Machine learning 2
Exercise sheet 3

FLEISCHMANN Kay, Matrnr: 352247
ROHRMANN Till, Matrnr: 343756

May 5, 2013

3 The SSA Cost Function

Given $X_1 \sim \mathcal{N}(\mu_1, \Sigma_1)$ and $X_2 \sim \mathcal{N}(\mu_2, \Sigma_2)$ be Gaussian random variables with values in $\mathbb{R}^n$. The probability density function of X_i is

$$p_i(x) \quad = \quad ((2\pi)^n \det \Sigma)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)\right) \tag{1}$$

Derive the explicit formula for the KL-divergence between X_1 and X_2:

$$\begin{aligned}
D_{KL}(X_2 \mid\mid X_1) &= \int p_2(x) \log\left(\frac{p_2(x)}{p_1(x)}\right) \, dx & (2)\\
&= \int p_2(x) \left(-\frac{1}{2}\log\left((2\pi)^n \det \Sigma_2\right) - \frac{1}{2}(x-\mu_2)^T \Sigma_2^{-1}(x-\mu_2)\right.\\
&\quad \left.+\frac{1}{2}\log\left((2\pi)^n \det \Sigma_1\right) + \frac{1}{2}(x-\mu_1)^T \Sigma_1^{-1}(x-\mu_1)\right) \, dx & (3)\\
&= \frac{1}{2}\log\left(\frac{\det \Sigma_1}{\det \Sigma_2}\right) + \int p_2(x)\left(-\frac{1}{2}(x-\mu_2)^T \Sigma_2^{-1}(x-\mu_2)\right.\\
&\quad \left.+\frac{1}{2}(x-\mu_2+\mu_2-\mu_1)^T \Sigma_1^{-1}(x-\mu_2+\mu_2-\mu_1)\right) \, dx & (4)\\
&= \frac{1}{2}\left(\log\left(\frac{\det \Sigma_1}{\det \Sigma_2}\right) - \mathbb{E}\left[(x-\mu_2)^T \Sigma_2^{-1}(x-\mu_2)\right]\right.\\
&\quad +\mathbb{E}\left[(x-\mu_2)^T \Sigma_1^{-1}(x-\mu_2)\right] + \mathbb{E}\left[(x-\mu_2)^T \Sigma_1^{-1}(\mu_2-\mu_1)\right]\\
&\quad \left.\mathbb{E}\left[(\mu_2-\mu_1)^T \Sigma_1^{-1}(x-\mu_2)\right] + (\mu_2-\mu_1)^T \Sigma_1^{-1}(\mu_2-\mu_1)\right) & (5)
\end{aligned}$$

Assume ϵ is a vector of n random variables. Let μ denote its mean and Σ its covariance matrix. Let Λ be an n-dimensional symmetric matrix.

$$\mathbb{E}\left[\epsilon^T \Lambda \epsilon\right] \quad = \quad \operatorname{tr}\left(\mathbb{E}\left[\epsilon^T \Lambda \epsilon\right]\right) \tag{6}$$

Due to the linearity of tr it holds $\mathbb{E} \circ \operatorname{tr} = \operatorname{tr} \circ \mathbb{E}$ and thus

$$\mathbb{E}\left[\epsilon^T \Lambda \epsilon\right] \quad = \quad \mathbb{E}\left[\operatorname{tr}\left(\epsilon^T \Lambda \epsilon\right)\right] \tag{7}$$

Due to the circular property of the trace operator $\operatorname{tr}(ABC) = \operatorname{tr}(BCA)$ the following equation holds

$$\begin{aligned}
\mathbb{E}\left[\epsilon^T \Lambda \epsilon\right] &= \mathbb{E}\left[\operatorname{tr}\left(\Lambda \epsilon \epsilon^T\right)\right] & (8)\\
&= \operatorname{tr}\left(\Lambda \mathbb{E}\left[\epsilon \epsilon^T\right]\right) & (9)\\
&= \operatorname{tr}\left(\Lambda\left(\Sigma + \mu\mu^T\right)\right) & (10)\\
&= \operatorname{tr}(\Lambda\Sigma) + \mu^T \Lambda \mu & (11)
\end{aligned}$$

