

Lyapunov Functions for Convergence of Principal Component Algorithms

Mark D Plumbley

Centre for Neural Networks,

Department of Mathematics, King's College London,

Strand, London WC2R 2LS, UK

mark@dcs.kcl.ac.uk

Abstract

Recent theoretical analyses of a class of unsupervised Hebbian principal component algorithms have identified its local stability conditions. The only locally stable solution for the subspace \mathbf{P} extracted by the network is the principal component subspace \mathbf{P}^* . In this paper we use the Lyapunov function approach to discover the *global* stability characteristics of this class of algorithms. The subspace projection error, least mean squared projection error, and mutual information I are all Lyapunov functions for convergence to the principal subspace, although the various domains of convergence indicated by these Lyapunov functions leave some of \mathbf{P} -space uncovered. A modification to I yields a ‘principal subspace information’ Lyapunov function I' with a domain of convergence which covers almost all of \mathbf{P} -space. This shows that this class of algorithms converges to the principal subspace from almost everywhere.

Keywords: Neural networks, unsupervised learning, principal component analysis, information theory, Hebbian algorithms, Lyapunov functions, Oja rule.

1 Introduction

Principal Component Analysis (PCA) is a popular statistical tool for linearly reducing the dimensionality of a set of measurements while retaining as much ‘information’ as possible about the original measurements. It appears under various names, including Karhunen-Loève Transform in signal processing, the Hotelling transform in image processing, and Factor Analysis in the statistical literature (see e.g. (Watanabe, 1985)).

Suppose we have a linear transform from an N -dimensional zero-mean input vector $\mathbf{x} = [x_1, \dots, x_N]$ to an M -dimensional output vector $\mathbf{y} = [y_1, \dots, y_M]$, and \mathbf{y} is related to \mathbf{x} by the expression

$$\mathbf{y} = \mathbf{W}\mathbf{x} \tag{1}$$

where \mathbf{W} is an $M \times N$ matrix, with $M < N$. Principal Component Analysis sets the M successive rows of \mathbf{W} to the M largest eigenvectors of the input covariance matrix $\Sigma_{\mathbf{x}} = E(\mathbf{x}\mathbf{x}^T)$. Thus y_1 represents the component of \mathbf{x} in the direction of its largest eigenvector of $\Sigma_{\mathbf{x}}$ (the *principal* component), y_2 the component in the direction of the second largest, and so on. PCA is an optimal linear transform in the sense that it minimizes the least mean squared reconstruction error of the input \mathbf{x} from the output \mathbf{y} (see e.g. (Hornik & Kuan, 1992)). It also maximizes the transmitted information from a Gaussian input \mathbf{x} to the output \mathbf{y} , given uncorrelated equal-variance Gaussian additive noise on the input (Linsker, 1988; Plumbley & Fallside, 1988).

In fact, PCA is sufficient, but not necessary to find this optimum. Provided the rows of \mathbf{W} span the same subspace as that spanned by the largest M eigenvectors of $\Sigma_{\mathbf{x}}$, the *principal subspace* for a given M , then the mean squared reconstruction error will be minimized and the transmitted information will be maximized (Plumbley & Fallside, 1988). We shall refer to this as *principal subspace analysis*.

Over the last decade or so, a number of neural network algorithms have been suggested to enable a one-layer linear neural network (Fig. 1) to perform principal subspace analysis. Most of these PCA algorithms are based on the Oja (1982)

principal component finder neuron (see e.g. (Oja, 1989)), and update their weights at the n th time step by a small change

$$\Delta \mathbf{W}_n = \eta_n (\mathbf{y}_n \mathbf{x}_n^T - \mathbf{K}_n \mathbf{W}_n) \quad (2)$$

where the matrix \mathbf{K}_n varies from from one algorithm to another, and is normally a function of \mathbf{W}_n and \mathbf{x}_n . Particular examples of these algorithms include the Williams (1985) Symmetric Error Correction algorithm, which uses

$$\mathbf{K}_n = \mathbf{y}_n \mathbf{y}_n^T \quad (3)$$

the Oja and Karhunen (1985) Symmetric Gradient Ascent algorithm which uses

$$\mathbf{K}_n = \text{diag}(\mathbf{y}_n \mathbf{y}_n^T) + 2\text{LT}^+(\mathbf{y}_n \mathbf{y}_n^T) \quad (4)$$

and the Sanger (1989) Generalized Hebbian Algorithm which uses

$$\mathbf{K}_n = \text{LT}(\mathbf{y}_n \mathbf{y}_n^T). \quad (5)$$

In the previous expressions, $\text{diag}(\cdot)$ is the matrix function which sets entries to zero which are not on the diagonal, $\text{LT}^+(\cdot)$ sets entries to zero which are not strictly below the diagonal, and $\text{LT}(\cdot)$ sets entries to zero which are not on or below the diagonal.

Several authors have analyzed the behaviour of these and related algorithms (Hornik & Kuan, 1992; Oja, 1992; Oja & Karhunen, 1985; Sanger, 1989), based on the stability or instability of various equilibrium points for the ordinary differential equation (o.d.e.) equivalent of the algorithms. Typically, depending on the particular algorithm chosen, these show that the equilibrium point corresponding to PCA is the only possible asymptotically stable solution for \mathbf{W} . However, these analyses only give *local* stability information, and it has so far proved problematical to identify the domain of attraction for the identified stable solution.

One popular method for identifying a domain of attraction for a system such as this is the *Lyapunov function* approach (Cook, 1986). If an energy-like function can be found which monotonically decreases (or monotonically increases) over time,

towards the stable point from anywhere within a given region, that region is a *domain of attraction* for the stable point.

In this article, we identify a number of Lyapunov functions for this class of PCA algorithms, with their associated domains of attraction. These include

- a) the subspace projection error (section 4.1);
- b) the least mean square reconstruction error (section 4.2); and
- c) the mutual information (section 4.3)

although all of these leave a significant section of \mathbf{W} -space outside of their respective domains of attraction. Finally, a modification to the latter mutual information function yields what we term

- d) the principal subspace information (section 5)

which has an associated domain of attraction which covers almost all values of \mathbf{W} : this is the main result of this article. This allows us to show that the class of PCA algorithms considered here converges to the principal subspace from almost everywhere, and consequently that a random initial value of \mathbf{W} will almost surely converge to the principal subspace. Most of the proofs will be relegated to the Appendix.

2 The Ordinary Differential Equation Approach

Rather than attempting a direct analysis of the difference equation (2), following other authors (see e.g. (Hornik & Kuan, 1992)) we consider instead the behaviour of the equivalent ordinary differential equation (o.d.e.).

Informally, if the update factor η_n is small enough to make only small changes to \mathbf{W}_n at each step, we can consider the mean weight update $\Delta\mathbf{W} = \langle \Delta\mathbf{W}_n \rangle$, assuming that the weight matrix \mathbf{W}_n remains approximately constant ($\approx \mathbf{W}$) over the period over which we are taking the mean. If η_n is also approximately constant we get

$$\Delta\mathbf{W} = \langle \eta_n \rangle \left(\langle \mathbf{W}_n \mathbf{x}_n \mathbf{x}_n^T \rangle - \langle \mathbf{K}_n \mathbf{W}_n \rangle \right)$$

$$\begin{aligned}
&\approx \eta(\mathbf{W} \langle \mathbf{x}_n \mathbf{x}_n^T \rangle - \langle \mathbf{K}_n \rangle \mathbf{W}) \\
&\approx \eta(\mathbf{W} \boldsymbol{\Sigma}_x - \mathbf{K} \mathbf{W})
\end{aligned} \tag{6}$$

where $\boldsymbol{\Sigma}_x = E(\mathbf{x} \mathbf{x}^T)$ is the the covariance matrix of \mathbf{x} (making the assumption that \mathbf{x} is zero mean), and $\mathbf{K} = E(\mathbf{K}_n)$ is the expected value of \mathbf{K}_n . In what follows, we assume that the covariance matrix $\boldsymbol{\Sigma}_x$ is nonsingular with distinct eigenvalues $\lambda_1 > \lambda_2 > \dots > \lambda_N$.

More formally, under certain conditions (see e.g. (Hornik & Kuan, 1992)) it can be shown that the path of \mathbf{W}_n in algorithm (2) will approximately follow the path of \mathbf{W} in the o.d.e.

$$\frac{d\mathbf{W}}{dt} = \mathbf{W} \boldsymbol{\Sigma}_x - \mathbf{K} \mathbf{W} \tag{7}$$

and in particular \mathbf{W}_n will converge with probability 1 (i.e. *almost surely*) to the roots of the o.d.e. (7).

We remarked in the introduction that only principal *subspace* analysis, and not full-blown principal *component* analysis, is necessary to optimize either the least mean squared reconstruction error of the transmitted information. This leads us to consider the behaviour of the subspace spanned by the rows of \mathbf{W} rather than \mathbf{W} itself. Oja (1992) showed that this subspace can be represented by the *orthogonal projection* matrix \mathbf{P} which projects vectors into the given subspace by left (or right) multiplication. If $\mathbf{W} \mathbf{W}^T$ has full rank, then this orthogonal projection is defined by

$$\mathbf{P} = \mathbf{W}^T (\mathbf{W} \mathbf{W}^T)^{-1} \mathbf{W}. \tag{8}$$

It is immediately evident that \mathbf{P} as defined satisfies the two conditions for an orthogonal projection, namely

$$\mathbf{P}^2 = \mathbf{P} \tag{9}$$

$$\mathbf{P}^T = \mathbf{P}. \tag{10}$$

If only the first of these conditions were satisfied, \mathbf{P} would be a *projection*, but not an *orthogonal* projection. Note that if \mathbf{P} is an orthogonal projection, then $(\mathbf{I}_N - \mathbf{P})$ where \mathbf{I}_N is the $N \times N$ identity matrix, is also an orthogonal projection,

A way to visualize the situation is to imagine \mathbf{P} as a operator which projects input vectors into the given subspace. Thus the vector $\tilde{\mathbf{x}} = \mathbf{P}\mathbf{x}$ is the component of \mathbf{x} which lies entirely within the subspace we are interested in. Note that after \mathbf{P} has been applied once to yield a vector within our subspace, another application of \mathbf{P} will have no further effect: thus $\mathbf{P}\tilde{\mathbf{x}} = \mathbf{P}(\mathbf{P}\mathbf{x}) = \mathbf{P}\mathbf{x} = \tilde{\mathbf{x}}$ since $\mathbf{P}^2 = \mathbf{P}$.

Before we proceed, we show that we can choose \mathbf{K} in (7) so that $\mathbf{W}\mathbf{W}^T$ has full rank for all t , so that P can continue to be formed as in (8).

Theorem 2.1 *Suppose that \mathbf{K} in (7) satisfies*

$$\mathbf{K} + \mathbf{K}^T = 2\mathbf{W}\Sigma_{\mathbf{x}}\mathbf{W}^T \quad (11)$$

and that $\mathbf{W}\mathbf{W}^T$ is finite and nonsingular at $t = 0$. Then $\mathbf{W}\mathbf{W}^T$ remains finite and nonsingular for all $t > 0$, and asymptotically converges to $\mathbf{W}\mathbf{W}^T = \mathbf{I}_M$.

Proof See Appendix.

Note that (11) is satisfied for the Williams (1985) SEC algorithm (3) and the Oja and Karhunen (1985) SGA algorithm (4), as well as the original single-output Oja (1982) algorithm, for which these two are slightly different generalizations. Therefore $\mathbf{W}\mathbf{W}^T$ will remain finite and non-singular for these algorithms. However, consider the simple Hebbian algorithm

$$d\mathbf{W}/dt = \mathbf{W}\Sigma_{\mathbf{x}} \quad (12)$$

with no ‘weight decay’ term. This is of the form given in (7) with $\mathbf{K} = \mathbf{0}$. There is nothing to prevent the rows of \mathbf{W} from becoming infinitely large in the direction of the single principal eigenvector of $\Sigma_{\mathbf{x}}$ as the algorithm progresses. As $t \rightarrow \infty$, the weight matrix \mathbf{W} becomes degenerate, and $\mathbf{W}\mathbf{W}^T$ becomes singular.

Other algorithms which do not satisfy (11) have to be approached individually. For example, using (7) with $\mathbf{K} = \mathbf{W}\mathbf{W}^T$ produces the novel algorithm

$$d\mathbf{W}/dt = \mathbf{W}\Sigma_{\mathbf{x}} - \mathbf{W}\mathbf{W}^T\mathbf{W} \quad (13)$$

for which $\mathbf{W}\Sigma_{\mathbf{x}}^{-1}\mathbf{W}^T$ remains finite and nonsingular if so initially, implying the same about $\mathbf{W}\mathbf{W}^T$. The proof is similar for the previous case, and shows that the

selected output components have the *square* of the input variance components at convergence. This algorithm will be covered in more detail in a later paper.

Another example is Sanger's (1989) Generalized Hebbian Algorithm (5). While this again does not satisfy (11), Sanger's direct proof of convergence does show that $\mathbf{W}\mathbf{W}^T$ does not become degenerate.

