LETTER Communicated by Erkki Oja

The Whitney Reduction Network: A Method for Computing Autoassociative Graphs

D. S. Broomhead
Department of Mathematics, University of Manchester Institute of Science and Technology, Manchester M60 1QD, U.K.

M. J. Kirby
Department of Mathematics, Colorado State University, Fort Collins, CO 80523, U.S.A.

This article introduces a new architecture and associated algorithms ideal for implementing the dimensionality reduction of an m-dimensional manifold initially residing in an n-dimensional Euclidean space where $n \gg m$. Motivated by Whitney's embedding theorem, the network is capable of training the identity mapping employing the idea of the graph of a function. In theory, a reduction to a dimension d that retains the differential structure of the original data may be achieved for some $d \leq 2m + 1$. To implement this network, we propose the idea of a *good*-projection, which enhances the generalization capabilities of the network, and an adaptive secant basis algorithm to achieve it. The effect of noise on this procedure is also considered. The approach is illustrated with several examples.

1 Data Parameterization

The application of analytical transforms, such as the Fourier or wavelet transform, is an established technique for investigating low-dimensional data sets. Analogously, constructing custom empirical transforms (and their inverses) is an attractive means for extracting and manipulating information in large, high-dimensional data sets. The subject of this investigation concerns the construction of (invertible) dimensionality-reducing tranformations of data sets for the case where the initial, or ambient, dimension is much greater than its intrinsic (topological) dimension.

There are two fundamental approaches for representing data in a reduced coordinate system: data parameterization and data encapsulation. The distinction between the two approaches lies in the differing assumptions concerning how the data set $\mathcal{A}$ occupies the data space. Data encapsulation is based on the assumption that the data are a subset of a linear subspace and seek a coordinate change that makes this apparent. Data parameterization, on the other hand, assumes that the data set is produced by a discrete sampling of an m-dimensional submanifold $\mathcal{M}$; in this case, we

seek to represent the data as a graph from a suitable linear subspace to its orthogonal compliment.

The parameterization of a data set $\mathcal{A} \subset \mathbb{R}^n$ may be achieved by determining a new coordinate system that is the result of a dimensionality-reducing mapping $y = G(x) \in \mathcal{B} \subset \mathbb{R}^d$ where $d < n$. The associated reconstruction mapping $x = H(y) \in \mathcal{A} \subset \mathbb{R}^n$ takes the data and maps them back to the original ambient coordinates. Thus, H provides a global d-dimensional parameterization of the data. By contrast, data encapsulation seeks to map, typically via a unitary transformation, the data into a subspace of reduced dimension. If the reduced data may be reconstructed via another linear transformation, then the new coordinate system may be viewed as encapsulating the entirety of the data. In this setting, every data point in $\mathcal{A}$ lies in the span of the encapsulating basis; if a component of a data point does not lie in this span, it appears as the residual, or error, when the data are reconstructed. For data that reside on a submanifold $\mathcal{M}$, the number of parameters required to encapsulate, or span, the data set is typically significantly larger than the number of dimensions required to parameterize it nonlinearly. Therefore, except in special cases, even an optimal linear parameterization (i.e., data encapsulation) will not produce a representation of the data that reflects the intrinsic dimensionality of $\mathcal{M}$.

This article presents a new method for determining nonlinear parameterizations of data sets, which are subsets of high-dimensional spaces. At the center of our approach is Whitney's theorem, which motivates the architecture of what we will refer to as the Whitney reduction network (WRN). In this architecture, the reduction mapping G is a projection designed to retain the (differential) structure of the data. In addition, this projection is empirically constructed to be good in the sense that its inverse will be well conditioned, a critical feature of networks with good generalization properties. A natural way to quantify the idea of a well conditioned nonlinear inverse is to require that H is Lipschitz with a small Lipschitz constant. Since G is also Lipschitz by virtue of being a projection, a useful consequence of this is that the topological dimension of $\mathcal{A}$ and its reduced form $\mathcal{B}$ are equal since bi-Lipschitz mappings are dimensionality preserving (see Falconer, 1990). Since the choice of G determines the Lipschitz constant of the inverse, we propose an adaptive basis algorithm that optimizes this. (Note that if the data set is not smooth but fractal, then it is necessary to replace the topological dimension by the Hausdorff dimension in the preceding discussion.)

In section 2 a discussion of Whitney's embedding theorem is presented; in section 3 the network architecture and implementation are presented; in section 4 the distinguishing features of the WRN are summarized; in section 5 we analyze the effects of noise on the procedure; in section 6 we present three illustrative examples of the methodology. Finally, in section 7, the main results are summarized, and some future work is outlined. More details concerning the mathematical theory of the reduction network proposed are presented in Broomhead and Kirby (2000).

2 Whitney's Embedding Theorem

This article is motivated by the Whitney embedding theorem, which shows that it is always possible to represent a compact, finite-dimensional, differentiable manifold as a submanifold of a vector space (Hirsch, 1976). Roughly speaking, given an m-dimensional differentiable manifold $\mathcal{M}$, we can find a mapping to the Euclidean space $\mathbb{R}^{2m+1}$, which is diffeomorphic onto its image. (A diffeomorphism is a differentiable map with a differentiable inverse.) In this sense it might be said that $\mathbb{R}^{2m+1}$ is large enough to contain a "diffeomorphic copy" of every m-dimensional differentiable manifold (Guillemin & Pollack, 1974).

The proof given in Hirsch (1976) has two stages. The first, and preliminary, stage shows that there is some sufficiently large q for which there exists an embedding in $\mathbb{R}^q$. So we concentrate on the situation where we have an m-dimensional submanifold, $\mathcal{M}$, of $\mathbb{R}^q$ and note that here q may be very large. The central idea of the rest of the proof is to show that there exists an "admissible" projection from $\mathbb{R}^q$ to $\mathbb{R}^{q-1}$ and then apply this argument recursively until we reach a value of q, beyond which the argument fails. Since we are dealing with compact manifolds, "admissible" here means that the projection is an injective immersion when restricted to $\mathcal{M}$. A projection that is an injective immersion has a smooth inverse when restricted to its image; it is this that is interesting to us since it suggests the possibility of a lossless data compression technique. This proof of the Whitney embedding theorem demonstrates that for each $q > 2m + 1$, there is an open dense set of such projections, which is to say that in a suitable topology, every neigborhood of a projection that is not admissible contains a projection that is (the dense part); and, moreover, every sufficiently small perturbation of an admissible projection is also admissible (the open part).

There is an appealing way to visualize these ideas. Imagine $\mathcal{M}$ as a submanifold of $\mathbb{R}^q$ and connect each pair of points in $\mathcal{M}$ with a straight line segment. We shall refer to these as the *secants* of $\mathcal{M}$. Now consider Σ, the set of unit vectors parallel to the secants—the *unit secants*. These can be thought of as points on the $q - 1$–dimensional unit sphere since this is the set of all unit vectors in $\mathbb{R}^q$. We can also associate projections from $\mathbb{R}^q$ to $\mathbb{R}^{q-1}$ with points on the $q - 1$–dimensional unit sphere by labeling each projection with the unit vector, which it maps to zero. Clearly, the set of unit secants constitutes the set of inadmissible projections in the sense that a projection that is also a unit secant must identify at least two distinct points in $\mathcal{M}$ and consequently is not invertible. The method of proof is then to show that the set of unit secants is nowhere dense in the $q - 1$–dimensional unit sphere provided that $q > 2m + 1$. A set that is nowhere dense has no subsets that are open, so that any point in a nowhere dense set must have neighboring points not in the set. This is enough to establish the existence of a projection that is an injection. A similar approach

is adopted to establish a similar result for the set of projections parallel to tangents of $\mathcal{M}$. These correspond to projections that are not immersions.

A subset that is nowhere dense is, from a topological point of view, a very small set. Although this remark should be treated with some caution—there are well-known examples of nowhere dense sets with positive measure—it suggests our approach to compression, since it says that an arbitrarily selected projection from $\mathbb{R}^q$ to $\mathbb{R}^{q-1}$ will have a smooth inverse whenever $q > 2m + 1$.

3 The WRN Architecture

The architecture of our network is driven by the decomposition of a data point $x \in \mathcal{A} \subset \mathcal{M}$ under the action of a projector $\mathbb{P}$,

$$x = \mathbb{P}x + (I - \mathbb{P})x, \tag{3.1}$$

where, by definition, $\mathbb{P}^2 = \mathbb{P}$. If we let $p = \mathbb{P}x$ and $q = \mathbb{Q}x$ (where $\mathbb{Q} = I - \mathbb{P}$), then we view any element x as being the sum of the portion of x in the range of $\mathbb{P}$, that is, $p \in \mathcal{R}(\mathbb{P})$ and the portion in the null space of $\mathbb{P}$, that is, $q \in \mathcal{N}(\mathbb{P})$. If the rank of $\mathbb{P}$ is d where $d > 2m$, then Whitney's theorem ensures the existence of a global map from the range of the projector to its null space, that is,

$$q = f(p).$$

This provides a parameterization of the data set $\mathcal{A}$ in terms of p as

$$x = p + f(p). \tag{3.2}$$

The inverse of $\mathbb{P}$ takes a projected data point $\mathbb{P}x \in \mathbb{P}\mathcal{A}$ and maps it back to x.

3.1 The Reduction Mapping. To begin, we need a good projector $\mathbb{P}$ of rank d. This will parameterize the data. The corresponding orthogonal projector $\mathbb{Q}$ gives the residual of the linear approximation and hence provides the target data for the nonlinear function approximation of f.

Given an appropriate orthonormal basis for $\mathbb{R}^n$, $U = [u_1|u_2|\cdots|u_n]$. The rank d projector is defined by $\mathbb{P} = \hat{U}_1 \hat{U}_1^{\,T}$ where the reduced $n \times d$ matrix $\hat{U}_1 = [u_1|, \ldots, |u_d]$. Similarly, the complementary projector is defined $\mathbb{Q} = \hat{U}_2 \hat{U}_2^{\,T}$ where $\hat{U}_2 = [u_{d+1}|, \ldots, |u_n]$. We note that the quantity $\mathbb{P}x \in \mathcal{R}(\mathbb{P})$ is an n-tuple in the ambient basis, that is, $\mathbb{P}x = (u_1^T x)u_1 + \cdots + (u_d^T x)u_d$. It is the expansion coefficients that provide the d-dimensional representation, and these are given as $\hat{p} = \hat{U}_1^T x \in \mathbb{R}^d$.

We eschew principal component analysis (PCA) because it is based on minimizing mean-square projection residuals rather than achieving a well-

conditioned inverse. Rather, we propose to optimize our projector $\mathbb{P}$ by requiring that the inequality

$$\|\mathbb{P}x - \mathbb{P}y\| \geq k^* \|x - y\| \tag{3.3}$$

be satisfied for all $x, y \in \mathcal{A}$ and some fixed *tolerance* $k^* > 0$. Note that this criterion is applied pointwise, whereas PCA is based on optimizing an average quantity. The tolerance k^* is a measure of the maximum permissible shortening of the distance between any two projected data points. Equivalently, as shown below, k^* is a lower bound on the norm of the projected secants. Note that by construction, $0 < k^* \leq 1$ with $k^* = 1$ when $\mathbb{P}$ represents a unitary transformation.

Thus, the goal of the first learning phase is to determine a basis that is good in the sense that the dimension d of the range of $\mathbb{P}$ is as small as possible for a given tolerance k^*. We employ the fact that given a projector $\mathbb{P}$ and a data set $\mathcal{A}$, the minimum projected interpoint distance k^* may be calculated directly from

$$\left\|\mathbb{P}\hat{k}\right\| = \frac{\|\mathbb{P}x - \mathbb{P}y\|}{\|x - y\|}, \tag{3.4}$$

where a unit secant is defined as $\hat{k} = (x - y)/\|x - y\|$ for any $x, y \in \mathcal{A}$; (see Broomhead & Kirby, 2000, for details). Given an n-dimensional basis, we may use the above formula for the minimum projected secant norm to determine the number of dimensions d required to satisfy the condition

$$\left\|\mathbb{P}\hat{k}\right\| \geq k^* \tag{3.5}$$

for all $\hat{k} \in \Sigma$, which is equivalent to the criterion of equation 3.3.

We now propose a procedure for determining a good projection in the sense of equation 3.3 or 3.5. Our approach is adaptive and is initialized using the principal components of the matrix of secants K. This is an $n \times N$ matrix where $N = P(P - 1)/2$ is the total number of secants and n is the ambient dimension (in practice, a subset of the secants is sufficient). We project the columns of K onto d-dimensional subspaces spanned by the principal components of K and look for the smallest value of d for which equation 3.5 is satisfied for all the secants. We denote the set of secants that do not satisfy equation 3.5 for $d - 1$ as "bad" secants. The matrix S whose columns consist of bad secants is used iteratively to update the initial covariance matrix $\Theta = KK^T$. The update involves reweighting the bad secants by an amount proportional to an adjustable parameter α:

$$\Theta' = \left(1 - \frac{\alpha}{N}\right)\Theta + \frac{\alpha}{m} SS^T, \tag{3.6}$$

where m is the number of columns in the matrix S. This iteration acts to reduce the number of dimensions d required such that $\|\mathbb{P}_d\hat{k}\| \geq k^*$ for all $\hat{k} \in \Sigma$. (Here, for clarity, we have indicated the dimension d of the projector as $\mathbb{P}_d$.)

Adaptive Basis Algorithm

1. Compute the initial basis via PCA of secants.
2. Determine the smallest dimension d such that $\|\mathbb{P}_d\hat{k}\| \geq k^*$ for all $\hat{k} \in \Sigma$.
3. Find the matrix S of bad secants defined by $\{\hat{k} \in \Sigma\colon \|\mathbb{P}_{d-1}\hat{k}\| < k^*\}$.
4. Update the covariance matrix via equation 3.6.
5. Compute the new basis consisting of the eigenvectors of Θ'.
6. Stop if basis satisfactory; else return to 2.

3.2 The Reconstruction Mapping. Given that the data have been parameterized via a good projection, the problem is now to determine a mapping for reconstructing the data. The need for an inverse mapping is in analogy with analytical transform methods. Our objective is the construction of an empirical dimension preserving transform of the data to facilitate its analysis. The existence of a (Lipschitz) inverse assures us that the projection has preserved the dimension of the data.

The reconstruction phase rebuilds a data point $x \in \mathbb{R}^n$ from its projection $p = \mathbb{P}x$ by learning the associated value $q = \mathbb{Q}x$, that is, the projection onto the orthogonal complement. The superposition of these values then rebuilds a data point; the identity mapping is $I = \mathbb{P} + \mathbb{Q}$ or

$$x = \mathbb{P}x + \mathbb{Q}x.$$

It is most natural to do these computations in the appropriate bases; the representations

$$\hat{p} = \hat{U}_1^T x, \quad \hat{q} = \hat{U}_2^T x$$

have dimensions d and $n - d$, respectively.

The (strictly) nonlinear portion of the reconstruction involves fitting a function f with the projected data set $\{\hat{p}\}$ as the input and the orthogonal complement data set $\{\hat{q}\}$ as the output. (It is this function f that is guaranteed to exist by Whitney's theorem.) This mapping f is approximated by $\tilde{f}$, providing an estimate $\tilde{\hat{q}}$ for $\hat{q}$:

$$\hat{p} \to \tilde{\hat{q}} = \tilde{f}\left(\hat{p}\right).$$

During the training phase, the target points $\{\hat{q}\}$ are required for the second internal layer. The dotted connections from the input layer to the second

internal layer represent the mapping

$$x \to \hat{q} = \hat{U}_2^T x \in \mathbb{R}^{n-d}.$$

This portion of the network is required for training purposes only.

Finally the data are reconstructed (i.e., transformed to the original ambient coordinates) in two parallel stages $x = p + q$. A mapping from the first internal layer to the output layer accomplishes the strictly linear reconstruction:

$$p = \hat{U}_1 \hat{p}.$$

This mapping is not learned by the network but comes from the linear inversion of the projection. The nonlinear component of the reconstruction is the result of the parameterization, which produces an approximation to $\hat{q}$, that is, $\tilde{\hat{q}} = \tilde{f}(\hat{p})$. Thus, the desired reconstruction $q = \hat{U}_2 \hat{q}$ is now approximated by a mapping from the second internal layer to the output layer:

$$\tilde{q} = \hat{U}_2 \tilde{\hat{q}}.$$

Hence, x is approximated as $x \approx \tilde{x} = p + \tilde{q}$. In summary, the network approximates the identity mapping as

$$\tilde{\mathbb{P}}^{-1} \circ \mathbb{P} \colon x \to \tilde{x},$$

where this mapping may be written

$$\tilde{x} = \hat{U}_1 \hat{p} + \hat{U}_2 \tilde{f}\left(\hat{p}\right). \tag{3.7}$$

The (well-conditioned) inverse $\hat{q} = f(\hat{p})$ may be fit using the radial basis function approach (Broomhead & Lowe, 1988). Here we employed the thin plate spline radial basis function (RBF) $\phi(r) = r^2 \ln r$ with randomly selected centers. The least-squares problem was solved using the singular value decomposition.

4 Features of the Whitney Reduction Network

Various features of the WRN that distinguish it from other techniques, such as linear and nonlinear principal components.

4.1 Decomposition of Reconstruction into Linear and Nonlinear Parts. The reconstruction of the data achieved via equation 3.2 is a decomposition into separate linear and nonlinear pieces. The term p is the linear reconstruction, while $q = f(p)$ is the nonlinear term fit by RBFs. Note that p is immediately available once a basis has been selected for projection since it is just the linear reconstruction of the projected data. In other words, it is

not necessary to use a neural algorithm to learn the weights associated with the linear component of the reconstruction.

Another interpretation of this decomposition is that the WRN automatically identifies dependent and independent variables in the function approximation process. This approach is distinct from nonlinear (as well as linear) PCA, which uses the original data as the target for the reconstruction. Note also that in nonlinear PCA, the bottleneck variables may change during the course of training the inverse, while the independent variables produced by the WRN are found and then fixed before the nonlinear inverse is approximated.

4.2 Projection Optimized for Producing Good Generalization. A mapping is said to be ill conditioned if small perturbations in the domain lead to large changes in the range. If an ill conditioned map is approximated (say, using RBFs or multilayer perceptrons) on a fixed training set, it will not generalize well to nontraining data.

The inverse mapping of the WRN is well conditioned as a result of the requirement that the projection $\mathbb{P}$ satisfy the optimality criterion given by equation 3.3, from which it follows that

$$\frac{\|\mathbb{P}^{-1}x - \mathbb{P}^{-1}y\|}{\|x - y\|} \leq \frac{1}{k^*}; \tag{4.1}$$

that is, $\mathbb{P}^{-1}$ is Lipschitz with constant $1/k^*$. Since the left-hand side of equation 4.1 is exactly that ratio of the change in the image to the associated change in the domain, we conclude that the absolute condition number $\hat{\kappa}$ (this is defined in Trefethen & David Bau, 1997) is bounded by our k^* as

$$\hat{\kappa} \leq \frac{1}{k^*}. \tag{4.2}$$

Hence by designing the projection such that equation 3.3 is satisfied for a reasonably large k^* the inverse mapping will be well conditioned. In addition it can be shown (see Broomhead & Kirby, 2000) that the RBF approximation is also Lipschitz. Neither standard nor nonlinear PCA constrains the inverse mappings to be well conditioned. Indeed, our examples show that the measure of conditioning k^* attained by the data PCA basis can be significantly inferior to that for secant bases.

4.3 Linear Optimization Problem. We fit the nonlinear portion of the inverse mapping using RBFs since this leads to a linear optimization problem for determining the unique weights. By contrast, the model parameters of the bottleneck network require nonlinear optimization methods, which are inherently nonunique.

4.4 Estimate for Dimension of Data Set. Our approach provides an objective means for estimating the dimension of the manifold from which the data are sampled. Standard PCA measures the distribution of data in a subspace of the ambient space and is thus an unreliable estimate of the manifold dimension. Also, the bottleneck, or autoassociative, network architecture for implementing nonlinear PCA does not provide a direct means to estimate the dimension of a data set.

4.5 Estimate for Complexity of the Approximation. The number of parameters required by the network may be estimated as a function of k^*. To this end, it can be shown (see Broomhead & Kirby, 2000) that

$$\nu \stackrel{\text{def}}{=} \sqrt{1 + k_\phi^2 \|W\|^2 (n_c + 1)} > \frac{1}{k^*},$$

where k_ϕ is the Lipschitz constant of ϕ, n_c is the number of RBF centers, and $\|W\|$ is the two-norm of the weight matrix W. This quantity ν is a characteristic of the RBF and may be a useful measure of the complexity of the network.

4.6 Application to Problems of Very High Dimension. Given that the reduction mapping $\mathbb{P}$ is linear and the nonlinear component of the inverse is approximated when $\mathbb{P}$ has been fixed, it follows that the size of ambient dimension that this network can reduce is far greater than fully nonlinear bottleneck networks. In addition, the nonlinear mapping has a d-dimensional domain and an $r - d$ dimensional range where r is the rank of the data matrix.

5 Projecting Noisy Data

Depending on the application, it can be the case that the data set to be compressed has been corrupted by noise. Noise that perturbs data points so that they are no longer on their manifold leads to uncertainty in the directions of the secants. As a result, direct application of the methodology to data with noise needs some care. Here we propose an approach for dealing with noise that amounts to filtering the smallest secants, the ones most affected by the noise.

The motivation for this approach is made clear by analyzing the effect of noise on a given secant. If we assume that the noise is isotropic with compact support (as would be the case with quantization noise, for example), then each point may be associated with a hypersphere of radius ϵ, which defines the region within which the point with the noise added is assumed to lie. Joining two points, each of which has added noise, forms a set of possible secants consisting of all lines between the hyperspheres associated with each point. For example, in two dimensions, the secant may now be visualized as being generated by a "dumbell" configuration, as shown in Figure 1.

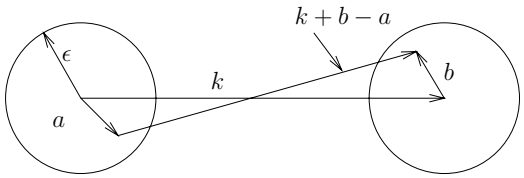


Figure 1: Let k be the secant between two data points. If the points have uniformly distributed added noise, they may be displaced in the directions a and b, respectively. The new vector $k+a-b$ represents the noise-perturbed secant. The points a, b may be any points in the corresponding hyperspheres, drawn here in two dimensions.

To determine the potentially destructive effect of the noise on our good projection, we estimate the maximum angle between possible secants in the dumbbell. Writing $\alpha = b - a$, we have the unit secant $(k+\alpha)/\|k+\alpha\|$. The largest possible angle between two secants would arise if $\alpha = 2\epsilon\hat{\epsilon}$ and $\alpha' = -2\epsilon\hat{\epsilon}$, as shown in Figure 2; here $\hat{\epsilon}$ is a unit vector perpendicular to k. From elementary trigonometry, it follows that

$$\sin\theta = \frac{4\epsilon\|k\|}{\|k\|^2 + 4\epsilon^2}. \tag{5.1}$$

For the small noise case we may assume $\epsilon \ll \|k\|$ and hence θ is near zero; there is essentially no uncertainty concerning the true direction of the secant. As the magnitude of the noise increases (i.e., the radius of the hypersphere about each data point gets larger), we see that the maximum angle between potential secants increases (see Figure 3). We conclude from equation 5.1

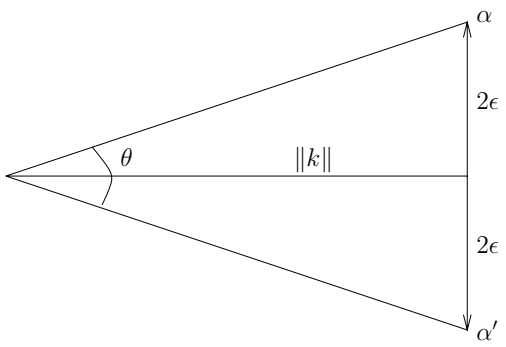


Figure 2: Maximum perturbation of the secant results may be represented geometrically as a triangle where the added noise vectors α and α' are in opposite directions.

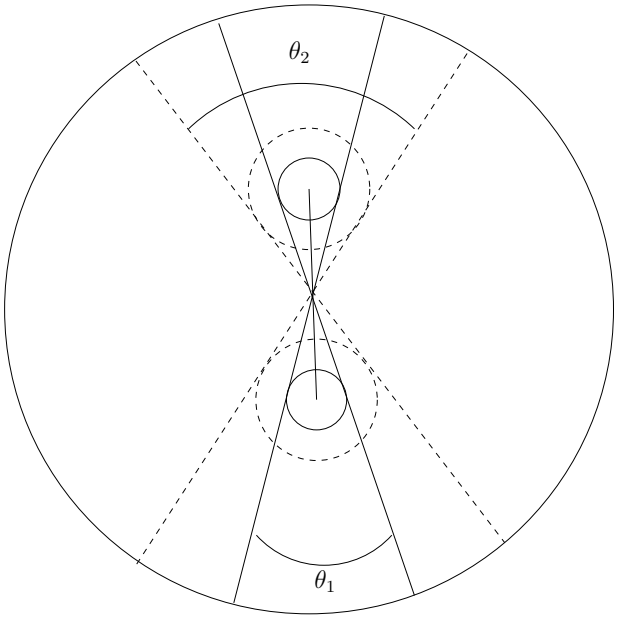


Figure 3: The two pairs of concentric circles represent differing levels of noise for two data points. The smaller noise level, represented by the solid lines, produces a maximum angle θ_1 between extreme secants. The dotted lines correspond to a larger noise level and show a greater uncertainty for the true direction of the secant. Here the noisy secant could differ from the actual by an angle of θ_2.

that in the worst case, that is, when $\|k\| = 2\epsilon$, the secants could actually be perpendicular.

It is apparent from the above arguments that the orientation of the shortest secants generated by the data is most sensitive to additive noise. Thus, we propose a filtering procedure, which amounts to removing the secants shorter than a cutoff value from the set of secants Σ. One might anticipate that as long as it is large enough, the actual value of the cutoff parameter is not critical given our goal is a good projection rather than an optimal one.

6 Illustrative Examples

6.1 Data on the Boundary of a Pringle. We begin with a simple, easily visualizable problem for which the PCA basis produces a bad (not one-to-one) projection of the data set. The boundary of a *pringle*, as shown in the left of Figure 4, is an embedding of the circle in $\mathbb{R}^3$ and is defined as the triplet $(\sin\theta, \cos\theta, \sin 2\theta)$. It may be argued analytically that the projection onto the two most energetic secant PCA basis vectors produces the circle in

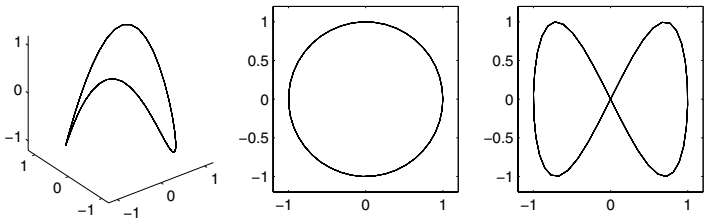


Figure 4: (Left) Example of a one-dimensional manifold embedded in $\mathbb{R}^3$. (Middle) Projection on the best two secant basis vectors, along the z-axis. (Right) Projection onto the best two data PCA vectors, along the y-axis.

the middle of the figure, while the projection onto the two most energetic data PCA vectors produces the lemniscate on the right (Broomhead & Kirby, 2000). Although the PCA projection is not invertible, the secant basis gives an embedding of the pringle in two dimensions.

Now we add uniform noise with compact support—points from the interval $[-0.1, 0.1]$, to the pringle (see top left of Figure 5). The top-right figure shows the set of unit secants for the pringle data with no added noise. These are confined to four segments, the complement of which—the region where permissible projections can be found—consists of two four-fold stars centered at the north and south poles. The bottom-left figure shows the set of unit secants for the noisy pringle data. The effect of adding noise is to spread the unit secants over the whole sphere, filling in the admissible regions for projections in the neighborhood of the poles. Our analysis in section 5 suggests that these troublesome unit secants are derived from secants with small norm. We filter the data by removing all secants with length smaller than 1.8. This number was determined empirically to ensure that the effect—the north and south poles becoming significantly devoid of secants—was visible in Figure 5. The principal component basis constructed from the filtered unit secants produced an admissible projection in the polar region that was skewed away from the pole due to nonuniformity in the distribution of the unit secants.

Next we applied the adaptive secant basis algorithm to determine an improved projection for this data. The adaptation procedure rotated this direction until it was essentially colinear with the z-axis. In this case, the adaptive algorithm actually found the optimum projection.

6.2 Case Study: A Noisy Circle in 32 Dimensions. In this section we present the results of applying the WRN to a synthetic data set consisting of a narrow pulse moving with unit velocity on a circle: $g(x - t)$ where $g(\theta) = g(\theta + 2\pi)$. In the absence of noise, these data are topologically a circle—that is, a one-dimensional manifold—in a high-dimensional function space. If

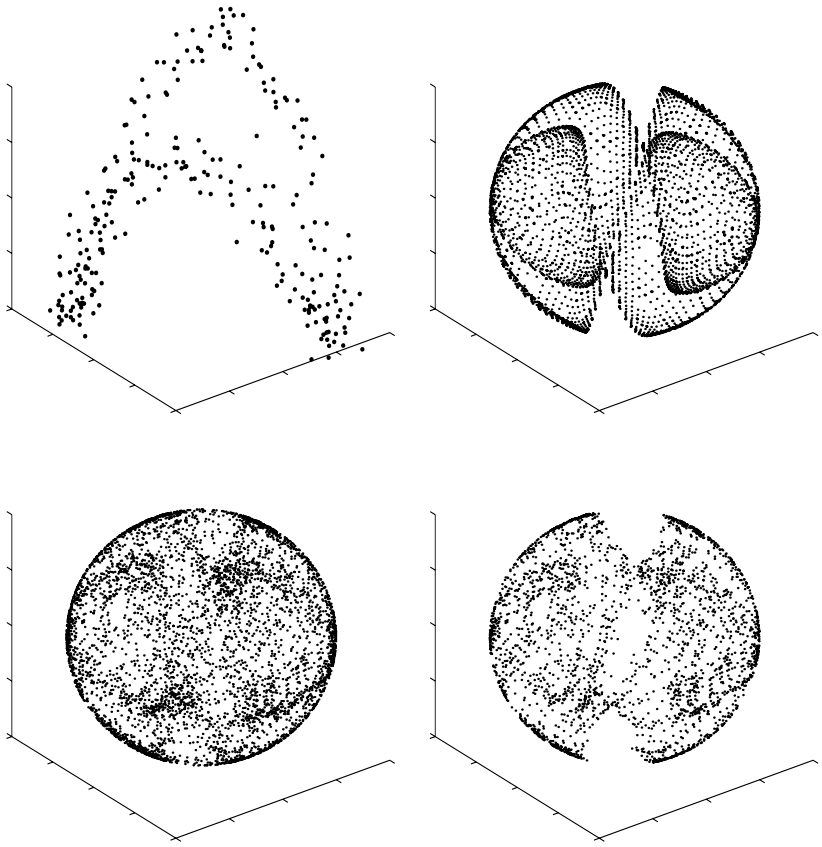


Figure 5: (Top left) Uniform noise on interval $[-.1, .1]$ added to curve. (Top right) Unit secant set for noise-free data. (Bottom left) Associated secants for curve with uniform noise added. (Bottom right) Filtered secant set.

we approximate g as a traveling gaussian pulse and add a noise term, we get a set of functions whose space-time translational symmetry has been broken:

$$g(x, t) = e^{-(t-x)^2/\gamma} + \eta(x, t).$$

The noise term $\eta(x, t)$ has been added to test the robustness of the WRN for data that do not reside exactly on a manifold. For this purpose, two data sets (labeled I and II) were generated by adding normally distributed noise with variances 0.025 and 0.05 (and zero mean) to the traveling pulse (see Figure 6). Taking $\gamma = 25$, the function was sampled at 32 points in the x-direction at half-integer intervals and at 128 points in the t-direction at integer intervals. The resulting data matrices have size 32×128. Thus,

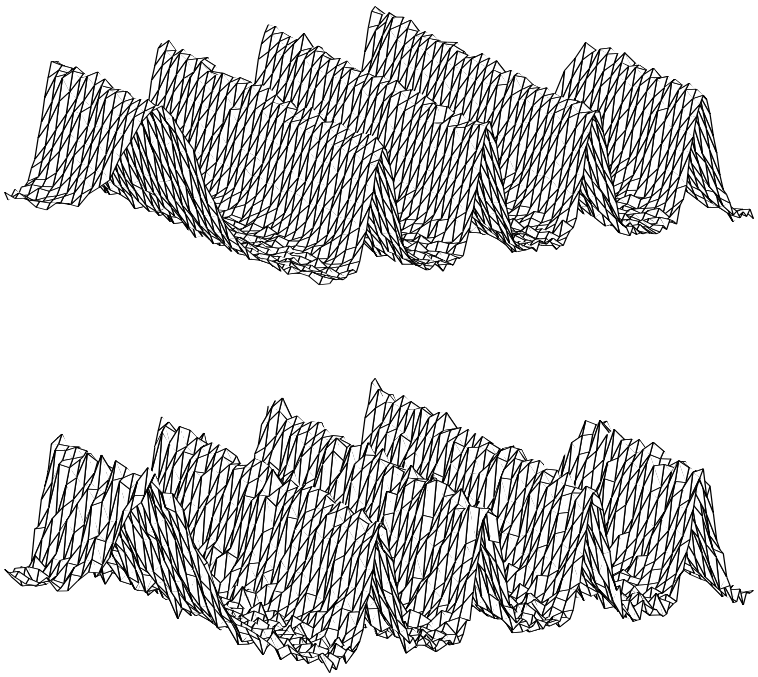


Figure 6: (Top) Right traveling wave data corrupted with normally distributed noise with zero mean and variance 0.025. (Bottom) Right traveling wave data corrupted with normally distributed noise with zero mean and variance 0.05.

the ambient space for the data is $\mathbb{R}^{32}$, and the data matrix has full rank $n = 32$, regardless of the level of the added noise. The noisy data sit only approximately on a one-dimensional manifold ($m = 1$), and we propose to determine the actual number of dimensions required in the WRN for reconstruction to within the level of the noise.[1]

6.2.1 Learning Phase I: Finding a Good Basis. The application of PCA to translationally invariant data produces sinusoidal eigenvectors. In addition, it has been shown that the principal components of the secants of translationally invariant data are also sinusoidal (Broomhead & Kirby, 2000). Hence, for translationally invariant data, the bases produced by PCA on the data and the secants are identical.

[1] In the absence of noise, these data reside on a one-dimensional manifold ($m = 1$), which may be reduced to $\mathbb{R}^3$ (and possibly $\mathbb{R}^2$, although now we cannot expect that suitable projections are open dense) and reconstructed without loss by Whitney's theorem.

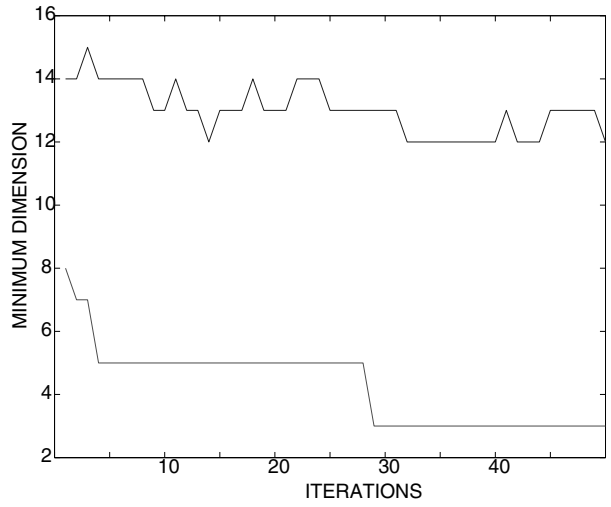
When noise is added to translationally invariant data, the result is data that are no longer translationally invariant. As a result, the adaptive secant algorithm now produces basis vectors that differ significantly from the sinusoids. (Note that the unadapted secant basis is still essentially sinusoidal until mode 7, and the PCA on the data produces similar eigenvectors.)

To compare the results of the adapted secant basis algorithm, with minimum secant norm filtering and without, see Figure 7a. This plot shows the smallest dimension for which the minimum norm of the projected secants is above the selected tolerance $k^* = 0.5$ as a function of the adaptation iteration. The secant basis was adapted with no secant filtering (top curve) and with a secant cutoff of 0.25 (bottom curve) for data set I. The adapted secant algorithm with filtering reduces the reduction dimension from 14 to 3. Without filtering, the adaptation reduces the required dimension only to 12. The results for data set II (not shown) show a similar improvement through adaptation. The filtering of the secants produced a 5-dimensional basis, while the unfiltered data set produced a 12-dimensional adapted basis. Thus, while added gaussian noise appears to increase the reduction dimension, the proposed secant filtering procedure does greatly mitigate this problem. Indeed, the 3-dimensional reduction suggested by Whitney's is obtained with tolerance to $k^* = 0.3$.

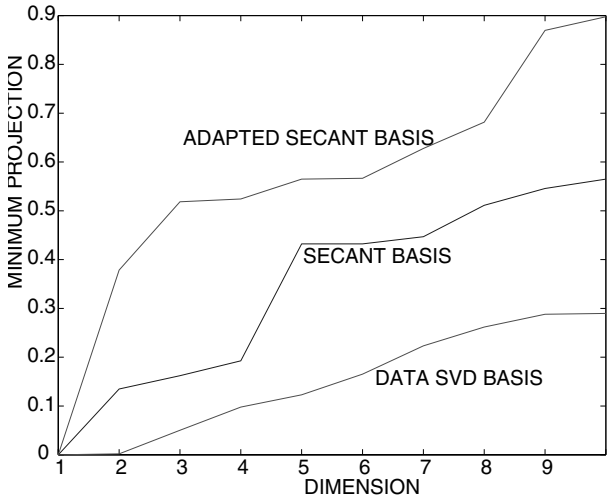
The performances of the PCA data basis, the PCA secant basis, and the adapted PCA secant bases are compared in Figure 7b. The adapted secant basis is clearly superior at maximizing the minimum projected norm. Thus, the projection onto this basis will produce the best-conditioned inverse. The basis produced by the PCA of the data actually will lead to a seriously ill-conditioned inverse with poor generalization capabilities.

To examine geometrically the difference between the adapted secant basis and the PCA data (and secant) basis of Fourier vectors, we have plotted the time evolution of the first four one-dimensional subspaces, or space-time modes, for the adapted secant basis. Specifically, Figure 8 shows the discrete space-time modes obtained by projecting the data onto the one-dimensional subspaces associated with the most important adapted secant basis vectors. These adapted space-time modes are seen to have markedly different profiles from those obtained with PCA on the data or secants (i.e., Fourier modes).

6.2.2 Learning Phase II: The Well-Conditioned Inverse. In this section, we examine the results of data reconstructions for several different projection dimensions and bases. According to equation 3.7, the reconstruction of the data consists of the sum of two components: the linear inverse, or demapping, $p = \hat{U}_1\hat{p}$, which is accomplished by reverting the projected data back to their original coordinates, and the nonlinear inverse, which approximates q by the RBF as $\tilde{q} = \hat{U}_2\tilde{f}(\hat{p})$. The results presented here serve to illustrate this linear-nonlinear decomposition, as well as to highlight the difference



(a)



(b)

Figure 7: (a) Minimum projection dimension that achieves tolerance $k^* = 0.5$ as a function of adaptation iteration for data set I. (b) Minimum norm of the projected unit secants of data set I as a function of dimension for several bases, For secant bases, a minimum secant cutoff = 0.25 was used. The results were similar for data set II.

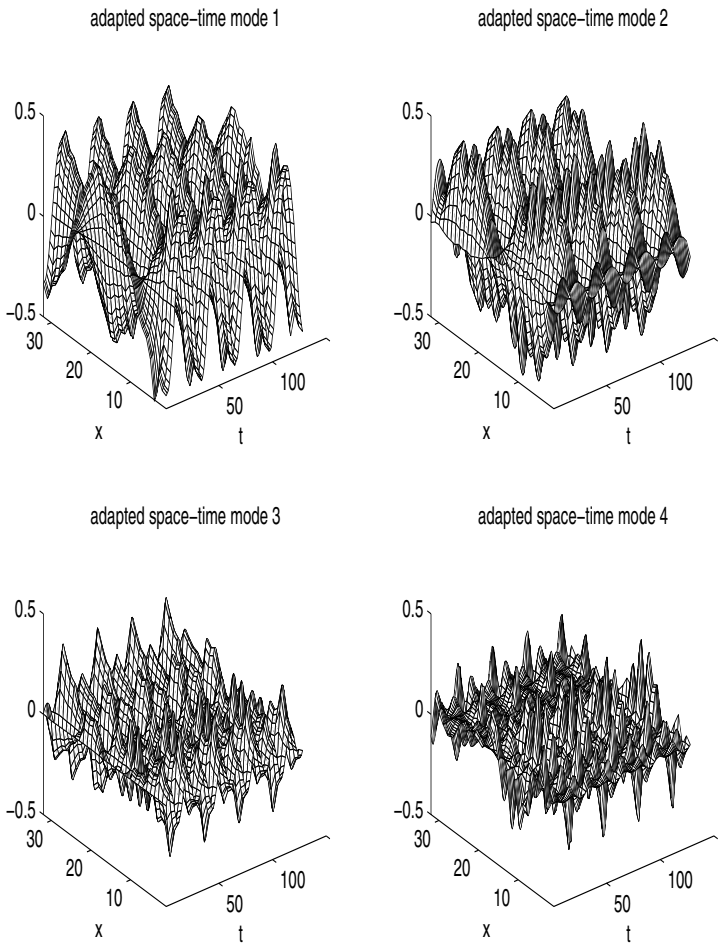


Figure 8: Projection of the right traveling wave onto the one-dimensional subspaces of the first four adapted secant singular vectors.

between the parameterization and encapsulation of the data set. The examples that follow pertain to data set I, as described above.

Adapted secant PCA basis $d = 2$. This example employs the adapted secant basis for which the minimum projected secant norm onto two dimensions is approximately 0.4. Thus, by construction, the dimensionality of the data will be preserved and the inverse mapping will be well conditioned with a Lipschitz contant less than 2.5. The linear reconstruction term $p = \hat{U}_1 \hat{p}$ is displayed at the top of Figure 9. Given the large deviation from the original

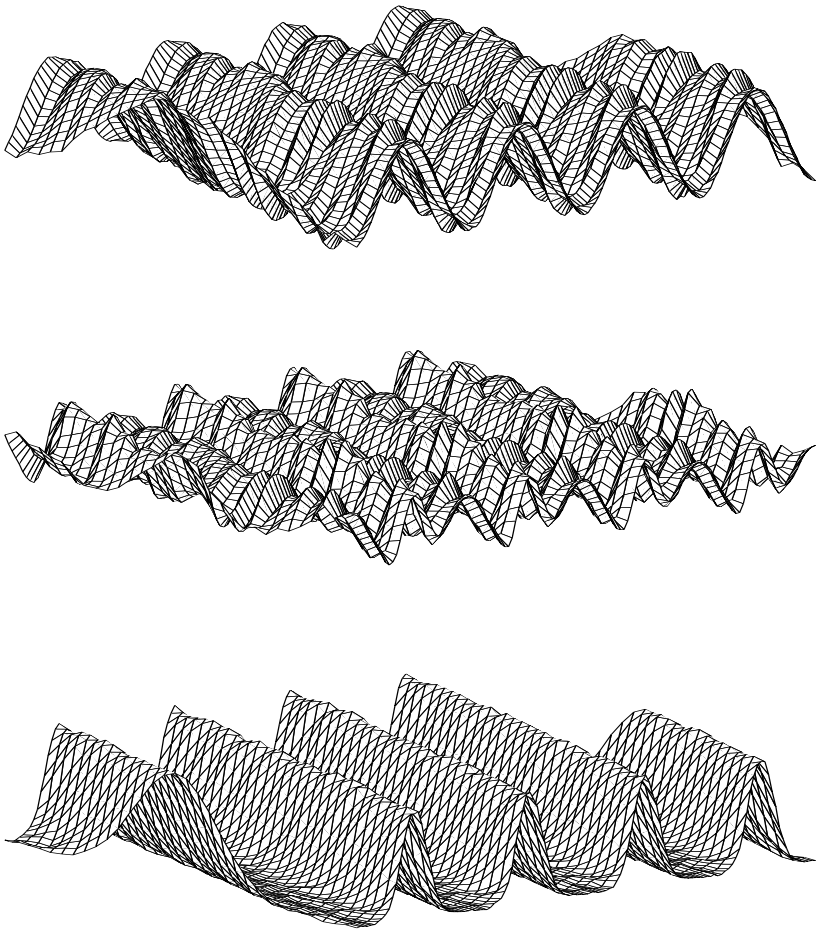


Figure 9: (Top) Two-dimensional linear demapping of the right traveling wave. Variance 0.025, secant basis with cutoff $= 0.025$. (Middle) Two-dimensional nonlinear demapping (only) of the right traveling wave. (Bottom) Two-dimensional full reconstruction of the right traveling wave, that is, the superposition of the linear and nonlinear demapping layers.

data, we see that two dimensions will not span, or encapsulate, the data without significant residual. The nonlinear component of the inverse—the RBF approximation to q defined as $\tilde{q} = \hat{U}_2 \tilde{f}(\hat{p})$—is shown in the middle of Figure 9. Here the RBF is a mapping from $\mathbb{P}X \in \mathbb{R}^2 \rightarrow \mathbb{R}^{30}$, so the nonlinear component of the reconstruction is not to the full original ambient space, but rather to the smaller null space of the projector $\mathbb{P}$. The full re-

construction $\tilde{x} = p + \tilde{q}$ is shown at the bottom of Figure 9, and the relative error $\|x - \tilde{x}\|/\|x\|$ was approximately 0.03, suggesting the data have been reconstructed roughly to the noise level. In fact, for data set I, $\|\eta\|/\|g\|$ was approximately 0.024. No attempts were made here to optimize the location of the centers in the RBF reconstruction. In fact, 12 centers were simply selected at random from the data.

Data PCA data basis $d = 3$. Our second example employs a basis produced by the PCA of the data matrix. Retaining three dimensions results in a linear reconstruction with a 38% relative error. However, the minimum projected secant norm onto this basis is approximately 0.04 (see Figure 7b) and significantly worse than for the adapted secant basis. This relatively poor conditioning manifested itself in the approximation of the nonlinear term, which was much more difficult to compute than for the adapted secant basis, and considerable effort was needed to achieve an error comparable to that obtained above.

6.3 The Rogues Gallery Problem. As another illustrative example, we consider the application of the WRN to the Rogues Gallery problem, the low-dimensional characterization of snapshots of human faces. The application of the Karhunen-Loève (KL) procedure, or data PCA, for the representation of digital images of faces was introduced in Sirovich and Kirby (1987). Since this time, this basic work has been extended in several directions, for example, the construction of symmetric eigenpictures (Kirby & Sirovich, 1990), three-dimensional eigenfaces (Atick, Griffin, & Redlich, 1996), and eigenpictures with incomplete data (Everson & Sirovich, 1995). The implicit assumption in these investigations is that the data reside in a subspace of significantly lower dimension than the ambient space. The purpose of (data) PCA is then to find an optimal set of spanning eigenvectors, or eigenpictures, to represent the data such that the mean square error is a minimum.

It is possible, however, that the points associated with a family of images lie on a submanifold rather than a subspace. In this situation, the representation of the data may be achieved via the WRN. Again, the goal is to construct a nonlinear parameterization for the images by computing a graph of the surface on which the data lie. To test the effectiveness of the Whitney architecture to the problem of the representation of high-dimensional images, we applied the procedure to an ensemble of 200 faces. Note that since we seek to represent the faces as the graph of a function, our approach is fundamentally different from the fully nonlinear approach in Cottrell and Metcalfe (1993).

The images were normalized for lighting as in Sirovich and Kirby (1987) and Kirby and Sirovich (1990), and the background was eliminated partially by constructing 310×380 pixel cameos. In Figure 10, we have shown the results, in terms of the original image coordinate system, of the linear, nonlinear, and full reconstruction, and compare them to the original

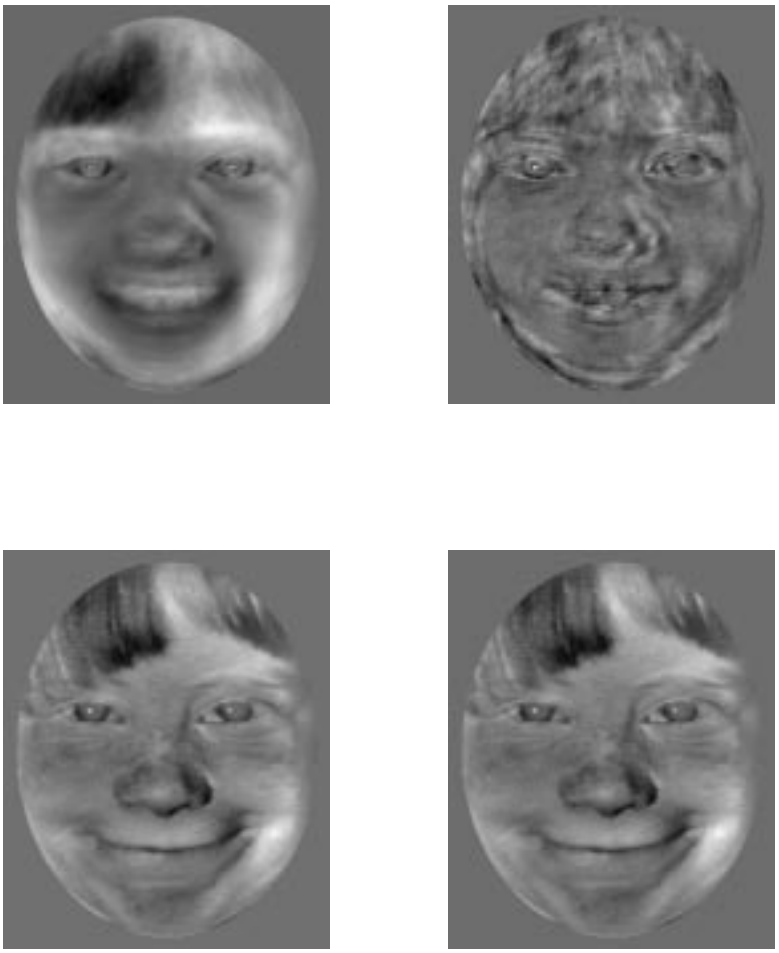


Figure 10: (Top left) Linear reconstruction p using a 10-dimensional secant basis projection. (Top right) Nonlinear RBF reconstruction $\tilde{q}$ from the 10-dimensional secant subspace to its 190-dimensional orthogonal complement. (Bottom left) Total reconstruction $\tilde{x} = p + \tilde{q}$. (Bottom right) Original mean-subtracted image. The data in this example came from the Vail schoolchildren data set (provided by Walter Bender of MIT). This figure origininally appeared in Kirby (2001) and is reprinted with permission.

image. Here the data are all mean subtracted. It is clear that a 10-dimensional projection onto the secant basis when linearly reconstructed is visually a poor approximation to the original data (this statement remains true for the projection onto the first 10 KL eigenpictures—data PCA). However, by

virtue of the fact that we have imposed the constraint equation 3.3 on the projection, we are assured of a lossless reconstruction. Indeed, the nonlinear mapping, again constructed using RBFs, provides the detail as the image of this projection (see the top right of Figure 10). The success of the full reconstruction confirms that the projection is indeed invertible (see the bottom left of Figure 10). The secant PCA basis performs only slightly better than the data PCA basis when compared in terms of the minimum projected secant norm criterion. The minimum projected secant norm for a 10-dimensional projection was 0.30 for the secant PCA basis and and 0.18 for the data PCA basis. These numbers indicate that 10 is a reasonable number of dimensions for the projection, as invertibility is guaranteed for the entire data set. The similar performance of these two bases is very likely due to the fact that such a small data set (200 images) was used. A 10-dimensional subspace is very large for just 200 points, and thus it is easy to find a projection that is invertible. Significantly more data are required to compare the performance of the different methods as well as to establish firmly whether the face data actually lie on a surface. Nonetheless, this example successfully illustrates the goal of obtaining an invertible projection in a high-dimensional setting.

7 Conclusion

We propose the WRN as a method for data reduction especially suited to problems where the data are sampled from an m-dimensional manifold residing in an high-dimensional ambient space. Theoretical insight is provided by Whitney's theorem, and the architecture is motivated by its constructive proof. A key idea is that a linear reduction mapping coupled with a nonlinear inverse is mathematically all that is required to compress the data on a manifold of dimension m to dimension $2m + 1$, although the ambient dimension may be very large.

In addition, a projection that optimizes the conditioning of the inverse mapping is proposed. This approach ensures good generalization properties of the inverse. The examples demonstrate that the method is effective for globally representing data on a manifold, even in the presence of noise. The preliminary results on digital images of faces indicate that this architecture scales well to problems of higher dimension but that, as expected, correspondingly large amounts of data are required.

Acknowledgments

This research has been supported in part by the Engineering and Physical Sciences Research Council, U.K., in the form of a visiting research fellowship to M. K. This research has also been supported by the National Science Foundation under grants DMS-9505863 and INT-9513880. We thank Jerry Huke, Doug Hundley, and Rick Miranda for their input concerning this

research and Walter Bender for providing the raw data used to compute Figure 10.

References

Atick, J. J., Griffin, P. A., & Redlich, A. N. (1996). The vocabulary of shape: Principal shapes for probing perception and neural response. *Network: Computation in Neural Systems, 7*, 1–5.

Broomhead, D., & Kirby, M. (2000). New approach for dimensionality reduction: Theory and algorithms. *SIAM J. of Applied Mathematics, 60*(6), 2114–2142.

Broomhead, D., & Lowe, D. (1988). Multivariable functional interpolation and adaptive networks. *Complex Systems, 2*, 321–355.

Cottrell, G. W., & Metcalfe, J. (1993). Empath: Face, emotion and gender recognition using holons. In R. Lippman, J. Moody, & D. Touretzky, (Eds.), *Advances in neural information processing systems, 3* (pp. 564–571). San Mateo, CA: Morgan Kaufmann.

Everson, R. M., & Sirovich, L. (1995). The Karhunen-Loève transform for incomplete data. *Journal of the Optical Society of America, A, 12*(8), 1657–1664.

Falconer, K. (1990). *Fractal geometry: Mathematical foundations and applications*. New York: Wiley.

Guillemin, V., & Pollack, A. (1974). *Differential topology*. Englewood Cliffs, NJ: Prentice Hall.

Hirsch, M. W. (1976). *Differential topology*. Berlin: Springer-Verlag.

Kirby, M. (2001). *Geometric data analysis*. New York: Wiley.

Kirby, M., & Sirovich, L. (1990). Application of the Karhunen-Loève procedure for the characterization of human faces. *IEEE Trans. PAMI, 12*(1), 103–108.

Sirovich, L., & Kirby, M. (1987). A low-dimensional procedure for the characterization of human faces. *J. of the Optical Society of America A, 4*, 529–524.

Trefethen, L. N., & David Bau, I. (1997). *Numerical linear algebra*. Philadelphia: SIAM.

Received July 14, 1998; accepted January 17, 2001.