LETTER Communicated by Gal Chechik

Predictive Coding and the Slowness Principle: An Information-Theoretic Approach

Felix Creutzig
felix@creutzig.de
Henning Sprekeler
h.sprekeler@biologie.hu-berlin.de
Institute for Theoretical Biology, Humboldt-Universität zu Berlin, 10115 Berlin, Germany, and Bernstein Center for Computational Neuroscience, 10115 Berlin, Germany

Understanding the guiding principles of sensory coding strategies is a main goal in computational neuroscience. Among others, the principles of predictive coding and slowness appear to capture aspects of sensory processing. Predictive coding postulates that sensory systems are adapted to the structure of their input signals such that information about future inputs is encoded. Slow feature analysis (SFA) is a method for extracting slowly varying components from quickly varying input signals, thereby learning temporally invariant features. Here, we use the information bottleneck method to state an information-theoretic objective function for temporally local predictive coding. We then show that the linear case of SFA can be interpreted as a variant of predictive coding that maximizes the mutual information between the current output of the system and the input signal in the next time step. This demonstrates that the slowness principle and predictive coding are intimately related.

1 Introduction

One outstanding property of sensory systems is the identification of invariances. The visual system, for example, can reliably identify objects after changes in distance (Kingdom, Keeble, & Moulden, 1995), translation (Hubel & Wiesel, 1962), and size and position (Ito, Tamura, Fujita, & Tanaka, 1995). Neuronal correlates of invariance detection range from phase-shift invariance in complex cells in primary visual cortex (Hubel & Wiesel, 1962) to high-level invariances related to face recognition (Quiroga, Reddy, Kreiman, Koch, & Fried, 2005). Hence, understanding the computational principles behind the identification of invariances is of considerable interest.

One approach for the self-organized formation of invariant representations is based on the observation that objects are unlikely to change or disappear completely from one moment to the next. Various paradigms for invariance learning have been proposed that exploit this observation

(Földiak, 1991; Wallis & Rolls, 1997; O'Reilly & Johnson, 1994; Stone & Bray, 1995; Einhäuser, Hipp, Eggert, Körner, & König, 2005). As these paradigms extract the slowly varying components of sensory signals, we will refer to this approach as the slowness principle (Wiskott & Sejnowski, 2002), in related literature also called temporal coherence or temporal stability principle (Einhäuser et al., 2005; Hurri & Hyvärinen, 2003; Wyss, König, & Verschure, 2006). One formulation of this principle is slow feature analysis (SFA; Wiskott & Sejnowski, 2002). SFA has been successfully applied to the learning of various invariances in a model of the visual system (Wiskott & Sejnowski, 2002) and reproduces a wide range of properties of complex cells in primary visual cortex (Berkes & Wiskott, 2005). In combination with a sparseness objective, SFA can also be used as a model for the self-organized formation of place cells in the hippocampus (Franzius, Sprekeler, & Wiskott, 2007; for related work, see Wyss et al., 2006).

A different approach to sensory processing is based on temporal prediction. For successful completion of many tasks, our brain has to predict future states of the environment from current or previous knowledge (Bialek, Nemenman, & Tishby, 2001). For example, when trying to catch a ball, it is not the current position of the ball that is relevant but its position at the moment of the catch. We will refer to processing strategies that aim at performing this prediction as predictive coding. Predictive coding is the precondition for certain forms of redundancy reduction that have been applied successfully to model receptive fields in primary visual cortex (Rao & Ballard, 1999) and surround inhibition in the retina (Srinivasan, Laughlin, & Dubs, 1982). Redundancy reduction has been proposed as the backbone of efficient coding strategies and inherently relates to information-theoretic concepts (Attneave, 1954; Barlow, 1961; Atick, 1992; Nadal & Parga, 1997). However, to our knowledge, an information-theoretic framework for predictive coding has not yet been formulated.

In this work, we use the information bottleneck method (Tishby, Pereira, & Bialek, 1999) to derive an information-theoretic objective function for predictive coding. The information about previous input is compressed into a variable such that this variable keeps information about the subsequent input. We focus on gaussian input signals and linear mapping. In this case, the optimization problem underlying the information bottleneck can be reduced to an eigenvalue problem (Chechik, Globerson, Tishby, & Weiss, 2005). We show that the solution to this problem is similar to linear slow feature analysis, thereby providing a link between the learning principles of slowness and predictive coding.

2 Linear SFA

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 $X_t = [X_1(t), \ldots, X_N(t)]^T$ with t indicating time, find a set of J real-valued instantaneous functions $g_j(X)$ of the input such that the output signals $(Y_j)_t := g_j(X_t)$ minimize

$$\Delta(Y_j) \equiv \left\langle \dot{Y}_j^2 \right\rangle_t \tag{2.1}$$

under the constraints

$$\langle Y_j \rangle_t = 0 \quad \text{(zero mean)} \tag{2.2}$$

$$\left\langle Y_j^2 \right\rangle_t = 1 \quad \text{(unit variance)} \tag{2.3}$$

$$\forall i < j : \langle Y_i Y_j \rangle_t = 0 \quad \text{(decorrelation and order)} \tag{2.4}$$

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

Equation 2.1 introduces the Δ-value, which is a measure of the slowness of the signal Y_t. The constraints 2.2 and 2.3 avoid the trivial constant solution. Constraint 2.4 ensures that different functions g_j code for different aspects of the input.