For the remainder of this article, we shall *assume* that \mathbf{K} in (7) is such that $\mathbf{W}\mathbf{W}^T$ does retain full rank for all t , to avoid any difficulties with \mathbf{P} .

The o.d.e. (7) defines an o.d.e. for the behaviour of the subspace \mathbf{P} itself (Oja, 1992), which can be written as

$$\frac{d\mathbf{P}}{dt} = (\mathbf{I}_N - \mathbf{P})\boldsymbol{\Sigma}_x\mathbf{P} + \mathbf{P}\boldsymbol{\Sigma}_x(\mathbf{I}_N - \mathbf{P}) \quad (14)$$

which is dependent on \mathbf{P} only, and not on the precise form of \mathbf{W} . The conditions for stationarity and stability of \mathbf{P} are given by the following two theorems (Williams, 1985; Oja, 1992).

Theorem 2.2 \mathbf{P} in (14) is stationary (is a fixed point of (14)) iff (if and only if) it is the subspace spanned by some M eigenvectors of $\boldsymbol{\Sigma}_x$.

Proof (See also (Oja, 1992) or (Williams, 1985).) If $d\mathbf{P}/dt$ in (14) is zero, multiplying on the right by \mathbf{P} we find that $(\mathbf{I}_N - \mathbf{P})\boldsymbol{\Sigma}_x\mathbf{P} = 0$, so $\boldsymbol{\Sigma}_x\mathbf{P} = \mathbf{P}\boldsymbol{\Sigma}_x\mathbf{P} = \mathbf{P}\boldsymbol{\Sigma}_x$. The converse is clearly also true (i.e. that \mathbf{P} is stationary if $(\mathbf{I}_N - \mathbf{P})\boldsymbol{\Sigma}_x\mathbf{P} = 0$), so we have that $d\mathbf{P}/dt = 0$ iff $\boldsymbol{\Sigma}_x\mathbf{P} = \mathbf{P}\boldsymbol{\Sigma}_x$, i.e. $\boldsymbol{\Sigma}_x$ and \mathbf{P} commute.

Now, $\boldsymbol{\Sigma}_x$ and \mathbf{P} commute iff they share the same eigenvectors (Strang, 1976), i.e. iff \mathbf{P} is the subspace spanned by some M of the eigenvectors of $\boldsymbol{\Sigma}_x$. (Note that \mathbf{P} has M eigenvectors with eigenvalue 1, and $N - M$ with eigenvalue 0). Therefore \mathbf{P} is stationary iff it is the subspace spanned by some M of the eigenvectors of $\boldsymbol{\Sigma}_x$. **QED.**

Theorem 2.3 \mathbf{P} in (14) is stable only when it is the principal subspace \mathbf{P}^* , i.e. the subspace spanned by the first M eigenvectors of $\boldsymbol{\Sigma}_x$. Any other fixed point is unstable.

Proof See (Oja, 1992) or (Williams, 1985).

This tells us that the principal subspace P^* is the unique stable point for \mathbf{P} , but it only tells us about the local behaviour around each of the stationary points. We cannot tell the domain of attraction for the stable solution, or whether it is possible to get caught by a saddle point ‘on the way’ to the stable point from any particular initial value of \mathbf{P} .

To get a more global view, we now proceed to identify a number of Lyapunov functions with associated domains of convergence. Any initial \mathbf{P} within a domain of convergence so identified will asymptotically converge to the stable point.

3 Energy Functions and Lyapunov Functions

3.1 Energy Functions

Energy functions, functions which decrease over time, have already been used in conjunction with PCA algorithms to visualize the behaviour of such algorithms. Baldi and Hornik (1989), for example, showed that the mean squared error of a linear auto-association network (trained with BackProp) has only one local minimum, in addition to a number of saddle points. The BackProp algorithm itself operates by performing a steepest-descent search in this energy function.

For the o.d.e. (7) it is possible to identify a number of different energy functions. For our purposes, an energy function is a function $S(\mathbf{W})$ for which its time derivative is non-positive, i.e. $dS/dt \leq 0$, when $d\mathbf{W}/dt$ is given by the o.d.e. (7). These typically arise from ‘error’ measures which decrease as an algorithm progresses. Of course, we could instead use an increasing function of t as an energy function, by simply taking the negative of the function. These functions sometimes arise as ‘signal’ or ‘output’ measures which increase as an algorithm progresses. We shall see that transmitted information is an increasing energy function for the o.d.e.s (7) and (14).

3.2 Lyapunov Functions

A Lyapunov function is like an energy function which can be used to estimate a domain of attraction for a stable point. In many cases, energy functions can be used as Lyapunov functions, although the estimated domain of attraction they define may only be subsets of the true domain of attraction. There is no general method available to construct a Lyapunov function which will identify the complete domain of convergence, so it is often a matter of trial-and-error to identify a good one.

The most general definition of a Lyapunov function is that it should be a function L which strictly decreases monotonically from any point a within a particular region D (the *domain of attraction*), except at a single point a^* within D , where it is stationary. If such an L can be found, then a^* is asymptotically stable (Cook, 1986). In what follows, it is convenient for us to use the following formulation.

Theorem 3.1 (Lyapunov) *Suppose that we have a function $L(a)$, defining a region D consisting of all points a such that $L(a) < c$ for some constant c , with the boundary of D given by the all points $L(a) = c$, such that*

L1 $dL(a)/dt < 0$ for all $a \neq a^*$ in D , and

L2 $dL(a^*)/dt = 0$.

Then the equilibrium point $a = a^$ is asymptotically stable, with a domain of attraction D .*

So, if we have an energy function with several equilibrium points $a^*, a^{[2]}, a^{[3]}, \dots$, we choose c to be the minimum value of $L(a)$ over all points (apart from a^*) where $dL(a)/dt = 0$. In other words, our domain of attraction will be all the points with energy below that of the second lowest stationary point or points, with the lowest stationary point being the stable point.

We can now consider a number of energy functions for our Hebbian PCA algorithms from this point of view, to see what domains of attractions are identified. These functions include the subspace projection error; least mean square reconstruction error; and mutual information. For the purposes of this article we shall call

these *non-maximal* Lyapunov functions, since their domains of attraction are not, in general, the largest which can be identified. The largest domain of attraction identified here is given by our final ‘principal subspace information’ measure: its domain of attraction covers almost all of \mathbf{P} -space.

4 Non-maximal Lyapunov functions

4.1 Subspace Projection Error

The first function we shall consider is the mean squared error $S_{\mathbf{P}}$ between the input vector \mathbf{x} and its projection $\tilde{\mathbf{x}} = \mathbf{P}\mathbf{x}$ into the subspace defined by \mathbf{P} . As we have already noted in the introduction, this is the subspace spanned by the rows of the weight matrix \mathbf{W} .

Lemma 4.1 *Suppose that \mathbf{P} varies according to the o.d.e. (14). Then the subspace projection error*

$$S_{\mathbf{P}} = E(|\mathbf{x} - \tilde{\mathbf{x}}|^2) = E(|\mathbf{x} - \mathbf{P}\mathbf{x}|^2) \quad (15)$$

is a nonincreasing function of t , and is stationary iff \mathbf{P} is the subspace spanned by some M eigenvectors of $\Sigma_{\mathbf{x}}$.

Proof See Appendix.

The function $S_{\mathbf{P}}$ is thus an energy function with stationary points corresponding to the stationary points of \mathbf{P} (see Theorem 2.2).

Theorem 4.2 *Let $S_{\mathbf{P}^*}$ be the value of $S_{\mathbf{P}}$ when \mathbf{P} is the principal subspace \mathbf{P}^* of $\Sigma_{\mathbf{x}}$, and let λ_i be the i th largest eigenvalue of $\Sigma_{\mathbf{x}}$. Then $S_{\mathbf{P}}$ is a Lyapunov function for convergence of \mathbf{P} to \mathbf{P}^* , with domain of attraction*

$$D_{\mathbf{P}} = \{\mathbf{P} | S_{\mathbf{P}} < S_{\mathbf{P}^*} + \lambda_M - \lambda_{M+1}\}. \quad (16)$$

Consequently, the principal subspace \mathbf{P}^ is asymptotically stable with domain of convergence $D_{\mathbf{P}}$.*

Proof See Appendix.

For a network with two inputs ($N = 2$) and a single output ($M = 1$), this domain of attraction $D_{\mathbf{P}}$ covers the whole of \mathbf{P} -space except the point $\mathbf{P}^{[2]}$ where \mathbf{P} is the subspace corresponding to the second eigenvector (i.e. the *minor component*) of the 2×2 input covariance matrix. This point $\mathbf{P}^{[2]}$ is the only stationary point for \mathbf{P} apart from the optimal solution \mathbf{P}^* itself.

For the more general case with $N > 2$, a significant proportion of \mathbf{P} -space is outside the domain of attraction $D_{\mathbf{P}}$. For example, significant regions near to the third-lowest stationary point $\mathbf{P}^{[3]}$ (and higher stationary points) will be outside $D_{\mathbf{P}}$, but simulations suggest that initial values of \mathbf{P} within these regions will also converge to \mathbf{P}^* . This suggests that $D_{\mathbf{P}}$ is not the true, or *maximal*, domain of attraction for \mathbf{P}^* . Theorem 4.2 tells us that \mathbf{P} converges to \mathbf{P}^* from any initial point within $D_{\mathbf{P}}$, but not what happens to points *outside* of $D_{\mathbf{P}}$.

4.2 Least Mean Square Reconstruction Error

The error measure $S_{\mathbf{P}}$ used in section 4.1 measures the mean squared error in the projection $\tilde{\mathbf{x}} = \mathbf{P}\mathbf{x}$ from \mathbf{x} . Since $\mathbf{P}\mathbf{x} = \mathbf{W}^T(\mathbf{W}\mathbf{W}^T)^{-1}\mathbf{W}\mathbf{x}$, we can construct the projected vector $\tilde{\mathbf{x}}$ as $\tilde{\mathbf{x}} = \mathbf{W}^T(\mathbf{W}\mathbf{W}^T)^{-1}\mathbf{y}$ where \mathbf{y} is the output from the network given in (1). However, this is not the best reconstruction of \mathbf{x} given \mathbf{y} , in the minimum mean squared error sense.

It is well known that the least mean squared reconstruction of \mathbf{x} given the network output $\mathbf{y} = \mathbf{W}\mathbf{x}$ is given by the vector

$$\hat{\mathbf{x}} = \Sigma_{\mathbf{x}}\mathbf{W}^T(\mathbf{W}\Sigma_{\mathbf{x}}\mathbf{W}^T)^{-1}\mathbf{y} = \mathbf{Q}\mathbf{x} \quad (17)$$

where \mathbf{Q} is given by the expression

$$\mathbf{Q} = \Sigma_{\mathbf{x}}\mathbf{W}^T(\mathbf{W}\Sigma_{\mathbf{x}}\mathbf{W}^T)^{-1}\mathbf{W}. \quad (18)$$

Now, we see that $\mathbf{Q}^2 = \mathbf{Q}$, so \mathbf{Q} is a projection. However, $\mathbf{Q}^T \neq \mathbf{Q}$ in general, so \mathbf{Q} is not an *orthogonal* projection.

The projection \mathbf{Q} is related in a number of ways to \mathbf{P} . From their definitions, for example, we can immediately write down the equations:

$$\mathbf{PQ} = \mathbf{P} \quad (19)$$

$$\mathbf{QP} = \mathbf{Q} \quad (20)$$

$$\mathbf{Q}\Sigma_{\mathbf{x}} = \Sigma_{\mathbf{x}}\mathbf{Q}^T \quad (21)$$

which suggest that \mathbf{P} and \mathbf{Q} are closely related. In fact, for a given covariance matrix $\Sigma_{\mathbf{x}}$ there is a one-to-one mapping from \mathbf{Q} to \mathbf{P} , as the following lemma demonstrates. In particular, it is not necessary to know \mathbf{W} in order to determine \mathbf{Q} from \mathbf{P} and $\Sigma_{\mathbf{x}}$.

Lemma 4.3 *\mathbf{P} is a function of \mathbf{Q} only, and \mathbf{Q} is a function of \mathbf{P} and $\Sigma_{\mathbf{x}}$ only. Furthermore, \mathbf{P} and \mathbf{Q} are equal iff they are the subspace spanned by some M eigenvectors of $\Sigma_{\mathbf{x}}$.*

Proof See Appendix.

Consider then, as our next tentative Lyapunov function, the least mean squared reconstruction error

$$S_{\mathbf{Q}} = E(|\mathbf{x} - \hat{\mathbf{x}}|^2) = E(|\mathbf{x} - \mathbf{Q}\mathbf{x}|^2) \quad (22)$$

which, for any given $\Sigma_{\mathbf{x}}$ is a function of \mathbf{P} only. Therefore we can view it as a function of \mathbf{P} instead of \mathbf{Q} if we wish.

Lemma 4.4 *The least mean square reconstruction error $S_{\mathbf{Q}}$ is a nonincreasing function of t , and is stationary iff \mathbf{Q} is the subspace spanned by some M eigenvectors of $\Sigma_{\mathbf{x}}$.*

Proof See Appendix.

