ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0. Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

Non-Euclidean Independent Component Analysis and Oja's Learning

M. Lange[1], M. Biehl[2], and T. Villmann[1]

1- University of Appl. Sciences Mittweida - Dept. of Mathematics
Mittweida, Saxonia - Germany

2- University Groningen - J.-Bernoulli-Inst. of Mathematics and Computer Sciences
Groningen, The Netherlands

Abstract. In the present contribution we tackle the problem of nonlinear independent component analysis by non-Euclidean Hebbian-like learning. Independent component analysis (ICA) and blind source separation originally were introduced as tools for the linear unmixing of the signals to detect the underlying sources. Hebbian methods became very popular and succesfully in this context. Many nonlinear ICA extensions are known. A promising strategy is the application of kernel mapping. Kernel mapping realizes an usually nonlinear but implicite data mapping of the data into a reproducing kernel Hilbert space. After that a linear demixing can be carried out there. However, explicit handling in this non-Euclidean kernel mapping space is impossible. We show in this paper an alternative using an isomorphic mapping space. In particular, we show that the idea of Hebbian-like learning of *kernel* ICA can be transferred to this non-Euclidean space realizing an non-Euclidean ICA.

1 Introduction

Independent component analysis (ICA) and blind source separation (BSS) constitute a paradigm to extract independent sources from a sequence of mixtures [5, 10, 15]. It can be seen as a generalization of principal component analysis (PCA). The generic problem consists in separating useful independent signals from noise and interferences [6]. Originally, respective methods were considered as linear models. There exists a whole bunch of methods to tackle this problem. They differ mainly on the *a priori* assumptions about the underlying mixing model. One of the key principles of ICA is 'non-Gaussianity' as an equivalent of independence together with the central limit theorem (CLT,[13]) [5, 10]. Different approaches were developed to contrast this property: approximations of the negentropy [19] were developed in [10, 14], mutual information and the infomax principle were considered in [4, 22, 25, 26, 9]. ICA and projection pursuit is addressed in [14]. A large family of Hebbian-like learning algorithms for ICA utilize implicitly the *kurtosis* as a contrast function [12, 11, 17, 18].

Nonlinear approaches of ICA and BSS are natural generalizations of the linear approaches and were addressed in [12, 16]. Several investigations consider kernel methods to deal with nonlinear separation [2, 7, 20]. However, these models are not based on Hebbian-like learning ICA. Recently, kernel PCA has been studied in terms of Hebbian learning by a kernelized variant of Oja-learning realizing a non-Euclidean PCA [3]. We transfer this idea to Hebbian-like ICA in the present contribution and provide the theoretical basis. We demonstrate the abilities of this new approach for exemplary source separation problems.

The paper is organized as follows: First, we briefly review ICA based on kurtosis contrast. Then, we reconsider the Hebbian-like algorithms for learning independent components. Finally we introduce our kernel based approach illustrate it in terms of an application example..

2 Linear Independent Component Analysis by Hebbian-like Learning Using the Kurtosis

As mentioned in the introduction, one key principle of ICA is to estimate independence by non-Gaussianity. In this context we consider sequences of n-dimensional mixture vectors $\mathbf{v}(t) \in V_{d_E} \subseteq \mathbb{R}^n$ with the Euclidean distance d_E. We assume that a pre-whitening $\mathbf{x}(t) = \mathbf{P}\mathbf{v}(t)$ takes place such that the expectation becomes $E\left[\mathbf{x}\mathbf{x}^\top\right] = \mathbf{I}$. Further, we suppose a linear mixing model

$$\mathbf{v}(t) = \mathbf{A}\mathbf{s}(t) \tag{1}$$

such that $\mathbf{x}(t) = \mathbf{M}\mathbf{s}(t)$ is valid with $\mathbf{M} = \mathbf{PA}$. It turns out that $\mathbf{M}$ has to be orthonormal to ensure $E\left[\mathbf{x}\mathbf{x}^\top\right] = \mathbf{I}$. Let

$$s_i = \left\langle \mathbf{m}_i^\top, \mathbf{x} \right\rangle = \sum_{j=1}^{n} m_{i,j} x_j \tag{2}$$

