Cognitive Sciences EPrint Archive (CogPrints) 6223, (2008)
http://cogprints.org/6223/

Understanding Slow Feature Analysis: A Mathematical Framework

Henning Sprekeler[*] H.SPREKELER@BIOLOGIE.HU-BERLIN.DE

Laurenz Wiskott L.WISKOTT@BIOLOGIE.HU-BERLIN.DE

Institute for Theoretical Biology
and Bernstein Center for Computational Neuroscience Berlin
Humboldt University Berlin
Unter den Linden 6
10099 Berlin, Germany

[*] *corresponding author; now at: Laboratory for Computational Neuroscience, École Polytechnique Fédérale de Lausanne, Station 15, 1015 Lausanne, Switzerland*

Abstract

Slow feature analysis is an algorithm for unsupervised learning of invariant representations from data with temporal correlations. Here, we present a mathematical analysis of slow feature analysis for the case where the input-output functions are not restricted in complexity. We show that the optimal functions obey a partial differential eigenvalue problem of a type that is common in theoretical physics. This analogy allows the transfer of mathematical techniques and intuitions from physics to concrete applications of slow feature analysis, thereby providing the means for analytical predictions and a better understanding of simulation results. We put particular emphasis on the situation where the input data are generated from a set of statistically independent sources. The dependence of the optimal functions on the sources is calculated analytically for the cases where the sources have Gaussian or uniform distribution.

Keywords: slow feature analysis, unsupervised learning, invariant representations, statistically independent sources, theoretical analysis

1. Introduction

Reliable recognition of objects in spite of changes in position, size, or illumination is a problem that, although highly relevant for computer vision, has not been solved in a general manner. The key problem is to establish representations of the data that are invariant with respect to typical changes in the appearance of the objects while remaining selective for object identity.

One approach for the unsupervised learning of such invariant representations is based on the exploitation of temporal correlations in the training data. The basic idea is that representations that are invariant with respect to typical transformations in a given set of time-dependent training data should lead to temporally slowly varying output signals. Using this observation as a rational, it is possible to learn invariant representations by favoring representations that generate slowly varying signals over others that generate quickly varying signals. There are several implementations of this principle, with approaches ranging from gradient descent (Stone and Bray, 1995; Kayser et al., 2001), over temporally nonlocal Hebbian learning (Mitchison, 1991; Földiàk, 1991; Wallis and Rolls, 1997; Sprekeler et al., 2007) to batch learning (Wiskott and Sejnowski, 2002). Here, we focus on Slow Feature Analysis (SFA) as introduced by Wiskott (1998).

SFA has been applied to a set of problems, ranging from biological modeling of receptive fields in early visual processing (Berkes and Wiskott, 2005) and hippocampal place and head direction cells (Franzius et al., 2007a) to technical applications such as invariant object recognition (Berkes, 2005; Franzius et al., 2007b). On the conceptual side, it has turned out that SFA is closely related to independent component analysis techniques that rely on second order statistics (Blaschke et al., 2006). SFA or variations thereof can therefore also be used for problems of blind source separation (Blaschke et al., 2007).

Previous studies have shown that SFA is amenable to analytical considerations (Wiskott, 2003; Franzius et al., 2007a). Here, we extend this line of research and present a mathematical framework for the case where SFA has access to an unlimited function space. The aim of this article is threefold. Firstly, the mathematical framework allows us to make analytical predictions for the input-output functions found by SFA in concrete applications. Parts of the theory have been presented and used for this purpose earlier (Franzius et al., 2007a). Secondly, the theory shows that SFA bears close analogies to standard systems in physics, which provide an intuitive interpretation of the functions found by SFA. Thirdly, the theory makes specific predictions for the case where the input data are generated from a set of statistically independent sources.

The structure of the paper is as follows. In section 2 we introduce the optimization problem that underlies SFA. In sections 3 and 4 we develop the mathematical framework and study the case of statistically independent sources. In section 5, we briefly discuss analogies to standard systems in physics and try to convey an intuitive understanding of the solutions of SFA based on these analogies.

2. Slow Feature Analysis

Slow Feature Analysis is based on the following learning task: Given a multidimensional input signal we want to find scalar input-output functions that generate output signals that vary as slowly as possible but carry significant information. To ensure the latter we require the output signals to be uncorrelated and have unit variance. In mathematical terms, this can be stated as follows:

Optimization problem 1: *Given a function space $\mathcal{F}$ and an N-dimensional input signal $\mathbf{x}(t)$ find a set of J real-valued input-output functions $g_j(\mathbf{x})$ such that the output signals $y_j(t) := g_j(\mathbf{x}(t))$ minimize*

$$\Delta(y_j) = \langle \dot{y}_j^2 \rangle_t \tag{1}$$

under the constraints

$$\begin{aligned} \langle y_j \rangle_t &= 0 \quad \textit{(zero mean)}, & (2)\\ \langle y_j^2 \rangle_t &= 1 \quad \textit{(unit variance)}, & (3)\\ \forall i < j : \langle y_i y_j \rangle_t &= 0 \quad \textit{(decorrelation and order)}, & (4) \end{aligned}$$

with $\langle \cdot \rangle_t$ and $\dot{y}$ indicating temporal averaging and the derivative of y, respectively.

Equation (1) introduces the Δ-value, which is a measure of the temporal slowness (or rather 'fastness') of the signal $y(t)$. The constraints (2) and (3) avoid the trivial constant solution. Constraint (4) ensures that different functions g_j code for different aspects of the input. Note that the decorrelation constraint is asymmetric: The function g_1 is the slowest function in $\mathcal{F}$, while the function g_2 is the slowest function that fulfills the constraint of generating a signal that is uncorrelated to the output signal of g_1. Therefore, the resulting sequence of functions is ordered according to the slowness of their output signals on the training data.

It is important to note that although the objective is the slowness of the output signal, the functions g_j are instantaneous functions of the input, so that slowness cannot be achieved by low-pass filtering. Slow output signals can only be obtained if the input signal contains slowly varying features that can be extracted by the functions g_j.

Depending on the dimensionality of the function space $\mathcal{F}$, the solution of the optimization problem requires different techniques. If $\mathcal{F}$ is finite-dimensional, the problem can be reduced to a (generalized) eigenvalue problem (Wiskott and Sejnowski, 2002; Berkes and Wiskott, 2005). If $\mathcal{F}$ is infinite-dimensional, the problem requires variational calculus and is in general difficult to solve. Here, we consider this second case for the special situation that there are no constraints on $\mathcal{F}$ apart from sufficient differentiability and integrability. Although strictly speaking this case cannot be implemented numerically, it has the advantage that it permits the analytical derivation of partial differential equations for the optimal functions and predictions for the behavior of systems that implement very high-dimensional function spaces such as hierarchical systems (Franzius et al., 2007a). It also yields an intution as to how the structure of the input signal is reflected by the optimal solutions.

3. A Mathematical Framework for SFA

In this section, we present a rigorous mathematical framework for SFA for the case of an unrestricted function space $\mathcal{F}$. The key results are that the output signals extracted by SFA are independent of the representation of the input signals and that the optimal functions for SFA are the solutions of a partial differential eigenvalue problem.

3.1 Representations of the input signals

The assumption that SFA has access to an unrestricted function space $\mathcal{F}$ has important theoretical implications. For restricted (but possibly still infinitely-dimensional) function spaces, coordinate changes in the space of the input data generally alter the results, because they effectively change the function space from which the solutions are taken. As an example, assume that the input signal $\mathbf{x} = (x, y)$ is two-dimensional and the function space is the space of linear functions. Then, a change of the coordinate system to $(x', y') = (x^3, y)$ if still allowing only linear functions in the new coordinates leads to a very different function space in the variables x and y. Thus the optimal functions generate different optimal output signals y_j for the different coordinate systems. The optimization problem with a restricted function space is generally not invariant with respect to coordinate changes of the input signals.

For an unrestricted function space, the situation is different, because the concatenation of any function in $\mathcal{F}$ with the inversion of the coordinate change is again an element of the function space. The set of output signals that can be generated by the function space is then invariant with respect to invertible coordinate changes of the input signals. Because the slowness of a function is measured in terms of its output signal, the optimal functions will of course depend on the coordinate system used, but the output signals will be the same.

This is particularly interesting in situations where the high-dimensional input signal does not cover the whole space of possible values, but lies on a low-dimensional manifold. For illustration, consider the example of a video showing a single rotating object. In this case, the set of possible images can be parameterized by three angles that characterize the orientation of the object in space. Therefore, these images lie on a three-dimensional manifold in the space of all possible images. Because there are no data *outside* the input manifold, we are generally only interested in the behavior of the optimal functions *within* the input manifold, that is, in the reaction of the system to all images that are possible within the given training scenario. The equivalence of different coordinate systems then implies that it is not important whether we take the (high-dimensional) video sequence or the (3-dimensional) time-dependent abstract angles as input signals. The output signal is the same. Of course the low-dimensional representation is much more amenable to analytical predictions and to intuitive interpretations of the system behavior. We have previously used this simplification to predict the behavior of a hierarchical model of visual processing that reproduces the behavior of several cell types in the hippocampal formation of rodents commonly associated with spatial navigation (Franzius et al., 2007a).

