A Variational EM Algorithm for the Separation of Time-Varying Convolutive Audio Mixtures

Notes on Forward Backward Algorithm

Kounades-Bastian D.
Girin L.
Alameda-Pineda X.
Horaud R.
Gannot S.

The Forward Backward Algorithm

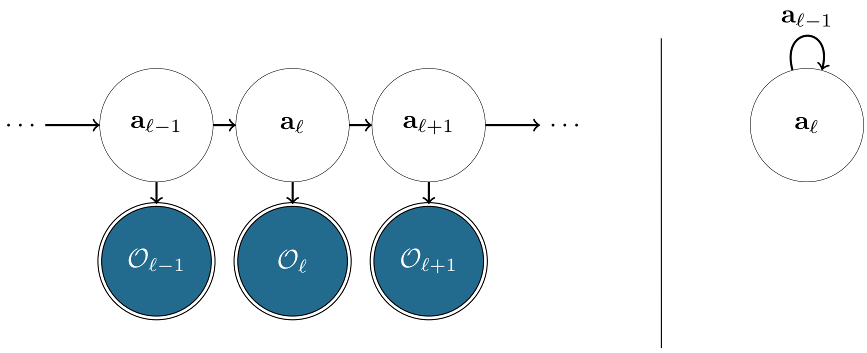

Figure 1: LEFT: Standard representation of a 1^{st} order LDS. RIGHT: The graphical abbreviation of a 1^{st} order LDS used in [1].

This report is supplementary material for [1], in essence it provides a detailed derivation of the E-A step.

The 1^{st} order Linear Dynamical System (LDS) in [1] has L hidden state vectors $\{\mathbf{a}_{:,f1}, .., \mathbf{a}_{:,f\ell}, .., \mathbf{a}_{:,fL}\}$ (for notation clarity we drop the index f and denote a hidden state $\mathbf{a}_{:,f\ell}$ as $\mathbf{a}_\ell$) with probability of the 1-st frame:[1]

$$p(\mathbf{a}_1) = \mathcal{N}_c\left(\mathbf{a}_1; \boldsymbol{\mu}^a, \boldsymbol{\Sigma}^a\right), \tag{1}$$

probability distribution of transition between succesive states:

$$p(\mathbf{a}_\ell|\mathbf{a}_{\ell-1}) = \mathcal{N}_c\left(\mathbf{a}_\ell; \mathbf{a}_{\ell-1}, \boldsymbol{\Sigma}^a\right). \tag{2}$$

and observation model

$$p\left(\mathcal{O}_\ell|\mathbf{a}_\ell\right) = \mathcal{N}_c\left(\boldsymbol{\mu}_\ell^{\iota a}; \mathbf{a}_\ell, \boldsymbol{\Sigma}_\ell^{\iota a}\right) \tag{3}$$

Note that $\mathcal{O}_l$ is an abstract notion of observation introduced only to enable us to define conditional probabilities over the hidden states. In cases where an LDS is embedded in a larger Bayesian network, as it is the case for [1], the observation probability of the LDS is calculated based on the surrounding latent variables and may not share a direct link with the observed data of the model that here are the STFT coefficients of the mixture $\mathbf{x}_{1:F1:L}$.

The forward-backward algorithm calculates the marginal posterior probability $p(\mathbf{a}_\ell|\mathcal{O}_{1:L})$ of a hidden state $\mathbf{a}_\ell$ conditioned on the sequence of all observations $\mathcal{O}_{1:L}$.[2] For any particular state $\mathbf{a}_\ell$ we have using Bayes:

$$p(\mathbf{a}_\ell|\mathcal{O}_{1:L}) \propto p(\mathbf{a}_\ell, \mathcal{O}_{1:L}) = \tag{4}$$

$$p(\mathcal{O}_{\ell+1:L}|\mathbf{a}_\ell, \cancel{\mathcal{O}_{1:\ell}})p(\mathbf{a}_{f\ell}, \mathcal{O}_{1:\ell}). \tag{5}$$

[1]The proper complex Gaussian distribution [2] is defined as $\mathcal{N}_c(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = |\pi\boldsymbol{\Sigma}|^{-1} \exp\big(-[\mathbf{x}-\boldsymbol{\mu}]^{\mathrm{H}}\boldsymbol{\Sigma}^{-1}[\mathbf{x}-\boldsymbol{\mu}]\big)$, with $\mathbf{x}, \boldsymbol{\mu} \in \mathbb{C}^I$ and $\boldsymbol{\Sigma} \in \mathbb{C}^{I\times I}$ being the argument, mean vector, and covariance matrix respectively.