be the ith source where $\mathbf{m}_i$ is a column vector of $\mathbf{M}$. Here x_j are stochastic quantities such that the central limit theorem (CLT) is valid, i.e. the quantity s_i is more Gaussian than the single summands. Thus, ICA can be performed by maximization of the absolute value of the *kurtosis* $\mathtt{kurt}(s_i)$ as a measure of Non-Gaussianity. The kurtosis is defined by $\mathtt{kurt}(y) = E\left[y^4\right] - 3\left(E\left[y^2\right]\right)^2$ using the fourth and the second moments $E\left[y^4\right]$ and $E\left[y^2\right]$, respectively. If we consider $\mathbf{w} = \mathbf{m}_i$, Hebbian-like learning interprets the vector $\mathbf{w}$ as a weight vector of a linear perceptron, which is trained by a sequence $\mathbf{x}(t)$ of input vectors [8, 23]. A it was shown in [11, 12, 17, 24], Hebbian-like ICA-learning can be achieved applying the weight vector dynamic

$$\mathbf{w}(t+1) = \mathbf{w}(t) + \varepsilon_t \left[-\sigma \cdot \mathbf{x}(t) \cdot g\left(\langle \mathbf{w}(t), \mathbf{x}(t)\rangle\right) - f\left(\langle \mathbf{w}, \mathbf{x}\rangle^2\right)\mathbf{w}(t)\right] \tag{3}$$

with a nonlinear modulation function $g(y) = ay - by^3$. The function f is a non-vanishing scalar and $\sigma = \pm 1$ is a sign that determines whether we are minimizing ($\sigma = -1$) or maximizing ($\sigma = +1$) the kurtosis. The constants are chosen to be $a \geq 0$ and $b > 0$. The maximization of the positive kurtosis is obtained for $a = 0$, the optimization of the negative kurtosis requires $a > 0$ [11]. The value $0 < \varepsilon_t \ll 1$ is the learning rate. The term $\mathbf{x}(t) \cdot g\left(\langle \mathbf{w}(t), \mathbf{x}(t)\rangle\right)$ reflects the Hebb-signal-enhancing idea with the perceptron output learning function $g\left(\langle \mathbf{w}(t), \mathbf{x}(t)\rangle\right)$ whereas $f\left(E\left[\langle \mathbf{w}, \mathbf{x}\rangle^2\right]\right)\mathbf{w}(t)$ defines a constraint term to prevent $\mathbf{w}(t)$ from infinite growing. The learning rule (3) performs a stochastic gradient descent on the cost function

$$J(\mathbf{w}) = \sigma\left(-\frac{a}{2} \cdot E\left[\langle \mathbf{w}, \mathbf{x}\rangle^2\right] - \frac{b}{4} \cdot E\left[\langle \mathbf{w}, \mathbf{x}\rangle^4\right]\right) + F\left(E\left[\langle \mathbf{w}, \mathbf{x}\rangle^2\right]\right) \tag{4}$$

with $F'(y) = f(y)$.

3 Nonlinear Kernel Independent Component Analysis by Hebbian-like Learning

Kernel ICA and BSS make use of the nonlinear kernel mapping to perform a nonlinear ICA $\mathbf{v}(t) = h(\mathbf{s}(t))$ for an unknown non-linear function h. Several approaches are discussed [2, 7, 20]. We concentrate on the approach presented by Harmeling et al. in [7]. We consider a generally *nonlinear* kernel data map

$$\Phi : V \ni \mathbf{v} \longmapsto \Phi(\mathbf{v}) \in \mathcal{H} \tag{5}$$

with a positive definite kernel

$$\kappa_\Phi(\mathbf{v}, \mathbf{w}) = \langle \Phi(\mathbf{v}), \Phi(\mathbf{w}) \rangle_{\mathcal{H}} \tag{6}$$

for all $\mathbf{v}, \mathbf{w} \in V$ and $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ such that the space $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ *uniquely* corresponding the kernel to a reproducing kernel κ_Φ a canonical manner [1, 21]. The norm $\|\Phi(\mathbf{v})\|_{\mathcal{H}} = \sqrt{\kappa_\Phi(\mathbf{\Phi}(\mathbf{v}), \mathbf{\Phi}(\mathbf{v}))}$ of this RKHS induces a metric $d_{\mathcal{H}}(\Phi(\mathbf{v}), \Phi(\mathbf{w})) = \sqrt{\kappa_\Phi(\mathbf{v}, \mathbf{v}) - 2\kappa_\Phi(\mathbf{v}, \mathbf{w}) + \kappa_\Phi(\mathbf{w}, \mathbf{w})}$ based on the kernel κ_Φ [27]. Steinwart has shown that continuous, universal kernels induce the continuity[1] and separability of the corresponding feature map Φ and the image $\mathcal{I}_{\kappa_\Phi} = span(\Phi(V))$ is a subspace of $\mathcal{H}$ [28]. Let $\mathbf{b}_i$ form an orthonormal basis in $\mathcal{I}_{\kappa_\Phi}$. Then