Another situation in which the coordinate invariance is useful is the case of nonlinear blind source separation. Here, the input data are assumed to be a nonlinear mixture of some underlying sources. The task is to reconstruct the sources from the data without knowledge of the mixture or the sources. A natural prerequisite for the reconstruction is that the mixture is invertible. The mixture can then be interpreted as a nonlinear coordinate change, which – due to the equivalence of different coordinate systems – is immaterial to the optimization problem above. From the theoretical perspective, we can thus simply assume that we had the sources themselves as input signals and try to make predictions about how they are mixed in the optimal output signals found by SFA. If we can infer the sources (or good representatives thereof) from the optimal output signals under this condition, we can infer the sources from the output signals, no matter how they are mixed in the input data. Thus, SFA may be an interesting way of solving certain nonlinear blind source separation problems.

It is important to bear in mind that the theory developed in the following is valid for an arbitrary choice of the input coordinate system, so that $\mathbf{x}(t)$ can stand for concrete input signals (e.g. video sequences) as well as abstract representations of the input (e.g. angles that denote the orientation of the object in the video). Note however, that as the input data (or the manifold they lie on) becomes very high-dimensional, the resulting equations may be tedious to solve.

3.2 Further assumptions and notation

We assume that the input signal $\mathbf{x}(t)$ is ergodic, so that we can replace time averages by ensemble averages with a suitable probability density. Because the optimization problem underlying SFA relies on the temporal structure of the training data as reflected by its derivative, a statistical description of the training signal $\mathbf{x}$ must incorporate not only the probability distribution for the values of $\mathbf{x}$, but rather the joint distribution $p_{\mathbf{x},\dot{\mathbf{x}}}(\mathbf{x},\dot{\mathbf{x}})$ of the input signal $\mathbf{x}$ and its derivative $\dot{\mathbf{x}}$. We assume that $p_{\mathbf{x},\dot{\mathbf{x}}}(\mathbf{x},\dot{\mathbf{x}})$ is known and that we can define the marginal and conditional probability densities

$$p_{\mathbf{x}}(\mathbf{x}) := \int p_{\mathbf{x},\dot{\mathbf{x}}}(\mathbf{x},\dot{\mathbf{x}})\,\mathrm{d}^N\dot{x}\,, \tag{5}$$

$$p_{\dot{\mathbf{x}}|\mathbf{x}}(\dot{\mathbf{x}}|\mathbf{x}) := \frac{p_{\mathbf{x},\dot{\mathbf{x}}}(\mathbf{x},\dot{\mathbf{x}})}{p_{\mathbf{x}}(\mathbf{x})}\,, \tag{6}$$

and the corresponding averages

$$\langle f(\mathbf{x},\dot{\mathbf{x}})\rangle_{\mathbf{x},\dot{\mathbf{x}}} := \int p_{\mathbf{x},\dot{\mathbf{x}}}(\mathbf{x},\dot{\mathbf{x}})f(\mathbf{x},\dot{\mathbf{x}})\,\mathrm{d}^N x\,\mathrm{d}^N\dot{x}\,, \tag{7}$$

$$\langle f(\mathbf{x})\rangle_{\mathbf{x}} := \int p_{\mathbf{x}}(\mathbf{x})f(\mathbf{x})\,\mathrm{d}^N x\,, \tag{8}$$

$$\langle f(\mathbf{x},\dot{\mathbf{x}})\rangle_{\dot{\mathbf{x}}|\mathbf{x}}(\mathbf{x}) := \int p_{\dot{\mathbf{x}}|\mathbf{x}}(\dot{\mathbf{x}}|\mathbf{x})f(\mathbf{x},\dot{\mathbf{x}})\,\mathrm{d}^N\dot{x}\,. \tag{9}$$

We assume throughout that all averages taken exist. This introduces integrability constraints on the functions of which the average is taken. The function space is thus not completely unrestricted. The functions are restricted to be integrable in the sense that the averages above exist. In addition, they should be differentiable, simply to assure that the temporal derivative of their output signal exists.

We use greek letters to denote the index of vector components. Partial derivatives with respect to a component x_μ are written as ∂_μ. For example, the divergence of a vector field $\mathbf{v}(\mathbf{x})$ takes the short form

$$\operatorname{div} \mathbf{v}(\mathbf{x}) := \sum_\mu \frac{\partial v_\mu(\mathbf{x})}{\partial x_\mu} = \sum_\mu \partial_\mu v_\mu(\mathbf{x})\,. \tag{10}$$

We use the convention that within products, ∂_μ acts on all functions to its right. If we want ∂_μ to act locally, we use square brackets. This convention can be illustrated by the product rule

$$\partial_\mu f(\mathbf{x}) g(\mathbf{x}) = [\partial_\mu f(\mathbf{x})] g(\mathbf{x}) + f(\mathbf{x})[\partial_\mu g(\mathbf{x})]\,. \tag{11}$$

3.3 Reformulation of the optimization problem

To describe the Δ-value in terms of the probability distribution $p_{\mathbf{x},\dot{\mathbf{x}}}(\mathbf{x},\dot{\mathbf{x}})$, we need to express the temporal derivative of the output signal $y(t) = g(\mathbf{x}(t))$ in terms of the input signals $\mathbf{x}(t)$ and their derivatives. This is readily done by the chain rule

$$\dot{y}(t) = \frac{\mathrm{d}}{\mathrm{d}t} g(\mathbf{x}(t)) = \sum_\mu \dot{x}_\mu(t) \partial_\mu g(\mathbf{x}(t))\,. \tag{12}$$

We can now rewrite the objective function (1) by replacing the time average $\langle\cdot\rangle_t$ by the ensemble average $\langle\cdot\rangle_{\mathbf{x},\dot{\mathbf{x}}}$

$$\begin{aligned}
\Delta(g_j) &\overset{(1)}{=} \langle \dot{y}_j(t)^2 \rangle_t && (13)\\
&\overset{(12)}{=} \sum_{\mu,\nu} \langle \dot{x}_\mu [\partial_\mu g_j(\mathbf{x})] \dot{x}_\nu [\partial_\nu g_j(\mathbf{x})] \rangle_{\mathbf{x},\dot{\mathbf{x}}} && (14)\\
&= \sum_{\mu,\nu} \langle \underbrace{\langle \dot{x}_\mu \dot{x}_\nu \rangle_{\dot{\mathbf{x}}|\mathbf{x}}}_{=:K_{\mu\nu}(\mathbf{x})} [\partial_\mu g_j(\mathbf{x})][\partial_\nu g_j(\mathbf{x})] \rangle_{\mathbf{x}} && (15)\\
&= \sum_{\mu,\nu} \langle K_{\mu\nu}(\mathbf{x}) [\partial_\mu g_j(\mathbf{x})][\partial_\nu g_j(\mathbf{x})] \rangle_{\mathbf{x}}\,, && (16)
\end{aligned}$$

where $K_{\mu\nu}(\mathbf{x})$ is the matrix of the second moments of the conditional velocity distribution $p_{\dot{\mathbf{x}}|\mathbf{x}}(\dot{\mathbf{x}}|\mathbf{x})$ and reflects the dynamical structure of the input signal.

An elegant reformulation of the optimization problem can be obtained by introducing the following scalar product (f, g) between functions $f, g \in \mathcal{F}$:

$$(f, g) := \langle f(\mathbf{x}) g(\mathbf{x}) \rangle_{\mathbf{x}}\,. \tag{17}$$

With this definition, the function space $\mathcal{F}$ becomes a Hilbert space and the slowness objective for a function g can be written as

$$\Delta(g_j) = \sum_{\mu,\nu} (\partial_\mu g, K_{\mu\nu} \partial_\nu g)\,. \tag{18}$$

Note that we restrict the action of the partial derivatives to the argument of the scalar product they appear in.

Replacing the temporal averages by ensemble averages and using the scalar product (17), the original optimization problem becomes

Optimization problem 2: *Given a function space $\mathcal{F}$ and a probability distribution $p_{\mathbf{x},\dot{\mathbf{x}}}(\mathbf{x},\dot{\mathbf{x}})$ for the input signal $\mathbf{x}$ and its derivative $\dot{\mathbf{x}}$, find a set of $J+1$ real-valued input-output functions $g_j(\mathbf{x}), j \in \{0,1,...,J\}$ that minimize*

$$\Delta(g_j) = \sum_{\mu,\nu}(\partial_\mu g_j, K_{\mu\nu}\partial_\nu g_j) \tag{19}$$

under the constraint

$$\forall i < j : \quad (g_i, g_j) = \delta_{ij} \qquad \textit{(orthonormality and order)}\,. \tag{20}$$

Here we dropped the zero mean constraint and allow the trivial constant solution $g_0 = \pm 1$ to occur. As any function whose scalar product with the constant vanishes must have zero mean, the constraint (20) implies zero mean for all functions with $j > 0$. For functions f and g with zero mean, in turn, the scalar product (17) is simply the covariance, so that the constraints (2-4) can be compactly written as the orthonormality constraint (20).