[2]The notation $\mathcal{O}_{1:L}$ is an abbrevation for the set $\{\mathcal{O}_\ell\}_{\ell=1}^{L}$.

The past observations $\mathcal{O}_{1:\ell}$ are canceled from (5) via the Markov property (i.e. the future observations $\mathcal{O}_{\ell+1:L}$ are linked with the current hidden state $\mathbf{a}_\ell$ but are not linked with the past observations $\mathcal{O}_{1:\ell}$ as it is seen from Fig. 1).

Therefore the posterior probability of a hidden state is proportional to the product of two quantities. The forward probability $p(\mathbf{a}_\ell, \mathcal{O}_{1:\ell})$ and the backward probability $p(\mathcal{O}_{\ell+1:L}|\mathbf{a}_\ell)$.

The Forward Recursion

The forward probability can be developed further by un-marginalising the previous state $\mathbf{a}_{\ell-1}$:

$$p(\mathbf{a}_\ell, \mathcal{O}_{1:\ell}) = \int_{\mathbf{a}_{\ell-1}} p(\mathbf{a}_\ell, \mathbf{a}_{\ell-1}, \mathcal{O}_{1:\ell-1}, \mathcal{O}_\ell)\mathbf{d}\mathbf{a}_{\ell-1} = \tag{6}$$

$$\int_{\mathbf{a}_{\ell-1}} p(\mathcal{O}_\ell|\mathbf{a}_\ell, \cancel{\mathbf{a}_{\ell-1}}, \cancel{\mathcal{O}_{1:\ell-1}})p(\mathbf{a}_\ell, \mathbf{a}_{\ell-1}, \mathcal{O}_{1:\ell-1})\mathbf{d}\mathbf{a}_{\ell-1} = \tag{7}$$

$$p(\mathcal{O}_\ell|\mathbf{a}_\ell) \int_{\mathbf{a}_{\ell-1}} p(\mathbf{a}_\ell|\mathbf{a}_{\ell-1}, \cancel{\mathcal{O}_{1:\ell-1}})p(\mathbf{a}_{\ell-1}, \mathcal{O}_{1:\ell-1})\mathbf{d}\mathbf{a}_{\ell-1}. \tag{8}$$

where all cancelings emerge from independencies seen in Fig. 1. In summary the forward recursion calculates the forward probability $p(\mathbf{a}_\ell, \mathcal{O}_{1:\ell})$ using the forward probability for frame $\ell - 1$ that is $p(\mathbf{a}_{\ell-1}, \mathcal{O}_{1:\ell-1})$, with:

$$p(\mathbf{a}_\ell, \mathcal{O}_{1:\ell}) = p(\mathcal{O}_\ell|\mathbf{a}_\ell) \int_{\mathbf{a}_{\ell-1}} p(\mathbf{a}_\ell|\mathbf{a}_{\ell-1})p(\mathbf{a}_{\ell-1}, \mathcal{O}_{1:\ell-1})\mathbf{d}\mathbf{a}_{\ell-1}. \tag{9}$$

To obtain that $p(\mathbf{a}_\ell, \mathcal{O}_{1:\ell})$ we use proof by induction, that is we assume that for frame $\ell - 1$ the forward probability is complex-Gaussian with:

$$p(\mathbf{a}_{\ell-1}, \mathcal{O}_{1:\ell-1}) = \mathcal{N}_c\left(\mathbf{a}_{f\ell-1}, \boldsymbol{\mu}^\phi_{\ell-1}, \boldsymbol{\Sigma}^\phi_{\ell-1}\right), \tag{10}$$

and by replacing (10), (2), (3) in (9), combining the Gaussian distributions[3] and integrating[4] we derive the forward distribution $p(\mathbf{a}_\ell, \mathcal{O}_{1:\ell})$:

$$p(\mathbf{a}_\ell, \mathcal{O}_{1:\ell}) = \mathcal{N}_c\left(\mathbf{a}_\ell, \boldsymbol{\mu}^\phi_\ell, \boldsymbol{\Sigma}^\phi_\ell\right), \tag{11}$$

with covariance matrix $\boldsymbol{\Sigma}^\phi_\ell$ and mean vector $\boldsymbol{\mu}^\phi_\ell$ given respectively from eqs. (13) and (14) in [1].