Applied to equation (5) we obtain

$$\begin{aligned}
(5) &= \frac{1}{2}\left(\log\left(\frac{\det \Sigma_1}{\det \Sigma_2}\right) - \operatorname{tr}\left(\Sigma_2^{-1}\Sigma_2\right) + \operatorname{tr}\left(\Sigma_1^{-1}\Sigma_2\right) + 0 + 0\right.\\
&\quad \left.+(\mu_2-\mu_1)^T \Sigma_1^{-1}(\mu_2-\mu_1)\right) & (12)\\
&= \frac{1}{2}\left(\log\left(\frac{\det \Sigma_1}{\det \Sigma_2}\right) - n + \operatorname{tr}\left(\Sigma_1^{-1}\Sigma_2\right) + (\mu_2-\mu_1)^T \Sigma_1^{-1}(\mu_2-\mu_1)\right) & (13)
\end{aligned}$$

Which is the final result.
Show that the explicit formula for the SSA cost function can be written:

$$\begin{aligned} L(R) &= \sum_{i=1}^{N} D_{KL}\left[\mathcal{N}(\hat{\mu}_i^s, \hat{\Sigma}_i^s) \mid\mid \mathcal{N}(0, I)\right] && (14) \\ &= \frac{1}{2}\sum_{i=1}^{N}\left(-\log\left(\det \hat{\Sigma}_i^s\right) + (\hat{\mu}_i^s)^T \hat{\mu}_i^s\right) - \frac{N-1}{2}d && (15) \end{aligned}$$

Proof.

$$\begin{aligned} D_{KL}\left[\mathcal{N}(\hat{\mu}_i^s, \hat{\Sigma}_i^s) \mid\mid \mathcal{N}(0, I)\right] &= \frac{1}{2}\left(\log\left(\frac{1}{\det \hat{\Sigma}_i^s}\right) + \mathrm{tr}(\hat{\Sigma}_i^s) + (\hat{\mu}_i^s)^T(\hat{\mu}_i^s) - d\right) && (16) \\ \sum_{i=1}^{N} D_{KL}\left[\mathcal{N}(\hat{\mu}_i^s, \hat{\Sigma}_i^s) \mid\mid \mathcal{N}(0, I)\right] &= \frac{1}{2}\sum_{i=1}^{N}\left(-\log\left(\det \hat{\Sigma}_i^s\right) + (\hat{\mu}_i^s)^T(\hat{\mu}_i^s)\right) \\ &\quad + \frac{1}{2}\sum_{i=1}^{N}\mathrm{tr}(\hat{\Sigma}_i^s) - \frac{N}{2}d && (17) \\ \sum_{i=1}^{N}\mathrm{tr}(\hat{\Sigma}_i^s) &= \mathrm{tr}\left(\sum_{i=1}^{N}\hat{\Sigma}_i^s\right) && (18) \\ &= \mathrm{tr}\left(\sum_{i=1}^{N} I^d R\hat{\Sigma}_i (I^d R)^T\right) && (19) \\ &= \mathrm{tr}\left(I^d R\left(\sum_{i=1}^{N}\hat{\Sigma}_i\right)(I^d R)^T\right) && (20) \\ &= \mathrm{tr}\left(I^d R I (I^d R)^T\right) && (21) \\ &= d && (22) \end{aligned}$$

By inserting this result into equation (17) we finally obtain:

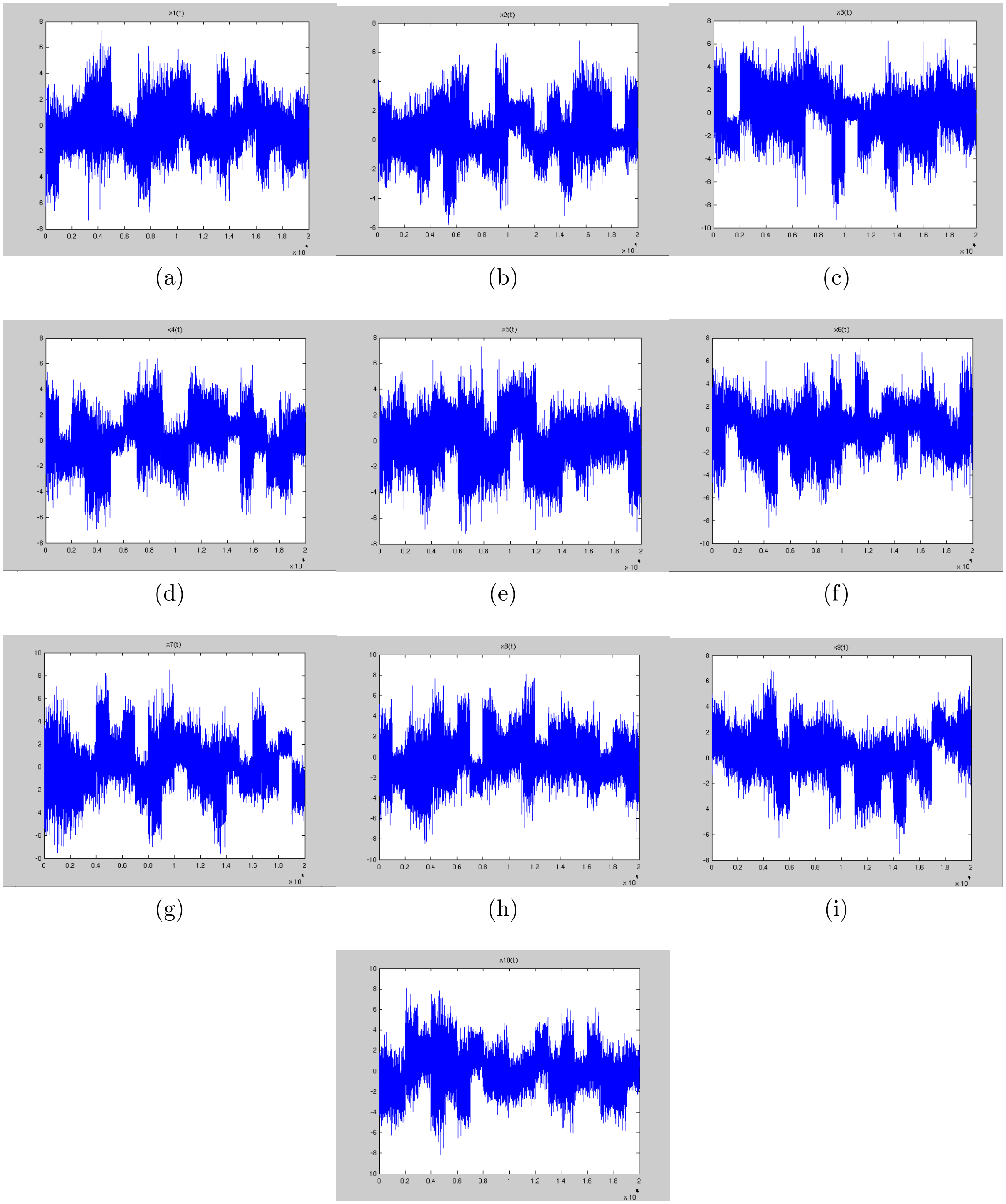
$$L(R) = \frac{1}{2}\sum_{i=1}^{N}\left(-\log\left(\det \hat{\Sigma}_i^s\right) + (\hat{\mu}_i^s)^T(\hat{\mu}_i^s)\right) - \frac{N-1}{2}d \qquad (23)$$

□

4 Finding the stationary subspace

4.1 Dataset

The following plots describe the dataset given for the exercise on which the SSA should be applied. In this dataset stationarity an non-stationarity are embedded.