3.4 A differential equation for the solutions

In this section we show that optimization problem 2 can be reduced to a partial differential eigenvalue equation. As some of the proofs are lengthy and not very illustrative, we state and motivate the main results while postponing the exact proofs to the appendix.

Under the assumption that all functions $g \in \mathcal{F}$ fulfill a boundary condition that is stated below, the objective function (19) can be written as

$$\Delta(g) = (g, \underbrace{\sum_{\mu,\nu}\partial_\mu^\dagger K_{\mu\nu}\partial_\nu}_{=:\mathcal{D}}\, g) = (g, \mathcal{D}g)\,. \tag{21}$$

Here, $A^\dagger$ denotes the adjoint operator to A with respect to the scalar product (17), i.e., the operator that fulfills the condition $(Af, g) = (f, A^\dagger g)$ for all functions $f, g \in \mathcal{F}$. By means of an integration by parts, it can be shown that $\partial_\mu^\dagger = -\frac{1}{p_{\mathbf{x}}}\partial_\mu p_{\mathbf{x}}$. Thus the operator $\mathcal{D}$ is the partial differential operator

$$\mathcal{D} = -\frac{1}{p_{\mathbf{x}}}\sum_{\mu,\nu}\partial_\mu p_{\mathbf{x}} K_{\mu\nu}\partial_\nu\,. \tag{22}$$

Because $K_{\mu\nu}$ is symmetric, $\mathcal{D}$ is self-adjoint, i.e. $(\mathcal{D}f, g) = (f, \mathcal{D}g)$ (see appendix, Lemma 2).

The main advantage of this reformulation is that the Δ-value takes a form that is common in other contexts, e.g., in quantum mechanics, where the operator $\mathcal{D}$ corresponds to the Hamilton operator (e.g. Landau and Lifshitz, 1977, §20). This analogy allows us to transfer the well-developed theory from these areas to our problem. As in quantum mechanics, the central role is played by the eigenfunctions of $\mathcal{D}$. This culminates in theorem 1, which we will briefly motivate. A rigorous proof can be found in the appendix.

Because the operator $\mathcal{D}$ is self-adjoint, it possesses a complete set of eigenfunctions g_i that are mutually orthogonal with respect to the scalar product (17) (spectral theorem, see e.g. Courant and Hilbert, 1989, chapter V, §14). The eigenfunctions of $\mathcal{D}$ are defined by the eigenvalue equation

$$\mathcal{D}g_i = \lambda_i g_i \tag{23}$$

and are assumed to be normalized according to

$$(g_i, g_i) = 1\,. \tag{24}$$

Because they are orthogonal, they fulfill the orthonormality constraint (20). Inserting these expressions into (21) immediately shows that the Δ-value of the eigenfunctions is given by their eigenvalue

$$\Delta(g_i) \overset{(21)}{=} (g_i, \mathcal{D} g_i) \overset{(23)}{=} (g_i, \lambda_i g_i) \overset{(24)}{=} \lambda_i \,. \tag{25}$$

Because of the completeness of the eigenfunctions g_i, any function g can be represented as a linear combination $g = \sum_i w_i g_i$ of the eigenfunctions g_i. The Δ-value of g can then be decomposed into a sum of the Δ-values of the eigenfunctions

$$\Delta(g) \overset{(21)}{=} (g, \mathcal{D} g) \overset{(23,20)}{=} \sum_i w_i^2 \lambda_i \,. \tag{26}$$

The unit variance constraint requires that the square sum of the coefficients w_i is unity: $\sum_i w_i^2 = 1$. It is then evident that the Δ-value (26) can be minimized by choosing $w_i = \delta_{0i}$, so that the slowest function is simply the eigenfunction g_0 with the smallest eigenvalue. The space of all functions that are orthogonal to g_0 is spanned by the remaining eigenfunctions g_i with $i > 0$. The slowest function in this space is the eigenfunction g_1 with the second smallest eigenvalue. Iterating this scheme makes clear that the optimal functions for SFA are simply the eigenfunctions g_i, ordered by their eigenvalue.

The eigenvalue problem (23) is a partial differential equation and thus requires boundary conditions. A detailed analysis of the problem shows that the optimal functions are those that fulfill von Neumann boundary conditions (see Appendix). Altogether this yields

Theorem 1 *The solution of optimization problem 2 is given by the $J+1$ eigenfunctions of the operator $\mathcal{D}$ with the smallest eigenvalues, i.e. the functions that fulfill*

$$\mathcal{D} g_j = \lambda_j g_j \tag{27}$$

with the boundary condition

$$\sum_{\mu,\nu} n_\mu p_{\mathbf{x}} K_{\mu\nu} \partial_\nu g_j = 0 \,, \tag{28}$$

and the normalization condition

$$(g_j, g_j) = 1 \,. \tag{29}$$

Here, $\mathbf{n}(\mathbf{x})$ is the normal vector on the boundary for the point $\mathbf{x}$. The Δ-value of the eigenfunctions is given by their eigenvalue

$$\Delta(g_j) = \lambda_j \,. \tag{30}$$

If the input data $\mathbf{x}$ are not bounded, the boundary condition has to be replaced by a limit, so that for parameterized boundaries that grow to infinity, the left hand side of equation (28) converges to zero for all points on the boundary. Note that we assumed earlier that all averages taken exist. This implies that the square of the functions and their first derivatives decay more quickly than $p_{\mathbf{x}}(\mathbf{x})$ as $||\mathbf{x}|| \to \infty$. Functions that do not fulfill the limit case of the boundary condition tend to have infinite variance or Δ-value.

The key advantage of the theory is that it converts the (global) optimization problem into a (local) partial differential eigenvalue equation. Moreover, the eigenvalue equation (27) belongs to a class that is known as Sturm-Liouville problems (see e.g., Courant and Hilbert, 1989), for which a well-developed theory exists. In the next chapter we use Sturm-Liouville theory to study the case of input data that are generated from a set of statistically independent sources.

4. Statistically Independent Sources

In this section we extend the theory to the scenario where the input signals are generated from a set of statistically independent sources. This case is particularly interesting in the light of nonlinear blind source separation and independent component analysis. We show that the eigenvalue equation for the optimal functions of SFA can be split into separate eigenvalue problems for a set of *harmonics*, each of which depends on only one of the sources. The optimal functions for the full optimization problem can be expressed as products of these harmonics. We then study the structure of the harmonics and show that the slowest non-constant harmonics are monotonic functions of the sources and therefore good representatives thereof. In the case of Gaussian or uniform statistics, the harmonics can be calculated analytically. In the Gaussian case, they are Hermite polynomials of the sources, which implied in particular that the slowest non-constant harmonics are simply the sources. Finally, we present a perturbartion-theoretical analysis to understand the effects of weakly inhomogeneous input statistics.

4.1 Statistically independent input data

In this section, we assume that the input signals $\mathbf{x}$ are generated from a set of statistically independent sources $s_\alpha, \alpha \in \{1, ..., S\}$ by means of an instantaneous, invertible function $\mathbf{F}$, i.e., $\mathbf{x}(t) = \mathbf{F}(\mathbf{s}(t))$. As discussed in section 3.1, the nonlinear function $\mathbf{F}$ can be regarded as a coordinate transformation of the input data and is thus immaterial to the output signals generated by SFA. Consequently, it makes no difference if we use the nonlinear mixture as the input signals or the sources. The output signals should be the same. In the following, we therefore assume that the input signals for SFA are simply the sources $\mathbf{s} \in \mathbb{R}^S$ themselves. To emphasize that the input signals are the sources and not the "mixture" $\mathbf{x}$, we use indices α and β for the sources instead of μ and ν for the components of the input signals. The statistical independence of the sources is formally reflected by the factorization of their joint probability density

$$p_{\mathbf{s},\dot{\mathbf{s}}}(\mathbf{s}, \dot{\mathbf{s}}) = \prod_\alpha p_{s_\alpha, \dot{s}_\alpha}(s_\alpha, \dot{s}_\alpha) \,. \tag{31}$$

Then, the marginal probability $p_\mathbf{s}$ also factorizes into the individual probabilities $p_\alpha(s_\alpha)$

$$p_\mathbf{s}(\mathbf{s}) = \prod_\alpha p_\alpha(s_\alpha) \tag{32}$$

and $K_{\alpha\beta}$ is diagonal

$$K_{\alpha\beta}(\mathbf{s}) = \delta_{\alpha\beta} K_\alpha(s_\alpha) \quad \text{with} \quad K_\alpha(s_\alpha) := \langle \dot{s}_\alpha^2 \rangle_{\dot{s}_\alpha | s_\alpha} \,. \tag{33}$$

The latter is true because the mean temporal derivative of 1-dimensional stationary and differentiable stochastic processes vanish for any s_α for continuity reasons, so that $K_{\alpha\beta}$ is not only the matrix of the second moments of the derivatives, but actually the conditional covariance matrix of the derivatives of the sources given the sources. As the sources are statistically independent, their derivatives are uncorrelated and $K_{\alpha\beta}$ has to be diagonal.