$$\Phi(\mathbf{v}) = \sum_i \langle \Phi(\mathbf{v}), \mathbf{b}_i \rangle_{\mathcal{H}} \cdot \mathbf{b}_i \tag{7}$$

is the representation of an image vector $\Phi(\mathbf{v})$ in $\mathcal{I}_{\kappa_\Phi}$. In analogy to the linear ICA/BSS we consider now the linear mixing problem in the Hilbert space $\mathcal{H}$:

$$\Phi(\mathbf{v}) = \mathbf{M}_{\mathcal{H}}[\Phi(\mathbf{s})] \tag{8}$$

with $\mathbf{M}_{\mathcal{H}}$ being a linear operator in $\mathcal{H}$. We denote by $\mathbf{M}^k_{\mathcal{H}}$ the kth component of $\mathbf{M}_{\mathcal{H}}$. Because $\mathcal{H}$ is a RKHS, each linear operator can be expressed in terms of the inner product, i.e.

$$\mathbf{M}^k_{\mathcal{H}}[\Phi(\mathbf{s})] = \langle \mathbf{M}^k_{\mathcal{H}}, \Phi(\mathbf{s}) \rangle_{\mathcal{H}}. \tag{9}$$

Using the basis representation form(7) we obtain

$$\Phi_k(\mathbf{v}) = \langle \mathbf{M}^k_{\mathcal{H}}, \sum_i \langle \Phi(\mathbf{s}), \mathbf{b}_i \rangle_{\mathcal{H}} \cdot \mathbf{b}_i \rangle_{\mathcal{H}} = \sum_i \langle \Phi(\mathbf{s}), \mathbf{b}_i \rangle_{\mathcal{H}} \cdot \langle \mathbf{M}^k_{\mathcal{H}}, \mathbf{b}_i \rangle_{\mathcal{H}} \,. \tag{10}$$

We remark at this point that $\langle \Phi(\mathbf{s}), \mathbf{b}_i \rangle_{\mathcal{H}}$ is a random variable because of the stochastic character of $\mathbf{s}$ and, hence, $\Phi(\mathbf{s})$ is random too. Therefore, we can state that the sum in (10) is more Gaussian than the single components according to the CLT. In consequence we can take $\Phi_k(\mathbf{v})$ as a quantity the absolute kurtosis of which has to be maximized for separating independent components in $\mathcal{I}_{\kappa_\Phi}$ by the same arguments as for linear ICA.

[1] Note that continuity ensures the existence of the inverse mapping Φ^{-1}.

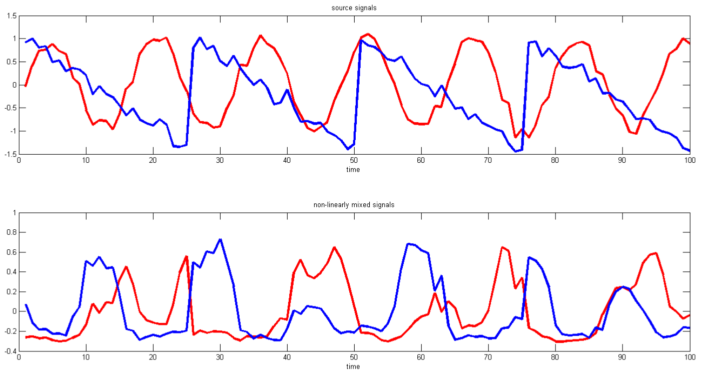

Figure 1: Visualization of the original signals (top) and the mixed sources (bottom).

For a finite number N of samples in V the number of basis elements $\mathbf{b}_i$ is $D \leq \min(n, N)$. In that case we can simplify (10) to

$$\Phi_k(\mathbf{v}) = D \cdot \kappa_\Phi(\Phi^{-1}\left(\mathbf{M}^k_\mathcal{H}\right), \mathbf{s}) \tag{11}$$

using the reproducing property of the inner product of a RKHS and keeping in mind definition (6) of the kernel.