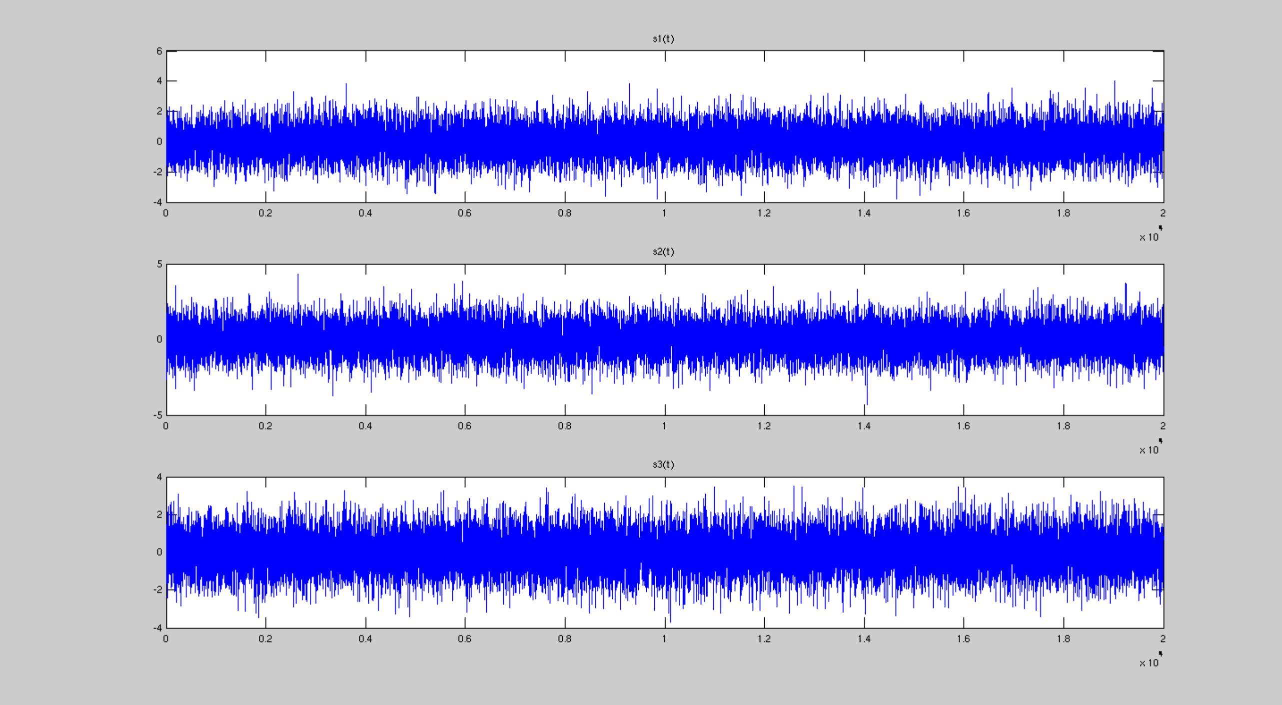


(a) (b) (c) (d) (e) (f) (g) (h) (i) (j)

Figure 1: Plots of each of the 10 (a)-(j) components with stationary and non-stationary included