4.2 Factorization of the output signals

In the case of statistically independent input signals, the operator $\mathcal{D}$ introduced in section 3.4 can be split into a sum of operators $\mathcal{D}_\alpha$, each of which depends on only one of the sources:

$$\mathcal{D}(\mathbf{s}) = \sum_\alpha \mathcal{D}_\alpha(s_\alpha) \tag{34}$$

with

$$\mathcal{D}_\alpha = -\frac{1}{p_\alpha}\partial_\alpha p_\alpha K_\alpha \partial_\alpha \,, \tag{35}$$

as follows immediately from equations (22) and (33). This has the important implication that the solution to the full eigenvalue problem for $\mathcal{D}$ can be constructed from the 1-dimensional eigenvalue problems associated with $\mathcal{D}_\alpha$:

Theorem 2 *Let $g_{\alpha i}$ ($i \in \mathbb{N}$) be the normalized eigenfunctions of the operators $\mathcal{D}_\alpha$, i.e., the set of functions $g_{\alpha i}$ that fulfill the eigenvalue equations*

$$\mathcal{D}_\alpha g_{\alpha i} = \lambda_{\alpha i} g_{\alpha i} \tag{36}$$

with the boundary conditions

$$p_\alpha K_\alpha \partial_\alpha g_{\alpha i} = 0 \tag{37}$$

and the normalization condition

$$(g_{\alpha i}, g_{\alpha i})_\alpha := \langle g_{\alpha i}^2 \rangle_{s_\alpha} = 1 \,. \tag{38}$$

Then, the product functions

$$g_{\mathbf{i}}(\mathbf{s}) := \prod_\alpha g_{\alpha i_\alpha}(s_\alpha) \tag{39}$$

form a complete set of (normalized) eigenfunctions to the full operator $\mathcal{D}$ with the eigenvalues

$$\lambda_{\mathbf{i}} = \sum_\alpha \lambda_{\alpha i_\alpha} \tag{40}$$

and thus those $g_{\mathbf{i}}$ with the smallest eigenvalues $\lambda_{\mathbf{i}}$ form a solution of optimization problem 2. Here, $\mathbf{i} = (i_1, ..., i_S) \in \mathbb{N}^S$ denotes a multi-index that enumerates the eigenfunctions of the full eigenvalue problem.

In the following, we assume that the eigenfunctions $g_{\alpha i}$ are ordered by their eigenvalue and refer to them as the *harmonics* of the source s_α. This is motivated by the observation that in the case where p_α and K_α are independent of s_α, i.e., for a uniform distribution, the eigenfunctions $g_{\alpha i}$ are harmonic oscillations whose frequency increases linearly with i (for a derivation see below). Moreover, we assume that the sources s_α are ordered according to slowness, in this case measured by the eigenvalue $\lambda_{\alpha 1}$ of their lowest non-constant harmonic $g_{\alpha 1}$. These eigenvalues are the Δ-value of the slowest possible nonlinear point transformations of the sources.

The main result of the above theorem is that in the case of statistically independent sources, the output signals are products of harmonics of the sources. Note that the constant function $g_{\alpha 0}(s_\alpha) = 1$ is an eigenfunction with eigenvalue 0 to all the eigenvalue problems (36). As a consequence, the harmonics $g_{\alpha i}$ of the single sources are also eigenfunctions to the full operator $\mathcal{D}$ (with the index $\mathbf{i} = (0, ..., 0, i_\alpha = i, 0, ..., 0)$) and can thus be found by SFA. Importantly, the lowest non-constant harmonic of the slowest source (i.e., $g_{(1,0,0,...)} = g_{11}$) is the function with the smallest overall Δ-value (apart from the constant) and thus the first function found by SFA. In the next sections, we show that the lowest non-constant harmonics $g_{\alpha 1}$ reconstruct the sources up to a monotonic and thus invertible point transformation and that in the case of sources with Gaussian statistics, they even reproduce the sources exactly.

4.3 Monotony of the first harmonic

Let us assume that the source s_α is bounded and takes on values on the interval $s_\alpha \in [a, b]$. The eigenvalue problem (36,37) can be rewritten in the standard form of a Sturm-Liouville problem:

$$\partial_\alpha p_\alpha K_\alpha \, \partial_\alpha g_{\alpha i} + \lambda_{\alpha i} p_\alpha g_{\alpha i} \overset{(36,35)}{=} 0\,, \tag{41}$$

$$\text{with} \quad p_\alpha K_\alpha \partial_\alpha g_{\alpha i} \overset{(37)}{=} 0 \quad \text{for} \quad s_\alpha \in \{a, b\}\,. \tag{42}$$

Note that both p_α and $p_\alpha K_\alpha$ are positive for all s_α. Sturm-Liouville theory states that the solutions $g_{\alpha i}, i \in \mathbb{N}^0$ of this problem are oscillatory and that $g_{\alpha i}$ has exactly i zeros on $]a, b[$ if the $g_{\alpha i}$ are ordered by increasing eigenvalue $\lambda_{\alpha i}$ (Courant and Hilbert, 1989, chapter IV, §6). All eigenvalues are positive. In particular, $g_{\alpha 1}$ has only one zero $\xi \in]a, b[$. Without loss of generality we assume that $g_{\alpha 1} < 0$ for $s_\alpha < \xi$ and $g_{\alpha 1} > 0$ for $s_\alpha > \xi$. Then equation (41) implies that

$$\partial_\alpha p_\alpha K_\alpha \partial_\alpha g_{\alpha 1} = -\lambda_\alpha p_\alpha g_{\alpha 1} < 0 \quad \text{for } s_\alpha > \xi \tag{43}$$

$$\Longrightarrow \quad p_\alpha K_\alpha \partial_\alpha g_{\alpha 1} \quad \text{is monotonic decreasing on }]\xi, b] \tag{44}$$

$$\overset{(42)}{\Longrightarrow} \quad p_\alpha K_\alpha \, \partial_\alpha g_{\alpha 1} > 0 \text{ on }]\xi, b[\tag{45}$$

$$\Longrightarrow \quad \partial_\alpha g_{\alpha 1} > 0 \quad \text{since } p_\alpha K_\alpha > 0 \text{ on }]\xi, b[\tag{46}$$

$$\Longleftrightarrow \quad g_{\alpha 1} \quad \text{is monotonic increasing on }]\xi, b[\,. \tag{47}$$

A similar consideration for $s < \xi$ shows that $g_{\alpha 1}$ is also monotonically increasing on $]a, \xi[$. Thus, $g_{\alpha 1}$ is monotonic and invertible on the whole interval $[a, b]$. Note that the monotony of $g_{\alpha 1}$ is important in the context of BSS, because it ensures that not only some of the output signals of SFA depend on only one of the sources (the harmonics), but that there should actually be some (the lowest non-constant harmonics) that are very similar to the source itself.

4.4 Gaussian sources

We now consider the situation that the sources are reversible Gaussian stochastic processes, (i.e., that the joint probability density of $s(t)$ and $s(t + dt)$ is Gaussian and symmetric with respect to $s(t)$ and $s(t + dt)$). In this case, the instantaneous values of the sources and their temporal derivatives are statistically independent, i.e., $p_{\dot{s}_\alpha | s_\alpha}(\dot{s}_\alpha | s_\alpha) = p_{\dot{s}_\alpha}(\dot{s}_\alpha)$. Thus, K_α is independent of s_α, i.e., $K_\alpha(s_\alpha) = K_\alpha =$ const. Without loss of generality we assume that the sources have unit variance. Then the probability density of the source is given by

$$p_\alpha(s_\alpha) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-s_\alpha^2/2} \tag{48}$$

and the eigenvalue equations (41) for the harmonics can be written as

$$\partial_\alpha e^{-s_\alpha^2/2} \partial_\alpha g_{\alpha i} + \frac{\lambda_{\alpha i}}{K_\alpha} e^{-s_\alpha^2/2} g_{\alpha i} = 0\,. \tag{49}$$

This is a standard form of Hermite's differential equation (see Courant and Hilbert, 1989, chapter V, § 10). Accordingly, the harmonics $g_{\alpha i}$ are given by the (appropriately normalized) Hermite polynomials H_i of the sources:

$$g_{\alpha i}(s_\alpha) = \frac{1}{\sqrt{2^i i!}} H_i \left(\frac{s_\alpha}{\sqrt{2}} \right)\,. \tag{50}$$

The Hermite polynomials can be expressed in terms of derivatives of the Gaussian distribution:

$$H_n(x) = (-1)^n \mathrm{e}^{x^2} \partial_x^n \mathrm{e}^{-x^2}\,. \tag{51}$$

It is clear that Hermite polynomials fulfill the boundary condition

$$\lim_{s_\alpha \to \infty} K_\alpha p_\alpha \partial_\alpha g_{\alpha i} = 0\,, \tag{52}$$

because the derivative of a polynomial is again a polynomial and the Gaussian distribution decays faster than polynomially as $|s_\alpha| \to \infty$. The eigenvalues are given by

$$\lambda_{\alpha i} = i/K_\alpha\,. \tag{53}$$

The most important consequence is that the lowest non-constant harmonics simply reproduce the sources: $g_{\alpha 1}(s_\alpha) = 1/\sqrt{2}H_1(s_\alpha/\sqrt{2}) = s_\alpha$. Thus, for Gaussian sources, SFA with an unrestricted function space reproduces the sources, although it still remains to determine which of the output signals are the sources and which are higher harmonics or products of the harmonics of the sources.