Theorem 4.5 *$S_{\mathbf{Q}}$ is a Lyapunov function for convergence of \mathbf{P} to \mathbf{P}^* , with domain of attraction*

$$D_{\mathbf{Q}} = \{\mathbf{P} | S_{\mathbf{Q}} < S_{\mathbf{P}^*} + \lambda_M - \lambda_{M+1}\} \supset D_{\mathbf{P}}. \quad (23)$$

Consequently, the principal subspace \mathbf{P}^ is asymptotically stable with domain of convergence $D_{\mathbf{Q}}$.*

Proof See Appendix.

Note that the boundary for $D_{\mathbf{Q}}$ is expressed in terms of $S_{\mathbf{P}}^*$ for simpler comparison with $D_{\mathbf{P}}$. We have therefore identified a Lyapunov function with a larger domain of attraction $D_{\mathbf{Q}}$ than $D_{\mathbf{P}}$. However, as for $D_{\mathbf{P}}$, $D_{\mathbf{Q}}$ still leaves a significant proportion of \mathbf{P} -space uncovered in general.

There is also a dual projection \mathbf{O} of \mathbf{Q} , given by

$$\mathbf{O} = \mathbf{W}^T (\mathbf{W}\Sigma_{\mathbf{x}}^{-1}\mathbf{W}^T)^{-1} \mathbf{W}\Sigma_{\mathbf{x}}^{-1} \quad (24)$$

which has similar properties to \mathbf{Q} . However, since the energy function $S_{\mathbf{O}} = \text{Tr}((\mathbf{I}_N - \mathbf{O})\Sigma_{\mathbf{x}})$ is the ‘greatest’ mean squared reconstruction error (under certain constraints), the domain of attraction $D_{\mathbf{O}}$ identified by $S_{\mathbf{O}}$ is smaller than $D_{\mathbf{P}}$, and hence smaller than $D_{\mathbf{Q}}$.

4.3 Mutual Information

For a multivariate Gaussian input signal \mathbf{x} corrupted by uncorrelated equal variance Gaussian noise, PCA or principal subspace analysis, maximizes mutual information (MI) between the input and the output of the network in Fig. 1 (Linsker, 1988; Plumbley & Fallside, 1988).

In a previous article, the author used MI as a Lyapunov function for a decorrelating network with noise on the output instead of the input (Plumbley, 1993). In this article, we take a similar approach for these PCA algorithms and attempt to use MI in a similar way. However, as we shall see, the situation in this case is not quite so straightforward, and the MI function has to be modified before yielding the final result here.

Suppose that input signal, with covariance matrix $\Sigma_{\mathbf{x}}$, contains noise with covariance matrix $\Sigma_{\phi} = \sigma_{\phi}^2 \mathbf{I}_N$. Then the total output and output noise have respective covariance matrices $\mathbf{W}\Sigma_{\mathbf{x}}\mathbf{W}^T$ and $\sigma_{\phi}^2 \mathbf{W}\mathbf{W}^T$. The mutual information I is given by (Linsker, 1988; Plumbley & Fallside, 1988)

$$I = \frac{1}{2} \left(\log \det(\mathbf{W}\Sigma_{\mathbf{x}}\mathbf{W}^T) - \log \det(\sigma_{\phi}^2 \mathbf{W}\mathbf{W}^T) \right). \quad (25)$$

Although this has a very different form to the previous Lyapunov functions considered, it can be used as a Lyapunov function in a very similar way.

Lemma 4.6 *The mutual information I in (25) is a nondecreasing function of t , and is stationary iff \mathbf{P} is the subspace spanned by some M eigenvectors of $\Sigma_{\mathbf{x}}$.*

Proof See Appendix.

Thus, in a similar way that both $S_{\mathbf{P}}$ and $S_{\mathbf{Q}}$ decrease monotonically over time, the mutual information I increases monotonically over time, except when \mathbf{P} is the subspace spanned by some M eigenvectors of $\Sigma_{\mathbf{x}}$. Consequently, I is an energy function for \mathbf{W} , albeit an increasing rather than a decreasing energy function. Furthermore, the following lemma shows that I can also be viewed as an energy function for \mathbf{P} .

Lemma 4.7 *For a given covariance matrix $\Sigma_{\mathbf{x}}$, the mutual information I in (25) is a function of \mathbf{P} only.*

Proof See Appendix.

We mentioned above that I is maximized when at the principal subspace solution $\mathbf{P} = \mathbf{P}^*$. This means that we can propose I (or, strictly speaking, $-I$) as a Lyapunov function.

Theorem 4.8 *Let $I_{\mathbf{P}^*}$ be the value of I at the point $\mathbf{P} = \mathbf{P}^*$. Then I is a Lyapunov function for convergence of \mathbf{P} to \mathbf{P}^* , with domain of attraction*

$$D_I = \left\{ \mathbf{P} \mid I > I_{\mathbf{P}^*} - \frac{1}{2} \log \frac{\lambda_M}{\lambda_{M+1}} \right\}. \quad (26)$$

Consequently, the principal subspace \mathbf{P}^ is asymptotically stable with domain of attraction D_I .*

Proof See Appendix.

For a single output, i.e. $M = 1$, it is easy to show that this reduces to the same domain of attraction to that found for $S_{\mathbf{P}}$ above, since both $S_{\mathbf{P}}$ and I are monotonic

in the output variance. Consequently, for a single output $M = 1$, $D_I = D_{\mathbf{P}}$, and $D_I \subset D_{\mathbf{Q}}$, so D_I is no better than $D_{\mathbf{Q}}$ for a single-output network at least.

This information function still does not identify a domain of attraction covering the as much of \mathbf{P} -space as we would like. However, the proof of Lemma 4.6 uses the fact that $\mathbf{W}\Sigma_{\mathbf{x}}\mathbf{W}^T$ is nonsingular. We can see from (25) that if $\mathbf{W}\Sigma_{\mathbf{x}}\mathbf{W}^T$ is singular (but $\mathbf{W}\mathbf{W}^T$ is nonsingular), then $\det(\mathbf{W}\Sigma_{\mathbf{x}}\mathbf{W}^T) = 0$ so $I \rightarrow -\infty$. This suggests a modification which can be made to the information function I which will finally yield a domain of attraction covering virtually all of \mathbf{P} -space.

5 Principal Subspace “Information”

It is well known that if the network (1) has precisely zero weight component in the direction of a particular eigenvector of $\Sigma_{\mathbf{x}}$, then there will be no tendency for a Hebbian algorithm to increase that weight component away from zero. This is easy to see for a system $y = wx$ with a single scalar weight w , since the weight update algorithm $\Delta w = yx = wx^2$ will be zero if the weight w is zero.

Let us tentatively suppose that this will generalize to \mathbf{P} in the following way. If a non-zero eigenvector \mathbf{e} of \mathbf{P} can be found for which is entirely perpendicular to the principal subspace \mathbf{P}^* , i.e. $\mathbf{e}^T\mathbf{P}^*\mathbf{e} = 0$, then it may be difficult or impossible for \mathbf{P} to converge to \mathbf{P}^* . This suggests that $\mathbf{W}\mathbf{P}^*\mathbf{W}^T$ should be nonsingular for convergence, since \mathbf{W} can be decomposed into the product of an $M \times N$ matrix formed from the M non-zero eigenvectors of \mathbf{P} on the right, and an $M \times M$ scaling and rotation matrix on the left.

Let us therefore construct an artificial ‘principal subspace information’ function I' as follows

$$I' = \frac{1}{2} \left(\log \det(\mathbf{W}\mathbf{P}^*\mathbf{W}^T) - \log \det(\mathbf{W}\mathbf{W}^T) \right) \quad (27)$$

which is the mutual information function in (25) with the covariance matrix $\Sigma_{\mathbf{x}}$ replaced by the principal subspace \mathbf{P}^* . In a sense, this is a weighted information function, so that information about any input component in the principal subspace is equal, but information about any input component which is not in the principal

subspace is ignored. As for I , this can be shown to be a function of \mathbf{P} (and \mathbf{P}^*) only. Let us now explore I' as a possible Lyapunov function.

Lemma 5.1 *Suppose that $\mathbf{W}\mathbf{P}^*\mathbf{W}^T$ is initially nonsingular. Then the principal subspace information I' is a nondecreasing function of t , and is stationary iff $\mathbf{P} = \mathbf{P}^*$.*

Proof See Appendix.

We note immediately that Lemma 5.1 differs in two important ways from the equivalent lemmas for the previous Lyapunov functions considered. Firstly, it is only valid when $\mathbf{W}\mathbf{P}^*\mathbf{W}^T$ is initially nonsingular. Secondly, with this condition, I' is *only* stationary when \mathbf{P} is the *principal* subspace \mathbf{P}^* .

What has happened is that the condition for $\mathbf{W}\mathbf{P}^*\mathbf{W}^T$ to be nonsingular has excluded the other stationary conditions for \mathbf{P} , but which were still present in the other energy functions $S_{\mathbf{P}}$, $S_{\mathbf{Q}}$, $S_{\mathbf{O}}$ and I . This allows us to state the main theorem of this article.

Theorem 5.2 *The principal subspace information I' is a Lyapunov function for convergence of \mathbf{P} to \mathbf{P}^* , with domain of attraction*

$$D_{I'} = \{\mathbf{P} | \mathbf{W}\mathbf{P}^*\mathbf{W}^T \text{ nonsingular}\} \quad (28)$$

Consequently, the principal subspace \mathbf{P}^ is asymptotically stable with domain of convergence $D_{I'}$.*

Proof See Appendix.

The region where \mathbf{P} is outside of $D_{I'}$, for which $\mathbf{W}\mathbf{P}^*\mathbf{W}^T$ is singular, has dimensionality 1 less than the whole of \mathbf{P} -space. Consequently, $D_{I'}$ covers almost everywhere in \mathbf{P} -space. Thus if an initial \mathbf{P} is chosen at random, it will be in the domain of attraction $D_{I'}$ with probability 1.

Finally, we note that if \mathbf{P} is initially within $D_{I'}$, it will only be stationary when $\mathbf{P} = \mathbf{P}^*$, i.e. when it has converged to the final solution. Since our previous Lyapunov functions $S_{\mathbf{P}}$, $S_{\mathbf{Q}}$, $S_{\mathbf{O}}$, and I are stationary iff \mathbf{P} is stationary, these will also be

stationary only at convergence. Thus in particular the subspace projection error $S_{\mathbf{P}}$ and least mean squared reconstruction error $S_{\mathbf{Q}}$ strictly decrease monotonically over time until convergence to $\mathbf{P} = \mathbf{P}^*$, and similarly the mutual information I strictly increases monotonically over time until convergence. Consequently, $S_{\mathbf{P}}$, $S_{\mathbf{Q}}$ and I are also Lyapunov functions with domain of convergence D' (according to the general definition), since they vary monotonically from any point within D' and are stationary only at convergence¹. The increase in information I over time was noted by Földiák (Földiák, 1989) for an algorithm related to (although not one of) the class considered here.

6 An Illustrative Example

As we have already mentioned, the 2-input case ($N = 2$) leads to a relatively trivial situation where even the domain of attraction $D_{\mathbf{P}}$ is sufficient to cover virtually the whole of \mathbf{P} -space. Also, with just a single output ($M = 1$), the domains of attraction $D_{\mathbf{P}}$ and D_I are identical. We therefore choose to illustrate the domains of convergence identified by these Lyapunov functions for a 3-input 2-output network ($N = 3$, $M = 2$) with an ordered diagonal input covariance matrix $\Sigma_{\mathbf{x}}$ with eigenvalues $\lambda_1 = 3$, $\lambda_2 = 2$, and $\lambda_3 = 1$.

In this system, each orthogonal projection \mathbf{P} is a 2-dimensional subspace of 3-space, which we can represent by the unit vector normal to \mathbf{P} (and the opposite unit vector). Thus \mathbf{P} -space is represented by the surface of a half-sphere. Figs. 2(a)-(d) show the domains of attraction $D_{\mathbf{P}}$, $D_{\mathbf{Q}}$, D_I , and $D_{I'}$ respectively. Note that a view of only half of the sphere is sufficient, since the reverse side of the sphere is identical but opposite to the front side.

In this special case where the number of outputs M is one less than the number of inputs N , several simplifications can be made. For any orthogonal projection \mathbf{P} of rank M can write $\mathbf{P} = \mathbf{U}_{\mathbf{P}}\mathbf{U}_{\mathbf{P}}^T$ where $\mathbf{U}_{\mathbf{P}}$ is an $N \times M$ matrix such that $\mathbf{U}_{\mathbf{P}}^T\mathbf{U}_{\mathbf{P}} = \mathbf{I}_M$.

¹This is perhaps a little academic, since it would not have been possible to state this without first using the Lyapunov function I' .