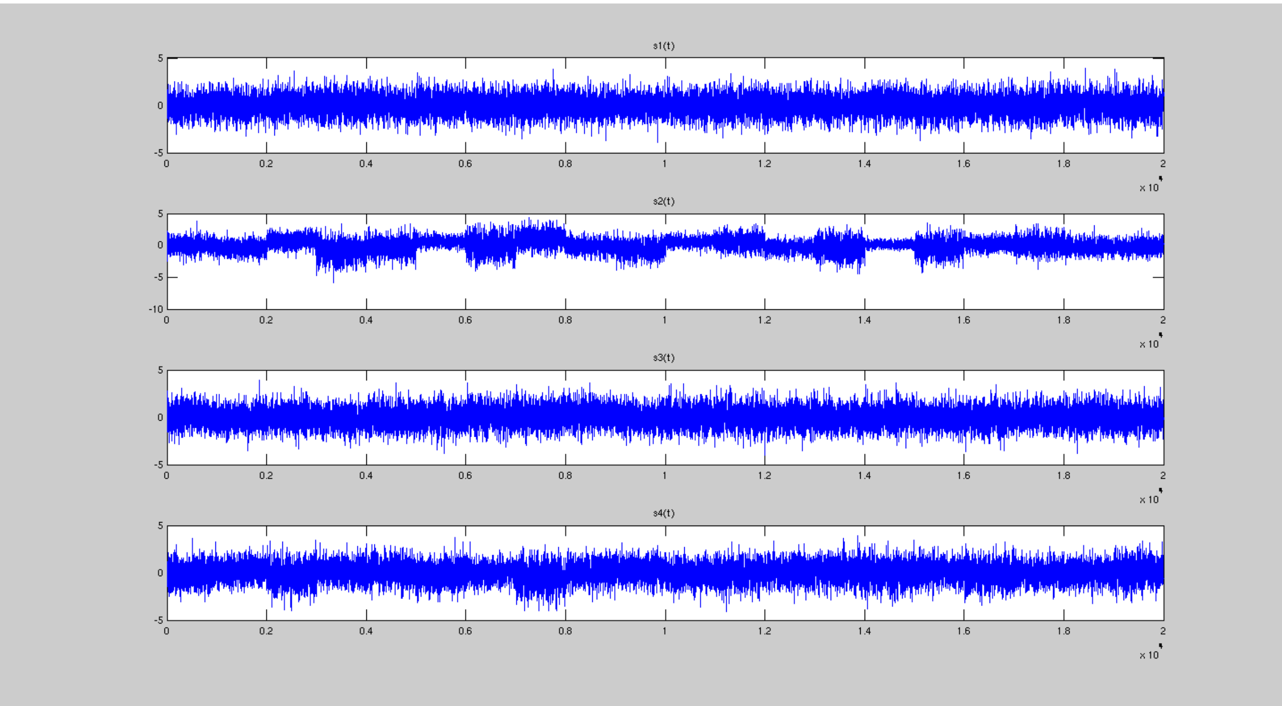
4.2 Experiments

For the next experiments the number of expoches is set to 25. If one assume the number stationary sources in the dataset is $dd = 3$, the plot of each component looks like:



(a)

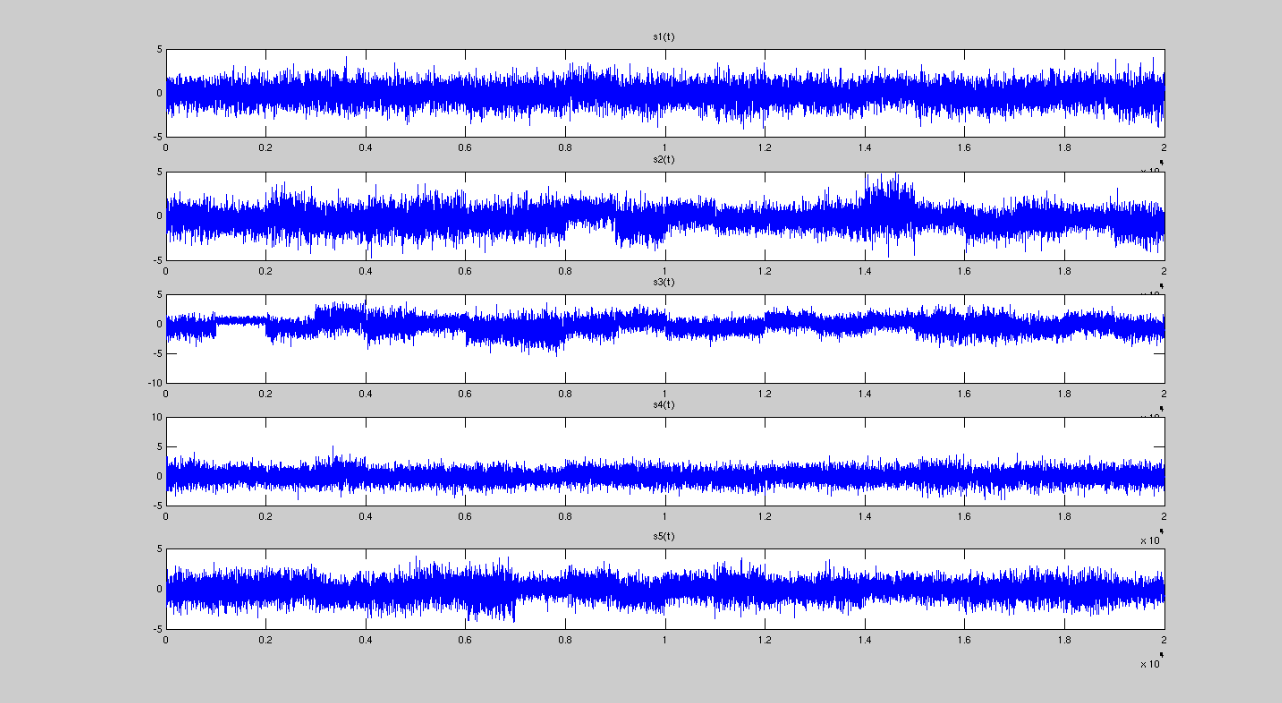
Figure 2: Plot if number of stationary sources is set to 3.