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 slowness cannot be enforced by low-pass filtering. Slow output signals can be achieved only if the input signal contains slowly varying features that can be extracted by the functions g_j.

If the function space $\mathcal{F}$ is finite-dimensional, the optimization problem can be reduced to a (generalized) eigenvalue problem (Wiskott & Sejnowski, 2002; Berkes & Wiskott, 2005). Here, we restrict $\mathcal{F}$ to the set of linear functions $Y_t = AX_t$, where A is a $J \times N$-dimensional matrix. In the following, we also assume that input signals X_t have zero mean. Then the optimal matrix obeys the generalized eigenvalue equation,

$$A\Sigma_{\dot{X}_t} = \Lambda A \Sigma_{X_t}. \tag{2.5}$$

Here, $\Sigma_{\dot{X}} := \langle \dot{X}\dot{X}^T \rangle_t$ denotes the matrix of the second moments of the temporal derivative of the input signals, and Σ_{X_t} is the covariance matrix of the input signals. Λ is a diagonal matrix that contains the eigenvalues λ_j on the diagonal. The solution of the optimization problem for SFA is given by the $J \times N$ matrix A that contains the eigenvectors to the smallest

eigenvalues λ_j as determined by the generalized eigenvalue equation, 2.5. For the mathematically interested reader, a derivation of equation 2.5 can be found in appendix A.

We assume that the covariance matrix of the input data has full rank and is thus invertible. The generalized eigenvalue problem, 2.5, can then be reduced to a standard left eigenvalue problem by multiplication with $\Sigma_{X_t}^{-1}$ from the right:

$$A\left[\Sigma_{\dot{X}_t}\Sigma_{X_t}^{-1}\right] = \Lambda A. \tag{2.6}$$

For discretized time, the temporal derivative is replaced by $X_{t+1} - X_t$, and $\Sigma_{\dot{X}}$ can be rewritten as $\Sigma_{\dot{X}} = 2\Sigma_X - [\Sigma_{X_{t+1};X_t} + \Sigma_{X_t;X_{t+1}}]$, where $\Sigma_{X_{t+1};X_t} = \langle X_{t+1}X_t\rangle_t$ is the matrix containing the covariance of the input signals with the input signal delayed by one time step 1 (Blaschke, Berkes, & Wiskott, 2006). Moreover, if the statistics of the input data are reversible, $\Sigma_{X_{t+1};X_t}$ is symmetric and $\Sigma_{X_{t+1};X_t} = \Sigma_{X_t;X_{t+1}}$. Using these relations in equation 2.6 yields

$$2A\left[I - \underbrace{\Sigma_{X_{t+1};X_t}\Sigma_{X_t}^{-1}}_{=:\Sigma}\right] = \Lambda A. \tag{2.7}$$

Note that the eigenvectors of the SFA problem are also the eigenvectors of the matrix Σ as defined in equation 2.7. Given the form of equation 2.7, we will be able to compare the eigenvalue problem with its counterpart from the information bottleneck ansatz of predictive coding.

3 The Information Bottleneck Method

The information bottleneck is a method for extracting relevant aspects of data (Tishby et al., 1999). One seeks to capture those components of a random variable X that can explain observed values of another variable R. This task is achieved by compressing the variable X into its compressed representation Y while preserving as much information as possible about R. The trade-off between these two targets is controlled by the trade-off parameter β. Hence, the information bottleneck problem can be formalized as minimizing the following Lagrangian:

$$\min \mathcal{L} : \mathcal{L} \equiv I(X;Y) - \beta I(Y;R). \tag{3.1}$$

The first term can be regarded as minimizing the complexity of the mapping, while the second term tries to increase the accuracy of the representation. From the point of view of clustering, the information bottleneck method finds a quantization, or partition, of X that preserves as much mutual information as possible about R. From the perspective of machine learning,

this corresponds to supervised learning. X is the input signal, and R tells what aspects of the input should be learned. The information bottleneck method has been applied successfully in different circumstances: for document clustering (Slonim & Tishby, 2000), neural code analysis (Dimitrov & Miller, 2001), gene expression analysis (Slonim, Friedman, & Tishby, 2006), and extraction of speech features (Hecht & Tishby, 2005). In particular, in case of a linear mapping between gaussian variables, the optimal functions are the solution of an eigenvalue problem (Chechik et al., 2005). The key point is that the entropy of gaussian variables can be written as the logarithm of the relevant covariance matrices between input and output. Minimizing the Lagrangian, finally, is equivalent to diagonalizing the covariance matrices; the eigenvector with the smallest respective eigenvalue gives the most informative part of the mapping between input and output.

In the following, we transfer the gaussian information bottleneck to sensory input data represented as a time series. We obtain a low-dimensional encoding of the current input, while maximizing the information about the subsequent input, and thus maximize predictive information.