Similarly, for the complementary projection $\bar{\mathbf{P}} = (\mathbf{I}_N - \mathbf{P})$ we can write $\bar{\mathbf{P}} = \mathbf{u}_{\bar{\mathbf{P}}}\mathbf{u}_{\bar{\mathbf{P}}}^T$ where $\mathbf{u}_{\bar{\mathbf{P}}}$ is a unit-length vector such that $\mathbf{U}_{\mathbf{P}}\mathbf{U}_{\mathbf{P}}^T + \mathbf{u}_{\bar{\mathbf{P}}}\mathbf{u}_{\bar{\mathbf{P}}}^T = \mathbf{I}_N$. (In the general case, $\mathbf{U}_{\bar{\mathbf{P}}}$ will be an $N \times (N - M)$ matrix such that $\mathbf{U}_{\bar{\mathbf{P}}}^T\mathbf{U}_{\bar{\mathbf{P}}} = \mathbf{I}_{N-M}$.) This vector $\mathbf{u}_{\bar{\mathbf{P}}}$ is the normal vector used to plot the surfaces in Figs. 2(a)-(d).

A number of the expressions for the energy functions become simplified, based on the vector $\mathbf{u}_{\bar{\mathbf{P}}}$. For example, it is easy to show that

$$S_{\mathbf{P}} = \text{Tr}(\mathbf{u}_{\bar{\mathbf{P}}}^T \Sigma_{\mathbf{x}} \mathbf{u}_{\bar{\mathbf{P}}}) = \mathbf{u}_{\bar{\mathbf{P}}}^T \Sigma_{\mathbf{x}} \mathbf{u}_{\bar{\mathbf{P}}} \quad (29)$$

with the first equality holding for any M . Similarly, we can verify that

$$\bar{\mathbf{Q}} = (\mathbf{I}_N - \mathbf{Q}) = \mathbf{u}_{\bar{\mathbf{P}}}(\mathbf{u}_{\bar{\mathbf{P}}}^T \Sigma_{\mathbf{x}}^{-1} \mathbf{u}_{\bar{\mathbf{P}}})^{-1} \mathbf{u}_{\bar{\mathbf{P}}}^T \Sigma_{\mathbf{x}}^{-1} \quad (30)$$

and that $S_{\mathbf{Q}}$ simplifies to

$$S_{\mathbf{Q}} = \text{Tr}((\mathbf{u}_{\bar{\mathbf{P}}}^T \Sigma_{\mathbf{x}}^{-1} \mathbf{u}_{\bar{\mathbf{P}}})^{-1}) = \frac{1}{\mathbf{u}_{\bar{\mathbf{P}}}^T \Sigma_{\mathbf{x}}^{-1} \mathbf{u}_{\bar{\mathbf{P}}}}. \quad (31)$$

Also, for this case, the boundaries of the domains of attraction have the following property, which is evident from the figures.

Theorem 6.1 *Suppose that $N = 3$ and $M = 2$, with $\mathbf{u}_{\bar{\mathbf{P}}}$ the normal vector to \mathbf{P} . Then in $\mathbf{u}_{\bar{\mathbf{P}}}$ -space, the domains of attraction $D_{\mathbf{P}}$ and $D_{\mathbf{Q}}$ are bounded by planar rings with normal vectors*

$$\begin{aligned} \mathbf{b}_{\mathbf{P}} &= [(\lambda_1 - \lambda_2)^{1/2} \quad 0 \quad \mp (\lambda_2 - \lambda_3)^{1/2}]^T \\ \mathbf{b}_{\mathbf{Q}} &= \left[\left(\frac{1}{\lambda_2} - \frac{1}{\lambda_1} \right)^{1/2} \quad 0 \quad \mp \left(\frac{1}{\lambda_3} - \frac{1}{\lambda_2} \right)^{1/2} \right]^T \end{aligned}$$

Proof See Appendix.

Comparing the angles $\theta_{\mathbf{P}}$ and $\theta_{\mathbf{Q}}$ of $\mathbf{b}_{\mathbf{P}}$ and $\mathbf{b}_{\mathbf{Q}}$ to the ‘vertical’ (\mathbf{e}_3), we find that

$$\cos \theta_{\mathbf{Q}} = \left(\frac{\lambda_1}{\lambda_2} \right)^{1/2} \cos \theta_{\mathbf{P}} \quad (32)$$

which confirms that $\mathbf{b}_{\mathbf{Q}}$ is closer to \mathbf{e}_3 than $\mathbf{b}_{\mathbf{P}}$ is, so the boundaries of $D_{\mathbf{Q}}$ are more ‘horizontal’ than those of $D_{\mathbf{P}}$.

It also appears from the diagrams that D_I and $D_{\mathbf{Q}}$ are identical, even though they were plotted from different functions. This is confirmed by the following slightly more general theorem.

Theorem 6.2 *Suppose that $M = N - 1$. Then there is a one-to-one relationship between I and $S_{\mathbf{Q}}$, and consequently $D_I = D_{\mathbf{Q}}$.*

Proof See Appendix.

We mentioned at the beginning of this section that for a single output $M = 1$, D_I and $D_{\mathbf{P}}$ are identical, so

$$D_{\mathbf{P}} = D_I \subset D_{\mathbf{Q}}$$

and from Theorem 6.2, for $M = N - 1$ we now have

$$D_{\mathbf{P}} \subset D_I = D_{\mathbf{Q}}.$$

Preliminary work suggests that $D_{\mathbf{P}} \subset D_I$ may hold more generally for $M > 1$, which leads us to *conjecture* that for any $1 \leq M \leq N$

$$D_{\mathbf{P}} \subset D_I \subset D_{\mathbf{Q}} \tag{33}$$

holds, allowing for set equality in the relations. Strict subsets may be possible for $M = 2, N = 4$.

7 Conclusions

We have considered the global convergence properties of a class of Oja-like Hebbian algorithms which either perform principal component analysis, or find the principal subspace, of an input covariance matrix. The algorithms, which are of the form given in o.d.e. (7), are known to be locally stable only at the principal subspace (or principal components) solution.

Initially, we showed that many of this class of algorithms, such as the original Oja (1982) PCA neuron, the Williams (1985) SEC algorithm, and the Oja and Karhunen (1985) SGA algorithm, ensure that the weight matrix \mathbf{W} remains finite and full rank if it is so initially, and converges to an orthonormal set such that $\mathbf{W}\mathbf{W}^T$ converges to \mathbf{I}_M .

We considered the behaviour of the orthogonal projection \mathbf{P} into the subspace spanned by the rows of the weight matrix \mathbf{W} . We identified a number of Lyapunov

functions for convergence of \mathbf{P} to the principal subspace \mathbf{P}^* , namely the subspace projection error $S_{\mathbf{P}}$, the least mean square reconstruction error $S_{\mathbf{Q}}$, the ‘greatest’ mean square reconstruction error $S_{\mathbf{O}}$, and the mutual information I . However, in general, each the domains of attraction $D_{\mathbf{Q}} \supset D_{\mathbf{P}} \supset D_{\mathbf{O}}$ and D_I for these Lyapunov functions excludes a significant proportion of \mathbf{P} -space.

By modifying the mutual information I to give what we term the ‘principal subspace information’ I' , we get a domain of attraction $D_{I'}$ which covers all of \mathbf{P} -space except a lower-dimensional subset of \mathbf{P} -space for which $\mathbf{W}\mathbf{P}^*\mathbf{W}^T$ is singular. We conjecture that this is the maximal domain of attraction for \mathbf{P} to \mathbf{P}^* .

Therefore, if \mathbf{W} is adapted according to the o.d.e. (7) such that $\mathbf{W}\mathbf{W}^T$ remains finite and nonsingular, then \mathbf{P} will asymptotically converge to the principal subspace $\mathbf{P} = \mathbf{P}^*$ from almost everywhere.

Acknowledgments

The author would like to thank John Taylor and Guido Bugmann for many useful discussions on this and related work, and the comments of an anonymous referee which helped to improve this article considerably. During part of this work the author was supported by a Temporary Lectureship from the Academic Initiative of the University of London.

References

- Baldi, P. & Hornik, K. (1989). Neural networks and principal component analysis: Learning from examples without local minima. *Neural Networks*, **2**, 53–58.
- Cook, P. A. (1986). *Nonlinear dynamical systems*. Prentice Hall.
- Földiák, P. (1989). Adaptive network for optimal linear feature extraction. *Proceedings of the International Joint Conference on Neural Networks, IJCNN-89, Washington, DC*, (pp. 401–405).

- Hornik, K. & Kuan, C.-M. (1992). Convergence analysis of local feature extraction algorithms. *Neural Networks*, **5**, 229–240.
- Linsker, R. (1988). Self-organization in a perceptual network. *IEEE Computer*, **21**(3), 105–117.
- Oja, E. (1982). A simplified neuron model as a principal component analyser. *Journal of Mathematical Biology*, **15**, 267–273.
- Oja, E. (1989). Neural networks, principal components, and subspaces. *International Journal of Neural Systems*, **1**(1), 61–68.
- Oja, E. (1992). Principal components, minor components, and linear neural networks. *Neural Networks*, **5**, 927–935.
- Oja, E. & Karhunen, J. (1985). On stochastic approximation of the eigenvectors and eigenvalues of the expectation of a random matrix. *Journal of Mathematical Analysis and Applications*, **106**, 69–84.
- Plumbley, M. D. (1991). On information theory and unsupervised neural networks. Technical Report CUED/F-INFENG/TR. 78. Cambridge, UK: Cambridge University Engineering Department.
- Plumbley, M. D. (1993). Efficient information transfer and anti-Hebbian neural networks. *Neural Networks*, **6**, 823–833.
- Plumbley, M. D. & Fallside, F. (1988). An information-theoretic approach to unsupervised connectionist models. In D. Touretzky, G. Hinton, & T. Sejnowski (Eds.), *Proceedings of the 1988 Connectionist Models Summer School*, (pp. 239–245). San Mateo, CA: Morgan-Kaufmann.
- Sanger, T. D. (1989). Optimal unsupervised learning in a single-layer feedforward neural network. *Neural Networks*, **2**, 459–473.
- Strang, G. (1976). *Linear Algebra and its Applications*. New York: Academic Press.

Watanabe, S. (1985). *Pattern Recognition: Human and Mechanical*. New York: John Wiley & Sons.

Williams, R. J. (1985). Feature discovery through error-correction learning. ICS Report 8501. San Diego, CA: University of California.

A Appendix: Proofs of Theorems

A.1 Preliminaries

The proofs in this appendix use a number of standard results concerning matrix expressions: for details see any standard linear algebra text, such as (Strang, 1976). Before proceeding with the proofs, we remind the reader of some of the results we use here. Matrices here are all real.

Suppose that $\mathbf{A} = [a_{ij}]$ an $n \times n$ square matrix. Then we say \mathbf{A} is *positive definite*, written $\mathbf{A} > 0$, if $\mathbf{v}\mathbf{A}\mathbf{v}^T > 0$ for any non-zero n -vector \mathbf{v} . This holds iff all its eigenvectors are positive. If \mathbf{A} is positive definite, then it is certainly nonsingular, since all of its eigenvectors are non-zero. If $\mathbf{B}\mathbf{B}^T$ is nonsingular, then it is positive definite: it cannot have negative eigenvalues. If \mathbf{A} is positive definite, and \mathbf{B} is an $m \times n$ matrix, such that $\mathbf{B}\mathbf{B}^T$ is also positive definite, then $\mathbf{B}\mathbf{A}\mathbf{B}^T$ is positive definite.

Two matrices \mathbf{A} and \mathbf{B} *commute*, i.e. $\mathbf{A}\mathbf{B} = \mathbf{B}\mathbf{A}$, iff they have the same eigenvectors. If one (or both) of \mathbf{A} or \mathbf{B} has repeated eigenvalues (such as the projection \mathbf{P} , which has M of value 1 and N of value 0), then it is sufficient that one set can be found which is identical to the set of eigenvectors of the other matrix.

The *trace* of \mathbf{A} , written $\text{Tr}(\mathbf{A})$, is the sum of its diagonal entries, $\sum_i a_{ii}$, and satisfies $\text{Tr}(\mathbf{A}) = \text{Tr}(\mathbf{A}^T)$. For any $m \times n$ matrix \mathbf{B} and $n \times m$ matrix \mathbf{C} , we have $\text{Tr}(\mathbf{B}\mathbf{C}) = \text{Tr}(\mathbf{C}\mathbf{B})$. The trace of a matrix is the sum of its eigenvalues, much as the determinant of a matrix $\det(\mathbf{A})$ is the product of its eigenvalues. Suppose now that \mathbf{A} is positive definite. Then $\text{Tr}(\mathbf{B}\mathbf{A}\mathbf{B}^T) \geq 0$ for any $m \times n$ matrix \mathbf{B} , with equality iff \mathbf{B} is identically zero. In particular, $\text{Tr}(\mathbf{A}) > 0$, and $\text{Tr}(\mathbf{B}\mathbf{B}^T) \geq 0$

with the same equality condition for the latter inequality. Trace is a linear map, so $d/dt(\text{Tr}(\mathbf{A})) = \text{Tr}(d/dt(\mathbf{A}))$.