In the last step of our argument we replace the kernel map Φ by

$$\Psi : V_{d_E} \longrightarrow V_{d_{\kappa_\Phi}} \tag{12}$$

with $d_{\kappa_\Phi}(\mathbf{v}, \mathbf{w}) = d_\mathcal{H}(\Phi(\mathbf{v}), \Phi(\mathbf{w}))$. For universal continuous kernels $V_{d_{\kappa_\Phi}}$ is a compact vector space with the kernel induced metric d_{κ_Φ}. Note that Ψ is formally the identity map but changing the metric, and, hence, generally nonlinear. Moreover, it turns out that the kernel space $V_{d_{\kappa_\Phi}}$ is isometric and isomorphic to $\mathcal{I}_{\kappa_\Phi}$ [29]. Under this assumption, the operator $\mathbf{M}_\mathcal{H}$ is equivalent to a conventional matrix $\mathbf{M}$ but $\mathbf{M}[\Psi(\mathbf{s})]$ is defined by

$$\mathbf{M}^k[\Psi(\mathbf{s})] = D \cdot \kappa_\Phi(\mathbf{m}_k, \mathbf{s}) \tag{13}$$

where $\mathbf{m}_k$ is the kth row vector of $\mathbf{M}$. Hence, we can replace the Euclidean inner product in (3) by $\kappa_\Phi(\mathbf{m}_k, \mathbf{s})$ to obtain an ICA in $V_{d_{\kappa_\Phi}}$ as a Hebbian-like *kernel* ICA (kICA) based of the original data, which realizes a non-linear de-mixing because of the non-linear kernel mapping Φ or its analagon Ψ.

4 Exemplary Simulations

In an illustrative simulation we consider a non-linear mixture of two signals obtained by linear mixing in the kernel space and subsequent back transformation into the data space. The sources and the mixed signals are depicted in Fig. 1. Then we applied the original linear Hebbian-like ICA according to (3). The kurtosis was estimated using the online learning

$$\mu_4(t+1) = (1-\xi)\,\mu_4(t) + \xi\,\langle \mathbf{w}(t), \mathbf{x}(t)\rangle^4 \tag{14}$$

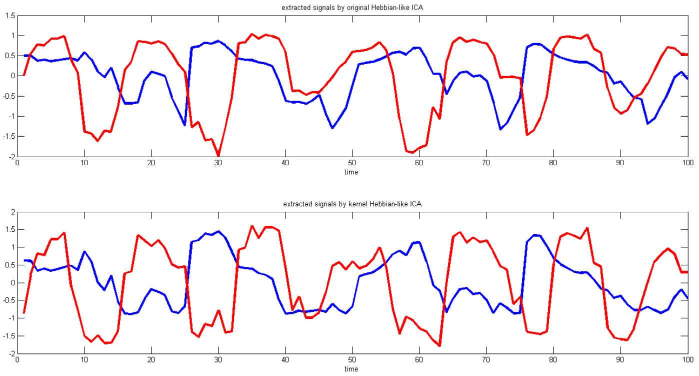

Figure 2: Estimated source signals by original Hebbian-like ICA based on the Euclidean inner product (top) and by the kernelized variant (bottom).

suggested in [11] with an averaging parameter of $\xi = 0.05$. In comparison we estimated the source signals by the kernelized variant replacing the inner product $\langle \mathbf{w}(t), \mathbf{x}(t) \rangle$ by the Gaussian kernel $G(\mathbf{w}(t), \mathbf{x}(t))$. The results of both approaches are visualized in Fig. 2. We observe that the kernelized variant is better able to de-mix the signals and to reconstruct the original signals than the original Euclidean approach. The obtained integrated squared errors oft the linear model are $e^1_{ICA} = 22.3$ and $e^2_{ICA} = 55.0$ for the two signals, respectively, whereas the kICA yields $e^1_{kICA} = 23.1$ and $e^2_{kICA} = 34.8$. The respective correlation coefficients are $\rho^1_{ICA} = 0.86$ and $\rho^2_{ICA} = 0.81$ for the linear model. The kICA yields $\rho^1_{kICA} = 0.88$ and $\rho^2_{kICA} = 0.84$. The improved performance of kICA is due to its non-linear character implicitly realized by the kernel trick. However, the kICA is very sensitive.

5 Conclusion

In this paper we introduced a kernelized variant of the original Hebbian-like ICA proposed by Hyvärinen&Oja. We showed that the Euclidean inner product in the original approach can be replaced by an universal and continuous kernel with an appropriate interpretation in the resulting generally non Euclidean kernelized data space. Thus, a non-linear demixing can be realized. We demonstrated in an exemplary application that this non-Euclidean variant of Hebbian-like ICA is able to extract non-linearly mixed signals, however, it is difficult to stabilize the model.