(a)

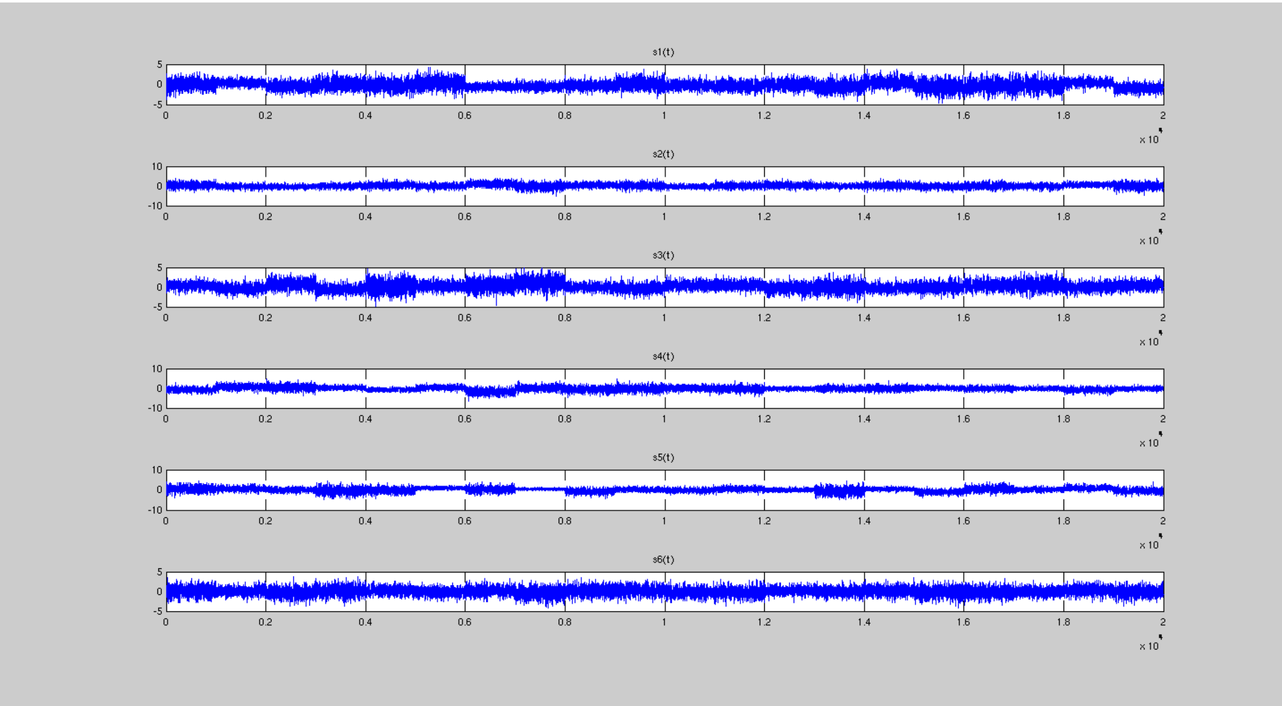
Figure 3: Plot if number of stationary sources is set to 4.

In the next plot the number of stationary sources is incremnted to $dd = 5$:



(a)