If \mathbf{A} and \mathbf{B} are two $n \times n$ square matrices, then $\det(\mathbf{AB}) = \det(\mathbf{A}) \det(\mathbf{B})$.

The matrix exponentiation function is written

$$e^{\mathbf{A}} = \mathbf{I} + \frac{\mathbf{A}^1}{1!} + \frac{\mathbf{A}^2}{2!} + \dots \quad (34)$$

with $\log(\mathbf{B}) = \mathbf{A}$ defined if $\mathbf{B} = e^{\mathbf{A}}$, where \mathbf{B} is positive definite if, for example, \mathbf{A} is real and symmetric. We have $e^{-\mathbf{A}} = (e^{\mathbf{A}})^{-1}$, and $\det e^{\mathbf{A}} = e^{\text{Tr}\mathbf{A}}$, so $\log \det \mathbf{B} = \text{Tr} \log \mathbf{B}$ if \mathbf{B} is positive definite.

Derivatives of functions of matrices are perhaps not so familiar. However, it is very easy to verify the identity

$$\text{Tr} d/dt (e^{\mathbf{A}}) = \text{Tr} (e^{\mathbf{A}} d\mathbf{A}/dt) \quad (35)$$

simply by differentiating the expansion (34) and rearranging terms within the trace to move the $d\mathbf{A}/dt$ terms to the end. In a similar way, we can show

$$\text{Tr} d\mathbf{A}/dt = \text{Tr}(\mathbf{B}^{-1} d\mathbf{B}/dt) \quad (36)$$

if $\mathbf{B} = e^{\mathbf{A}}$, so

$$\begin{aligned} d/dt(\log \det \mathbf{B}) &= d/dt(\text{Tr} \log \mathbf{B}) \\ &= \text{Tr}(\mathbf{B}^{-1} d\mathbf{B}/dt) \end{aligned} \quad (37)$$

if \mathbf{B} is positive definite. Equation (37) is particularly important for our treatment of mutual information functions.

A.2 Proof of Theorem 2.1

Since $\mathbf{W}\mathbf{W}^T$ is initially finite and nonsingular, then both $\mathbf{W}\mathbf{W}^T$ and $(\mathbf{W}\mathbf{W}^T)^{-1}$ are initially bounded towards \mathbf{I}_M .

Consider the following squared Frobenius norm cost functions:

$$\begin{aligned} J_1 &= \left\| \mathbf{I}_M - \mathbf{W}\mathbf{W}^T \right\|_F^2 \\ &= \text{Tr} \left((\mathbf{I}_M - \mathbf{W}\mathbf{W}^T)^2 \right) \end{aligned} \quad (38)$$

and

$$\begin{aligned}
J_2 &= \left\| \mathbf{I}_M - (\mathbf{W}\mathbf{W}^T)^{-1} \right\|_F^2 \\
&= \text{Tr} \left((\mathbf{I}_M - (\mathbf{W}\mathbf{W}^T)^{-1})^2 \right).
\end{aligned} \tag{39}$$

Taking the time derivative of J_1 and substituting for (7) and (11) when appropriate, we find

$$\begin{aligned}
dJ_1/dt &= -2\text{Tr} \left((\mathbf{I}_M - \mathbf{W}\mathbf{W}^T)(\mathbf{W}(d\mathbf{W}^T/dt) + (d\mathbf{W}/dt)\mathbf{W}^T) \right) \\
&= -2\text{Tr} \left((\mathbf{I}_M - \mathbf{W}\mathbf{W}^T)(2\mathbf{W}\Sigma_{\mathbf{x}}\mathbf{W}^T - (\mathbf{K}\mathbf{W}\mathbf{W}^T + \mathbf{W}\mathbf{W}^T\mathbf{K})) \right) \\
&= -4\text{Tr} \left((\mathbf{I}_M - \mathbf{W}\mathbf{W}^T)\mathbf{W}\Sigma_{\mathbf{x}}\mathbf{W}^T(\mathbf{I}_M - \mathbf{W}\mathbf{W}^T) \right) \\
&\leq 0.
\end{aligned} \tag{40}$$

Similarly, the time derivative of J_2 is given by

$$\begin{aligned}
dJ_2/dt &= -2\text{Tr} \left((\mathbf{I}_M - (\mathbf{W}\mathbf{W}^T)^{-1}) \left((\mathbf{W}\mathbf{W}^T)^{-1}(\mathbf{W}(d\mathbf{W}^T/dt) \right. \right. \\
&\quad \left. \left. + (d\mathbf{W}/dt)\mathbf{W}^T)(\mathbf{W}\mathbf{W}^T)^{-1} \right) \right) \\
&= 2\text{Tr} \left((\mathbf{I}_M - (\mathbf{W}\mathbf{W}^T)^{-1})(\mathbf{W}\mathbf{W}^T)^{-1}(2\mathbf{W}\Sigma_{\mathbf{x}}\mathbf{W}^T \right. \\
&\quad \left. - (\mathbf{K}\mathbf{W}\mathbf{W}^T + \mathbf{W}\mathbf{W}^T\mathbf{K}))(\mathbf{W}\mathbf{W}^T)^{-1} \right) \\
&= 4\text{Tr} \left((\mathbf{I}_M - (\mathbf{W}\mathbf{W}^T)^{-1})(\mathbf{W}\mathbf{W}^T)^{-1}\mathbf{W}\Sigma_{\mathbf{x}}\mathbf{W}^T(\mathbf{I}_M - (\mathbf{W}\mathbf{W}^T)^{-1}) \right) \\
&= -4\text{Tr} \left(\mathbf{W}^T(\mathbf{W}\mathbf{W}^T)^{-1}(\mathbf{I}_M - (\mathbf{W}\mathbf{W}^T)^{-1})\mathbf{W}\Sigma_{\mathbf{x}}\mathbf{W}^T \right. \\
&\quad \left. \times (\mathbf{I}_M - (\mathbf{W}\mathbf{W}^T)^{-1})(\mathbf{W}\mathbf{W}^T)^{-1}\mathbf{W} \right) \\
&\leq 0.
\end{aligned} \tag{41}$$

Thus the norms of both $(\mathbf{I}_M - \mathbf{W}\mathbf{W}^T)$ and $(\mathbf{I}_M - (\mathbf{W}\mathbf{W}^T)^{-1})$ are bounded for all t by their initial values, so both $\mathbf{W}\mathbf{W}^T$ and $(\mathbf{W}\mathbf{W}^T)^{-1}$ remain finite, so $\mathbf{W}\mathbf{W}^T$ remains finite and nonsingular for all $t > 0$.

Finally, since $\Sigma_{\mathbf{x}}$ and $\mathbf{W}\mathbf{W}^T$ are nonsingular for all t , $\mathbf{W}\Sigma_{\mathbf{x}}\mathbf{W}^T$ is also nonsingular for all t . Consequently, equality in (40) holds iff $\mathbf{W}\mathbf{W}^T = \mathbf{I}_M$. Therefore J_1 is a Lyapunov function for the convergence of $\mathbf{W}\mathbf{W}^T$ to the point $\mathbf{W}\mathbf{W}^T = \mathbf{I}_M$, with the domain of attraction consisting of all nonsingular $\mathbf{W}\mathbf{W}^T$. **QED.**

A.3 Proof of Lemma 4.1

We have

$$S_{\mathbf{P}} = E \left(|(\mathbf{I}_N - \mathbf{P})\mathbf{x}|^2 \right) \quad (42)$$

$$= E \left(\text{Tr} \left((\mathbf{I}_N - \mathbf{P})\mathbf{x}\mathbf{x}^T(\mathbf{I}_N - \mathbf{P}) \right) \right) \quad (43)$$

$$= \text{Tr} \left((\mathbf{I}_N - \mathbf{P})\boldsymbol{\Sigma}_{\mathbf{x}} \right) \quad (44)$$

using the identities $\text{Tr}(\mathbf{AB}) = \text{Tr}(\mathbf{BA})$ and $(\mathbf{I}_N - \mathbf{P})^2 = (\mathbf{I}_N - \mathbf{P})$. Differentiating with respect to t and using the identity $\text{Tr}(\mathbf{A}) = \text{Tr}(\mathbf{A}^T)$ we get

$$\frac{dS_{\mathbf{P}}}{dt} = -2\text{Tr} \left((\mathbf{I}_N - \mathbf{P})\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{P}\boldsymbol{\Sigma}_{\mathbf{x}} \right) \quad (45)$$

$$= -2\text{Tr} \left((\mathbf{I}_N - \mathbf{P})\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{P}\mathbf{P}\boldsymbol{\Sigma}_{\mathbf{x}}(\mathbf{I}_N - \mathbf{P}) \right) \quad (46)$$

$$\leq 0 \quad (47)$$

so $S_{\mathbf{P}}$ is a nonincreasing² function of t .

Now, equality in (47) holds iff $(\mathbf{I}_N - \mathbf{P})\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{P} = 0$, which is true iff $\mathbf{P}\boldsymbol{\Sigma}_{\mathbf{x}} = \boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{P}$. So $S_{\mathbf{P}}$ is stationary iff \mathbf{P} is the subspace spanned by some M eigenvectors of $\boldsymbol{\Sigma}_{\mathbf{x}}$. **QED.**

A.4 Proof of Theorem 4.2

When \mathbf{P} is the subspace spanned by some M of the eigenvectors of $\boldsymbol{\Sigma}_{\mathbf{x}}$, we have

$$S_{\mathbf{P}} = \text{Tr}(\boldsymbol{\Sigma}_{\mathbf{x}}) - \sum_{j=1}^M \lambda_{i_j} \quad (48)$$

where $\lambda_{i_1} \geq \dots \geq \lambda_{i_M}$ are the M eigenvalues of $\boldsymbol{\Sigma}_{\mathbf{x}}$ selected by \mathbf{P} . Thus $S_{\mathbf{P}}$ is minimized when the M *principal* eigenvalues are selected by \mathbf{P} , giving a minimal value of

$$S_{\mathbf{P}^*} = \min_{\mathbf{P}} S_{\mathbf{P}} = \sum_{i=M+1}^N \lambda_i \quad (49)$$

which is the sum of the $N - M$ smallest eigenvalues of $\boldsymbol{\Sigma}_{\mathbf{x}}$.

²With a little more manipulation it can be shown (Plumbley, 1991) that $dS_{\mathbf{P}}/dt = -\text{Tr} \left((d\mathbf{P}/dt)(d\mathbf{P}/dt)^T \right)$, so the o.d.e. (14) represents a *steepest descent* search in \mathbf{P} -space for a minimum of $S_{\mathbf{P}}$.

At the next lowest energy point $S_{\mathbf{P}^{[2]}} = \min_{\mathbf{P}} \{S_{\mathbf{P}} | \mathbf{P} \neq \mathbf{P}^*\}$, the ‘second best’ projection $\mathbf{P}^{[2]}$ selects the subspace corresponding to the largest $M - 1$ eigenvalues and the $M + 1$ st eigenvalue, so

$$\begin{aligned} S_{\mathbf{P}^{[2]}} &= \lambda_M + \sum_{i=M+2}^N \lambda_i \\ &= S_{\mathbf{P}^*} + \lambda_M - \lambda_{M+1}. \end{aligned} \quad (50)$$

If we therefore use $c = S_{\mathbf{P}^*} + \lambda_M - \lambda_{M+1}$ in Theorem 3.1, the conditions for $S_{\mathbf{P}}$ to be a Lyapunov function with domain of attraction $D_{\mathbf{P}}$ are satisfied. **QED**.

A.5 Proof of Lemma 4.3

We proceed by a ‘continuous induction’ method. We demonstrate (a) that a one-to-one mapping exists for a particular value of \mathbf{P} , and (b) that any infinitesimal change $d\mathbf{P}$ (resp. $d\mathbf{Q}$) in the variable \mathbf{P} is a function of itself, \mathbf{Q} (resp. \mathbf{P} and $\Sigma_{\mathbf{x}}$), and $d\mathbf{Q}$ (resp. $d\mathbf{P}$) only.

Consider (a) first. Suppose that \mathbf{P} is a subspace spanning some M eigenvectors of $\Sigma_{\mathbf{x}}$, so that $\mathbf{P}\Sigma_{\mathbf{x}} = \Sigma_{\mathbf{x}}\mathbf{P}$. Using this with (19), (20) and (21) as appropriate, we get

$$\begin{aligned} \mathbf{Q} &= \mathbf{Q}\mathbf{P} = \mathbf{Q}\mathbf{P}\Sigma_{\mathbf{x}}\Sigma_{\mathbf{x}}^{-1} = \mathbf{Q}\Sigma_{\mathbf{x}}\mathbf{P}\Sigma_{\mathbf{x}}^{-1} \\ &= \Sigma_{\mathbf{x}}\mathbf{Q}^T\mathbf{P}\Sigma_{\mathbf{x}}^{-1} = \Sigma_{\mathbf{x}}\mathbf{P}\Sigma_{\mathbf{x}}^{-1} \\ &= \mathbf{P}. \end{aligned} \quad (51)$$

Conversely, suppose that $\mathbf{Q} = \mathbf{Q}^T$ as would be the case in (51) above (since $P = P^T$). Then from (19) and (20) we have

$$\mathbf{P} = \mathbf{P}^T = \mathbf{Q}^T\mathbf{P} = \mathbf{Q}\mathbf{P} = \mathbf{Q} \quad (52)$$

so \mathbf{P} and \mathbf{Q} have a one-to-one relation (equality) for a particular value of \mathbf{P} and \mathbf{Q} . Furthermore, if $\mathbf{P} = \mathbf{Q}$ ($= \mathbf{Q}^T$), then

$$\mathbf{P}\Sigma_{\mathbf{x}} = \mathbf{Q}\Sigma_{\mathbf{x}} = \Sigma_{\mathbf{x}}\mathbf{Q} = \Sigma_{\mathbf{x}}\mathbf{P} \quad (53)$$

which proves the final part of the theorem.