4.5 Homogeneously Distributed Sources

Another canonical example for which the eigenvalue equation (36) can be solved analytically is the case of homogeneously distributed sources, i.e., the case where the probability distribution $p_{s,\dot{s}}$ is independent of s. Consequently, neither $p_\alpha(s_\alpha)$ nor $K_\alpha(s_\alpha)$ can depend on s_α, i.e., they are constants. Note that such a distribution may be difficult to implement by a real differentiable process, because the velocity distribution should be different at boundaries that cannot be crossed. Nevertheless, this case provides an approximation to cases, where the distribution is close to homogeneous.

Let s_α take values in the interval $[0, L_\alpha]$. The eigenvalue equation (41) for the harmonics is then given by

$$K_\alpha \partial_\alpha^2 g_{\alpha i} + \lambda_{\alpha i} g_{\alpha i} = 0\,. \tag{54}$$

and readily solved by harmonic oscillations:

$$g_{\alpha i}(s_\alpha) = \sqrt{2}\cos\left(i\pi \frac{s_\alpha}{L_\alpha}\right)\,. \tag{55}$$

The Δ-value of these functions is given by

$$\Delta(g_{\alpha i}) = \lambda_{\alpha i} = K_\alpha \left(\frac{\pi}{L_\alpha} i\right)^2\,. \tag{56}$$

Note the similarity of these solutions with the optimal free responses derived by Wiskott (2003).

4.6 Weakly Inhomogeneous Sources

For homogeneous distributions, the optimal functions for SFA are harmonic oscillations. It is reasonable to assume that this behavior is preserved qualitatively if p_α and K_α are no longer homogeneous but depend weakly on the source s_α. In particular, if the wavelength of the oscillation is much shorter than the typical scale on which p_α and K_α vary, it can be expected that the oscillation "does not notice" the change. Of course, we are not principally interested in quickly varying functions, but they can provide insights into the effect of variations of p_α and K_α.

To examine this further, we can derive an approximate solution to the eigenvalue equation (36,41) by treating $\epsilon = 1/\sqrt{\lambda_{\alpha i}} = 1/\sqrt{\Delta}$ as a small perturbation parameter. This corresponds to large Δ-values, i.e., quickly varying functions. For this case we can apply a perturbation theoretical approach that follows the Wentzel-Kramers-Brillouin approximation used in quantum mechanics. For a more detailed description of the approach we refer the reader to the quantum mechanics literature (e.g., Davydov, 1976). Knowing that the solution shows oscillations, we start with the complex ansatz

$$g_\alpha(s_\alpha) = A \, \exp\left(\frac{i}{\epsilon}\Phi(s_\alpha)\right) \, , \tag{57}$$

where $\Phi(s_\alpha)$ is a complex function that needs to be determined. Treating ϵ as a small number, we can expand Φ in ϵ

$$\Phi(s_\alpha) = \Phi_0(s_\alpha) + \epsilon\Phi_1(s_\alpha) + \, ... \, , \tag{58}$$

where the ellipses stand for higher-order terms. We insert this expression into the eigenvalue equation (41) and collect terms of the same order in ϵ. Requiring each order to vanish separately and neglecting orders of ϵ^2 and higher, we get equations for Φ_0 and Φ_1:

$$(\partial_\alpha\Phi_0)^2 = \frac{1}{K_\alpha} \, , \tag{59}$$

$$\partial_\alpha\Phi_1 = \frac{i}{2}\frac{\partial_\alpha(p_\alpha K_\alpha \partial_\alpha \Phi_0)}{p_\alpha K_\alpha \partial_\alpha \Phi_0} \, . \tag{60}$$

These equations are solved by

$$\Phi_0(s_\alpha) = \int\limits_{s_0}^{s_\alpha} \sqrt{\frac{1}{K_\alpha(s)}} \, \mathrm{d}s \, , \tag{61}$$

$$\Phi_1(s_\alpha) = \frac{i}{2}\ln\left(p_\alpha K_\alpha^{1/2}\right) \, , \tag{62}$$

where s_0 is an arbitrary reference point. Inserting this back into equation (57), we get the approximate solution

$$g_\alpha(s_\alpha) = \frac{A}{\sqrt[4]{p_\alpha^2 K_\alpha}} \exp\left(i \int\limits_{s_0}^{s_\alpha} \sqrt{\frac{\Delta}{K_\alpha(s)}} \mathrm{d}s \right) \, . \tag{63}$$

This shows that the solutions with large Δ-values show oscillations with local frequency $\sqrt{\Delta/K_\alpha}$ and amplitude $\sim 1/\sqrt[4]{p_\alpha^2 K_\alpha}$. Large values of K_α indicate that the source changes quickly, i.e., where its "velocity" is high. This implies that the local frequency of the solutions is smaller for values of the sources where the source velocity is high, whereas small source velocities lead to higher frequencies than expected for homogeneous movement. The functions compensate for high source velocities with smaller spatial frequencies such that the effective temporal frequency of the output signal is kept constant.

Understanding the dependence of the amplitude on p_α and K_α is more subtle. Under the assumption that K_α is independent of s_α, the amplitude decreases where p_α is large and increases where p_α is small. Intuitively, this can be interpreted as an equalization of the fraction of the total variance that falls into a small interval of length $\Delta s_\alpha \gg \sqrt{K_\alpha/\Delta}$. This fraction is roughly given by the product of the probability $p_\alpha \Delta s_\alpha$ of being in this section times the squared amplitude $1/\sqrt{p_\alpha^2 K_\alpha}$ of the oscillation. For constant K_α, this fraction is also constant, so the amplitude is effectively rescaled to yield the same "local variance" everywhere. If p_α is constant and K_α varies, on the other hand, the amplitude of the oscillation is small for values of the sources where they change quickly and large where they change slowly. This corresponds to the intuition that there are two ways of treating regions where the sources change quickly: decreasing spatial frequency to generate slower output signals and/or decreasing the amplitude of the oscillation to "pay less attention" to these regions. There is also a strong formal argument why the amplitude should depend on $p_\alpha^2 K_\alpha$. As the optimization problem is invariant under arbitrary invertible coordinate transformations, the amplitude of the oscillation should depend on a function of p_α and K_α that is independent of the coordinate system. This constrains the amplitude to depend on $p_\alpha^2 K_\alpha$, as this is the only combination of these quantities that is invariant under coordinate transformations.

5. Analogies in Physics

The last two sections as well as previous studies (Wiskott, 2003) have illustrated that SFA allows a rich repertoire of analytical considerations. Why is that? The main reason is that both the Δ-value and the constraints are quadratic functionals of the output signals. As long as the output signal is linearly related to the parameters of the input-output functions (as is the case for the nonlinear expansion approach that underlies the SFA algorithm), both the Δ-value and the constraint quantities are quadratic forms of the parameters. The gradients involved in finding the optima are thus linear functions of the parameters, so that the solution can be found by means of linear methods, typically eigenvalue problems.

Eigenvalue problems have a long history in mathematical physics. They describe electron orbitals in atoms, acoustic resonances, vibrational modes in solids and light propagation in optical fibers. Whenever wave phenomena are involved, the associated theory makes use of eigenvalue problems in one way or another. Consequently, there is a well-developed mathematical theory for eigenvalue problems, including the infinite-dimensional case, which can be applied to SFA.

Interestingly, the occurence of eigenvalue problems in SFA is not the only analogy to physics. In the following, we point out three more analogies that may help in getting an intuitive understanding for the behavior of SFA.

5.1 Slow Feature Analysis and Hamilton's Principle

SFA aims at minimizing the mean square of the temporal derivative of the output signal y. Let us assume for a moment that we were only interested in the first, i.e., the slowest output signal of SFA. Then the only constraint that applies is that of unit variance. According to the technique of Lagrange multipliers, we are searching for stationary points (i.e., points, where the derivative with the respect to the parameters vanishes) of the objective function

$$\mathcal{L}(y) = \langle \dot{y}^2 \rangle_t - \lambda \langle y^2 \rangle_t = \frac{1}{T}\int \dot{y}(t)^2 \mathrm{d}t - \frac{\lambda}{T}\int y(t)^2 \mathrm{d}t\,, \tag{64}$$

where λ is a Lagrange multiplier, which has to be determined such that the constraint is fulfilled.