---

[3]From Sec. 8.1.7 at [3] we have that the product of two Gaussian distributions $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = \left[\boldsymbol{\Sigma}_1^{-1} + \boldsymbol{\Sigma}_2^{-1}\right]^{-1}$ and $\boldsymbol{\mu} = \boldsymbol{\Sigma}\left[\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2\right]$.

[4]Eq. (2.115) in [4] states that $\int_{\mathbf{x}} \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}_1)\mathcal{N}(\mathbf{y}; \mathbf{x}, \boldsymbol{\Sigma}_2)\mathbf{d}\mathbf{x} = \mathcal{N}(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)$.

The Backward Recursion

The backward distribution $p(\mathcal{O}_{\ell+1:L}|\mathbf{a}_\ell)$ is developed by un-marginalising the future state $\mathbf{a}_{\ell+1}$:

$$p(\mathcal{O}_{\ell+1:L}|\mathbf{a}_\ell) = \int_{\mathbf{a}_{\ell+1}} p(\mathcal{O}_{\ell+1}, \mathcal{O}_{\ell+2:L}, \mathbf{a}_{\ell+1}|\mathbf{a}_\ell)\mathbf{da}_{\ell+1} = \tag{12}$$

$$\int_{\mathbf{a}_{\ell+1}} p(\mathcal{O}_{\ell+1}|\cancel{\mathcal{O}_{\ell+2:L}}, \mathbf{a}_{\ell+1}, \cancel{\mathbf{a}_\ell})p(\mathcal{O}_{\ell+2:L}, \mathbf{a}_{\ell+1}|\mathbf{a}_\ell)\mathbf{da}_{\ell+1} = \tag{13}$$

$$\int_{\mathbf{a}_{\ell+1}} p(\mathcal{O}_{\ell+1}|\mathbf{a}_{\ell+1})p(\mathcal{O}_{\ell+2:L}|\mathbf{a}_{\ell+1}, \cancel{\mathbf{a}_\ell})p(\mathbf{a}_{\ell+1}|\mathbf{a}_\ell)\mathbf{da}_{\ell+1}. \tag{14}$$

In summary the backward recursion calculates $p(\mathcal{O}_{\ell+1:L}|\mathbf{a}_\ell)$ using the next frame's backward $p(\mathcal{O}_{\ell+2:L}|\mathbf{a}_{\ell+1})$ with:

$$p(\mathcal{O}_{\ell+1:L}|\mathbf{a}_\ell) = \int_{\mathbf{a}_{\ell+1}} p(\mathcal{O}_{\ell+1}|\mathbf{a}_{\ell+1})p(\mathcal{O}_{\ell+2:L}|\mathbf{a}_{\ell+1})p(\mathbf{a}_{\ell+1}|\mathbf{a}_\ell)\mathbf{da}_{\ell+1}. \tag{15}$$

All distributions are known, except of $p(\mathcal{O}_{\ell+2:L}|\mathbf{a}_{\ell+1})$. We use again induction and assume that the backward probability of future frame $\ell+1$ is:

$$p(\mathcal{O}_{\ell+2:L}|\mathbf{a}_{\ell+1}) = \mathcal{N}_c\left(\mathbf{a}_{\ell+1}; \boldsymbol{\mu}^\beta_{\ell+1}, \boldsymbol{\Sigma}^\beta_{\ell+1}\right). \tag{16}$$

Then, replacing (16), (2), (3) in (15) combining and integrating as in the forward recursion we arrive that:

$$p(\mathcal{O}_{\ell+1:L}|\mathbf{a}_\ell) = \mathcal{N}_c\left(\mathbf{a}_\ell; \boldsymbol{\mu}^\beta_\ell, \boldsymbol{\Sigma}^\beta_\ell\right), \tag{17}$$

with covariance matrix $\boldsymbol{\Sigma}^\beta_\ell$ and mean vector $\boldsymbol{\mu}^\beta_\ell$ computed respectively with eqs. (16) and (17) in [1].