Now consider (b). Differentiating (19) and rearranging we get

$$\mathbf{P}d\mathbf{Q} = d\mathbf{P}(\mathbf{I}_N - \mathbf{Q}) \quad (54)$$

which, postmultiplied by $(\mathbf{I}_N - \mathbf{P})$ gives us

$$\begin{aligned} \mathbf{P}d\mathbf{Q}(\mathbf{I}_N - \mathbf{P}) &= d\mathbf{P}(\mathbf{I}_N - \mathbf{Q})(\mathbf{I}_N - \mathbf{P}) \\ &= d\mathbf{P}(\mathbf{I}_N - \mathbf{P}) \end{aligned} \quad (55)$$

by substituting in (20). Also, differentiating (9) we get

$$0 = d\mathbf{P} - \mathbf{P}d\mathbf{P} - d\mathbf{P}\mathbf{P} \quad (56)$$

which, adding $d\mathbf{P}$ to both sides and substituting in (55) gives us

$$\begin{aligned} d\mathbf{P} &= d\mathbf{P}(\mathbf{I}_N - \mathbf{P}) + (\mathbf{I}_N - \mathbf{P})d\mathbf{P} \\ &= \mathbf{P}d\mathbf{Q}(\mathbf{I}_N - \mathbf{P}) + (\mathbf{I}_N - \mathbf{P})d\mathbf{Q}^T\mathbf{P} \end{aligned} \quad (57)$$

which defines any infinitesimal change $d\mathbf{P}$ in \mathbf{P} in terms of \mathbf{P} and $d\mathbf{Q}$ only.

For the converse direction, differentiating (19) and rearranging gives us

$$\mathbf{P}d\mathbf{Q} = d\mathbf{P}(\mathbf{I}_N - \mathbf{Q}) \quad (58)$$

which, together with (20) allows us to write

$$\mathbf{Q}d\mathbf{Q} = \mathbf{Q}\mathbf{P}d\mathbf{Q} = \mathbf{Q}d\mathbf{P}(\mathbf{I}_N - \mathbf{Q}). \quad (59)$$

Also, differentiating (21) for fixed $\Sigma_{\mathbf{x}}$ gives us

$$d\mathbf{Q}\Sigma_{\mathbf{x}} = \Sigma_{\mathbf{x}}d\mathbf{Q}^T \quad (60)$$

so

$$\begin{aligned} d\mathbf{Q}\mathbf{Q} &= d\mathbf{Q}\mathbf{Q}\Sigma_{\mathbf{x}}\Sigma_{\mathbf{x}}^{-1} \\ &= \Sigma_{\mathbf{x}}(\mathbf{Q}d\mathbf{Q})^T\Sigma_{\mathbf{x}}^{-1} \\ &= \Sigma_{\mathbf{x}}(\mathbf{I}_N - \mathbf{Q}^T)d\mathbf{P}\mathbf{Q}^T\Sigma_{\mathbf{x}}^{-1}. \end{aligned} \quad (61)$$

Finally, differentiating the relation $\mathbf{Q} = \mathbf{Q}^2$, and substituting in (59) and (61) we can now write

$$\begin{aligned} d\mathbf{Q} &= d\mathbf{Q}\mathbf{Q} + \mathbf{Q}d\mathbf{Q} \\ &= \Sigma_{\mathbf{x}}(\mathbf{I}_N - \mathbf{Q}^T)d\mathbf{P}\mathbf{Q}^T\Sigma_{\mathbf{x}}^{-1} + \mathbf{Q}d\mathbf{P}(\mathbf{I}_N - \mathbf{Q}). \end{aligned} \quad (62)$$

which shows that any infinitesimal change $d\mathbf{Q}$ to \mathbf{Q} is a function of \mathbf{Q} , $d\mathbf{P}$ and $\Sigma_{\mathbf{x}}$ only. **QED.**

A.6 Proof of Lemma 4.4

As for \mathbf{P} , the o.d.e. (7) defines the behaviour of \mathbf{Q} to be governed by the o.d.e.

$$d\mathbf{Q}/dt = (\mathbf{I}_N - \mathbf{Q})\Sigma_{\mathbf{x}}\mathbf{Q} + \mathbf{Q}\Sigma_{\mathbf{x}}(\mathbf{I}_N - \mathbf{Q}) \quad (63)$$

which is remarkable for being of identical form to the o.d.e. (14) for \mathbf{P} .

Using the identities $\Sigma_{\mathbf{x}}\mathbf{Q}^T = \mathbf{Q}\Sigma_{\mathbf{x}}$ and $(\mathbf{I}_N - \mathbf{Q})^2 = (\mathbf{I}_N - \mathbf{Q})$, we can write

$$\begin{aligned} S_{\mathbf{Q}} &= \text{Tr}\left((\mathbf{I}_N - \mathbf{Q})\Sigma_{\mathbf{x}}(\mathbf{I}_N - \mathbf{Q}^T)\right) \\ &= \text{Tr}\left((\mathbf{I}_N - \mathbf{Q})\Sigma_{\mathbf{x}}\right) \end{aligned} \quad (64)$$

which, taking the time derivative and substituting in (63), gives us

$$\begin{aligned} dS_{\mathbf{Q}}/dt &= -\text{Tr}\left((d\mathbf{Q}/dt)\Sigma_{\mathbf{x}}\right) \\ &= -2\text{Tr}\left(\Sigma_{\mathbf{x}}^{1/2}(\mathbf{I}_N - \mathbf{Q})^T\mathbf{Q}\Sigma_{\mathbf{x}}\mathbf{Q}^T(\mathbf{I}_N - \mathbf{Q})\Sigma_{\mathbf{x}}^{1/2}\right) \\ &\leq 0 \end{aligned} \quad (65)$$

with equality iff $(\mathbf{I}_N - \mathbf{Q})^T\mathbf{Q} = 0$. But $(\mathbf{I}_N - \mathbf{Q})^T\mathbf{Q} = 0$ implies that $\mathbf{Q} = \mathbf{Q}^T\mathbf{Q} = \mathbf{Q}^T$, and also conversely since $\mathbf{Q}^2 = \mathbf{Q}$. Therefore $S_{\mathbf{Q}}$ is stationary iff $\mathbf{Q} = \mathbf{Q}^T$ is an *orthogonal* projection.

If $\mathbf{Q} = \mathbf{Q}^T$ then $\mathbf{Q} = \mathbf{Q}\mathbf{P} = \mathbf{Q}^T\mathbf{P} = \mathbf{P}$, and conversely if $\mathbf{Q} = \mathbf{P}$ then $\mathbf{Q}^T = \mathbf{P} = \mathbf{Q}$, so $\mathbf{Q} = \mathbf{Q}^T$ iff $\mathbf{P} = \mathbf{Q}$. From Lemma 4.3 this holds iff they are the subspace spanned by some M of the eigenvectors of $\Sigma_{\mathbf{x}}$. **QED.**

A.7 Proof of Theorem 4.5

From Lemma 4.3, for any given $\Sigma_{\mathbf{x}}$ there is a one-to-one relationship between \mathbf{P} and \mathbf{Q} , so $S_{\mathbf{Q}}$ can be considered to be a function of \mathbf{P} only for any given $\Sigma_{\mathbf{x}}$. Also, from Lemma 4.4, whenever $S_{\mathbf{Q}}$ is stationary, we have $\mathbf{P} = \mathbf{Q}$, so $S_{\mathbf{Q}} = S_{\mathbf{P}}$.

In particular, we have

$$\min_{\mathbf{P}} S_{\mathbf{Q}} = \min_{\mathbf{P}} S_{\mathbf{P}} = S_{\mathbf{P}^*} \quad (66)$$

and

$$\min_{\mathbf{P} \neq \mathbf{P}^*} S_{\mathbf{Q}} = S_{\mathbf{P}^*} + \lambda_M - \lambda_{M+1}. \quad (67)$$

If we therefore use $c = S_{\mathbf{P}^*} + \lambda_M - \lambda_{M+1}$ in Theorem 3.1, the conditions for $S_{\mathbf{Q}}$ to be a Lyapunov function (for \mathbf{P}) with domain of attraction $D_{\mathbf{Q}}$ are satisfied.

Finally, note that since $S_{\mathbf{Q}}$ is the *least* mean squared reconstruction error, $S_{\mathbf{Q}} \leq S_{\mathbf{P}}$, so

$$S_{\mathbf{P}} < S_{\mathbf{P}^*} + \lambda_M - \lambda_{M+1} \quad (68)$$

implies

$$S_{\mathbf{Q}} < S_{\mathbf{P}^*} + \lambda_M - \lambda_{M+1} \quad (69)$$

so $D_{\mathbf{Q}} \supset D_{\mathbf{P}}$. **QED.**

A.8 Proof of Lemma 4.6

Since $\Sigma_{\mathbf{x}}$ and $\mathbf{W}\mathbf{W}^T$ are nonsingular for all t , $\mathbf{W}\Sigma_{\mathbf{x}}\mathbf{W}^T$ is also nonsingular for all t . Differentiating (25) with respect to t and using the identity (37) gives us

$$\begin{aligned} dI/dt &= \text{Tr} \left((\mathbf{W}\Sigma_{\mathbf{x}}\mathbf{W}^T)^{-1} \mathbf{W}\Sigma_{\mathbf{x}} (d\mathbf{W}^T/dt) - (\sigma_{\phi}^2 \mathbf{W}\mathbf{W}^T)^{-1} \sigma_{\phi}^2 \mathbf{W} (d\mathbf{W}^T/dt) \right) \\ &= \text{Tr} \left((\mathbf{W}\Sigma_{\mathbf{x}}\mathbf{W}^T)^{-1} \mathbf{W}\Sigma_{\mathbf{x}} (\mathbf{I}_N - \mathbf{P}) (d\mathbf{W}^T/dt) \right) \\ &= \text{Tr} \left((\mathbf{W}\Sigma_{\mathbf{x}}\mathbf{W}^T)^{-1} \mathbf{W}\Sigma_{\mathbf{x}} (d\mathbf{P}/dt) \mathbf{W}^T \right) \end{aligned} \quad (70)$$

using the identity $(d\mathbf{P}/dt)\mathbf{W}^T = (\mathbf{I}_N - \mathbf{P})(d\mathbf{W}^T/dt)$. Substituting in (14) we get

$$\begin{aligned} dI/dt &= \text{Tr} \left((\mathbf{W}\Sigma_{\mathbf{x}}\mathbf{W}^T)^{-1} \mathbf{W}\Sigma_{\mathbf{x}} (\mathbf{I}_N - \mathbf{P}) \Sigma_{\mathbf{x}} \mathbf{W}^T \right) \\ &= \text{Tr} \left((\mathbf{I}_N - \mathbf{P}) \Sigma_{\mathbf{x}} \mathbf{W}^T (\mathbf{W}\Sigma_{\mathbf{x}}\mathbf{W}^T)^{-1} \mathbf{W}\Sigma_{\mathbf{x}} (\mathbf{I}_N - \mathbf{P}) \right) \\ &\geq 0 \end{aligned} \quad (71)$$

with equality iff $(\mathbf{I}_N - \mathbf{P})\boldsymbol{\Sigma}_x \mathbf{W}^T = 0$, i.e. iff $(\mathbf{I}_N - \mathbf{P})\boldsymbol{\Sigma}_x \mathbf{P} = 0$, which holds iff \mathbf{P} is the subspace spanned by some M eigenvectors of $\boldsymbol{\Sigma}_x$. **QED.**

A.9 Proof of Lemma 4.7

Let us write a singular value decomposition (SVD) of \mathbf{W} ,

$$\mathbf{W} = \mathbf{V}_W \mathbf{G}_W^{1/2} \mathbf{U}_W^T \quad (72)$$

where \mathbf{V}_W is an $M \times M$ matrix such that $\mathbf{V}_W \mathbf{V}_W^T = \mathbf{V}_W^T \mathbf{V}_W = \mathbf{I}_M$, $\mathbf{G}_W^{1/2}$ is an $M \times M$ diagonal matrix, and \mathbf{U}_W is an $N \times M$ matrix such that $\mathbf{U}_W^T \mathbf{U}_W = \mathbf{I}_M$.