References

[1] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950.

[2] F. Bach and M. Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3:1–48, 2002.

[3] M. Biehl, M. Kästner, M. Lange, and T. Villmann. Non-euclidean principal component analysis and Oja's learning rule – theoretical aspects. In P. Estevez, J. Principe, and P. Zegers, editors, *Advances in Self-Organizing Maps: 9th International Workshop*

WSOM 2012 Santiage de Chile, volume 198 of *Advances in Intelligent Systems and Computing*, pages 23–34, Berlin, 2012. Springer.

[4] J.-F. Cardoso. Infomax and maximum likelihood for source separation. *IEEE Letters on Signal Processing*, 4:112–114, 1997.

[5] A. Cichocki and S. Amari. *Adaptive Blind Signal and Image Processing*. John Wiley, 2002.

[6] P. Comon and C. Jutten. *Handbook of Blind Source Separation*. Academic Press, 2010.

[7] S. Harmeling, A. Ziehe, and M. K. and. K.-R. Müller. Kernel-based nonlinear blind source separation. *Neural Computation*, 15(5):1089–1124, 2003.

[8] D. Hebb. *The Organization of Behavior. A Neuropsychological Theory*. John Wiley, New York, 1949.

[9] A. Hyvärinen. The fixed-point algorithm and maximum likelihood estimation for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3):626–634, 1999.

[10] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. J. Wiley & Sons, 2001.

[11] A. Hyvärinen and E. Oja. Simple neuron models for independent component analysis. *International Journal of Neural Systems*, 7(6):671–687, 1996.

[12] A. Hyvärinen and E. Oja. Independent component analysis by general nonlinear hebbian-like learning rules. *Signal Processing*, 64:301–313, 1998.

[13] I. Joliffe. *Principal Component Analysis*. Springer, 2nd edition, 2002.

[14] M. Jones and R. Sibson. What is projection pursuit? *Journal of the Royal Statistical Society, Series A*, 150:1–36, 1987.

[15] C. Jutten and J. Hérault. Blind separation of sources, Part I: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24:1–10, 1991.

[16] C. Jutten and J. Karhunen. Advances in blind source separation (BSS) and independent component analysis (ICA) for nonlinear mixtures. *International Journal of Neural Systems*, 14(5):267–292, 2004.

[17] J. Karhunen, E. Oja, L. Wang, R. Vigário, and J. Joutsensalo. A class of neural networks for independnet component analysis. *IEEE Transactions on Neural Networks*, 8(3):486–504, 1997.

[18] J. Karhunen, P. Pajunen, and E. Oja. The nonlinear PCA criterion in blind source separation: Relations with other approaches. *Neurocomputing*, 22:5–20, 1998.

[19] D. Mackay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2003.

[20] D. Martinez and A. Bray. Nonlinear blind source separation using kernels. *IEEE Transactions on Neural Networks*, 14(1):228–235, 2003.

[21] J. Mercer. Functions of positive and negative type and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society, London, A*, 209:415–446, 1909.

[22] J.-P. Nadal and N. Parga. Non-linear neurons in the low noise limit: a factorial code maximizes information transfer. *Netw*, 5:565–581, 1994.

[23] E. Oja. Neural networks, principle components and subspaces. *International Journal of Neural Systems*, 1:61–68, 1989.

[24] E. Oja. The nonlinear PCA learning rule in independent component analysis. *Neurocomputing*, 17:25–45, 1997.

[25] D. Pham. Mutual information approach to blind separation of stationary sources. *IEEE Transactions on Information Theory*, 48:1935–1946, 2002.

[26] D.-T. Pham, P. Garat, and C. Jutten. Separation of a mixture of independent sources through a maximum likelihood approach. In J. Vandewalle, R. Boite, M. Moonen, and A. Oosterlinck, editors, *Signal Processing VI: Theories and Applications (EUSIPCO)*, pages 771–774, 1997.

[27] B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, 2002.

[28] I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2:67–93, 2001.

[29] T. Villmann and S. Haase. A note on gradient based learning in vector quantization using differentiable kernels for Hilbert and Banach spaces. *Machine Learning Reports*, 6(MLR-02-2012):1–29, 2012. ISSN:1865-3960, http://www.techfak.uni-bielefeld.de/~fschleif/mlr/mlr_02_2012.pdf.