To interpret this objective function let us for a moment act as if the output signal y was the position of a physical particle. Then the square of the temporal derivative of y is proportional to the kinetic energy of the particle. We can thus interpret $K = \dot{y}^2/T$ as the *kinetic energy* of the output signal y. Consequently, it is only natural to interpret the second term in equation (64) in terms of a *potential energy* $U = \lambda y^2/T$. Then, the objective function $\mathcal{L}$ is the integral over the difference between the kinetic and the potential energy of the output signal, a quantity that is known as *action* in Lagrange mechanics:

$$\mathcal{L}(y) = \int [K(t) - U(t)]\,\mathrm{d}t\,. \tag{65}$$

One of the most important principles of Lagrange mechanics is Hamilton's principle of stationary action, which states that out of all possible trajectories, physical systems "choose" those for which the action $\mathcal{L}$ is stationary. It is immediately clear that with the above reinterpretation of the quantities appearing in SFA, the two problems are formally very similar.

Moreover, since the potential energy of the physical system corresponding to SFA is quadratic in the "position y of the particle", the problem is in essence that of a harmonic oscillator. From this perspective, it is not surprising that the optimal output signals for SFA are generally harmonic oscillations, as shown by Wiskott (2003).

5.2 Standing Waves

The optimal solutions for SFA are given by the eigenfunctions of the operator $\mathcal{D}$, which is a quadratic form in the partial derivatives ∂_μ. Hence, $\mathcal{D}$ belongs to the same class of operators as the Laplace operator. This implies that equation (27) has the form of a stationary wave equation, which describes oscillatory eigenmodes of fields.

An intuitive picture can be sketched for the exemplary case that the input data $\mathbf{x}$ lies on a 2-dimensional manifold, embedded in a 3-dimensional space. We can then interpret this manifold as an oscillating membrane. Equation (27) describes the vibrational eigenmodes or standing waves on the membrane. The boundary condition (28) means that the boundary of the membrane is open, i.e., the borders of the membrane can oscillate freely. The solutions $g_i(\mathbf{x})$ of SFA correspond to the amplitude of an eigenmode with frequency $\omega = \sqrt{\lambda_i}$ at position $\mathbf{x}$. For a constant probability distribution $p_\mathbf{x}$, the matrix $K_{\mu\nu}$ can moreover be interpreted as the "surface tension" of the membrane. In a given direction $\mathbf{d}$ ($\mathbf{d}$ tangential to the membrane and $|\mathbf{d}| = 1$), the "tension" of the membrane is given by $\kappa = \mathbf{d}^T\mathbf{K}\mathbf{d} = d_\mu K_{\mu\nu} d_\nu$. If the input changes quickly in the direction of $\mathbf{d}$, the surface tension κ is large. For large surface tension, however, oscillations with a given wavelength have a high frequency, that is, a large Δ-value. Thus, slow functions (solutions with small Δ-values corresponding to oscillations with low frequency) tend to be oscillatory in directions with small input velocity (low surface tension) and remain largely constant in directions of large input velocity (high surface tension). Directions with high surface tension correspond to input directions in which SFA learns invariances.

5.3 Quantum Mechanics

An intuition for the factorization of the solutions for independent sources can be gained by interpreting $\mathcal{D}$ as a formal equivalent of the Hamilton operator in quantum mechanics. Equation (27) then corresponds to the stationary Schrödinger equation and the Δ-values λ_i to the energies of stationary states of a quantum system. For statistically independent sources, the operator $\mathcal{D}$ decomposes into a sum of operators $\mathcal{D}_\mu$, which depend on only one of the sources each. The decomposition corresponds to the situation of a quantum system with "Hamilton operator" $\mathcal{D}$ that consists of a set of independent quantum systems with "Hamilton operators" $\mathcal{D}_\mu$. For readers who are familiar with quantum mechanics, it is then no longer surprising that the eigenvalue equation for $\mathcal{D}$ can be solved by means of a separation ansatz. The solutions of SFA (stationary states of the full quantum system) are thus products of the harmonics of the sources in isolation (the stationary states of the independent subsystems). Similarly, it is clear that the Δ-value of the product states (the energy of the full system) is the sum of the Δ-values of the harmonics (the energies of the subsystems).

The dependence of the Δ-value (energy) $\lambda_\mathbf{i}$ on the index (quantum number) $\mathbf{i}$ also has a counterpart in physics. As a function of the source s_μ, the harmonics $g_{\mu i_\mu}$ show oscillations with i_μ zeros. Thus, the index i_μ is a measure of the spatial frequency (or, in quantum mechanics, the momentum) of the harmonic. From this perspective, the dependence of the Δ-value (energy) on the index (frequency or momentum) $\mathbf{i}$ plays the role of a dispersion relation. For homogeneously distributed sources, the dispersion is quadratic, while for Gaussian sources it is linear.

Wave equations of the type of equation (27) are ubiquitous in physics and there are probably more formally equivalent physical systems. We believe that these analogies can help substantially in getting an intuitive understanding of the behavior of SFA in the limit case of very rich function spaces.

6. Discussion

In this article, we have presented a mathematical framework for SFA for the case of unrestricted function spaces. The theory shows that the solutions of SFA obey a partial differential eigenvalue problem that bears strong similarities with systems common in physics. This analogy leads to an intuitive interpretation of the functions found by SFA, e.g., as vibrational eigenmodes on the manifold spanned by the input data.

The predictive power of the presented framework is particularly strong in applications where high-dimensional training data are taken from low-dimensional manifolds. One example are videos of a single object that rotates or translates in space. In such a scenario, the images lie on a manifold whose points can be uniquely parametrized by the position and orientation of the object. By use of the mathematical framework presented here, analytical predictions for the dependence of the solutions of SFA on these parameters are straightforward. A reduced version of such a training scheme has previously been used to learn invariant representations of objects by means of SFA (Wiskott and Sejnowski, 2002). The application of SFA for learning place and head direction codes (Franzius et al., 2007a) from quasi-natural images also belongs to this class of problems. The input manifold for this scenario can be parametrized by the position and head direction of the simulated rat.

The application of the theory to the case of input signals generated from statistically independent sources has shown that the optimal output of SFA consists of products of signals, each of which depend on a single source only and that some of these harmonics should be monotonic functions of the sources themselves. In a subsequent article, we use these results to propose a new algorithm for nonlinear independent component analysis.

References

P. Berkes. Pattern recognition with slow feature analysis. *Cognitive Sciences EPrint Archive (CogPrints)*, 4104, 2005.

P. Berkes and L. Wiskott. Slow feature analysis yields a rich repertoire of complex cells. *Journal of Vision*, 5(6):579–602, 2005.

T. Blaschke, P. Berkes, and L. Wiskott. What is the relation between slow feature analysis and independent component analysis? *Neural Computation*, 18(10):2495–2508, 2006.

T. Blaschke, T. Zito, and L. Wiskott. Independent slow feature analysis and nonlinear blind source separation. *Neural Computation*, 19(4):994–1021, 2007.

R. Courant and D. Hilbert. *Methods of mathematical physics Part I.* Wiley, 1989.

A. S. Davydov. *Quantum mechanics.* Pergamon Press New York, 1976.

P. Földiàk. Learning invariance from transformation sequences. *Neural Computation*, 3: 194–200, 1991.

M. Franzius, H. Sprekeler, and L. Wiskott. Slowness and sparseness lead to place, head-direction, and spatial-view cells. *PLoS Computationl Biology*, 3(8):e166, 2007.

M. Franzius, N. Wilbert, and L. Wiskott. Unsupervised learning of invariant 3D-object representations with slow feature analysis. In *Proc. 3rd Bernstein Symposium for Computational Neuroscience, Göttingen, September 24–27*, page 105, 2007b.

C. Kayser, W. Einhäuser, O. Dümmer, K.P. Körding, and P. König. Extracting slow subspaces from natural videos leads to complex cells. In *Proc. Int. Conf. on Artif. Neural Networks (ICANN) Springer: Lecture Notes in Computer Science*, volume 2130, pages 1075–1079, 2001.

L. D. Landau and E. M. Lifshitz. *Quantum Mechanics: Non-relativistic theory*, volume 3 of *Course of Theoretical Physics*. Pergamon Press, 1977.

G. Mitchison. Removing time variation with the anti-Hebbian differential synapse. *Neural Computation*, 3:312–320, 1991.

H. Sprekeler, C. Michaelis, and L. Wiskott. Slowness: An objective for spike-timing-plasticity? *PLoS Computational Biology*, 3(6):e112, 2007.

J. V. Stone and A. Bray. A learning rule for extracting spatio-temporal invariances. *Network: Computation in Neural Systems*, 6:429–436, 1995.