Also, since \mathbf{P} has rank M we can also decompose \mathbf{P} into the product

$$\mathbf{P} = \mathbf{U}_P \mathbf{U}_P^T \quad (73)$$

where \mathbf{U}_P is an $N \times M$ matrix such that $\mathbf{U}_P^T \mathbf{U}_P = \mathbf{I}_M$ (since $\mathbf{P}^2 = \mathbf{P}$).

However, from the definition of \mathbf{P} , we can also write

$$\begin{aligned} \mathbf{P} &= \mathbf{U}_W \mathbf{G}_W^{1/2} \mathbf{V}_W^T \left(\mathbf{V}_W \mathbf{G}_W^{1/2} \mathbf{U}_W^T \mathbf{U}_W \mathbf{G}_W^{1/2} \mathbf{V}_W^T \right)^{-1} \mathbf{V}_W \mathbf{G}_W^{1/2} \mathbf{U}_W^T \\ &= \mathbf{U}_W \mathbf{U}_W^T \end{aligned} \quad (74)$$

so for any \mathbf{U}_P chosen in (73) we have

$$\mathbf{U}_W = \mathbf{U}_P \mathbf{R}_P \quad (75)$$

for some $M \times M$ matrix \mathbf{R}_P such that $\mathbf{R}_P \mathbf{R}_P^T = \mathbf{I}_M$. (Note that we cannot simply state that \mathbf{U}_P and \mathbf{U}_W are equal.)

Now expanding I in terms of our SVD of \mathbf{W} , we get

$$\begin{aligned} I &= \frac{1}{2} \left(\log \det(\mathbf{V}_W \mathbf{G}_W^{1/2} \mathbf{U}_W^T \boldsymbol{\Sigma}_x \mathbf{U}_W \mathbf{G}_W^{1/2} \mathbf{V}_W^T) \right. \\ &\quad \left. - \log \det(\sigma_\phi^2 \mathbf{V}_W \mathbf{G}_W^{1/2} \mathbf{U}_W^T \mathbf{U}_W \mathbf{G}_W^{1/2} \mathbf{V}_W^T) \right) \\ &= \frac{1}{2} \left(\log \det(\mathbf{G}_W^{1/2} \mathbf{U}_W^T \boldsymbol{\Sigma}_x \mathbf{U}_W \mathbf{G}_W^{1/2}) - \log \det(\sigma_\phi^2 \mathbf{G}_W) \right) \\ &= \frac{1}{2} \left(\log \det(\mathbf{G}_W) + \log \det(\mathbf{U}_W^T \boldsymbol{\Sigma}_x \mathbf{U}_W) - \log \sigma_\phi^{2M} - \log \det(\mathbf{G}_W) \right) \\ &= \frac{1}{2} \left(\log \det(\mathbf{R}_P^T \mathbf{U}_P^T \boldsymbol{\Sigma}_x \mathbf{U}_P \mathbf{R}_P) - \log \sigma_\phi^{2M} \right) \\ &= \frac{1}{2} \left(\log \det(\mathbf{U}_P^T \boldsymbol{\Sigma}_x \mathbf{U}_P) - \log \sigma_\phi^{2M} \right) \end{aligned} \quad (76)$$

which is a function of \mathbf{P} only, given $\boldsymbol{\Sigma}_x$ and σ_ϕ^2 . **QED.**

A.10 Proof of Theorem 4.8

The mutual information at the stationary points is given by

$$I = \frac{1}{2} \sum_{j=1}^M \log \frac{\lambda_{i_j}}{\sigma_\phi^2} \quad (77)$$

where $\lambda_{i_1} \geq \dots \geq \lambda_{i_M}$ are the eigenvalues of $\Sigma_{\mathbf{x}}$ selected by \mathbf{P} as before, and will have maximum value

$$I_{\mathbf{P}^*} = \max_{\mathbf{P}} I = \frac{1}{2} \sum_{i=1}^M \log \frac{\lambda_i}{\sigma_\phi^2} \quad (78)$$

when \mathbf{P}^* is the subspace spanned by the M principal eigenvectors of $\Sigma_{\mathbf{x}}$. Thus if I is to be used as a Lyapunov function, the identified domain of attraction will be the set D_I of points \mathbf{P} such that

$$\begin{aligned} I &> I_{\mathbf{P}^{[2]}} \\ &= \frac{1}{2} \log \frac{\lambda_{M+1}}{\sigma_\phi^2} + \frac{1}{2} \sum_{i=1}^{M-1} \log \frac{\lambda_i}{\sigma_\phi^2} \\ &= I_{\mathbf{P}^*} - \frac{1}{2} \log \frac{\lambda_M}{\lambda_{M+1}}. \end{aligned} \quad (79)$$

QED.

A.11 Proof of Lemma 5.1

Let us tentatively suppose that $\mathbf{W}\mathbf{P}^*\mathbf{W}^T$ is nonsingular for all t . The $N \times N$ orthogonal projection \mathbf{P}^* is of rank M , so it can be expressed as $\mathbf{P}^* = \mathbf{U}^*(\mathbf{U}^*)^T$ where \mathbf{U}^* is an $M \times N$ matrix. Therefore $\mathbf{W}\mathbf{P}^*\mathbf{W}^T = (\mathbf{W}\mathbf{U}^*)(\mathbf{W}\mathbf{U}^*)^T$ where $(\mathbf{W}\mathbf{U}^*)$ is an $M \times M$ nonsingular matrix.

Differentiating (27) with respect to t as in the proof of Theorem 4.8, we get

$$dI'/dt = \text{Tr} \left((\mathbf{W}\mathbf{P}^*\mathbf{W}^T)^{-1} \mathbf{W}\mathbf{P}^* (d\mathbf{P}/dt) \mathbf{W}^T \right). \quad (80)$$

Substituting (14) into (80) and noting that $\mathbf{P}^*\Sigma_{\mathbf{x}} = (\mathbf{P}^*)^2\Sigma_{\mathbf{x}} = \mathbf{P}^*\Sigma_{\mathbf{x}}\mathbf{P}^*$ since \mathbf{P}^* is a projection which shares its eigenvectors with $\Sigma_{\mathbf{x}}$, we get

$$\begin{aligned} dI'/dt &= \text{Tr} \left((\mathbf{W}\mathbf{P}^*\mathbf{W}^T)^{-1} \mathbf{W}\mathbf{P}^* (\mathbf{I}_N - \mathbf{P}) \Sigma_{\mathbf{x}} \mathbf{W}^T \right) \\ &= \text{Tr} \left((\mathbf{W}\mathbf{P}^*\mathbf{W}^T)^{-1} \mathbf{W}\mathbf{P}^* \Sigma_{\mathbf{x}} \mathbf{P}^* \mathbf{W}^T \right) \end{aligned}$$

$$\begin{aligned}
& -\text{Tr} \left((\mathbf{W}\mathbf{P}^*\mathbf{W}^T)^{-1} \mathbf{W}\mathbf{P}^*\mathbf{W}^T (\mathbf{W}\mathbf{W}^T)^{-1} \mathbf{W}\boldsymbol{\Sigma}_x \mathbf{W}^T \right) \\
&= \text{Tr} \left(((\mathbf{U}^*)^T \mathbf{W}^T)^{-1} (\mathbf{W}\mathbf{U}^*)^{-1} (\mathbf{W}\mathbf{U}^*) (\mathbf{U}^*)^T \boldsymbol{\Sigma}_x \mathbf{U}^* ((\mathbf{U}^*)^T \mathbf{W}^T) \right) - \text{Tr} (\mathbf{P}\boldsymbol{\Sigma}_x) \\
&= \text{Tr} (\mathbf{P}^* \boldsymbol{\Sigma}_x) - \text{Tr} (\mathbf{P}\boldsymbol{\Sigma}_x) \\
&= \text{Tr} ((\mathbf{I}_N - \mathbf{P})\boldsymbol{\Sigma}_x) - \text{Tr} ((\mathbf{I}_N - \mathbf{P}^*)\boldsymbol{\Sigma}_x) \\
&= S_{\mathbf{P}} - S_{\mathbf{P}^*} \\
&\geq 0
\end{aligned} \tag{81}$$

with equality iff $\mathbf{P} = \mathbf{P}^*$, since $S_{\mathbf{P}^*}$ is the unique minimum of $S_{\mathbf{P}}$.

It remains to prove that $\mathbf{W}\mathbf{P}^*\mathbf{W}^T$ is nonsingular for all t . From the proof of Theorem 2.1, we know that the norm of $(\mathbf{I}_M - (\mathbf{W}\mathbf{W}^T)^{-1})$ for all t is bounded above by its initial value. Therefore, the smallest eigenvector of $\mathbf{W}\mathbf{W}^T$ is bounded below, so $\det(\mathbf{W}\mathbf{W}^T)$ is bounded below for all t . Now, $\mathbf{W}\mathbf{P}^*\mathbf{W}^T$ is initially nonsingular, so $\log \det(\mathbf{W}\mathbf{P}^*\mathbf{W}^T)$ is initially finite, giving a finite initial value for I' . The term $\mathbf{W}\mathbf{P}^*\mathbf{W}^T$ cannot become singular without $\log \det(\mathbf{W}\mathbf{P}^*\mathbf{W}^T)$ first going below the lower bound fixed by $\det(\mathbf{W}\mathbf{W}^T)$ and I' (which is nondecreasing as long as $\mathbf{W}\mathbf{P}^*\mathbf{W}^T$ is nonsingular). Therefore $\mathbf{W}\mathbf{P}^*\mathbf{W}^T$ is forced to remain nonsingular for all t . **QED.**

A.12 Proof of Theorem 5.2

I' has no secondary stationary points which concern us, so we only need ensure that $\mathbf{W}\mathbf{P}^*\mathbf{W}^T$ is initially nonsingular to find our domain of attraction. This condition is equivalent to the condition for I' to be finite, i.e. $I' > -\infty$, so I' the domain of attraction is $D_{I'} = \{\mathbf{P} | I' > -\infty\}$. However, it is probably clearer to write this more simply in terms of the nonsingularity of $\mathbf{W}\mathbf{P}^*\mathbf{W}^T$, as $D_{I'} = \{\mathbf{P} | \mathbf{W}\mathbf{P}^*\mathbf{W}^T \text{ is nonsingular}\}$. **QED.**

A.13 Proof of Theorem 6.1

Considering first the boundaries of $D_{\mathbf{P}}$, suppose without loss of generality that the input covariance is diagonal, equal to

$$\boldsymbol{\Sigma}_{\mathbf{x}} = \begin{bmatrix} \lambda_1 & & 0 \\ & \lambda_2 & \\ 0 & & \lambda_3 \end{bmatrix}. \quad (82)$$

The the boundary condition that $S_{\mathbf{P}} = \lambda_2$ is equivalent to

$$0 = \begin{bmatrix} \lambda_1 - \lambda_2 & & 0 \\ & 0 & \\ 0 & & -(\lambda_2 - \lambda_3) \end{bmatrix} \quad (83)$$

$$= (\mathbf{u}_{\mathbf{P}} \cdot \mathbf{b}_{\mathbf{P}}^+) (\mathbf{u}_{\mathbf{P}} \cdot \mathbf{b}_{\mathbf{P}}^-) \quad (84)$$

where

$$\mathbf{b}_{\mathbf{P}}^+ = [(\lambda_1 - \lambda_2)^{1/2} \quad 0 \quad (\lambda_2 - \lambda_3)^{1/2}]^T \quad (85)$$

$$\mathbf{b}_{\mathbf{P}}^- = [(\lambda_1 - \lambda_2)^{1/2} \quad 0 \quad -(\lambda_2 - \lambda_3)^{1/2}]^T \quad (86)$$

by superposition of the antisymmetric forms of the expression. Therefore $S_{\mathbf{P}} = \lambda_2$ iff $\mathbf{u}_{\mathbf{P}} \cdot \mathbf{b}_{\mathbf{P}}^+ = 0$ or $\mathbf{u}_{\mathbf{P}} \cdot \mathbf{b}_{\mathbf{P}}^- = 0$, giving the two planes we are looking for. In terms of the projection \mathbf{P} itself, on the boundary of $D_{\mathbf{P}}$ one of either $\mathbf{b}_{\mathbf{P}}^+$ or $\mathbf{b}_{\mathbf{P}}^-$ must be in the plane of \mathbf{P} .