4 Temporally Local Predictive Coding

The predictive coding hypothesis states that an organism extracts information from its sensory input that is predictive for the future (see, e.g., Bialek et al., 2001). Information theoretically, this corresponds to mapping the data from the past into an internal state variable such that information between that state and the future data is maximized. To enforce a compact mapping, we introduce an additional penalty term that restricts the complexity of the mapping:

$$\max \mathcal{L} : \mathcal{L} \equiv I(\mathit{state}; \mathit{future}) - \beta^{-1} I(\mathit{past}; \mathit{state}). \tag{4.1}$$

Obviously the state variable cannot contain more information about the future than about the past, so for $\beta^{-1} \geq 1$, the objective function $\mathcal{L}$ is negative: $\mathcal{L} \leq 0$. In this case, $\mathcal{L}$ is optimized by the trivial solution, where the state variable does not contain any information at all because then $\mathcal{L} = 0$. Thus, to obtain nontrivial solutions, the trade-off parameter should be chosen such that $0 < \beta^{-1} < 1$ or, equivalently, $1 < \beta < \infty$.

The optimization problem above can also be formulated as an equivalent minimization problem that has the form of an information bottleneck as introduced in the previous section:

$$\min \mathcal{L} : \mathcal{L} \equiv I(\mathit{past}; \mathit{state}) - \beta I(\mathit{state}; \mathit{future}). \tag{4.2}$$

Here, we restrict ourselves to the special case of only one time step and a linear mapping. An extension to more time steps is possible with similar

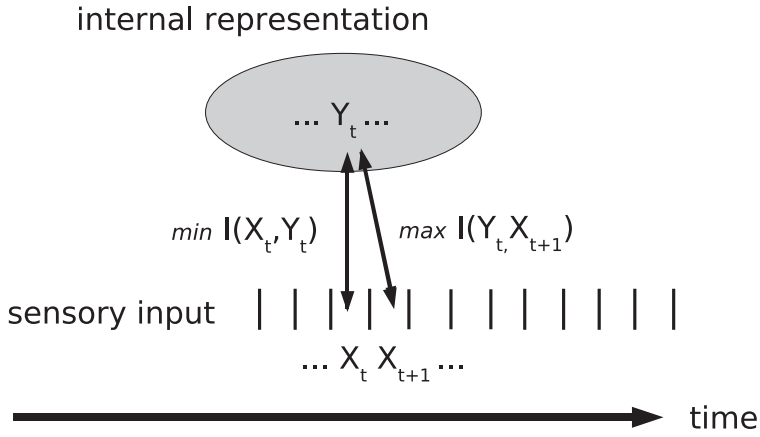

Figure 1: Temporally local predictive coding. The sensory system compresses information of the current input X_t into Y_t such that the mutual information between Y_t and the next input X_t is maximized.

techniques as presented here but exceeds the scope of this letter. Let us assume a discrete input signal X_t that is mapped to an output signal Y_t such that Y_t is most predictive about the next input signal X_{t+1} while minimizing the complexity in the information bottleneck sense, as illustrated in Figure 1.

We assume that the input signal X_t is an n-dimensional gaussian vector and that the output signal Y_t is generated by a noisy linear transformation:

$$Y_t = AX_t + \xi. \tag{4.3}$$

The gaussian white process noise ξ is introduced for reasons of regularization; otherwise, information-theoretic quantities would diverge. For simplicity, we will assume that the noise is isotropic and normalized—that $\Sigma_\xi = \langle \xi \xi^T \rangle_t = I$, where I denotes the unit matrix. This is no limitation, as it has been shown that every pair of (A, Σ_ξ) can be mapped into another pair $(\hat{A}, I)$ such that the value of the target function $\mathcal{L}$ remains the same (Chechik et al., 2005).

Optimization problem 1 can now be stated in information-theoretic terms:

Optimization problem 2. Temporally local predictive coding (TLPC). Given input signal X_t and output signal $Y_t = AX_t + \xi$ where X_t and ξ are gaussian with $\langle \xi_t \xi_{t+1} \rangle_t = 0$, find the matrix $A(\beta)$ that minimizes

$$\min \mathcal{L} : \mathcal{L}_{TLPC} \equiv I(X_t; Y_t) - \beta\, I(Y_t; X_{t+1}) \tag{4.4}$$

with $\beta > 1$.

The general solution to this problem has been derived in Chechik et al. (2005). For completeness, a sketch of the derivation can be found in appendix B. Here we state the solution:

Proposition 1. *The solution to optimization problem 2 is given by*

$$A(\beta) = \begin{cases} [\mathbf{0}; \ldots; \mathbf{0}] & 0 \leq \beta \leq \beta_1^c \\ [\alpha_1 W_1; \mathbf{0}; \ldots; \mathbf{0}] & \beta_1^c \leq \beta \leq \beta_2^c \\ [\alpha_1 W_1; \alpha_2 W_2; \mathbf{0}; \ldots; \mathbf{0}] & \beta_2^c \leq \beta \leq \beta_3^c \\ \vdots & \end{cases} \tag{4.5}$$

where W_i and λ_i (assume $\lambda_1 \leq \lambda_2 \leq \ldots$) are the left eigenvectors and eigenvalues of $\Sigma_{X_t|X_{t+1}} \Sigma_{X_t}^{-1}$, α_i are coefficients defined by $\alpha_i \equiv \sqrt{\frac{\beta(1-\lambda_i)-1}{\lambda_i r_i}}$, $r_i \equiv W_i \Sigma_{X_t} W_i^T$, $\mathbf{0}$ is an m-dimensional column vector of zeros, and semicolons separate columns in the matrix $A(\beta)$. The critical β-values are $\beta_i^c = \frac{1}{1-\lambda_i}$.

