymbol{A}\boldsymbol{\Lambda}^{-1/2}\tilde{\boldsymbol{B}}'_{n,W} \\ &= \boldsymbol{W}\tilde{\boldsymbol{B}}'_{n,W}\end{aligned}$$

and therefore computing the estimate $\hat{\boldsymbol{S}}_{\text{sig}}$ does not require to compute the estimate $\hat{\boldsymbol{B}}_n$ of $\boldsymbol{B}$.

Remark: Estimating $\boldsymbol{S}_{\text{sig}}$ by $\hat{\boldsymbol{S}}_{\text{sig}} = \boldsymbol{W}\tilde{\boldsymbol{B}}'_{n,W}$ is natural. Indeed, under the ICA model we have $W = \boldsymbol{B}_W S$ and thus, as we are interpreting the w_i's as realizations of W, this implies that $w_i = \boldsymbol{B}_W s_i$ for all i, and thus that $\boldsymbol{W} = \boldsymbol{S}_{\text{sig}}\boldsymbol{B}_W^\top \Leftrightarrow \boldsymbol{S}_{\text{sig}} = \boldsymbol{W}\boldsymbol{B}_W$.

Identifiability of the ICA model

Assume that $\mathrm{Var}(X) = \boldsymbol{I}_p$ and that the ICA model is correct, so that for some $\boldsymbol{B} \in O(p)$ we have $\boldsymbol{X} = \boldsymbol{S}_{\text{sig}}\boldsymbol{B}^\top$ where the rows of $\boldsymbol{S}_{\text{sig}}$ are n realizations of S.

Then,

- If all the components of S are $\mathcal{N}_1(0,1)$ random variables then $\boldsymbol{G}S \stackrel{\text{dist.}}{=} S$ for any matrix $\boldsymbol{G} \in O(p)$[a]. Therefore, in this case we have

$$X = \boldsymbol{B}S \stackrel{\text{dist.}}{=} \boldsymbol{B}\boldsymbol{G}S = (\boldsymbol{B}\boldsymbol{G})S, \quad \boldsymbol{B}\boldsymbol{G} \in O(p), \quad \forall \boldsymbol{G} \in O(p)$$

  showing that the matrix $\boldsymbol{B}$, and thus the signals $\boldsymbol{S}_{\text{sig}}$, can only be estimated up to an orthogonal transformation.

- The columns of $\mathbf{B}$, and thus the rows of $\boldsymbol{S}_{\text{sig}}$, can only be estimated up to a multiplicative sign. To see this let $\boldsymbol{D}$ be a $p \times p$ diagonal matrix with such that $d_{jj} \in \{-1, 1\}$ for all j and let $\tilde{S} = \boldsymbol{D}S$. Then,

$$X = \boldsymbol{B}S = \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{D}S = (\boldsymbol{B}\boldsymbol{D})\tilde{S}$$

  where $\boldsymbol{B}\boldsymbol{D} \in O(p)$ and $\mathrm{Var}(\tilde{S}) = \boldsymbol{I}_d$.

Remark: It can be shown that if at most one component of S is Gaussian and $\mathrm{rank}(\boldsymbol{B}) = p$ then the columns of the matrix $\boldsymbol{B} \in \mathcal{O}(p)$ are unique up to a multiplicative sign (see [**?**]).

[a] Use the change of variable formula to show this.

Illustration of ICA

We let $p = 3$, $n = 1\,000$ and simulate the true signals matrix $\boldsymbol{S}_{\text{sig}} = [s_{ij}] \in \mathbb{R}^{n\times p}$ using

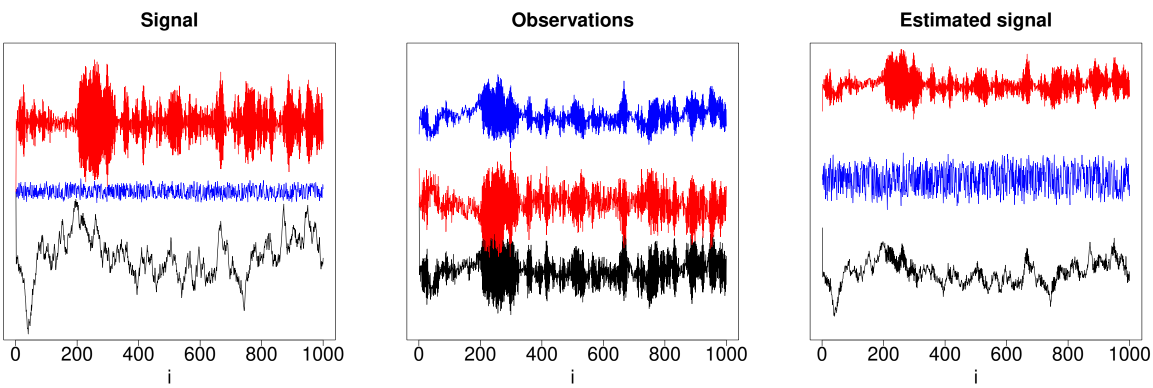
$$S_{ij} = \rho_j S_{(i-1)j} + \epsilon_{ij}, \quad \epsilon_{ij} \overset{\text{iid}}{\sim} \frac{1}{2}\mathcal{N}_1(-1, 0.25) + \frac{1}{2}\mathcal{N}_1(1, 0.25)$$

where $S_{(i-1)j} = 0$ for $i = 1$ and where $\rho_2 = -\rho_1 = 0.98$ and $\rho_3 = 0.2$. Remark that each row of the matrix $\boldsymbol{S}_{\text{sig}}$ is a trajectory of a Markov chain. The three Markov chains (i.e. the three signals) are mixed using the matrix

$$\boldsymbol{B} = \begin{pmatrix} 1 & -1 & -3 \\ 1 & 1 & 2 \\ -1 & 3 & -3 \end{pmatrix}.$$

Remark: In this example the distribution of S_i is not exactly the same for all i.

The true signals $\boldsymbol{S}_{\text{sig}}$ as well as the observations $\boldsymbol{X} = \boldsymbol{S}_{\text{sig}}\boldsymbol{B}^\top$ and the ICA estimate $\hat{\boldsymbol{S}}_{\text{sig}}$ of $\boldsymbol{S}_{\text{sig}}$ (obtained with $\varphi(u) = \log\cosh(u)$) are shown in the following plots.



Remark: In the two plots the signals are shifted and coloured to facilitate the comparisons. In addition, the y-axis has been removed since, as shown above, we cannot estimate the variance of the components of S.