A Dummy Distribution

To simplify the notation in [1] we've introduced an intermediate distribution:

$$\mathcal{N}_c\left(\mathbf{a}_{\ell+1}; \boldsymbol{\mu}^\zeta_\ell, \boldsymbol{\Sigma}^\zeta_\ell\right) = p(\mathcal{O}_{\ell+1}|\mathbf{a}_{\ell+1})p(\mathcal{O}_{\ell+2:L}|\mathbf{a}_{\ell+1}), \tag{18}$$

with covariance matrix $\boldsymbol{\Sigma}^\zeta_\ell$ (eq. (15) of [1]) and mean vector $\boldsymbol{\mu}^\zeta_\ell$ (eq. (17) of [1]):

$$\boldsymbol{\Sigma}^\zeta_\ell = \left({\boldsymbol{\Sigma}^\beta_{\ell+1}}^{-1} + {\boldsymbol{\Sigma}^\iota_{\ell+1}}^{-1}\right)^{-1}, \tag{19}$$

$$\boldsymbol{\mu}^\zeta_\ell = \boldsymbol{\Sigma}^\zeta_\ell\left({\boldsymbol{\Sigma}^\iota_{\ell+1}}^{-1}\boldsymbol{\mu}^\iota_{\ell+1} + {\boldsymbol{\Sigma}^\beta_{\ell+1}}^{-1}\boldsymbol{\mu}^\beta_{\ell+1}\right). \tag{20}$$

Because the integration does not change the mean, we have $\boldsymbol{\mu}^\beta_\ell = \boldsymbol{\mu}^\zeta_\ell$. Note also that there are only $L-1$ ζ distributions, but there are L β distributions.

The Probability of Two Successive States

Eq.(21) of [1] is incorrect and is correctly derived here!

The joint posterior probability of two succesive states is defined in terms of the forward and backward probabilities in eq, (13.43) in [4]:

$$p(\mathbf{a}_{\ell+1}, \mathbf{a}_\ell | \mathcal{O}_{1:L}) \propto p(\mathcal{O}_{\ell+2:L}|\mathbf{a}_{\ell+1}) p(\mathcal{O}_{\ell+1}|\mathbf{a}_{\ell+1}) p(\mathbf{a}_{\ell+1}|\mathbf{a}_\ell) p(\mathbf{a}_\ell, \mathcal{O}_{1:\ell}) \propto \tag{21}$$

$$\underbrace{\mathcal{N}_c(\mathbf{a}_{\ell+1}; \boldsymbol{\mu}_\ell^\zeta, \boldsymbol{\Sigma}_\ell^\zeta)}_{1} \underbrace{\mathcal{N}_c(\mathbf{a}_{\ell+1}; \mathbf{a}_\ell, \boldsymbol{\Sigma}^a)}_{2} \underbrace{\mathcal{N}_c(\mathbf{a}_\ell; \boldsymbol{\mu}_\ell^\phi, \boldsymbol{\Sigma}_\ell^\phi)}_{3}. \tag{22}$$

In (22) both the $\mathbf{a}_\ell$ and $\mathbf{a}_{\ell+1}$ are free variables, hence the joint distribution is a function of the joint state variable $\mathbf{a}_{\{\ell+1,\ell\}} = \left[\mathbf{a}_{\ell+1}^\top, \mathbf{a}_\ell^\top\right]^\top \in \mathbb{C}^{2I}$.

To write the (22) as a single functional we must express all three factors: 1, 2, 3 in terms of the joint variable $\mathbf{a}_{\{\ell+1,\ell\}}$. In practice we re-structure the exponents (quadratic terms) of 1, 2, 3.