The eigenvalues of $\Sigma_{X_t|X_{t+1}} \Sigma_{X_t}^{-1}$, α_i are guaranteed to be real and nonnegative, as full-rank covariance matrices are positive definite. The key observation is that with increasing β additional eigenvectors appear (second-order phase transitions), corresponding to the detection of additional features of decreasing information content.

5 Relationship Between Slow Feature Analysis and Temporally Local Predictive Coding

How does this solution relate to slow feature analysis? We can rewrite $\Sigma = \Sigma_{X_t; X_{t+1}} \Sigma_{X_t}^{-1}$ in a more convenient form using Schur's formula:

$$\Sigma_{X_t|X_{t+1}} \Sigma_{X_t}^{-1} = \left(\Sigma_{X_t} - \Sigma_{X_t; X_{t+1}} \Sigma_{X_t}^{-1} \Sigma_{X_{t+1}; X_t}\right) \Sigma_{X_t}^{-1} \tag{5.1}$$

$$= I - \Sigma_{X_t; X_{t+1}} \Sigma_{X_t}^{-1} \Sigma_{X_{t+1}; X_t} \Sigma_{X_t}^{-1} \tag{5.2}$$

$$= I - \left(\Sigma_{X_t; X_{t+1}} \Sigma_{X_t}^{-1}\right)^2 \tag{5.3}$$

$$\overset{(7)}{=} I - \Sigma^2 , \tag{5.4}$$

where we used the fact that time-delayed covariance matrices of reversible processes are symmetric. Note that the matrix $\Sigma = \Sigma_{X_t; X_{t+1}} \Sigma_{X_t}^{-1}$ also appears in the eigenvalue problem for linear SFA in the case of discrete time series 2.7, and hence, the optimal eigenvectors are the same for temporally local

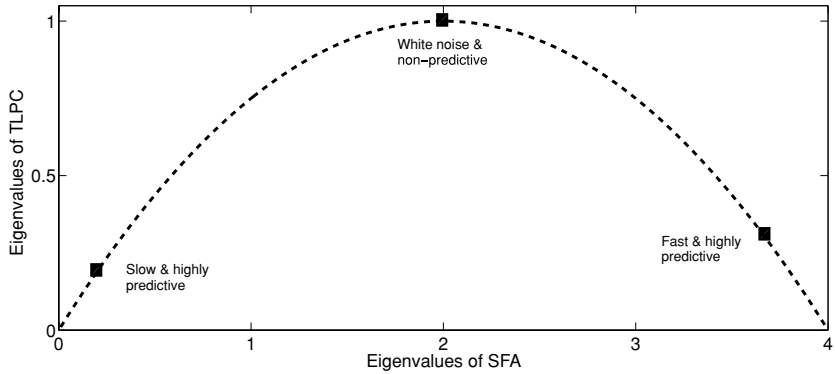

Figure 2: Relationship between eigenvalues of slow feature analysis and temporally local predictive coding. For discrete time series, fast components can be equally predictive as slow components. Only white noise is nonpredictive.

predictive coding (TLPC) and SFA. From equation 2.7, we know that that the matrix to diagonalize in SFA is

$$\Sigma_{SFA} = 2I - 2\Sigma, \tag{5.5}$$

with eigenvalues λ_i^{SFA}, whereas in TLPC, the target matrix is

$$\Sigma_{TLPC} = I - \Sigma^2, \tag{5.6}$$

with eigenvalues λ_i^{TLPC}. Solving equation 5.5 for Σ and substituting the solution into equation 5.6, we obtain the relationship between the eigenvalues:

$$\lambda_i^{TLPC} = \lambda_i^{SFA} - \frac{1}{4}\left(\lambda_i^{SFA}\right)^2. \tag{5.7}$$

SFA is guaranteed to find the slowest components first, whereas TLPC finds the most predictive components first. For example, a very fast component can be very predictive, for example, if the value at $t + 1$ is the negative of the current value (see Figure 2). Hence, from the TLPC point of view, the absolute deviation from random fluctuations rather than slowness is relevant. This may be important for the analysis of discrete time series with high-frequency components. However, this is true only for temporally discrete data: for continuous data, one would expect a monotonous relation between eigenvalues of an information bottleneck approach and SFA eigenvalues.