A similar argument gives a related result for $D_{\mathbf{Q}}$, with

$$\mathbf{b}_{\mathbf{Q}}^+ = \left[\left(\frac{1}{\lambda_2} - \frac{1}{\lambda_1} \right)^{1/2} \quad 0 \quad \left(\frac{1}{\lambda_3} - \frac{1}{\lambda_2} \right)^{1/2} \right]^T \quad (87)$$

$$\mathbf{b}_{\mathbf{Q}}^- = \left[\left(\frac{1}{\lambda_2} - \frac{1}{\lambda_1} \right)^{1/2} \quad 0 \quad - \left(\frac{1}{\lambda_3} - \frac{1}{\lambda_2} \right)^{1/2} \right]^T \quad (88)$$

as the two vectors normal to the boundary ring planes. **QED.**

A.14 Proof of Theorem 6.2

First, let us hypothesize that, for any finite invertible input covariance matrix $\boldsymbol{\Sigma}$, we have

$$I_{\text{TOT}}(\boldsymbol{\Sigma}) = I(\boldsymbol{\Sigma}) + \bar{I}(\boldsymbol{\Sigma}) \quad (89)$$

where

$$\begin{aligned}
2I_{\text{TOT}}(\boldsymbol{\Sigma}) &= \log \det(\boldsymbol{\Sigma}) \\
2I(\boldsymbol{\Sigma}) &= \log \det(\mathbf{U}_{\mathbf{P}}^T \boldsymbol{\Sigma} \mathbf{U}_{\mathbf{P}}) \\
2\bar{I}(\boldsymbol{\Sigma}) &= \log S_{\mathbf{Q}} = -\log(\mathbf{u}_{\mathbf{P}}^T \boldsymbol{\Sigma}^{-1} \mathbf{u}_{\mathbf{P}}).
\end{aligned}$$

The proof proceeds by ‘continuous induction’: i.e. we proceed to show that (89) is satisfied for a particular value of $\boldsymbol{\Sigma}$, and that the derivatives of both sides with respect to an infinitesimal offset $d\boldsymbol{\Sigma}$ of $\boldsymbol{\Sigma}$ are equal.

Consider first $\boldsymbol{\Sigma} = \mathbf{I}_N$. Clearly

$$I_{\text{TOT}}(\mathbf{I}_N) = \log \det \mathbf{I}_N = 0$$

and

$$I(\mathbf{I}_N) + \bar{I}(\mathbf{I}_N) = \log \det(\mathbf{U}_{\mathbf{P}}^T \mathbf{I}_N \mathbf{U}_{\mathbf{P}}) - \log(\mathbf{u}_{\mathbf{P}}^T \mathbf{I}_N \mathbf{u}_{\mathbf{P}}) = 0$$

so (89) is satisfied at $\boldsymbol{\Sigma} = \mathbf{I}_N$.

Now differentiating with respect to $\boldsymbol{\Sigma}$, and using the identity (37), for the derivative of I_{TOT} we get

$$d(I_{\text{TOT}}(\boldsymbol{\Sigma})) = \text{Tr}(\boldsymbol{\Sigma}^{-1} d\boldsymbol{\Sigma}) \quad (90)$$

while for I we get

$$\begin{aligned}
d(I(\boldsymbol{\Sigma})) &= \text{Tr}((\mathbf{U}_{\mathbf{P}}^T \boldsymbol{\Sigma} \mathbf{U}_{\mathbf{P}})^{-1} \mathbf{U}_{\mathbf{P}}^T d\boldsymbol{\Sigma} \mathbf{U}_{\mathbf{P}}) \\
&= \text{Tr}(\boldsymbol{\Sigma}^{-1} \mathbf{Q} d\boldsymbol{\Sigma})
\end{aligned} \quad (91)$$

and for \bar{I} we get

$$\begin{aligned}
d(\bar{I}(\boldsymbol{\Sigma})) &= -\text{Tr}\left(\left(\mathbf{u}_{\mathbf{P}}^T \boldsymbol{\Sigma}^{-1} \mathbf{u}_{\mathbf{P}}\right)^{-1} \mathbf{u}_{\mathbf{P}}^T (-\boldsymbol{\Sigma}^{-1} d\boldsymbol{\Sigma} \boldsymbol{\Sigma}^{-1}) \mathbf{u}_{\mathbf{P}}\right) \\
&= \text{Tr}\left(\boldsymbol{\Sigma}^{-1} \mathbf{u}_{\mathbf{P}} (\mathbf{u}_{\mathbf{P}}^T \boldsymbol{\Sigma}^{-1} \mathbf{u}_{\mathbf{P}})^{-1} \mathbf{u}_{\mathbf{P}}^T (-\boldsymbol{\Sigma}^{-1} d\boldsymbol{\Sigma})\right) \\
&= \text{Tr}(\boldsymbol{\Sigma}^{-1} \bar{\mathbf{Q}} d\boldsymbol{\Sigma}).
\end{aligned} \quad (92)$$

Combining (91) and (92) we get

$$\begin{aligned}
d(I(\boldsymbol{\Sigma})) + d(\bar{I}(\boldsymbol{\Sigma})) &= \text{Tr}(\boldsymbol{\Sigma}^{-1} (\mathbf{Q} + \bar{\mathbf{Q}}) d\boldsymbol{\Sigma}) \\
&= \text{Tr}(\boldsymbol{\Sigma}^{-1} d\boldsymbol{\Sigma}) \\
&= d(I_{\text{TOT}}(\boldsymbol{\Sigma}))
\end{aligned} \quad (93)$$

so the hypothesis is proved.

Since I_{TOT} is not a function of \mathbf{P} , and we have $I = I_{\text{TOT}} - \log S_{\mathbf{Q}}$, then there is a one-to-one relation between I and $S_{\mathbf{Q}}$, so $D_I = D_{\mathbf{Q}}$. **QED**.

Nomenclature

N, M	Number of inputs, outputs
\mathbf{x}, \mathbf{y}	Input, output vector
x_i, y_i	i th input, output component
\mathbf{W}	$M \times N$ weight matrix
$\Sigma_{\mathbf{x}}$	Covariance matrix of \mathbf{x} , ($= E(\mathbf{x}\mathbf{x}^T)$)
$E(\cdot)$	Expectation
$\mathbf{x}_n, \mathbf{y}_n, \mathbf{W}_n, \dots$	$\mathbf{x}, \mathbf{y}, \mathbf{W}, \dots$ at time step n
$\mathbf{x}^T, \mathbf{A}^T$	Transpose of vector, matrix
\mathbf{K}	Weight decay matrix in Hebbian algorithms
η_n	Update factor
$\text{diag}(\cdot)$	Set off-diagonal matrix entries to zero
$\text{LT}(\cdot)$ ($\text{LT}^+(\cdot)$)	Make matrix (strictly) lower triangular
$\langle \cdot \rangle$	Mean (in informal anal.)
t	Continuous time
\mathbf{P}	Orthogonal projection, ($= \mathbf{W}^T(\mathbf{W}\mathbf{W}^T)^{-1}\mathbf{W}$)
$\mathbf{I}_N, \mathbf{I}_M$	N -, M -dimensional identity matrix
$\tilde{\mathbf{x}}$	Orthogonal projection of \mathbf{x} ($= \mathbf{P}\mathbf{x}$)
$S, S(\mathbf{W})$	Energy function
$L, L(a)$	Lyapunov function of a
D	Lyapunov domain of attraction
c	Upper limit for Lyapunov function defining D
a^*	Attractor for a
$a^{[2]}, a^{[3]}, \dots$	Equilibrium points for a
$S_{\mathbf{P}}$ ($S_{\mathbf{P}^*}$)	Subspace projection error (at $P = P^*$)
$ \mathbf{v} $	Length of vector \mathbf{v}
$\text{Tr}(\mathbf{A})$	Trace of matrix \mathbf{A} ($= \sum_i a_{ii}$)
λ_i	i th largest eigenvalue of $\Sigma_{\mathbf{x}}$
\mathbf{P}^*	Point of convergence for \mathbf{P} (principal subspace of $\Sigma_{\mathbf{x}}$)
$S_{\mathbf{P}^{[2]}}$	Second-lowest equilibrium value of $S_{\mathbf{P}}$

$D_{\mathbf{P}}$	Domain of attraction defined by Lyapunov function $S_{\mathbf{P}}$
$\ \mathbf{A}\ _F$	Frobenius norm of \mathbf{A} $\left(= \left(\sum_{ij} a_{ij}^2 \right)^{1/2} \right)$
$\hat{\mathbf{x}}$	Least mean square projection of \mathbf{x} ($= \mathbf{Q}\mathbf{x}$)
\mathbf{Q}	Least mean square projection ($= \Sigma_{\mathbf{x}}\mathbf{W}^T(\mathbf{W}\Sigma_{\mathbf{x}}\mathbf{W}^T)^{-1}\mathbf{W}$)
$S_{\mathbf{Q}}$	Least mean square projection error
\mathbf{O}	Dual projection to \mathbf{Q} ('greatest' MSE projection)
$S_{\mathbf{O}}$	Projection error due to \mathbf{O}
I	Mutual information
$\Sigma_{\phi}, \sigma_{\phi}^2$	Noise covariance matrix, variance
$\det(\cdot)$	Determinant of matrix
I_{P^*}	Value of I at principal subspace $P = P^*$
D_I	Domain of attraction defined by Lyapunov function I
$x, y, w, \Delta w$	Scalar input, output, weight, weight update
\mathbf{U}^*	$N \times M$ matrix such that $\mathbf{U}^*(\mathbf{U}^*)^T = \mathbf{P}^*$, $(\mathbf{U}^*)^T\mathbf{U}^* = \mathbf{I}_M$
$\mathbf{W} = \mathbf{V}_{\mathbf{W}}\mathbf{G}_{\mathbf{W}}^{1/2}\mathbf{U}_{\mathbf{W}}^T$	Singular value decomposition (SVD) of \mathbf{W}
I'	Principal subspace 'information'
$D_{I'}$	Domain of attraction defined by Lyapunov function I'
J_1, J_2	Cost functions for convergence of $\mathbf{W}\mathbf{W}^T$
$\mathbf{U}_{\mathbf{P}}$	$N \times M$ matrix such that $\mathbf{U}_{\mathbf{P}}\mathbf{U}_{\mathbf{P}}^T = \mathbf{P}$, $\mathbf{U}_{\mathbf{P}}^T\mathbf{U}_{\mathbf{P}} = \mathbf{I}_M$
$\overline{\mathbf{P}}, \overline{\mathbf{Q}}$	Complement projections $(\mathbf{I}_N - \mathbf{P})$, $(\mathbf{I}_N - \mathbf{Q})$
\mathbf{e} (\mathbf{e}_i)	Eigenvector (i th eigenvector)
$\mathbf{u}_{\overline{\mathbf{P}}}$	Unit length vector if $M = N - 1$ such that $\mathbf{u}_{\overline{\mathbf{P}}}\mathbf{u}_{\overline{\mathbf{P}}}^T = \overline{\mathbf{P}}$
$\mathbf{b}_{\mathbf{P}}, \mathbf{b}_{\mathbf{P}}^+, \mathbf{b}_{\overline{\mathbf{P}}}^-$	Boundary rings for $S_{\mathbf{P}}$ when $N = 3, M = 2$
$\mathbf{b}_{\mathbf{Q}}, \mathbf{b}_{\mathbf{Q}}^+, \mathbf{b}_{\overline{\mathbf{Q}}}^-$	Boundary rings for $S_{\mathbf{Q}}$ when $N = 3, M = 2$
$\theta_{\mathbf{P}}, \theta_{\mathbf{Q}}$	Angle of $\mathbf{b}_{\mathbf{P}}, \mathbf{b}_{\mathbf{Q}}$ to the 'vertical'
$I_{\text{TOT}}, \overline{I}$	'Total', lost information such that $I = I_{\text{TOT}} - \overline{I}$
Σ	Some covariance matrix
$\mathbf{R}_{\mathbf{P}}$	$\mathbf{U}_{\mathbf{W}}$ 'rotation' matrix ($\mathbf{U}_{\mathbf{W}} = \mathbf{U}_{\mathbf{P}}\mathbf{R}_{\mathbf{P}}$, $\mathbf{R}_{\mathbf{P}}\mathbf{R}_{\mathbf{P}}^T = \mathbf{I}_M$)

Figure Captions

Figure 1: Linear network with input \mathbf{x} , weight matrix \mathbf{W} , and output $\mathbf{y} = \mathbf{W}\mathbf{x}$.

Figure 2: Domains of attraction (a) $D_{\mathbf{P}}$, (b) $D_{\mathbf{Q}}$, (c) D_I and (d) $D_{I'}$ for a 3-input 2-output network with input component eigenvalues λ_1 , λ_2 , and λ_3 or value 3, 2 and 1 respectively. The eigenvectors of the input covariance matrix are indicated by \mathbf{e}_1 , \mathbf{e}_2 , and \mathbf{e}_3 . Note that $D_{\mathbf{Q}}$ is larger than $D_{\mathbf{P}}$, D_I and $D_{\mathbf{Q}}$ are identical, and $D_{I'}$ covers the whole of \mathbf{P} -space except the equator of the sphere, where $\det(\mathbf{W}\mathbf{P}^*\mathbf{W}^T) = 0$.

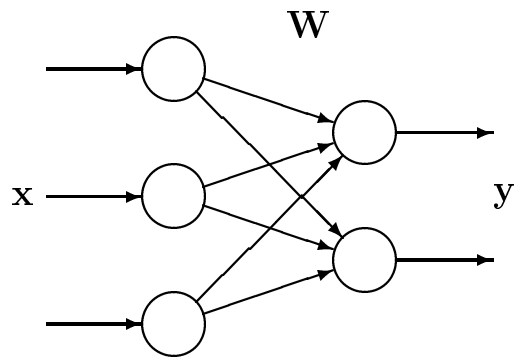


Fig. 1