G. Wallis and E. T. Rolls. Invariant face and object recognition in the visual system. *Progress in Neurobiology*, 51(2):167–194, February 1997.

L. Wiskott. Learning invariance manifolds. In L. Niklasson, M. Bodén, and T. Ziemke, editors, *Proceedings of the 8th International Conference on Artificial Neural Networks, ICANN'98, Skövde*, Perspectives in Neural Computing, pages 555–560, 1998.

L. Wiskott. Slow feature analysis: A theoretical analysis of optimal free responses. *Neural Computation*, 15(9):2147–2177, September 2003.

L. Wiskott and T. Sejnowski. Slow feature analysis: Unsupervised learning of invariances. *Neural Computation*, 14(4):715–770, 2002.

Appendix: Proof of Theorem 1

Theorem 1 *The solution of optimization problem 2 is given by the J eigenfunctions of the operator $\mathcal{D}$ with the smallest eigenvalues, i.e. the functions that fulfill the eigenvalue equation*

$$\mathcal{D}g_j = \lambda_j g_j \tag{66}$$

with the boundary condition

$$\sum_{\mu,\nu} n_\mu p_{\mathbf{x}} K_{\mu\nu} \partial_\nu g_i = 0\,. \tag{67}$$

Here, the operator $\mathcal{D}$ is given by

$$\mathcal{D} := -\frac{1}{p_{\mathbf{x}}} \sum_{\mu,\nu} \partial_\mu p_{\mathbf{x}} K_{\mu\nu} \partial_\nu \tag{68}$$

and the eigenfunctions are assumed to be normalized accoding to

$$(g_j, g_j) = 1\,. \tag{69}$$

$\mathbf{n}(\mathbf{x})$ *denotes the normal vector on the boundary for the point* $\mathbf{x}$*. The Δ-value of the eigenfunctions is given by their eigenvalue*

$$\Delta(g_j) = \lambda_j\,. \tag{70}$$

Preliminary Lemmas

For reasons of clarity, we first prove several lemmas that help to prove Theorem 1. The first lemma shows that the optimal functions for SFA fulfill an Euler-Lagrange equation that is similar to the eigenvalue equation for the operator $\mathcal{D}$.

Lemma 3 *For a particular choice of the parameters λ_{ij}, the solutions g_j of optimization problem 2 obey the Euler-Lagrange equation*

$$\mathcal{D}g_j(\mathbf{x}) - \lambda_{j0} - \lambda_{jj} g_j(\mathbf{x}) - \sum_{i<j} \lambda_{ji} g_i(\mathbf{x}) = 0 \tag{71}$$

with the boundary condition (67) and the operator $\mathcal{D}$ according to equation (68).

Proof

Optimization problem 2 is in essence a constrained optimization problem. The standard technique for such constrained optimization problems is that of Lagrange multipliers. This technique states that the solutions of the optimization problem have to fulfill the necessary condition to be stationary points of an objective function Ψ that incorporates the constraints

$$\Psi(g_j) = \frac{1}{2}\Delta(g_j) - \lambda_{j0}\langle g_j(\mathbf{x})\rangle_{\mathbf{x}} - \frac{1}{2}\lambda_{jj}\langle g_j(\mathbf{x})^2\rangle_{\mathbf{x}} - \sum_{i<j} \lambda_{ji}\langle g_i(\mathbf{x}) g_j(\mathbf{x})\rangle_{\mathbf{x}}\,, \tag{72}$$

where λ_{ij} are Lagrange multipliers that need to be chosen such that the stationary points fulfill the constraints.

The objective (72) is a functional of the function g_j we want to optimize. Because a gradient is not defined for functionals, we cannot find the stationary points by simply setting the gradient to zero. Instead, the problem requires variational calculus.

The technique of variational calculus can be illustrated by means of an expansion in the spirit of a Taylor expansion. Let us assume that we know the function

g_j that optimizes the objective function Ψ. The effect of a small change δg of g_j on the objective function Ψ can be written as

$$\Psi(g_j + \delta g) - \Psi(g_j) = \int \frac{\delta \Psi}{\delta g_j}(\mathbf{x})\, \delta g(\mathbf{x})\, \mathrm{d}^N x + \dots , \tag{73}$$

where the ellipses stand for higher order terms in δg. The function $\frac{\delta\Psi}{\delta g_j}$ is the *variational derivative* of the functional Ψ and usually depends on the input signal $\mathbf{x}$, the optimal function g_j, and possibly derivatives of g_j. Its analogue in finite-dimensional calculus is the gradient.

We now derive an expression for the variational derivative of the objective function (72). To keep the calculations tidy, we split the objective in two parts and omit the dependence on the input signal $\mathbf{x}$:

$$\Psi(g_j) =: \frac{1}{2}\Delta(g_j) - \tilde{\Psi}(g_j)\,. \tag{74}$$

The expansion of $\tilde{\Psi}$ is straightforward:

$$\begin{aligned}
\tilde{\Psi}(g_j + \delta g) - \tilde{\Psi}(g_j) &= \langle \delta g\, [\lambda_{j0} + \lambda_{jj} g_j + \sum_{i<j} \lambda_{ji} g_i] \rangle_{\mathbf{x}} + \dots &(75)\\
&= \int \delta g\, p_{\mathbf{x}}\, [\lambda_{j0} + \lambda_{jj} g_j + \sum_{i<j} \lambda_{ji} g_i] \mathrm{d}^N x + \dots &(76)
\end{aligned}$$

The expansion of $\Delta(g_j)$ is done after expressing the Δ-value in terms of probability density $p_{\mathbf{x}}$ and the matrix $K_{\mu\nu}$ (cf. equation (16)):

$$\begin{aligned}
\frac{1}{2}\left[\Delta(g_j + \delta g) - \Delta(g_j)\right] &\overset{(16)}{=} \frac{1}{2}\sum_{\mu,\nu} \langle K_{\mu\nu} [\partial_\mu (g_j + \delta g)][\partial_\nu (g_j + \delta g)] \rangle_{\mathbf{x}} &(77)\\
&\quad -\frac{1}{2}\sum_{\mu,\nu} \langle K_{\mu\nu} [\partial_\mu g_j][\partial_\nu g_j] \rangle_{\mathbf{x}} &(78)\\
&= \frac{1}{2}\sum_{\mu,\nu} \langle K_{\mu\nu} [\partial_\mu g_j][\partial_\nu \delta g] + K_{\mu\nu} [\partial_\mu \delta g][\partial_\nu g_j] \rangle_{\mathbf{x}} + \dots \\
&= \sum_{\mu,\nu} \langle K_{\mu\nu} [\partial_\mu \delta g][\partial_\nu g_j] \rangle_{\mathbf{x}} + \dots &(79)\\
&\quad \text{(since } K_{\mu\nu} \text{ is symmetric)}\\
&\overset{(8)}{=} \sum_{\mu,\nu} \int p_{\mathbf{x}} K_{\mu\nu} [\partial_\mu \delta g][\partial_\nu g_j]\, \mathrm{d}^N x &(80)\\
&= \sum_{\mu,\nu} \int \partial_\mu \delta g\, p_{\mathbf{x}}\, K_{\mu\nu} \partial_\nu g_j\, \mathrm{d}^N x &(81)\\
&\quad -\sum_{\mu,\nu} \int \delta g\, \partial_\mu p_{\mathbf{x}} K_{\mu\nu} \partial_\nu g_j\, \mathrm{d}^N x + \dots &(82)\\
&= \sum_{\mu,\nu} \int_{\partial V} n_\mu\, \delta g\, p_{\mathbf{x}}\, K_{\mu\nu} \partial_\nu g_j\, \mathrm{d}A &(83)\\
&\quad -\sum_{\mu,\nu} \int \delta g\, \partial_\mu p_{\mathbf{x}} K_{\mu\nu} \partial_\nu g_j\, \mathrm{d}^N x + \dots &(84)\\
&\quad \text{(Gauss' theorem)}\\
&\overset{(68)}{=} \int_{\partial V} \delta g \sum_{\mu,\nu} n_\mu\, p_{\mathbf{x}} K_{\mu\nu} \partial_\nu g_j\, \mathrm{d}A &(85)\\
&\quad + \int \delta g\, p_{\mathbf{x}}\, (\mathcal{D} g_j)\, \mathrm{d}^N x + \dots . &(86)
\end{aligned}$$

Here, $\mathrm{d}A$ is an infinitesimal surface element of the boundary ∂V of V and $\mathbf{n}$ is the normal vector on $\mathrm{d}A$. To get the expansion of the full objective function, we add (76) and (86):

$$\begin{aligned} \Psi(g_j + \delta g) - \Psi(g_j) &= \int_{\partial V} \delta g \sum_{\mu,\nu} n_\mu \, p_{\mathbf{x}} K_{\mu\nu} \partial_\mu g_j \, \mathrm{d}A \qquad (87) \\ &\quad + \int \delta g \, p_{\mathbf{x}} \, (\mathcal{D} g_j - \lambda_{j0} - \lambda_{jj} g_j - \sum_{i<j} \lambda_{ji} g_i) \, \mathrm{d}^N x + \dots . \end{aligned}$$

In analogy to the finite-dimensional case, g_j can only be an optimum of the objective function Ψ if any small change δg leaves the objective unchanged up to linear order. As we employ a Lagrange multiplier ansatz, we have an unrestricted optimization problem, so we are free in choosing δg. From this it is clear that the right hand side of (87) can only vanish if the integrands of both the volume and the boundary integral vanish separately. This leaves us with the differential equation (71) and the boundary condition (67). ■

Next, we show that the operator $\mathcal{D}$ is self-adjoint with respect to the scalar product (17) when restricted to the set of functions that fulfill the boundary condition (67).