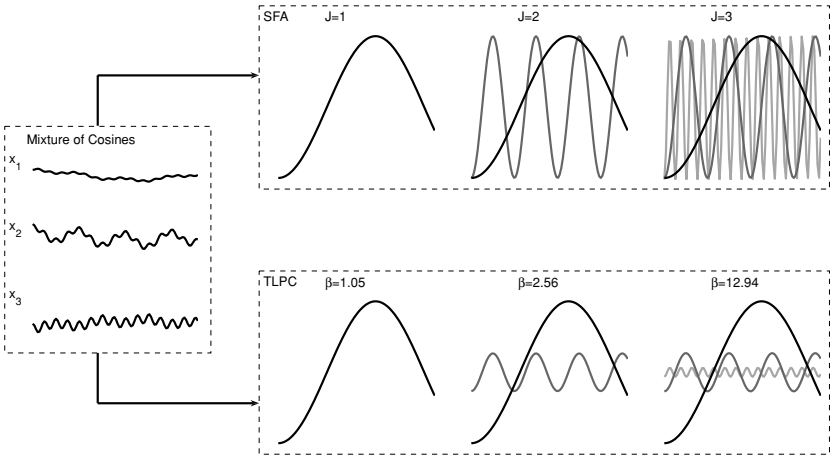

Figure 3: Temporally local predictive coding and SFA differ in weighting of filtered components. Both algorithms find the original cosines underlying the linear mixtures (x_1, x_2, x_3). SFA discovers features in order of slowness only. TLPC assigns weights to individual components according to their predictive information. For TLPC, relative, not absolute, weightings are shown.

Temporally local predictive coding and SFA find the same components in the same order. The difference is that TLPC allows quantifying the components in terms of predictive information. For example, take a three-dimensional signal that consists of a mixture of cosines with different frequencies. Both methods can separate the original signals successfully (see Figure 3). Slow feature analysis and temporally local predictive coding reveal components in the same order, that is, according to slowness. However, slow feature analysis accredits the same amplitude to all components, while TLPC gives higher weights to slower components according to their predictive power.

6 Discussion

In this work, we relate slowness in signals to predictability. We have shown that predictive coding and slow feature analysis correspond to each other for the restrictions of gaussianity, linearity, and one-time-step prediction. Both principles can explain some properties of visual receptive fields (Berkes & Wiskott, 2005; Einhäuser et al., 2005; Rao & Ballard, 1999). On the one hand, our approach indicates that results from SFA studies such as the findings on complex cell properties (Berkes & Wiskott, 2005) and hippocampal place cells (Franzius et al., 2007) can be seen in terms of predictive coding. On the other hand, predictive coding by surround inhibition (Srinivasan et al.,

1982) and feedback connections (Rao & Ballard, 1999) may be interpreted from the viewpoint of the slowness principle.

We have also shown that linear slow feature analysis can be motivated by information-theoretic principles. It is interesting to note that this linear, discrete case is also related to an implementation of second-order independent component analysis (Blaschke et al., 2006).

The relationship between predictive coding and temporal invariance learning has also been suggested in other work, for example, by Shaw (2006), who argued that temporal invariance learning is equivalent to predictive coding if the input signals are generated from Ornstein-Uhlenbeck processes.

In one regard, temporally local predictive coding differs from slow feature analysis. The information bottleneck approach is continuous in terms of the trade-off parameter β, and new eigenvectors appear as second-order phase transitions. The weighting of the eigenvectors is different in that it depends on their eigenvalue (see Figure 3). This can be important when analyzing or modeling sensory systems where available bandwidth and, hence, resulting signal-to-noise ratio, is a limiting factor. For temporally local predictive coding, available bandwidth, such as number of neurons, should be attributed according to relative amplitude, whereas slow feature analysis accredits the same bandwidth to all features.

We emphasize that our approach is not directly applicable to many real-world problems. Our derivation is restricted to gaussian variables and linear mappings. Both restrictions are not needed for SFA. Note that an extension of linear local predictive coding to nongaussian input signals would also capture the case of nonlinear processing, because after a nonlinear expansion, the problem can be treated in a linear fashion. Usually nonlinear SFA corresponds to linear SFA after a nonlinear expansion of the input signals. In this sense, nonlinear SFA can be regarded as the gaussian approximation to the full nongaussian local predictive coding problem on the nonlinearly expanded input. This argument, together with effective nonlinear SFA models of the visual system (Berkes & Wiskott, 2005; Franzius et al., 2007), indicates that sensory systems are tailored to extract (relevant) predictive information. For further research, we suggest comparing temporally local predictive coding and slow feature analysis to generative hierarchical models for learning nonlinear statistical regularities (Karklin & Lewicki, 2005; Schwartz, Sejnowski, & Dayan, 2006).

The restriction on the immediate past implies that SFA does not maximize predictive information for non-Markovian processes. The generalization—relating the infinite past with the infinite future—can be best framed in terms of linear dynamical systems. Work on this topic is in preparation. Finally, predictive coding is not a stationary property of the evolved sensory system but dynamic and adapts with input statistics (Hosoya, Baccus, & Meister, 2005). A plausible extension of our work would aim to incorporate dynamic properties.