Restructuring Quadratic Forms

Let us start with 1. Discarding any terms that do not depend on $\mathbf{a}_{\ell+1}$ we have:

$$\log(1) = -\mathbf{a}_{\ell+1}^{\mathrm{H}} {\boldsymbol{\Sigma}_\ell^\zeta}^{-1} \mathbf{a}_{\ell+1} + \mathbf{a}_{\ell+1}^{\mathrm{H}} {\boldsymbol{\Sigma}_\ell^\zeta}^{-1} \boldsymbol{\mu}_\ell^\zeta + {\boldsymbol{\mu}_\ell^\zeta}^{\mathrm{H}} {\boldsymbol{\Sigma}_\ell^\zeta}^{-1} \mathbf{a}_{\ell+1} =$$

$$-\mathbf{a}_{\{\ell+1,\ell\}}^{\mathrm{H}} \begin{bmatrix} {\boldsymbol{\Sigma}_\ell^\zeta}^{-1} & \mathbf{0}_{I\times I} \\ \mathbf{0}_{I\times I} & \mathbf{0}_{I\times I} \end{bmatrix} \mathbf{a}_{\{\ell+1,\ell\}} + \mathbf{a}_{\{\ell+1,\ell\}}^{\mathrm{H}} \begin{bmatrix} {\boldsymbol{\Sigma}_\ell^\zeta}^{-1} \boldsymbol{\mu}_\ell^\zeta \\ \mathbf{0}_I \end{bmatrix} + \begin{bmatrix} {\boldsymbol{\Sigma}_\ell^\zeta}^{-1} \boldsymbol{\mu}_\ell^\zeta \\ \mathbf{0}_I \end{bmatrix}^{\mathrm{H}} \mathbf{a}_{\{\ell+1,\ell\}} \tag{23}$$

with $\mathbf{0}_I$ the zero vector of dimension I, and $\mathbf{0}_{I\times I}$ the zero matrix of dimension $I \times I$.

Again, for 2, and by discarding terms independent on both $\mathbf{a}_{\ell+1}, \mathbf{a}_\ell$:

$$\log(2) = -\left(\mathbf{a}_{\ell+1} - \mathbf{a}_\ell\right)^{\mathrm{H}} {\boldsymbol{\Sigma}^a}^{-1} \left(\mathbf{a}_{\ell+1} - \mathbf{a}_\ell\right) =$$

$$-\mathbf{a}_{\{\ell+1,\ell\}}^{\mathrm{H}} \begin{bmatrix} {\boldsymbol{\Sigma}^a}^{-1} & -{\boldsymbol{\Sigma}^a}^{-1} \\ -{\boldsymbol{\Sigma}^a}^{-1} & {\boldsymbol{\Sigma}^a}^{-1} \end{bmatrix} \mathbf{a}_{\{\ell+1,\ell\}}. \tag{24}$$

Then, for 3, discarding terms independent on $\mathbf{a}_\ell$ we write:

$$\log(3) = -\mathbf{a}_\ell^{\mathrm{H}} {\boldsymbol{\Sigma}_\ell^\phi}^{-1} \mathbf{a}_\ell + \mathbf{a}_\ell^{\mathrm{H}} {\boldsymbol{\Sigma}_\ell^\phi}^{-1} \boldsymbol{\mu}_\ell^\phi + {\boldsymbol{\mu}_\ell^\phi}^{\mathrm{H}} {\boldsymbol{\Sigma}_\ell^\phi}^{-1} \mathbf{a}_\ell =$$

$$-\mathbf{a}_{\{\ell+1,\ell\}}^{\mathrm{H}} \begin{bmatrix} \mathbf{0}_{I\times I} & \mathbf{0}_{I\times I} \\ \mathbf{0}_{I\times I} & {\boldsymbol{\Sigma}_\ell^\phi}^{-1} \end{bmatrix} \mathbf{a}_{\{\ell+1,\ell\}} + \mathbf{a}_{\{\ell+1,\ell\}}^{\mathrm{H}} \begin{bmatrix} \mathbf{0}_I \\ {\boldsymbol{\Sigma}_\ell^\phi}^{-1} \boldsymbol{\mu}_\ell^\phi \end{bmatrix} + \begin{bmatrix} \mathbf{0}_I \\ {\boldsymbol{\Sigma}_\ell^\phi}^{-1} \boldsymbol{\mu}_\ell^\phi \end{bmatrix}^{\mathrm{H}} \mathbf{a}_{\{\ell+1,\ell\}}. \tag{25}$$