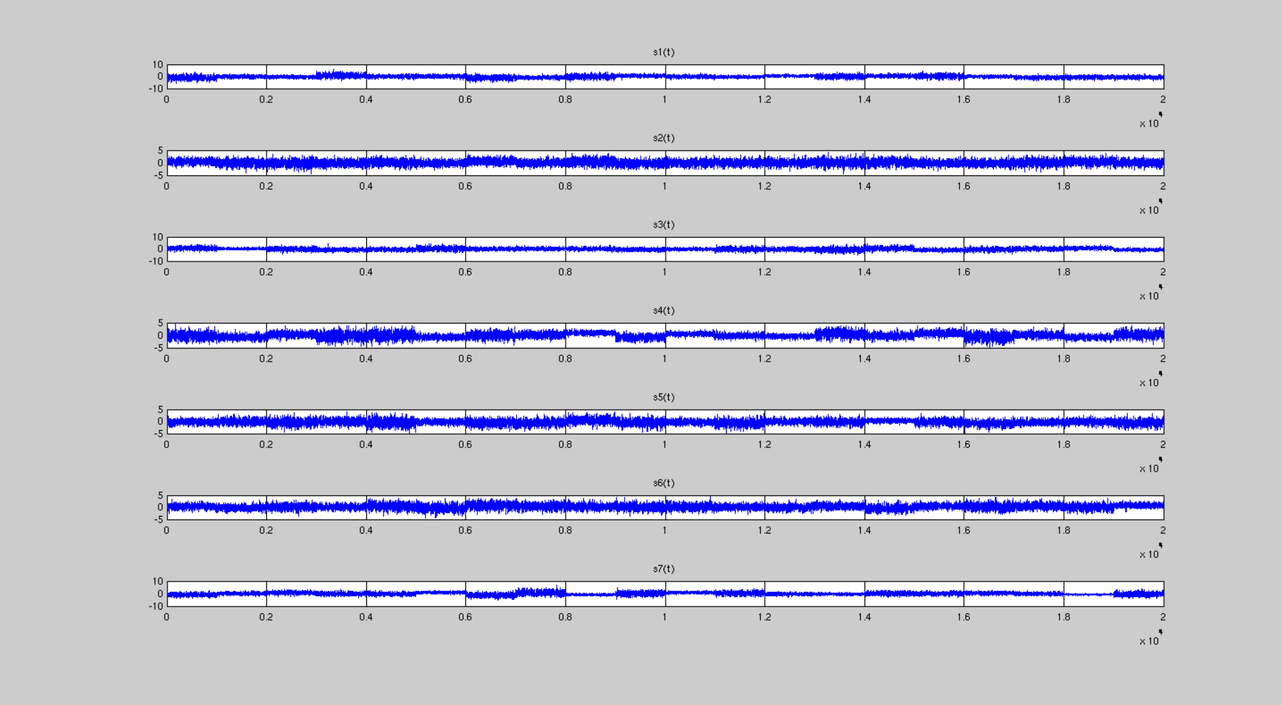
Figure 4: Plot if number of stationary sources is set to 5.



(a)

Figure 5: Plot if number of stationary sources is set to 6.

In the next experiment, the number of stationary sources is incremnted to $dd = 7$:



(a)

Figure 6: Plot if number of stationary sources is set to 7.

4.3 Number of stationary sources and projection

If one compare the different plots from $dd = 3$ to $dd = 7$ the non-stationariy sources appear more and more. Assuming the stationary sources coming from the same underlying distribution over the time-series, the best choice of the number of stationary sources ist 3.

The projectionmatrix B^s

$$\begin{matrix} 0.6581 & -0.6022 & -0.0495 & 0.6484 & -0.1115 & 0.3517 & 0.6469 & -0.3847 & 0.5758 & 0.0258 \\ -0.0397 & 0.4733 & 0.2207 & -0.1916 & -0.1813 & -0.3004 & -0.1542 & -0.2849 & -0.6388 & -0.4793 \\ 0.2932 & -0.2916 & -0.4973 & 0.7134 & 0.5037 & -0.3822 & 0.0260 & -0.3462 & 0.5083 & 0.1465 \end{matrix}$$

5 Uniqueness and identifiability of the SSA model

$(a^{\star})$

The mixing matrix A can in general not be identified exactly. The problem of identifying the mixing matrix A uniquely is that any linear combination of a stationary time series is again stationary. Furthermore, any linear combination of a stationary and non-stationary time series is non-stationary as-well. Thus, even if we find an unmixing matrix $\hat{B}$ which separates the signal into its stationary and non-stationary part, one cannot be sure to have found the true unmixing matrix, because the basis of the stationary and non-stationary spaces can only be identified up to an arbitrary linear transformation within these spaces. Having said that, a short example follows:

Assume a 3-dimensional source with 2 stationary sources and as mixing matrix the identity $A = I$:

$$\begin{pmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \end{pmatrix} = A \begin{pmatrix} s_1^s(t) \\ s_2^s(t) \\ s_1^n(t) \end{pmatrix} \tag{24}$$

Obviously, the identity matrix would also be a perfect unmixing matrix $\hat{B} = I$. However, we cannot distinguish that solution from another solution where the 2 stationary sources are permuted:

$$\begin{pmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \end{pmatrix} = A \cdot \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} s_1^s(t) \\ s_2^s(t) \\ s_1^n(t) \end{pmatrix} \tag{25}$$

$\hat{B} = I$ would again separate the stationary and non-stationary space perfectly but not get the order right. Consequently, matrix A cannot be identified exactly.