Appendix A: Derivation of the Generalized Eigenvalue Equation for SFA

Let W_j denote the row vector that is formed by the jth row of the weight matrix A. The output signal Y_j is then given by $Y_j = W_j X$. Accordingly, the slowness objective, equation 2.1, is given by

$$\Delta(Y_j) = \left\langle \dot{Y}_j^2 \right\rangle_t \tag{A.1}$$

$$= \langle (W_j \dot{X})^2 \rangle_t \tag{A.2}$$

$$= W_j \langle \dot{X} \dot{X}^T \rangle_t W_j^T = W_j \Sigma_{\dot{X}_t} W_j^T . \tag{A.3}$$

A similar calculation yields that the variance of the output signal Y_j is given by

$$\mathrm{var}(Y_j) \equiv \left\langle Y_j^2 \right\rangle_t = W_j \langle X X^T \rangle_t W_j^T = W_j \Sigma_{X_t} W_j^T \overset{(3)}{=} 1 . \tag{A.4}$$

The task is to minimize equation A.1 under the constraint A.4 and the decorrelation constraint, which we will neglect for now, as it will turn out to be fulfilled automatically. The method of Lagrange multipliers states the necessary condition that

$$\Psi = \Delta(Y_j) - \lambda \left\langle Y_j^2 \right\rangle_t \tag{A.5}$$

is stationary for some value of the Lagrange multiplier λ, that is, that the gradient of Ψ with respect to the weight vector W_j vanishes. When equations A.1 and A.4 are used, this gradient can be calculated analytically, yielding the following necessary condition for the weight vector W_j:

$$W_j \Sigma_{\dot{X}_t} - \lambda W_j \Sigma_{X_t} = 0. \tag{A.6}$$

Note that condition A.6 has the structure of a generalized eigenvalue problem, where the Lagrange multiplier λ plays the role of the eigenvalue. Multiplication with W_j^T from the right and using the unit variance constraint A.4 yields that the Δ-value of a solution of equation A.6 is given by its eigenvalue λ:

$$\underbrace{W_j \Sigma_{\dot{X}_t} W_j^T}_{\overset{(21)}{=} \Delta(Y_j)} - \lambda \underbrace{W_j \Sigma_{X_t} W_j^T}_{\overset{(24)}{=} \langle Y^2 \rangle_t = 1} = 0 \quad \Rightarrow \quad \Delta(Y_j) = \lambda . \tag{A.7}$$

From this, it is immediately clear that the slowest possible output signal is provided by the linear function associated with the eigenvector W_1 with the smallest eigenvalue λ_1. It can be shown that eigenvectors

W_i, W_j with different eigenvalues λ_i, λ_j are orthogonal in the sense that $\langle Y_i Y_j \rangle_t = W_i \Sigma_{X_t} W_j = 0$, so they yield decorrelated output signals. For eigenvectors with identical eigenvalues, any linear combination of them is still an eigenvector. Hence, it is always possible to choose a basis of the subspace that still consists of eigenvectors and yields decorrelated output signals (e.g., by Gram-Schmidt orthogonalization).

Combining these properties of the eigenvectors, it is clear that the optimization problem of linear SFA can be solved by choosing the functions associated with the J eigenvectors W_j with the smallest eigenvalues, ordered by their eigenvalue. Reinserting the eigenvectors W_j into the matrix A and the eigenvalues in a diagonal matrix Λ, the eigenvalue problem, A.6, takes the form of equation 2.5:

$$A\Sigma_{\dot{X}_t} = \Lambda A \Sigma_{X_t}. \tag{A.8}$$

Appendix B: Derivation of the Optimal Weight Matrix for Local Predictive Coding

We first rewrite the mutual information quantities in the objective function for TLPC in terms of differential entropies:

$$\mathcal{L}_{TLPC} = I(Y_t, X_t) - I(Y_t, X_{t+1}) \tag{B.1}$$

$$= h(Y_t) - h(Y_t \mid X_t) - \beta h(Y_t) + \beta h(Y_t \mid X_{t+1}). \tag{B.2}$$

Here, the differential entropy of a stochastic variable Z is given by $h(Z) = -\int_Z f(z) \log f(z)\, dz$ with $f(z)$ denoting the probability density of Z. In particular, for gaussian variables, the differential entropy becomes

$$h(Z) = \frac{1}{2} \log (2\pi e)^d |\Sigma_Z|, \tag{B.3}$$

where $|\Sigma_Z|$ denotes the determinant of Σ_Z and $\Sigma_Z := \langle Z Z^T \rangle_t$ is the covariance matrix of Z (Cover & Thomas, 1991). Hence, we have to find the covariance matrices of the quantities in equation B.2. As $Y_t = AX_t + \xi$, we have $\Sigma_{Y_t} = A\Sigma_{X_t} A^T + \Sigma_\xi$ and $\Sigma_{Y_t|X_t} = \Sigma_\xi$. The last covariance matrix is obtained as follows:

$$\Sigma_{Y_t|X_{t+1}} = \Sigma_{Y_t} - \Sigma_{Y_t;X_{t+1}} \Sigma^{-1}_{X_{t+1}} \Sigma_{X_{t+1};Y_t} \tag{B.4}$$

$$= A\Sigma_{X_t} A^T + \Sigma_\xi - A\Sigma_{X_t;X_{t+1}} \Sigma^{-1}_{X_{t+1}} \Sigma_{X_{t+1};X_t} A^T \tag{B.5}$$

$$= A\Sigma_{X_t|X_{t+1}} A^T + \Sigma_\xi, \tag{B.6}$$

where we used Schur's formula, i.e. $\Sigma_{X|Y} = \Sigma_X - \Sigma_{X;Y}\Sigma_Y^{-1}\Sigma_{Y;X}$, in the first and last step (Magnus & Neudecker, 1988). Neglecting irrelevant constants and using that the noise is isotropic, the objective function, equation B.2, becomes

$$\mathcal{L} = (1-\beta)\log|A\Sigma_{X_t}A^T + I| + \beta\log|A\Sigma_{X_t|X_{t+1}}A^T + I|. \tag{B.7}$$