Lemma 4 *Let $\mathcal{F}_b \subset \mathcal{F}$ be the space of functions obeying the boundary condition (28,67). Then $\mathcal{D}$ is self-adjoint on $\mathcal{F}_b$ with respect to the scalar product*

$$(f, g) := \langle f(\mathbf{x}) g(\mathbf{x}) \rangle_{\mathbf{x}}, \qquad (88)$$

i.e.

$$\forall f, g \in \mathcal{F}_b : (\mathcal{D} f, g) = (f, \mathcal{D} g). \qquad (89)$$

Proof The proof can be carried out in a direct fashion. Again, we omit the explicit dependence on $\mathbf{x}$.

$$\begin{aligned} (f, \mathcal{D} g) &\overset{(88,68,8)}{=} -\int p_{\mathbf{x}} f \frac{1}{p_{\mathbf{x}}} \sum_{\mu,\nu} \partial_\mu p_{\mathbf{x}} K_{\mu\nu} \partial_\nu g \, \mathrm{d}^N x \qquad (90) \\ &= -\sum_{\mu,\nu} \int \partial_\mu p_{\mathbf{x}} f K_{\mu\nu} \partial_\nu g \, \mathrm{d}^N x + \int p_{\mathbf{x}} \sum_{\mu,\nu} K_{\mu\nu} [\partial_\mu f][\partial_\nu g] \, \mathrm{d}^N x \qquad (91) \\ &= -\int_{\partial V} f \underbrace{\sum_{\mu,\nu} n_\mu p_{\mathbf{x}} K_{\mu\nu} \partial_\nu g}_{\overset{(67)}{=}0} \, \mathrm{d}A + \int p_{\mathbf{x}} \sum_{\mu,\nu} K_{\mu\nu} [\partial_\mu f][\partial_\nu g] \mathrm{d}^N x \\ &\quad \text{(Gauss' theorem)} \qquad (92) \\ &= \int p_{\mathbf{x}} \sum_{\mu,\nu} K_{\mu\nu} [\partial_\mu g][\partial_\nu f] \, \mathrm{d}^N x \qquad (93) \\ &\quad \text{(since } K_{\mu\nu} \text{ is symmetric)} \\ &\overset{(93-90)}{=} (\mathcal{D} f, g) \, . \qquad (94) \end{aligned}$$

■

This property is useful, because it allows the application of the spectral theorem known from functional analysis (Courant and Hilbert, 1989), which states that any self-adjoint operator possesses a complete set of eigenfunctions $f_j(\mathbf{s}) \in \mathcal{F}_b$ with real eigenvalues Δ_j, which are pairwise orthogonal, i.e. a set of functions that fulfill the following conditions:

$$\mathcal{D}f_j = \Delta_j f_j \quad \text{with } \Delta_j \in \mathbb{R} \qquad \text{(eigenvalue equation)}\,, \tag{95}$$

$$(f_i, f_j) = \delta_{ij} \qquad \text{(orthonormality)}\,, \tag{96}$$

$$\forall f \in \mathcal{F}_b \,\exists\, \alpha_k : f = \sum_{k=0}^{\infty} \alpha_k f_k \qquad \text{(completeness)}\,. \tag{97}$$

The eigenfunctions, normalized according to (96), thus fulfill the unit variance and decorrelation constraints (96). If we set $\lambda_{0j} = \lambda_{ji} = 0$ for $i \neq j$, the eigenfunctions also solve the Euler-Lagrange equation (71), which makes them good candidates for the solution of optimization problem 2. To show that they indeed minimize the Δ-value we need

Lemma 5 *The Δ-value of the normalized eigenfunctions f_j is given by their eigenvalue Δ_j.*

Proof

$$\Delta(f_j) \overset{(16,8,93-90)}{=} (f_j, \mathcal{D}f_j) \overset{(95)}{=} (f_j, \Delta_j f_j) = \Delta_j \underbrace{(f_j, f_j)}_{=1} \overset{(96)}{=} \Delta_j\,. \tag{98}$$

■

Proof of Theorem 1

At this point, we have everything we need to prove Theorem 1.

Proof Without loss of generality we assume that the eigenfunctions f_j are ordered by increasing eigenvalue, starting with the constant $f_0 = 1$. There are no negative eigenvalues, because according to Lemma 3, the eigenvalue is the Δ-value of the eigenfunction, which can only be positive by definition. According to Lemma 3, the optimal responses g_j obey the boundary condition (67) and are thus elements of the subspace $\mathcal{F}_b \subset \mathcal{F}$ defined in Lemma 2. Because of the completeness of the eigenfunctions on $\mathcal{F}_b$ we can do the expansion

$$g_j = \sum_{k=1}^{\infty} \alpha_{jk} f_k \tag{99}$$

where we may omit f_0 because of the zero mean constraint. We can now prove by complete induction that $g_j = f_j$ solves the optimization problem.
Basis (j=1): Inserting g_1 into equation (71) we find

$$0 = \mathcal{D}g_1 - \lambda_{10} - \lambda_{11} g_1 \tag{100}$$

$$\overset{(99,95)}{=} -\lambda_{10} + \sum_{k=1}^{\infty} \alpha_{1k}(\Delta_k - \lambda_{11}) f_k \tag{101}$$

$$\Longrightarrow \quad \begin{array}{c} \lambda_{10} = 0 \\ \wedge \quad (\alpha_{1k} = 0 \vee \Delta_k = \lambda_{11})\, \forall k\,, \end{array} \tag{102}$$

because f_k and the constant are linearly independent and (100) must be fulfilled for all $\mathbf{x}$. Equation (102) implies that the coefficients α_{1k} have to vanish unless the Δ-value of the associated eigenfunction is equal to λ_{11}. Thus, only eigenfunctions that have the same Δ-value can have non-vanishing coefficients. Therefore, the optimal response g_1 must also be an eigenfunction of $\mathcal{D}$. Since the Δ-value of the eigenfunctions is given by their eigenvalue, it is obviously optimal to chose $g_1 = f_1$. Note that although this choice is optimal, it is not necessarily unique, since there may be several eigenfunctions with the same eigenvalue. In this case any linear combination of these functions is also optimal.

Induction step: Given that $g_i = f_i$ for $i < j$, we prove that $g_j = f_j$ is optimal. Because of the orthonormality of the eigenfunctions the decorrelation constraint (20) yields

$$0 \overset{(20)}{=} (g_i, g_j) = (f_i, \sum_{k=1}^{\infty} \alpha_{jk} f_k) = \sum_{k=1}^{\infty} \alpha_{jk} \underbrace{(f_i, f_k)}_{\overset{(96)}{=} \delta_{ik}} = \alpha_{ji} \quad \forall i < j\,. \tag{103}$$

Again inserting the expansion (99) into (71) yields

$$\begin{aligned} 0 \quad &\overset{(71,99)}{=} \quad (\mathcal{D} - \lambda_{jj}) \sum_{k=1}^{\infty} \alpha_{jk} f_k - \lambda_{j0} - \sum_{i<j} \lambda_{ji} f_i && (104) \\ &\overset{(103)}{=} \quad (\mathcal{D} - \lambda_{jj}) \sum_{k=j}^{\infty} \alpha_{jk} f_k - \lambda_{j0} - \sum_{i<j} \lambda_{ji} f_i && (105) \\ &\overset{(95)}{=} \quad \sum_{k=j}^{\infty} \alpha_{jk} (\Delta_k - \lambda_{jj}) f_k - \lambda_{j0} - \sum_{i<j} \lambda_{ji} f_i && (106) \\ &\Longrightarrow \quad \begin{array}{lcl} & \lambda_{j0} = 0 & \\ \wedge & \lambda_{ji} = 0 & \forall i < j \\ \wedge & (\alpha_{jk} = 0 \vee \Delta_k = \lambda_{jj}) & \forall k \geq j\,, \end{array} && (107) \end{aligned}$$

because the eigenfunctions f_i are linearly independent. The conditions (107) can only be fulfilled if g_j is an eigenfunction of $\mathcal{D}$. Because of Lemma 3 an optimal choice for minimizing the Δ-value without violating the decorrelation constraint is $g_j = f_j$.

■