Restriction on solutions of unmixing matrix. Assuming that we have found an unmixing matrix $\hat{B} = \hat{A}^{-1}$, then we also get an approximation of the stationary and non-stationary space $\hat{A}^s$ and $\hat{A}^n$. Since $\hat{A}$ has full rank we can express the true mixing matrix A the following way:

$$A^s = \hat{A}^s M_1 + \hat{A}^n M_2 \tag{26}$$

$$A^n = \hat{A}^s M_3 + \hat{A}^n M_4 \tag{27}$$

$$\begin{pmatrix} \hat{s}^s \\ \hat{s}^n \end{pmatrix} = \hat{B} A \begin{pmatrix} s^s \\ s^n \end{pmatrix} \tag{28}$$

$$= \begin{pmatrix} M_1 & M_3 \\ M_2 & M_4 \end{pmatrix} \cdot \begin{pmatrix} s^s \\ s^n \end{pmatrix} \tag{29}$$

Since $\hat{s}^s$ shall be stationary, we mustn't mix in some non-stationary parts. Thus, $M_3 = 0$. The other matrices can be arbitrarily chosen. That implies:

$$A^s = \hat{A}^s M_1 + \hat{A}_n M_2 \tag{30}$$

$$A^n = \hat{A}^n M_4 \tag{31}$$

Consequently, the stationary space cannot be retrieved but it is possible to retrieve the non-stationary space up to an linear transformation of its basis.

$(b^\star)$

To show: Given the column span of A^n the mixing model is uniquely determined.

Proof. The idea behind the proof is that we can only solve the unmixing problem up to an multiplicative factor $\begin{pmatrix} M_1 & 0 \\ M_2 & M_4 \end{pmatrix}$. By solving the mixing problem with the column span A^n and some additionally chosen vectors instead of A^s, we also obtain a solution for the original mixing problem $A = [A^s \mid A^n]$. Given the column span A^n we can extend it to a basis of dimension D. Let $\tilde{A}^s$ denote the newly added basis vectors. $\tilde{A} = \left[\tilde{A}^s \mid A^n\right]$ is then used as mixing matrix:

$$x(t) \quad = \quad \tilde{A} \begin{pmatrix} s^s(t) \\ s^n(t) \end{pmatrix} \tag{32}$$

For this problem we find, we find a unmixing matrix $\hat{B}$ such that:

$$\hat{B}x(t) = \begin{pmatrix} M_1 & 0 \\ M_2 & M_4 \end{pmatrix} \cdot \begin{pmatrix} s^s(t) \\ s^n(t) \end{pmatrix} \tag{33}$$

By showing that $\hat{B}$ is also a solution for the original problem, we prove the equivalancy of the two problems and thus show that the mixing model is uniquely determined by its column span.

Since the columns of $\tilde{A}$ are a basis we can find a solution to the following problem:

$$A^s \quad = \quad \tilde{A}^s Q_1 + A^n Q_2 \tag{34}$$

And now

$$\begin{aligned}
\hat{B}A \quad &= \quad \begin{pmatrix} M_1 & 0 \\ M_2 & M_4 \end{pmatrix} \left[\frac{(\tilde{A}^s)^{-1}}{(A^n)^{-1}}\right] \cdot \left[\tilde{A}^s Q_1 + A^n Q_2 \mid A^n\right] && (35)\\
&= \quad \begin{pmatrix} M_1 & 0 \\ M_2 & M_4 \end{pmatrix} \cdot \begin{pmatrix} Q_1 & 0 \\ Q_2 & I \end{pmatrix} && (36)\\
&= \quad \begin{pmatrix} M_1 \cdot Q_1 & 0 \\ M_2 \cdot Q_1 + M_4 \cdot Q_2 & M_4 \end{pmatrix} && (37)\\
&= \quad \begin{pmatrix} \tilde{M}_1 & 0 \\ \tilde{M}_2 & M_4 \end{pmatrix} && (38)
\end{aligned}$$

Consequently, $\hat{B}$ is also a solution for the mixing matrix A and thus the problem is completely determined by the cloumn span of A^n. $\square$