The derivative of the objective function with respect to the weight matrix is given by

$$\frac{d\mathcal{L}}{dA} = (1-\beta)\big(A\Sigma_{X_t}A^T + I\big)^{-1}2A\Sigma_{X_t} + \beta\big(A\Sigma_{X_t|X_{t+1}}A^T + I\big)^{-1}2A\Sigma_{X_t|X_{t+1}}. \tag{B.8}$$

Equating this to zero and rearranging, we obtain a necessary condition for the weight matrix A:

$$\frac{\beta-1}{\beta}\underbrace{(A\Sigma_{X_t|X_{t+1}}A^T + I)(A\Sigma_{X_t}A^T + I)^{-1}}_{=:M}A = A\Sigma_{X_t|X_{t+1}}\Sigma_{X_t}^{-1}. \tag{B.9}$$

We will prove that this equation can be solved by filling the rows of A with adequately scaled versions of the solutions W_j of the following generalized (left) eigenvalue problem:

$$W_j\Sigma_{X_t|X_{t+1}} = \lambda_j W_j\Sigma_{X_t}. \tag{B.10}$$

We first make some considerations on the solutions of the eigenvalue equation, B.10, and then insert them into equation B.9 to show that this yields M diagonal. It then becomes clear that there are scaling factors for the eigenvectors such that equation B.9 is solved.

1. W_j is a left eigenvector of $\Sigma_{X_t|X_{t+1}}\Sigma_{X_t}^{-1}$:

$$W_j\Sigma_{X_t|X_{t+1}} = \lambda W_j\Sigma_{X_t} \tag{B.11}$$

$$\Leftrightarrow\ W_j\Sigma_{X_t|X_{t+1}}\Sigma_{X_t}^{-1} = \lambda W_j. \tag{B.12}$$

2. M is diagonal: The crucial observation for this statement is that the eigenvectors W_j need not be orthogonal because $\Sigma_{X_t|X_{t+1}}\Sigma_{X_t}^{-1}$ is not necessarily symmetric. The structure of the generalized eigenvalue equation is such that solutions of equation B.10 with different eigenvalues λ are orthogonal with respect to the positive definite bilinear form induced by Σ_{X_t}:

$$(W_i, W_j) = W_i\Sigma_{X_t}W_j^T = r_i\delta_{ij} \quad \text{with} \quad r_i > 0. \tag{B.13}$$

When there are several eigenvectors with the same eigenvalue, it is always possible to choose eigenvectors W_i that are orthogonal in the sense above. Assume that the rows of A are filled with the eigenvectors W_j, scaled by a factor α_j. With this choice, $A\Sigma_{X_t}A^T + I$ is diagonal with diagonal elements $r_j\alpha_j^2 + 1$. Right multiplication of equation B.10 with W_j^T yields that $A\Sigma_{X_t|X_{t+1}}A^T + I$ is also diagonal with diagonal elements $r_j\lambda_j\alpha_j^2 + 1$. Thus, M is diagonal with diagonal elements $M_{jj} = \frac{r_j\alpha_j^2\lambda_j+1}{r_j\alpha_j^2+1}$.

3. Using the above results, equation B.10 becomes

$$\left[\frac{\beta-1}{\beta}\frac{\lambda_j\alpha_j^2 r_j + 1}{\alpha_j^2 r_j + 1} - \lambda_j\right]\alpha_j W_j = 0\,. \tag{B.14}$$

This equation can be solved only if either $\alpha_j = 0$ or

$$\frac{\beta-1}{\beta}\frac{\lambda_j\alpha_j^2 r_j + 1}{\alpha_j^2 r_j + 1} = \lambda_j\,. \tag{B.15}$$

Rearranging for α_j^2 yields the normalization stated in proposition 1:

$$\alpha_j^2 = \frac{\beta(1-\lambda_j)-1}{\lambda_j r_j}\,. \tag{B.16}$$

Of course, this equation can be solved only if the right-hand side is positive. Because r_j and λ_j are positive, this reduces to a relation between the β-value and the eigenvalues:

$$\beta \geq \frac{1}{1-\lambda_j}\,. \tag{B.17}$$

For the eigenvalues that do not fulfill this condition for a given β, equation B.14 can be solved only by $\alpha_j = 0$. This shows that the critical β-values as stated in proposition 1 are those where a new eigenvector becomes available. Moreover, we have now demonstrated that $A(\beta)$ as stated in proposition 1 is a solution of equation B.9. Note that in line with the fact that the objective function of optimization problem 1 is invariant with respect to orthogonal transformations of the output signals, any matrix $\tilde{A} = UA$ with $U^{-1} = U^T$ is also a solution of equation B.9. We refer the reader to Chechik et al. (2005) for the proof that $A(\beta)$ is not only a stationary point of equation B.9 but also minimizes the objective function, equation B.7.

Acknowledgments

Most of all, we thank Laurenz Wiskott for carefully discussing all details. We are grateful to Martin Stemmler, Jan Benda, and Thomas Creutzig for helpful comments on the manuscript. Naftali Tishby and Amir Globerson provided F.C. with many insights on the information bottleneck method.