Composition of the Final Quadratic

Now, by replacing (23), (24), (25) in (22) and summing we have:

$$p(\mathbf{a}_{\ell+1}, \mathbf{a}_\ell|\mathcal{O}_{1:L}) \propto \exp\Bigg(- \mathbf{a}^{\mathrm{H}}_{\{\ell+1,\ell\}} \begin{bmatrix} \boldsymbol{\Sigma}^{\zeta\,-1}_\ell + \boldsymbol{\Sigma}^{a\,-1} & -\boldsymbol{\Sigma}^{a\,-1} \\ -\boldsymbol{\Sigma}^{a\,-1} & \boldsymbol{\Sigma}^{a\,-1} + \boldsymbol{\Sigma}^{\phi\,-1}_\ell \end{bmatrix} \mathbf{a}^{\mathrm{H}}_{\{\ell+1,\ell\}} + \mathbf{a}^{\mathrm{H}}_{\{\ell+1,\ell\}} \begin{bmatrix} \boldsymbol{\Sigma}^{\zeta\,-1}_\ell \boldsymbol{\mu}^{\zeta}_\ell \\ \boldsymbol{\Sigma}^{\phi\,-1}_\ell \boldsymbol{\mu}^{\phi}_\ell \end{bmatrix} + \begin{bmatrix} \boldsymbol{\Sigma}^{\zeta\,-1}_\ell \boldsymbol{\mu}^{\zeta}_\ell \\ \boldsymbol{\Sigma}^{\phi\,-1}_\ell \boldsymbol{\mu}^{\phi}_\ell \end{bmatrix}^{\mathrm{H}} \mathbf{a}_{\{\ell+1,\ell\}} \Bigg). \tag{26}$$

The last remaining step is to identify that (26) is a complex-Gaussian

$$p(\mathbf{a}_{\ell+1}, \mathbf{a}_\ell|\mathcal{O}_{1:L}) = \mathcal{N}_c\left(\mathbf{a}_{\{\ell+1,\ell\}}; \boldsymbol{\mu}^{\xi}_\ell, \boldsymbol{\Sigma}^{\xi}_\ell\right), \tag{27}$$

with covariance matrix $\boldsymbol{\Sigma}^{\xi}_\ell$ and mean vector $\boldsymbol{\mu}^{\xi}_\ell$ given respectively with:

$$\boldsymbol{\Sigma}^{\xi}_\ell = \begin{bmatrix} \boldsymbol{\Sigma}^{\zeta\,-1}_\ell + \boldsymbol{\Sigma}^{a\,-1} & -\boldsymbol{\Sigma}^{a\,-1} \\ -\boldsymbol{\Sigma}^{a\,-1} & \boldsymbol{\Sigma}^{\phi a\,-1}_{f\ell} + \boldsymbol{\Sigma}^{a\,-1} \end{bmatrix}^{-1}, \tag{28}$$

$$\boldsymbol{\mu}^{\xi}_\ell = \boldsymbol{\Sigma}^{\xi}_\ell \left[\left(\boldsymbol{\Sigma}^{\zeta\,-1}_\ell \boldsymbol{\mu}^{\beta}_\ell\right)^{\top}, \left(\boldsymbol{\Sigma}^{\phi\,-1}_\ell \boldsymbol{\mu}^{\phi}_\ell\right)^{\top} \right]^{\top}. \tag{29}$$

Notice that in eq. (22) in [1] it is incorrectly written $\boldsymbol{\mu}^{\beta a}_{f\ell+1}$ where it should be $\boldsymbol{\mu}^{\beta a}_{f\ell}$.

Bibliography

[1] D. Kounades-Bastian, L. Girin, X. Alameda-Pineda, S. Gannot, and R. Horaud, "A variational EM algorithm for the separation of time-varying convolutive audio mixtures," *IEEE/ACM Trans. Audio, Speech and Language Process.*, vol. 24, no. 8, pp. 1408–1423, 2016.

[2] F. Neeser and J. Massey, "Proper complex random processes with applications to information theory," *IEEE Trans. Info. Theory*, vol. 39, no. 4, pp. 1293–1302, 1993.

[3] K. B. Petersen and M. S. Pedersen, *The matrix cookbook.* Version. Nov 15, 2012.

[4] C. Bishop, *Pattern Recognition and Machine Learning.* Springer, 2006.