We thank Andreas V. M. Herz and Laurenz Wiskott for excellent working environments. F.C. is supported by Boehringer Ingelheim Fonds and the German National Merit Foundation, H.S. by the Volkswagen Foundation.

References

Atick, J. J. (1992). Could information theory provide an ecological theory of sensory processing? *Network: Computation in Neural Systems*, *3*, 213–251.

Attneave, F. (1954). Some informational aspects of visual perception. *Psychol. Rev.*, *61*, 183–193.

Barlow, H. B. (1961). Possible principles underlying the transformation of sensory messages. In W. A. Rosenblith (Ed.), *Sensory communications* (pp. 217–234). Cambridge, MA: MIT Press.

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

Bialek, W., Nemenman, I., & Tishby, N. (2001). Predictability, complexity and learning. *Neural Computation*, *13*(11), 2409–2463.

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

Chechik, G., Globerson, A., Tishby, N., & Weiss, Y. (2005). Information bottleneck for gaussian variables. *Journal of Machine Learning Research*, *6*, 165–188.

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory*. New York: Wiley.

Dimitrov, A. G., & Miller, J. P. (2001). Neural coding and decoding: Communication channels and decoding. *Network: Computation in Neural Systems*, *12*, 441–472.

Einhäuser, W., Hipp, J., Eggert, J., Körner, E., & König, P. (2005). Learning viewpoint invariant object representations using a temporal coherence principle. *Biological Cybernetics*, *93*(1), 79–90.

Földiak, P. (1991). Learning invariance from transformation sequences. *Neural Computation*, *3*, 194–200.

Franzius, M., Sprekeler, H., & Wiskott, L. (2007). Slowness leads to place, head-direction, and spatial-view cells. *PLoS Comput. Biol.*, *3*, e166.

Hecht, R. M., & Tishby, N. (2005). Extraction of relevant speech features using the information bottleneck method. In *Proceedings of InterSpeech*. Available online @ http://www.isca-speech.org_order.html.

Hosoya, T., Baccus, S. A., & Meister, M. (2005). Dynamic predictive coding by the retina. *Nature*, *436*, 71–77.

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction, and functional interaction in cat's visual cortex. *J. Physiol.*, *160*, 106–154.

Hurri, J., & Hyvärinen, A. (2003). Simple-cell-like receptive fields maximize temporal coherence in natural video. *Neural Computation*, *15*(3), 663–691.

Ito, M., Tamura, H., Fujita, I., & Tanaka, K. (1995). Size and position invariance of neuronal responses in monkey inferotemporal cortex. *Journal of Neurophysiology*, *73*, 218–226.

Karklin, Y., & Lewicki, M. S. (2005). A hierarchical Bayesian model for learning nonlinear statistical regularities in non-stationary natural signals. *Neural Computation*, *17*(2), 397–423.

Kingdom, F. A. A., Keeble, D. R. T., & Moulden, B. (1995). The perceived orientation of aliased lines. *Vision Research*, *35*(19), 2759–2766.

Magnus, J. R., & Neudecker, H. (1988). *Matrix differential calculus with applications in statistics and econometrics*. New York: Wiley.

Nadal, J.-P., & Parga, N. (1997). Redundancy reduction and independent component analysis: Conditions on cumulants and adaptive approaches. *Neural Computation*, *9*(7), 1421–1456.

O'Reilly, R. C., & Johnson, M. H. (1994). Object recognition and sensitive periods: A computational analysis of visual imprinting. *Neural Computation*, *6*(3), 357–389.

Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C., & Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature*, *435*, 1102–1107.

Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*, 79–87.

Schwartz, O., Sejnowski, T. J., & Dayan, P. (2006). Soft mixer assignment in a hierarchical generative model of natural scene statistics. *Neural Computation*, *18*(11), 2680–2718.

Shaw, J. (2006). *Unifying perception and curiosity*. Unpublished doctoral dissertation, University of Rochester.

Slonim, N., Friedman, N., & Tishby, N. (2006). Multivariate information bottleneck. *Neural Computation*, *18*, 1739–1789.

Slonim, N., & Tishby, N. (2000). Document clustering using word clusters via the information bottleneck method. In N. Belkin, P. Ingwersen, & M.-K. Leong (Eds.), *Proc. Research and Development in Information Retrieval (SIGIR-00)* (pp. 208–215). New York: ACM Press.

Srinivasan, M. V., Laughlin, S. B., & Dubs, A. (1982). Predictive coding: A fresh view of inhibition in the retina. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, *216*(1205), 427–459.

Stone, J., & Bray, A. (1995). A learning rule for extracting spatio-temporal invariances. *Network: Computation in Neural Systems*, *6*(3), 429–436.

Tishby, N., Pereira, F. C., & Bialek, W. (1999). The information bottleneck method. In *Proc. of 37th Allerton Conference on Communication and Computation*. Monticello, IL.

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

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

Wyss, R., König, P., & Verschure, P. F. M. (2006). A model of the ventral visual system based on temporal stability and local memory. *Plos Biology*, *4*(5), e120.

Received January 30, 2007; accepted June 29, 2007.

This article has been cited by: