SIAM J. Appl. Math.
Vol. 60, No. 6, pp. 2114–2142

A NEW APPROACH TO DIMENSIONALITY REDUCTION: THEORY AND ALGORITHMS[*]

D. S. BROOMHEAD[†] AND M. KIRBY[‡]

Abstract. This paper applies Whitney's embedding theorem to the data reduction problem and introduces a new approach motivated in part by the (constructive) proof of the theorem. The notion of a good projection is introduced which involves picking projections of the high-dimensional system that are optimized such that they are easy to invert. The basic theory of the approach is outlined and algorithms for finding the projections are presented and applied to several test cases. A method for constructing the inverse projection is detailed and its properties, including a new measure of complexity, are discussed. Finally, well-known methods of data reduction are compared with our approach within the context of Whitney's theorem.

Key words. dimensionality reduction, radial basis functions, Whitney's embedding theorem, secant basis

AMS subject classifications. 53-04, 62H99, 65D99, 65C99

PII. S0036139998338583

1. The data reduction problem. Assume that the data set of interest, $\mathcal{A}$, is a discrete sampling of $\mathcal{U}$, a compact m-dimensional submanifold of an ambient vector space $\mathbb{R}^q$, where minimally $q > 2m+1$ but it is generally much larger. Given that the q-dimensional ambient space is redundant and an overparameterization of the data, the aim of reduction is to determine a new parameterization which more closely reflects the intrinsic dimension of the data. Our basic approach is to find a mapping $G : \mathcal{A} \to \mathcal{B}$ such that $\mathcal{B} \subset \mathcal{V}$ and $\mathcal{V}$ is a submanifold of $\mathbb{R}^p$ with $p < q$. Thus, we seek a diffeomorphism $G : \mathcal{U} \to \mathcal{V}$ which is an embedding of $\mathcal{U}$ in $\mathbb{R}^p$. In practice the mapping G will be determined by the data and specifically will be chosen such that the inverse mapping H is especially *well-conditioned*. The resulting composition of mappings $H \circ G$ should closely approximate the identity mapping. The ability to construct an inverse H is central to our approach in that it assures us that no data have been lost in the procedure and that our embedded manifold $\mathcal{V}$ possesses all of the information contained in $\mathcal{U}$.

The approach to the reduction problem presented in this paper is motivated by the (easy) Whitney embedding theorem [9]. This theorem demonstrates that generically[1] a projection of any m-dimensional manifold is invertible provided the dimension of the range of the projection is no smaller that $2m+1$. Given this flexibility, we propose to examine the issue of how to determine especially good projections. A projection is acceptable according to Whitney if it is not along the direction of a secant connecting any pair of points on the manifold. We will consider a projection to be *good* if it is far

[*] Received by the editors May 13, 1998; accepted for publication (in revised form) December 21, 1999; published electronically June 15, 2000. Partial support for this research was provided by National Science Foundation travel grant INT-9513880 and AFOSR grant F49620-99-1-0034.
http://www.siam.org/journals/siap/60-6/33858.html

[†] Department of Mathematics, UMIST, P.O. Box 88, Manchester M60 1QD, U.K. (D.S.Broomhead@umist.ac.uk).

[‡] Department of Mathematics, Colorado State University, Fort Collins, CO 80523 (kirby@math.colostate.edu). Additional support for this research was provided to this author by the Engineering and Physical Sciences Research Council, United Kingdom.

[1] That is, there is a set which is open and dense in the set of all projections for which the following is true.

away from the set of all unit secants. Specifically, a good projection is based on the idea that distinct points in the ambient space should not be mapped close together in the reduced space. It is this quantity, i.e., the minimum distance between the reduced data points, which we propose to optimize.

Determining good projections, as described above, has several significant benefits. For example, optimizing the projection in this fashion can be shown to directly improve the absolute condition number of the (nonlinear) inverse to the projection. Consequently, the number of basis functions required to approximate this inverse is significantly reduced and the accuracy of interpolation improved. In this investigation we present both theoretical and numerical results which strongly suggest that the search for good projections is both purposeful and feasible.

This paper is organized as follows. In section 2 we describe the data sets used to test the theory and algorithms proposed in the paper. In section 3 we present the main motivating theorem and the basic idea of a good projection. In section 4 we propose a basis derived from the SVD on the unit secant data as a candidate for constructing a good orthogonal projector. In addition, an iterative method for improving upon this basis is proposed and tested. In section 5 the radial basis function (RBF) approximation procedure is outlined and applied to the reconstruction problem; some interesting properties of this inverse are also discussed. In section 6 we present a general framework for the reduction problem and compare other standard methods with the approach proposed in this paper. Lastly, in section 7, we conclude with a summary of our results.

2. Test applications.

2.1. The Kuramoto–Sivashinsky equation.

The Kuramoto–Sivashinsky (KS) equation

$$\frac{\partial u}{\partial t} + 4\frac{\partial^4 u}{\partial x^4} + \alpha\left(\frac{\partial^2 u}{\partial x^2} + \frac{1}{2}\left(\frac{\partial u}{\partial x}\right)^2\right) = 0 \tag{2.1}$$

is an appealing model equation for studying the reduction problem given the fact that its numerical solutions are attracted to (both smooth and fractal) sets of moderate dimension; see [13, 11]. In addition, there is strong evidence that in many instances the actual, or intrinsic, dimensionality of the numerical solutions is far smaller than the ambient dimension of the simulation. This study extends our previous investigations of the KS equation using linear reduction methods [17, 15] in addition to nonlinear reduction methods in [18, 19, 20].

In this investigation the raw numerical data are generated by a 20-dimensional Fourier Galerkin simulation (10 complex modes), as described in [15]. We test the methodology proposed in this paper on several canonical problems for which the dynamics are well understood.

For example, taking $\alpha = 84.25$ in the KS equation results in a numerical solution which corresponds to a limit cycle whose Fourier decomposition has significant energy in at least seven Fourier modes. Topologically this attractor is 1-dimensional and in fact it is diffeomorphic to the unit circle. The numerical experiments in this study employ a data set representing the limit cycle solution at 500 points in $\mathbb{R}^{20}$ sampled uniformly in time over approximately 10 periods of the oscillation.

For $\alpha = 87$ the numerical solutions to the KS equation are somewhat more complex. The solution corresponding to the previous parameter value of $\alpha = 84$ has undergone a Hopf bifurcation to a traveling wave with the localized oscillation persisting. Topologically this attractor is now 2-dimensional and is diffeomorphic to the canonical torus. In this paper we examine a data set consisting of 1000 points representing this torus in $\mathbb{R}^{20}$ sampled over five periods. We remark that it is not possible to (rigidly) rotate, i.e., via an orthogonal transformation, the whole torus into a 3-dimensional subspace [20].

Taking $\alpha = 91$ the spatiotemporal behavior has become chaotic and the solutions have been estimated to reside on a roughly 5-dimensional set. This data set is distinct from the previous two cases where the attractors were smooth manifolds. We collected a data set consisting of 5000 samples evenly spaced in time.

2.2. Digital image reduction. In [26] and [21] we initiated the application of optimal linear transformations to a collection of digital images of human faces. This is a very interesting prototype problem given the large dimensionality of the input data points. In the current investigation we examine a data set consisting of 200 images each of size 256×256. The data are initially projected to a 200-dimensional subspace using the Karhunen–Loève decomposition subspace without loss of any information. As a result of this preprocessing, in this study we examine a data set of 200 images each residing in $\mathbb{R}^{200}$. Our goal is to determine good projections for these data and to establish the importance of finding such projections for facilitating the reconstruction problem.

3. Good projections: Theory.

The main tool for this analysis is the following theorem from differential topology [9] (see also [8]).

> WHITNEY'S (EASY) EMBEDDING THEOREM. *Let $\mathcal{U}$ be a compact Hausdorff C^r m-dimensional manifold, $2 \le r \le \infty$. Then there is a C^r embedding of $\mathcal{U}$ in $\mathbb{R}^{2m+1}$.*

Paraphrasing, such m-dimensional manifolds may be diffeomorphically mapped to the Euclidean space $\mathbb{R}^{2m+1}$. Hence the Euclidean space $\mathbb{R}^{2m+1}$ is large enough to contain a "diffeomorphic copy" of these m-dimensional manifolds [8].

The proof of Whitney's theorem demonstrates that, generically, a projection (which is, of course, a linear mapping) has an inverse, provided the dimension of the target space is at least $2m+1$. We propose that this projection should be constructed such that its inverse—a potentially complicated nonlinear mapping—is especially easy to find. We now make this more precise.

Define a projection $\pi_{\hat{v}} : \mathbb{R}^q \to \mathbb{R}^{q-1}$ which is parallel to the unit vector $\hat{v}$ and acts to eliminate any component in the direction $\hat{v}$ (see Figure 3.1). We can write this explicitly as the linear map

$$\pi_{\hat{v}} x = (I - \hat{v}\hat{v}^T)x,$$

where $\hat{v} \in S_{q-1} = \{\hat{v} \in \mathbb{R}^q : \|\hat{v}\| = 1\}$. Note that $\pi_{-\hat{v}} \equiv \pi_{\hat{v}}$ so we will be better off identifying the antipodes of S_{q-1}, i.e., we can think of $\hat{v} \in P_{q-1}$, the projective $(q-1)$-space.

If $\pi_{\hat{v}}$ restricted to $\mathcal{U}$ is an embedding of $\mathcal{U}$ in $\mathbb{R}^{q-1}$, we know there is an inverse $\pi_{\hat{v}}^{-1} : \mathbb{R}^{q-1} \to \mathbb{R}^q$ in the form $\pi_{\hat{v}}^{-1}x = (x, g(x))$ where $g : \pi_{\hat{v}}\mathcal{U} \subset \mathbb{R}^{q-1} \to \mathbb{R}$ and we think of $\mathbb{R}$ as $\{\alpha\hat{v} \,|\, \alpha \in \mathbb{R}\}$, the null space of $\pi_{\hat{v}}$. In this picture $\mathcal{U}$ is the graph of g in the space $\mathbb{R}^{q-1} \times \mathbb{R}$, i.e.,

$$\mathcal{U} = \{(x, g(x)) \,|\, x \in \pi_{\hat{v}}\mathcal{U} \subset \mathbb{R}^{q-1}\}. \tag{3.1}$$

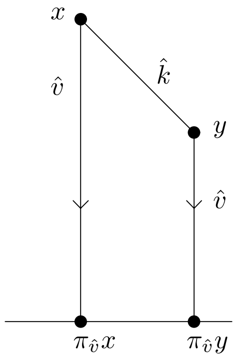

FIG. 3.1. *The projection of the data along direction $\hat{v}$.*

In practice, we shall find this by fitting data, using, for example, radial basis functions (RBFs), as described in section 5.

The ease with which this inverse can be computed is directly dependent on how ill-conditioned the mapping is (see section 5.2 for more details). As a measure of how singular the projection is we shall establish a strict lower bound on the proximity of the images of distinct points under the projection $\pi_{\hat{v}}$. Specifically, we restrict the set of admissible projections to be those which satisfy the inequality

$$\|\pi_{\hat{v}}x - \pi_{\hat{v}}y\| \geq \kappa_{\hat{\pi}}\|x - y\| \tag{3.2}$$

for all $x, y \in \mathcal{A}$ and some fixed *tolerance* $\kappa_{\hat{\pi}}$. If $\kappa_{\hat{\pi}} > 0$ the mapping is nonsingular; the size of $\kappa_{\hat{\pi}}$ can be taken as a measure of how well-conditioned we expect the inverse $\pi_{\hat{v}}^{-1}$ to be.

LEMMA 3.1. *For all $x, y \in \mathcal{U}$, $\|\pi_{\hat{v}}x - \pi_{\hat{v}}y\| \geq \kappa_{\hat{\pi}}\|x - y\|$ and $\kappa_{\hat{\pi}} > 0$ except on a set of singular $\hat{v}$ which is nowhere dense in S_{q-1}.*

We shall outline a proof of this result because it directly motivates our proposed algorithm for reduction; the discussion follows that in [9]. We require the following lemma which we state without proof.

LEMMA 3.2. *Let $f : P \to Q$ be any C^1 map. If $\dim Q > \dim P$ the fP is nowhere dense in Q.*

To begin, consider the set of normalized secants Σ consisting of the unit vectors

$$\hat{k} = \frac{x - y}{\|x - y\|}, \tag{3.3}$$

where $x, y \in \mathcal{U}, x \neq y$. It is also useful to define the set

$$\Delta = \{(a, b) \in \mathcal{U} \times \mathcal{U} \, : \, a = b\}$$

which is referred to as the *diagonal* of $\mathcal{U} \times \mathcal{U}$.

In [9] it is shown that the map

$$\sigma : \mathcal{U} \times \mathcal{U} - \Delta \to \Sigma \subset S_{q-1},$$

where $\sigma(x, y) = \hat{k}$ (see (3.3)), is a C^r map (assuming $\mathcal{U}$ is C^r). We note that in theory the set Σ is the set of all unit secants of the manifold $\mathcal{U}$. In practice our data are a finite sampling of points on $\mathcal{U}$ and we only have access to a subset of Σ.

After projecting and computing the norm, it follows from (3.3) that

$$\|\pi_{\hat{v}}x - \pi_{\hat{v}}y\| = \|\pi_{\hat{v}}\hat{k}\|\|x-y\|.$$

Now

$$\begin{aligned}\|\pi_{\hat{v}}\hat{k}\|^2 &= ((I-\hat{v}\hat{v}^T)\hat{k})^T(I-\hat{v}\hat{v}^T)\hat{k}\\ &= \hat{k}^T(I-\hat{v}\hat{v}^T)(I-\hat{v}\hat{v}^T)\hat{k}\end{aligned}$$

and since $\pi_{\hat{v}}^2 = \pi_{\hat{v}}$

$$\|\pi_{\hat{v}}\hat{k}\|^2 = \hat{k}^T(I-\hat{v}\hat{v}^T)\hat{k}$$

which implies

$$\|\pi_{\hat{v}}\hat{k}\|^2 = 1-(\hat{v}\cdot\hat{k})^2.$$

So we have

$$\|\pi_{\hat{v}}x - \pi_{\hat{v}}y\| = (1-(\hat{v}\cdot\hat{k})^2)^{1/2}\|x-y\|.$$

Using Lemma 3.2 it is shown in [9] that $\Sigma = \sigma(\mathcal{U}\times\mathcal{U}-\Delta)$ is nowhere dense in S_{q-1} if

$$q > 2m+1$$

but Σ is the set of secants of $\mathcal{U}$, so $\hat{v}$ such that $(\hat{v}\cdot\hat{k})^2 \neq 1$ are dense in S_{q-1}. Hence we can complete the proof by choosing

$$\kappa_{\hat{\pi}} = \min_{\hat{k}\in\Sigma}(1-(\hat{v}\cdot\hat{k})^2)^{1/2}. \tag{3.4}$$

A small extension of these arguments will allow us to show there exists an optimum projection. Consider now the closure, $\bar{\Sigma}$, of Σ. By considering Cauchy sequences of points $x_i \in \mathcal{U}$ and the corresponding sequences of images $\sigma(x_i, x_j) \in S_{q-1}$ we see that $\bar{\Sigma}$ contains, in addition to the unit secants, the unit tangent bundle of $\mathcal{U}$.[2] For our purposes it turns out that $\bar{\Sigma}$ is the natural object to study since if we choose a projection with $\hat{v}\in\bar{\Sigma}$, it fails to be an embedding of $\mathcal{U}$ either because it is not an injection ($\hat{v}$ is parallel to a secant) or because it is not an immersion ($\hat{v}$ is parallel to a tangent vector). Like Σ, $\bar{\Sigma}$ is nowhere dense in S_{q-1}. In addition, it is a closed subset of a compact manifold and is therefore itself compact.

If $\hat{v}\notin\bar{\Sigma}$ it follows $(1-(\hat{v}\cdot\hat{k})^2)^{1/2} > 0$, so we can redefine $\kappa_{\hat{\pi}}$ as

$$\kappa_{\hat{\pi}} = \min_{\hat{k}\in\bar{\Sigma}}(1-(\hat{v}\cdot\hat{k})^2)^{1/2}. \tag{3.5}$$

We now need to study in a little more detail the geometry of $\bar{\Sigma}$ as a subset of the set of all projections. First, we introduce the relation $\sim$ such that

$$a \sim b \Leftrightarrow (a\cdot b)^2 = 1,$$

[2] In practice, this means that in our finite sampling of $\mathcal{U}$, nearby data points will give secants that more or less approximate unit tangent vectors.

$a, b \in S_{q-1}$. In terms of this the set of projections is $P_{q-1} = S_{q-1} \backslash \sim$, the projective $(q-1)$-space.

To make contact with the foregoing arguments we define a metric on P_{q-1} according to

$$d(a,b) = (1-(a\cdot b)^2)^{1/2}. \tag{3.6}$$

Clearly

$$\text{(i)}\quad d(a,b)=0 \quad \text{iff} \quad a\sim b \tag{3.7}$$

$$\text{(ii)}\quad d(a,b)=d(b,a) \quad \text{for all } a,b\in S_{q-1}(P_{q-1}) \tag{3.8}$$

$$\text{(iii)}\quad d(a,b)\le d(a,c)+d(c,b). \tag{3.9}$$

Proof of (iii). Let the angle subtended by a, b be θ_{ab}. Then $d(a,b) = |\sin\theta_{ab}|$. Assume that c is coplanar with a and b. Then $\theta_{ab}+\theta_{bc}+\theta_{ca} = 2\pi n$, for some $n\in\mathbb{Z}$, implies

$$\sin\theta_{ab} = -\sin(\theta_{bc}+\theta_{ca}) = \sin(\theta_{ac}+\theta_{cb})$$

and hence

$$|\sin\theta_{ab}| = |\sin(\theta_{ac}+\theta_{cb})| = |\sin\theta_{ac}\cos\theta_{cb}+\sin\theta_{cb}\cos\theta_{ac}|,$$

$$|\sin\theta_{ab}| \le |\sin\theta_{ac}\cos\theta_{cb}| + |\sin\theta_{cb}\cos\theta_{ac}| \le |\sin\theta_{ac}| + |\sin\theta_{cb}|.$$

So (3.9) holds if c is in the same plane as a, b. But in that case it holds for all c since $d(a,c)+d(c,b)$ can only increase as c moves out of the a, b plane. So d is a metric on P_{q-1}. □

Having defined this metric we see that $\kappa_{\hat\pi}$ is just the distance from $\hat v$ to $\bar\Sigma$ (for a definition of distances between points and sets and discussion of the various notions of distance between sets, see [1]):

$$\kappa_{\hat\pi} = \min_{\hat k\in\bar\Sigma}(1-(\hat v\cdot\hat k)^2)^{1/2} = \min_{\hat k\in\bar\Sigma} d(\hat v,\hat k). \tag{3.10}$$

So

$$\kappa_{\hat\pi} = d(\hat v,\bar\Sigma). \tag{3.11}$$

This argument can be developed further if we consider maximizing $\kappa_{\hat\pi}$, i.e., making $\pi_{\hat v}$ as nonsingular as possible. Now we have

$$\kappa_{\text{opt}} = \max_{\hat v\in P_{q-1}} d(\hat v,\bar\Sigma) \tag{3.12}$$

but this is just the asymmetrically defined distance from P_{q-1} to $\bar\Sigma$,

$$\kappa_{\text{opt}} = d(P_{q-1},\bar\Sigma). \tag{3.13}$$

Since $\bar\Sigma\subset P_{q-1}$, the asymmetric distance from $\bar\Sigma$ to P_{q-1} vanishes. It follows that $d_H(P_{q-1},\bar\Sigma) = d(P_{q-1},\bar\Sigma)$ where d_H is the Hausdorff metric defined on the set of compact subsets of P_{q-1}. This gives a rather intuitive picture of how well we can do. If $\bar\Sigma$ were to be a large subset of P_{q-1} in the sense that it is close to every point in

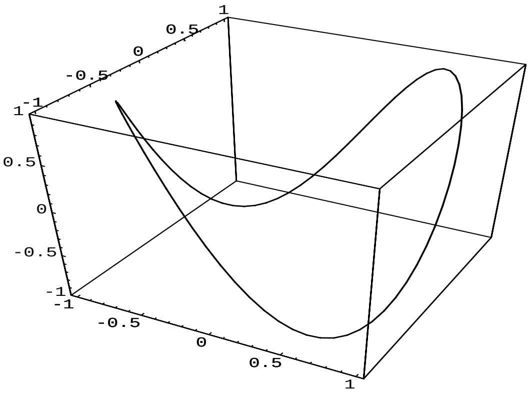

FIG. 3.2. *The image of $\mathcal{S}^1$ under the action of G.*

P_{q-1}, its Hausdorff distance from P_{q-1} would be small. Conversely, if $\bar{\Sigma}$ is small and localized, then its Hausdorff distance from P_{q-1} would tend to unity, the diameter of P_{q-1}.

We note that the compactness of $\bar{\Sigma}$ implies that there actually are pairs of points, say $\hat{v}_* \in P_{q-1}$ and $\hat{k}_* \in \bar{\Sigma}$, such that

$$\begin{aligned}\kappa_{\text{opt}} &= d(\hat{v}_*, \hat{k}_*)\\ &= d_H(P_{q-1}, \bar{\Sigma}),\end{aligned}$$

that is, we know that at least one optimum projection exists.

Example. We conclude this section with a simple example. Consider the following map of the circle $G : \mathcal{S}^1 \to \mathbb{R}^3$ where $\theta \mapsto (\sin\theta, \cos\theta, \sin 2\theta)$. The set $G(\mathcal{S}^1) \subset \mathbb{R}^3$, which is shown in Figure 3.2, is an embedding of the circle and will be $\mathcal{U}$ in this example.

The unit secants of $G(\mathcal{S}^1)$ are the image of the map $\sigma : \mathcal{S}^1 \times \mathcal{S}^1 \to \mathcal{S}^2$ where

$$\sigma(\theta_1, \theta_2) = \pm\frac{(\cos\frac{1}{2}\Theta, -\sin\frac{1}{2}\Theta, 2\cos\Theta\cos\frac{1}{2}\Delta)}{[1 + 4\cos^2\Theta\cos^2\frac{1}{2}\Delta]^{\frac{1}{2}}}. \tag{3.14}$$

Here $\Theta = \theta_1 + \theta_2$, $\Delta = \theta_1 - \theta_2$, and either choice of sign is valid since the unit secants are really points in the projective plane. We have included the diagonal in the domain of definition of σ and so the set $\sigma(\mathcal{S}^1 \times \mathcal{S}^1)$ plotted in Figure 3.3 is actually $\bar{\Sigma}$, the closure of the set of secants of $G(\mathcal{S}^1)$. In this example we see explicitly that the boundary of $\bar{\Sigma}$ is given by the unit tangent bundle of $G(\mathcal{S}^1)$ because, as $\theta_1 \to \theta_2 = \theta$, we have the following limit:

$$\sigma(\theta_1, \theta_2) \to \pm\frac{(\cos\theta, -\sin\theta, 2\cos 2\theta)}{[1 + 4\cos^2 2\theta]^{\frac{1}{2}}} = \pm\frac{D_\theta G}{\|D_\theta G\|}. \tag{3.15}$$

This formula has been used to plot Figure 3.4 which can be compared with Figure 3.3.

As we have discussed, when considering the possible projections of $G(\mathcal{S}^1)$ into $\mathbb{R}^2$, $\bar{\Sigma}$ would correspond to bad projections. We cannot invoke the Whitney embedding theorem here because the dimension of $\mathbb{R}^2$ is too small; indeed Figure 3.3 indicates that, rather than being nowhere dense, $\bar{\Sigma}$ is a 2-dimensional subset of $\mathcal{S}^2$. Nevertheless, there are projections which are not in $\bar{\Sigma}$ and are thus invertible. Some of

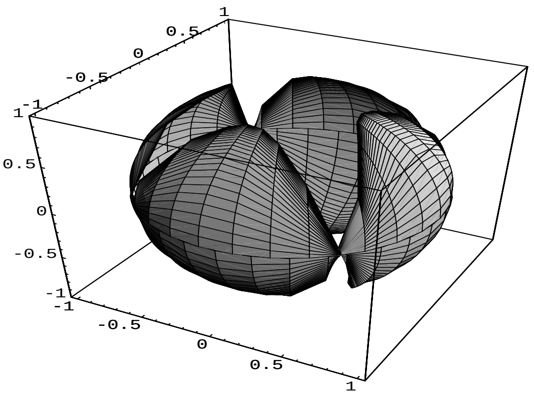

FIG. 3.3. $\bar{\Sigma}$*, the image of* $\mathcal{S}^1 \times \mathcal{S}^1$ *under the action of* σ.

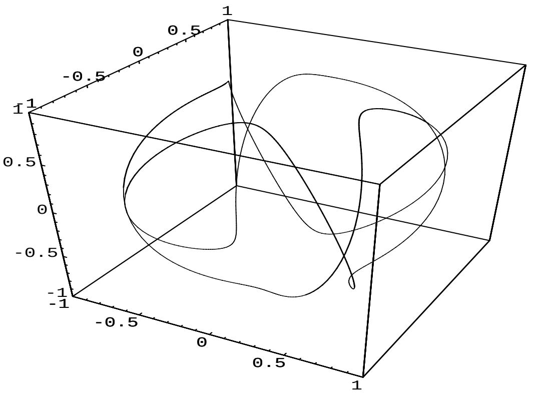

FIG. 3.4. *The unit tangent bundle of* $G(\mathcal{S}^1)$*. Comparison with Figure* 3.3 *demonstrates that it is the boundary of* $\bar{\Sigma}$.

the possibilities are shown in Figure 3.5. The figure shows projections of $G(\mathcal{S}^1)$ into planes plotted as 2-dimensional linear subspaces of $\mathbb{R}^3$. In each case the vector along which the projection is taken can be imagined as normal to the linear subspace in which the curve lies. The top row of the figure shows two invertible projections. The top left-hand plot shows a projection along the z-axis while the top right shows a projection closer to the boundary of $\bar{\Sigma}$ showing that part of the image is rather tightly pinched. The center image is of a projection taken from the boundary of $\bar{\Sigma}$. Here we see that the pinched part of the image has developed into a cusp, that is, the image has no well-defined tangent at this point. The bottom row of the figure shows two projections taken from within Σ—the left-hand plot shows a projection closer to the boundary, while the right-hand plot shows a projection along the x-axis. Now we see that both images selfintersect and hence that the projections are noninvertible.

Our example is sufficiently simple that we can find the optimum projection of $G(\mathcal{S}^1)$. Let's look more closely at $\bar{\Sigma}$ as shown in Figure 3.3. To find the optimum projection we need the point in the projective plane[3] which is furthest from $\bar{\Sigma}$ in the

[3] We have been using $\mathcal{S}^2$ so far in this discussion. Because we have chosen a suitable metric, in the following the optimization will actually take place on the projective plane.

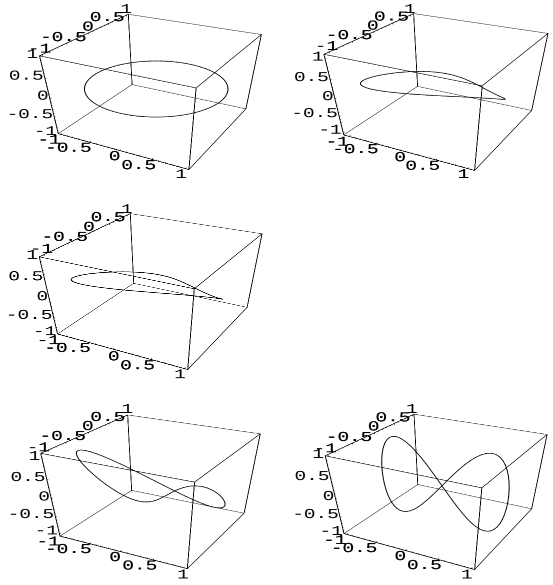

FIG. 3.5. *Examples of projections of $G(\mathcal{S}^1)$ onto 2-dimensional linear subspaces of $\mathbb{R}^3$. The top row of the figure shows two invertible projections. The center image is of a projection taken from the boundary of $\bar{\Sigma}$. The bottom row of the figure shows two noninvertible projections taken from within Σ.*

metric d defined by (3.6). As we described in the previous section this is a double optimization problem. We first choose a point in the projective plane and then find the point in $\bar{\Sigma}$ which is closest. We then find the point in the projective plane where this minimum distance is maximized. It is easy to see that the closest point in $\bar{\Sigma}$ to any external point must lie on the boundary. Moreover, a symmetry argument—or a brief study of Figure 3.3—shows that the points furthest from the boundary must lie on a great circle which passes through the pole and a point where the boundary crosses the equator. (Actually, there are two different—but for these purposes, equivalent—great circles which satisfy these conditions: one passing through the point parameterized by $\theta = \frac{\pi}{4}$ in (3.15), and the other through the point parameterized by $\theta = \frac{3\pi}{4}$). Figure 3.6 shows a contour plot of the distance, d, between points on one of these great circles (the one corresponding to $\theta = \frac{\pi}{4}$) and all the points on the boundary of $\bar{\Sigma}$. The boundary is parameterized by $\theta \in [0, 2\pi]$—this coordinate is plotted horizontally—while the position on the great circle—plotted vertically—is parameterized by an angle taking values in the interval $[0, \frac{\pi}{2}]$ (0 corresponding to the pole and $\frac{\pi}{2}$ corresponding to the equator). Starting at the equator—the top of the figure—there are two minima at $\theta = \frac{\pi}{4}$ and $\theta = \frac{5\pi}{4}$. These are the points where the great circle of interest intersects the boundary of $\bar{\Sigma}$. We now need to maximize the minima. Moving along the great circle toward the pole corresponds to moving down the figure. The minima at $\theta = \frac{\pi}{4}$ and $\theta = \frac{5\pi}{4}$ are at the ends of two valleys; as we move down the figure the horizontal minimization keeps us in the valley bottom but the value of d at the minimum increases until, as we approach the pole, it attains its maximum value of $\frac{1}{\sqrt{5}}$. Note that the positions of the minima have changed to $\theta = 0$ and $\frac{3\pi}{2}$ when the point is at the pole. In addition, two extra—equally deep—minima have formed at $\theta = \frac{\pi}{2}$ and π. These four angles parameterize points on the boundary which are closest to the pole. They

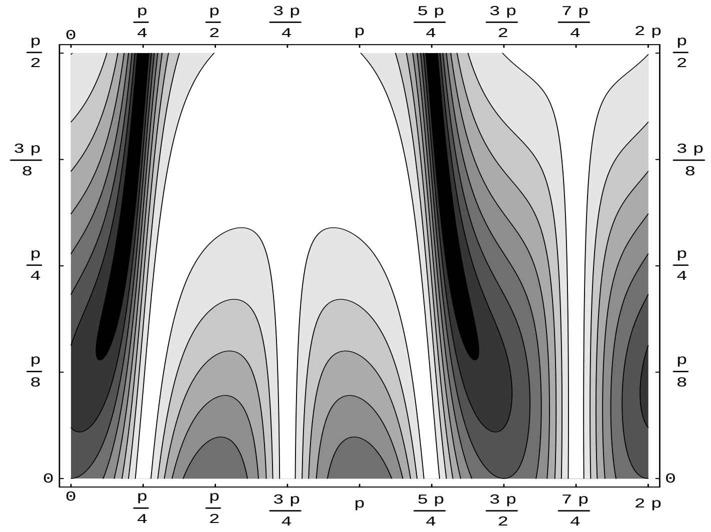

FIG. 3.6. *Contour plot of the distance, d, between points on the great circle described in the text and points on the boundary of $\bar{\Sigma}$. The abscissae parameterize the position on the boundary and the ordinates parameterize the position on the great circle. The darker the shade the smaller the value of d.*

can be readily identified by looking at Figures 3.3 and 3.4. We conclude that the optimum projection is along the z-axis. The effect of this on $G(\mathcal{S}^1)$ is shown in the top left-hand plot of Figure 3.5.

4. Good projections: Algorithms.

The emphasis in this section will be more pragmatic. Here we shall concentrate on finding “good”—as opposed to “optimum”—projections based on data. Our basic premise is that we should find a projection which is in some sense as far away as possible from any secant of the data. Since we are dealing with a finite set of distinct data points it would be unnecessarily sophisticated to speak of $\bar{\Sigma}$, so from now on we shall refer to the set of secants constructed from the data as Σ.

In the previous section we considered the particular case of a projection which reduces the dimension of the space by one. In practice we seek to determine a single projection which removes components which lie in an r-dimensional subspace which has no direction colinear with a secant of $\mathcal{U}$. We now denote this projection by $\pi : \mathbb{R}^q \to \mathbb{R}^{q-r}$ and define it as

$$\pi x = (I - \hat{v}_1\hat{v}_1^T - \hat{v}_2\hat{v}_2^T - \cdots - \hat{v}_r\hat{v}_r^T)x. \tag{4.1}$$

By induction, if $q - r \geq 2m + 1$, such a projection exists such that $\kappa_\pi > 0$ (again, by construction $\kappa_\pi \leq 1$) and

$$\|\pi x - \pi x'\| \geq \kappa_\pi \|x - x'\| \tag{4.2}$$

for all $x, x' \in \mathcal{U}$. Note that in the previous section we denoted this minimum by $\kappa_{\hat{\pi}}$ to indicate its dependence on $\hat{\pi}$, a 1-dimensional projection. Again, we seek to determine a projection as defined by (4.1) which is good, i.e., such that κ_π is relatively large, e.g., $\kappa_\pi \in [0.1, 0.5]$.

We propose an approach to this problem based on the SVD. First, construct all the unit secants

$$\hat{k}_j = \frac{x - x'}{\|x - x'\|}, \tag{4.3}$$

where x, x' are the jth pair of data points in $\mathcal{A}$. From these secants form the matrix K which has $\hat{k}_j^T$ as its jth row. K is $N \times q$, where N is the number of distinct data pairs, counting each pair only once, and q is the ambient dimension of the space in which the data are found. We assume that q is finite but possibly large and further that $N > q$. Also, given that the number of secant pairs grows as the square of the number of points we propose that, for large data sets, a manageable number of cluster points may be used in place of data points for estimating the secant set.

The SVD of this matrix will be written

$$K = U\mathbb{K}V^T,$$

where U is an $N \times q$ matrix with orthonormal columns, V is a $q \times q$ orthogonal matrix, and $\mathbb{K}$ is a $q \times q$ diagonal matrix containing the singular values of K. We note that if rank$(K)< q$, then the projection along the unit vectors contained in ker(K) will effectively result in $\kappa_\pi = 1$ since such projections will leave the $\hat{k}_j$ invariant. By extension, projection along singular vectors of K which correspond to small singular values might be expected to result in large values of κ_π. In fact, there are potential pitfalls here, since singular vectors give projections which are good in the L_2 sense, but we need to satisfy the pointwise criterion

$$\kappa_\pi = \min_{\hat{k} \in \Sigma} \|\pi\hat{k}\|$$

so there might be a small number of secants that seriously reduce the minimum value of $\|\pi\hat{k}\|$.

Let us proceed on the basis that the SVD of K is pointing us in the right direction and determine the minimum rank of the projection such that for a given tolerance κ_π the inequality (4.2) is achieved. Now the columns of V, which we shall call $\{\hat{v}_j\}_{j=1}^q$, will be candidate projections. So for each row $\hat{k}_l^T$ of K (where $\hat{k}_l \in \Sigma$) we construct the following partial sums:

$$s_d^{(l)} = \sum_{j=1}^{d} (\hat{v}_j^T \cdot \hat{k}_l)^2.$$

For any l the sequence $s_d^{(l)}$ is clearly a nondecreasing function of d, and since the $\{\hat{k}_l\}$ are unit vectors we have $s_q^{(l)} = 1$ for any l.

We can interpret $s_d^{(l)}$ as the squared norm of the lth secant projected onto the space $\mathbb{R}^d$ spanned by the first d singular vectors of K. So we can, for a given κ_π, look through all $\hat{k}_j$ and compute the corresponding $s_d^{(l)}$ to determine the smallest d such that

$$s_d^{(l)} \geq (\kappa_\pi)^2 \qquad \text{for all } \hat{k}_l \in \Sigma.$$

Let's consider further how this procedure works in practice. Imagine that we plot, as in Figure 4.1, the integer valued function $d_\pi(l)$ which is defined as the minimum

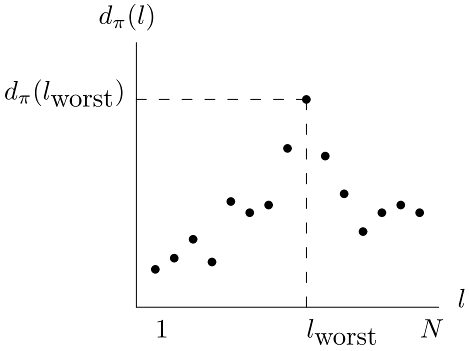

FIG. 4.1. *A representative plot of the minimum dimension required to achieve tolerance as a function of the secant pair index.*

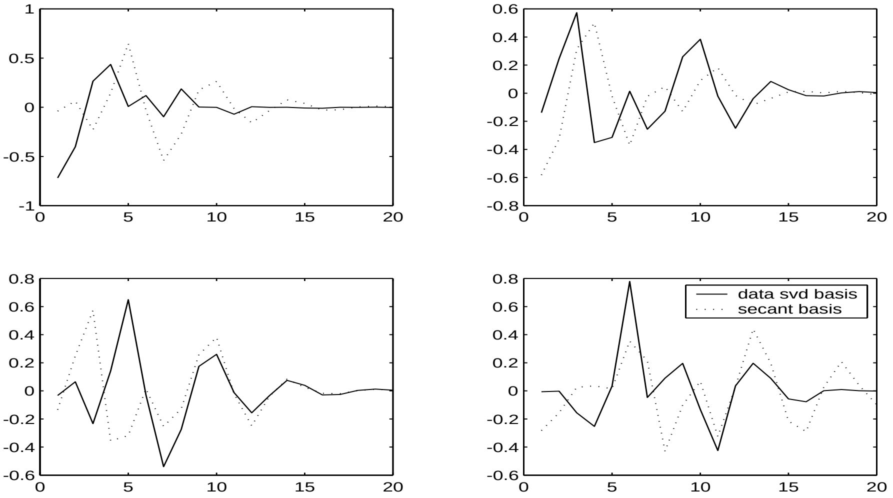

FIG. 4.2. *First four right singular vectors of the secant matrix (dashed) and the data matrix (solid) from the KS equation at $\alpha = 84$. The singular vectors are ordered according to the magnitude of their singular values decreasing along top left, bottom left, top right, bottom right.*

dimension d such that $s_d^{(l)} \geq (\kappa_\pi)^2$. In other words, the secant $\hat{k}_l$ requires $d_\pi(l)$ dimensions for the projection π to be good. The secant indexed by l_{worst} (which we shall denote as $\hat{k}_{\text{worst}}$) is a barrier to projection since it must have a significant projection onto a singular vector, $\hat{v}$, which has a small singular value. By definition of the SVD such occurrences should be rare (in the L_2 sense), so we might expect to improve the situation by weighting $\hat{k}_{\text{worst}}$ more in the matrix K. If we weight $\hat{k}_{\text{worst}}$ to an extreme degree then evidently $\hat{k}_{\text{worst}}$ will itself be a singular vector of the weighted K and it will have a large singular value (by virtue of the large weight) and hence we will have removed the peak in the graph $d_\pi(l)$.

The above procedure for computing good bases is readily implemented. To begin, a set of 4950 secants was computed from 100 points, each in $\mathbb{R}^{20}$, sampled from the KS equation limit cycle data set collected at $\alpha = 84$. The resulting *secant basis*, i.e., right singular vectors of the 4950×20 secant matrix K, were computed and compared to the *data SVD basis*, i.e., the right singular vectors of the 100×20 raw data matrix; see Figure 4.2. The coordinates of the eigenvectors are displayed with respect to the real

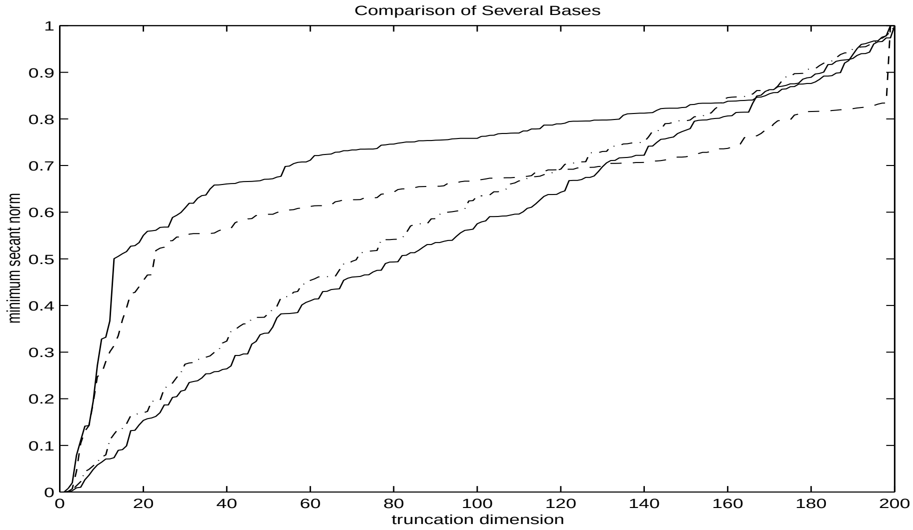

Fig. 4.3. *Tolerance achieved as a function of dimension for the face data set. For each dimension the minimum projected secant norm is computed for several different bases. The lower solid line corresponds to the SVD analysis on the data; the dot-dash line immediately above this corresponds to a reordering of the data SVD basis based on directions which optimize the minimum projected secant norm (note that an ordered random basis fits in between this and the preceding curve); the dashed line is the result of projecting on the basis computed from SVD on the data secants; the upper solid line is the secant basis adapted to* $\kappa_\pi = .5$.

Fourier basis so the domain value ranges from 1–20 and represents the index of the real Fourier vector $(\cos x, \sin x, \ldots, \cos 10x, \sin 10x)$ while the range value corresponds to the amplitude of the mode in the singular vector. Thus, we see that the eigenvectors for both the data basis and secant basis have significant energy across many frequency modes. However, when the secants are projected onto the 2-dimensional subspaces corresponding to the two largest singular values in each basis, the minimum projected secant norm for the secant basis is greater than 0.9 while for the data SVD basis is less than 0.03. These results indicate that the distances between the raw data points are not decreased significantly by the 2-dimensional projection onto the secant basis. On the other hand, projection onto two dimensions spanned by the data SVD basis significantly collapses for at least one pair of points. If the dimension of the projection is augmented to three, then both minima are over 0.95 with the secant basis being slightly better. As we see later, this significant difference in minimum projected secant norm has important ramifications in the calculation of the inverse of the projection.

The larger minimum projected secant norm characteristic of the secant basis is also readily demonstrated using our data set consisting of an ensemble of 200 digital images of faces. Now the secant matrix K has size 19900×200 and the secant basis consisting of the right singular vectors of K spans $\mathbb{R}^{200}$. For purposes of comparison we also generated a random basis and the standard data SVD basis, i.e., the right singular vectors of the data matrix. The minimum projected secant norms attained as a function of the number of (ordered) dimensions of the projection are plotted in Figure 4.3. It is clear from this figure that the secant basis provides a much better projection than the data SVD basis according to the criterion (4.2). The superiority of the secant SVD basis remains even if the data SVD basis is reordered such that the vectors onto which the secants (rather than the data) have maximum projection come first. Note that, by the criterion (4.2), a random basis performed as well as the data SVD basis in this example. For any given dimension the minimum distance separating the projected points is substantially greater for the secant basis than for the other bases shown. This example demonstrates that the selection of the basis is critical in distinguishing good projections from admissible projections in the sense of Whitney's theorem. As we have already suggested, the secant SVD basis my be adapted to improve the reduction dimension. We will refer to this new basis as the *adapted secant SVD basis*.

4.1. The adapted secant SVD basis. Here we develop the improvement on the secant basis algorithm described above. Given that we seek to optimize a pointwise value, as opposed to an average value, we propose an adaptive approach to give additional weight to rare but problematic secants.

Assume that we are applying the SVD using a diagonalization routine on the covariance matrix of secants $\Theta = K^T K$ ($q \times q$ matrix). We use the fact that $\mathrm{Tr}\Theta = \mathrm{Tr}(K^T K) = \mathrm{Tr}(KK^T)$. Now $\mathrm{Tr}\Theta = \sum_{j=1}^{q} \sigma_j^2$ where the σ_j are the singular values of K. However, the diagonal of KK^T contains terms like $\hat{k}_j^T \cdot \hat{k}_j = 1$. So since KK^T is $N \times N$ we find that $\mathrm{Tr}KK^T = N$. So we write

$$\Theta' = \Theta\left(1 - \frac{\alpha}{N}\right) + \alpha \hat{k}_{\text{worst}} \hat{k}_{\text{worst}}^T, \tag{4.4}$$

$$\mathrm{Tr}\Theta' = \left(1 - \frac{\alpha}{N}\right) \mathrm{Tr}\Theta + \alpha \mathrm{Tr} \hat{k}_{\text{worst}} \hat{k}_{\text{worst}}^T \tag{4.5}$$

$$= N - \alpha + \alpha, \tag{4.6}$$

and hence

$$\mathrm{Tr}\Theta' = N,$$

and this weighting leaves the trace invariant.

So we perform a rank-1 update of Θ in such a way as to preserve the sum of squared singular values and so as to increase the weighting of the vector $\hat{k}_{\text{worst}}$. The parameter α determines how much the weighting of $\hat{k}_{\text{worst}}$ is to be increased.

We could perform this process recursively choosing the secant index l each time to be such that $d_\pi(l)$ is largest (i.e., $l = l_{\text{worst}}$) and continue the process until the maximum value of $d_\pi(l)$ starts to decrease. This will happen inevitably because as the weight associated with a given secant increases, the secant will have more influence on the basis and hence will tend to be projected better. However, numerical results show that generally more than one secant will obstruct the projection, so we must modify the argument to account for this.

We can use the function $d_\pi(l)$ to put an equivalence relation on the set of secant vectors Σ:

$$\hat{k}_j \sim \hat{k}_l \Leftrightarrow d_\pi(l) = d_\pi(j).$$

Hence we partition the secants according to

$$\Sigma = \sum_{d=1}^{n} \Sigma_d,$$

where

$$\Sigma_d = \{\hat{k}_l \in \Sigma | d_\pi(l) = d\}.$$

Using this partition we can write matrices K_d which have rows $\hat{k}_j^T \in \Sigma_d$ and hence

$$\Theta = \sum_{d=1}^{n} \Theta_d \quad \text{where} \quad \Theta_d = K_d^T K_d.$$

Using similar arguments to those given above it is apparent that $\mathrm{Tr}\Theta_d = |\Sigma_d|$ so we can, by analogy with the rank-1 update method, construct a new matrix Θ':

$$\Theta' = \left(1 - \frac{\alpha}{N}\right)\Theta + \frac{\alpha}{|\Sigma_{d_{\max}}|}\Theta_{d_{\max}}, \tag{4.7}$$

where $d_{\max}$ is the largest d such that Σ_d is nonempty.

In Figures 4.3, 4.4, and 4.5 we show the results of applying the adaptive secant algorithm using the standard secant basis as an initial condition. We set $\kappa_\pi = 0.5$ and compute the number of dimensions required to achieve this tolerance using the eigenvectors of the covariance matrix—iterated according to (4.7)—as the basis. The improvement in the minimum projected secant norm for the adapted secant basis is shown in Figure 4.3. Note that the largest increase in the minimum projected secant norm is in the neighborhood of 0.5, the imposed tolerance. The adapted secant basis achieves this tolerance, i.e., all projected secant norms are greater than 0.5, using 13 dimensions, while 22 dimensions are required by the unadapted secant basis to achieve the same tolerance. Note that the data SVD basis and the random basis require on the order of 100 dimensions to achieve the prescribed tolerance. In Figure 4.4 we observe that as the algorithm proceeds the number of bad secants decreases (nonmonotonically) until all the secants are deemed good for that dimension. The dimension is then reduced and newly bad secants are included in the additional covariance matrix. As shown in Figure 4.5, the algorithm identifies secants which are initially "bad," i.e., their projected secant norm is less than 0.5 and improves them to the point where they are now bad only for a much smaller dimension projection. None of the secants projected onto 13 or more of the final adapted basis vectors are bad, i.e., fail the tolerance criterion.

While the spirit of our approach does not require us to determine that absolutely smallest basis to achieve a given tolerance, this adaptive scheme does appear to be quite successful in improving the basis. Further applications of this adaptive algorithm are presented in the next section.

4.2. Translational invariance.

Each of the KS data sets we consider can be thought of as a sequence of "snapshots" $\{x_i\}$ of the given time-dependent solution. Each x_i is a finite-dimensional vector obtained by sampling the spatial structure of the solution uniformly in space and time. For this study we have used a periodic spatial domain and so the KS equation together with boundary conditions is equivariant with respect to both spatial and temporal translations. This symmetry can be reflected directly in the symmetry of the solutions or—where a solution has less symmetry—in the interrelationship of different solutions under the action of symmetry operations. In particular, we have used some data sets which have a space-time translational symmetry (as in the case of, for example, a traveling wave of constant shape) and others which do not. In the case of data having this translational invariance, computation of the singular vectors based on the secants reveals an interesting fact: the singular vectors are sinusoids just as in the better-known case where the computation is based directly on the data.

This can be understood using the following argument. In matrix notation we can consider the (rectangular) matrix X whose rows are the x_i^T. There is a group

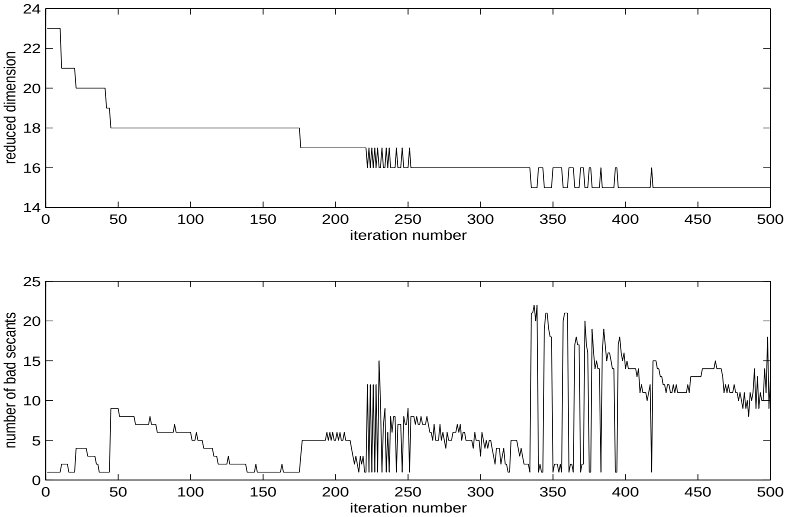

FIG. 4.4. *Results of adaptive secant algorithm on the face data for tolerance* $\kappa_\pi = 0.5$*. The top graph displays the dimension required to achieve the specified tolerance as a function of iterations of the adaptive algorithm. The number of bad secants (associated with the limiting dimension above) is plotted below as a function of the same iteration number.*

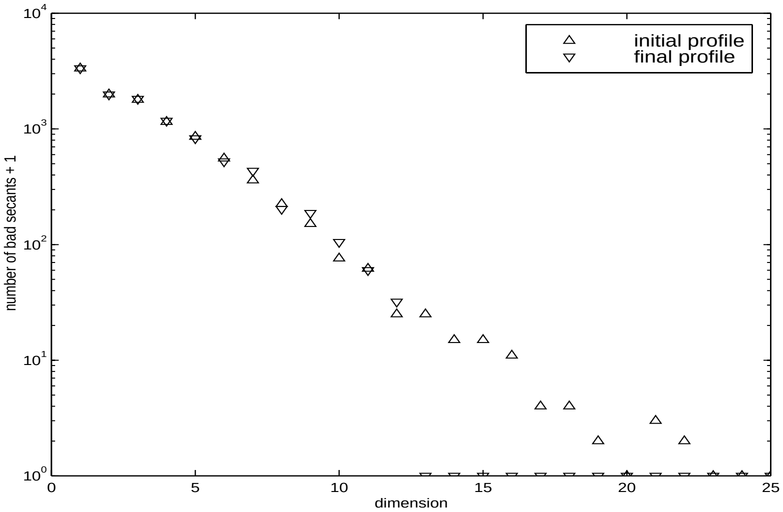

FIG. 4.5. *Results of the adaptive secant algorithm on the face data after* 2500 *iterations. The abscissae denote the dimension of the projection. The number of secants which are bad at the given dimension but which become satisfactory, i.e., satisify the tolerance* $\kappa_\pi = 0.5$*, at one greater dimension is plotted as the ordinates. (To distinguish between one and zero the number of bad secants is shifted up one before taking the log.)*

of cyclic permutations of the columns of X—that is, the components of x_i^T—which we represent by $\{C_i\}$, i.e., the appropriately sized circulant matrices. Corresponding to each of the C_i there is a permutation of the rows of X, denoted by P_i, which rearranges the rows of XC_i into their original order

$$P_i X C_i = X.$$

Note that $P_i^{-1} = P_i^T$. The right singular vectors of X may be determined by forming $X^T X$, i.e.,

$$X^T X = C_i^T X^T X C_i,$$

hence

$$X^T X C_i = C_i X^T X.$$

So the group of circulant matrices commutes with $X^T X$. Given that the C_i and $X^T X$ are *simple*, it follows that they share the same eigenvectors [23]. Since the eigenvectors of the circulant matrices are sinusoids we conclude that the right singular vectors are also sinusoids.

Now we address the right singular vectors of the secant data. Let K be the matrix whose rows consist of all the unit secants, i.e.,

$$\hat{k} = \frac{x_i^T - x_j^T}{\|x_i - x_j\|},$$

where $i \neq j$. Since we are assuming K consists of all pairs of secants we have

$$\hat{k}_l C_i = \hat{k}_{l'}$$

and we again find by permutation of the rows that

$$P_i K C_i = K.$$

Hence it follows that K has the same right singular vectors as X and that these are the Fourier modes, the eigenvectors of the circulant matrix. We remark that this analysis says nothing about the spectrum of singular values of either X or K, and in particular, how they differ.

To explore these results further numerically we consider the data produced by a simulation of the KS equation with $\alpha = 87$. For this parameter value the solutions consist of traveling beating waves, i.e., the solutions are translationally invariant and reside on a 2-dimensional torus which requires many Fourier modes to describe it accurately. Again, the data SVD basis for this case consists of Fourier modes. Similarly, as predicted by the above analysis, the eigenvectors for the unit secant data Σ are also sinusoids as shown by the dotted lines in Figure 4.6. Note that the double eigenvalues of the secant and data covariance matrices, shown in Figure 4.7, indicate that each complex Fourier mode forms a degenerate 2-dimensional eigenspace. Hence the basis vectors consist of orthogonal combinations of sinusoids of the same frequency.

It is interesting to consider the effect of the adaptation of the secant basis along the lines described in section 4.1. Now the action of the weighting is to break the translational symmetry of the data set. Consequently, the adapted secant basis vectors are no longer sinusoidal; see Figure 4.6. Another consequence of this adaptation is that the spectrum of the iterated covariance matrix is converging such that there is an eigenvalue of multiplicity eight; see Figure 4.7. The adaptation has weighted the data such that there is no preferred direction in the resulting 8-dimensional eigenspace.

Another example of a numerical implementation of the adaptive secant basis is provided by computing the adapted secant basis for the chaotic solutions of the KS equation corresponding to $\alpha = 91$. Similar to the above example for the traveling wave, the adapted secant basis algorithm produces an 8-dimensional eigenspace with

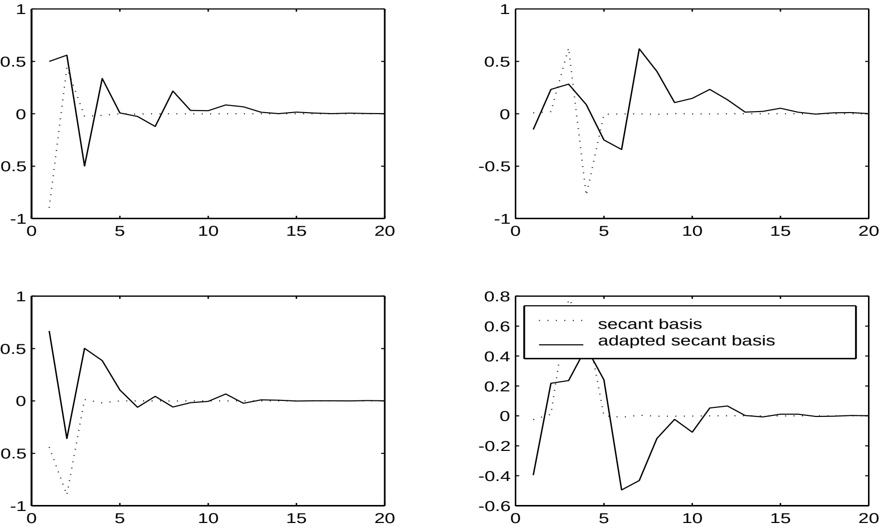

FIG. 4.6. *First four eigenvectors of the secant covariance matrix (dashed) and the adapted secant covariance matrix (solid) from the KS equation at $\alpha = 87$. The eigenvectors are ordered according to the magnitude of their singular values decreasing along top left, bottom left, top right, bottom right.*

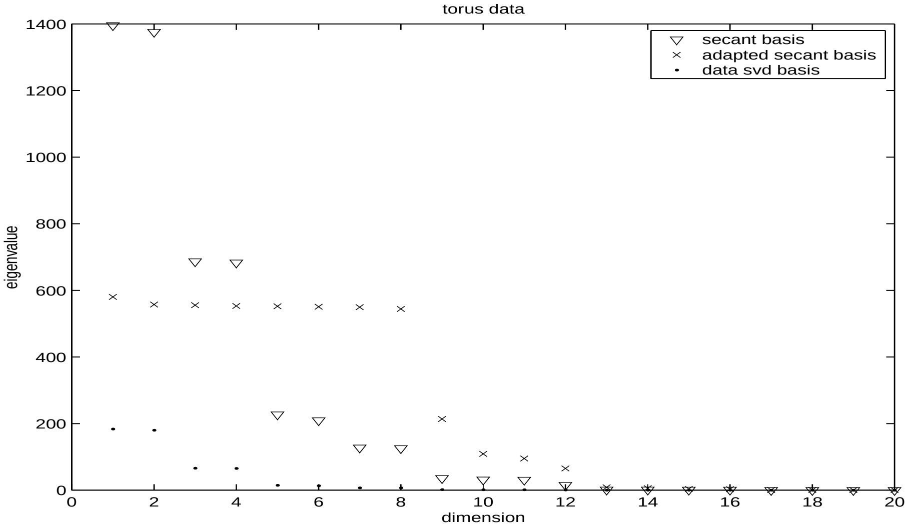

FIG. 4.7. *Eigenvalue spectra for data from the KS equation at $\alpha = 87$. The spectra are calculated from covariance matrices consisting of raw data, data secants and weighted data secants.*

equal variances in all directions; see Figure 4.8. Note that eight dimensions are required by the adapted secant basis to achieve a minimum projected secant norm greater than 0.5; see Figure 4.9. As is clearly indicated by this graph of minimum projected secant norms, the most improvement occurs in dimension seven where the minimum projected secant norm for the adapted basis is roughly twice that of the calculated secant basis and the analytical Fourier basis.

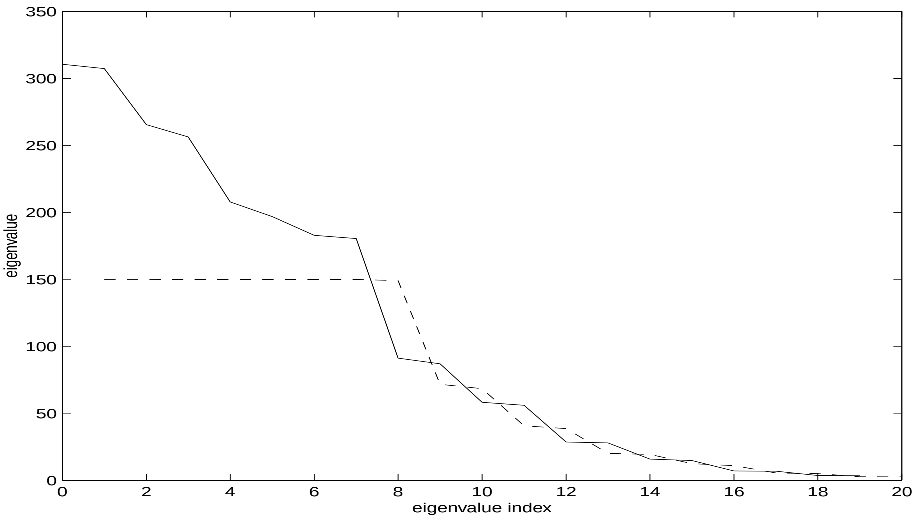

FIG. 4.8. *Eigenvalue spectra for the secant and weighted secant covariance matrices for data from the KS equation at $\alpha = 91$. The solid line corresponds to the nonadapted secant basis while the dashed line corresponds to the adapted secant basis. (Note that in this particular run we did not employ the weighting that conserves the trace of the spectrum.)*

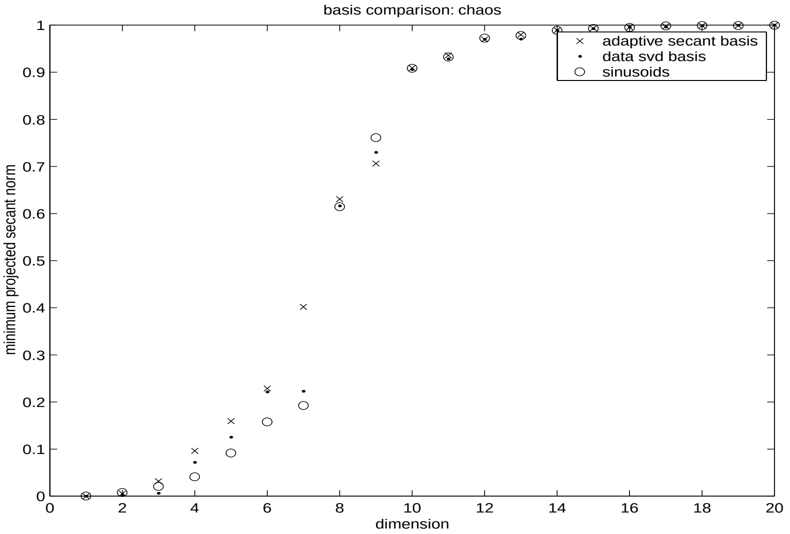

FIG. 4.9. *Minimum projected secant norm for the data SVD basis, the secant SVD basis, and the adapted secant SVD basis. The data consisted of chaotic solutions of the KS equation at $\alpha = 91$.*

5. Inverting the projection. Having constructed our good projection $\pi : \mathbb{R}^q \to \mathbb{R}^{2m+1}$, we are now interested in constructing the inverse π^{-1} which acts to reconstruct the data. As described in section 3, Whitney's theorem guarantees that almost any projection π restricted to $\mathcal{U}$ is an embedding of $\mathcal{U}$ in $\mathbb{R}^{2m+1}$. Hence there must exist an inverse $\pi^{-1} : \pi\mathcal{U} \subset \mathbb{R}^{2m+1} \to \mathbb{R}^q$. We can think of this as $\pi^{-1} = (id, g)$ where id denotes the identity map on $\mathbb{R}^{2m+1}$ and $g : \pi\mathcal{U} \to \mathbb{R}^r$. Here $r = q - 2m - 1$ and $\mathbb{R}^r$ is the null space of π. In this setting $\mathcal{U}$ is the graph of g in the space $\mathbb{R}^q = \mathbb{R}^{2m+1} \times \mathbb{R}^{q-2m-1}$, i.e.,

$$\mathcal{U} = \{(x, gx) \mid x \in \pi\mathcal{U} \subset \mathbb{R}^{2m+1}\}. \tag{5.1}$$

5.1. Radial basis function approximations.

The basic problem that we have to solve—that of constructing the graph of g from data—is known as the interpolation problem. Specifically, we can state this as follows:

Given an ensemble of P input vectors $\{x^{(i)}\}_{i=1}^P$, with each $x^{(i)} \in \mathbb{R}^p$, and an associated ensemble of P output vectors, $\{y^{(i)}\}_{i=1}^P$, with each $y^{(i)} \in \mathbb{R}^r$, find a function $y : \mathbb{R}^p \to \mathbb{R}^r$ such that the interpolation condition

$$y(x^{(i)}) = y^{(i)} \tag{5.2}$$

is satisfied for all $i = 1 \dots P$.

It has been shown that an adaptive basis, in contrast with a fixed set of basis functions, can overcome the curse of dimensionality [2, 3]. For adaptive bases the number of functions required is proportional to the volume of space occupied by the data; for fixed bases the number of required functions grows exponentially with the dimension of the domain since essentially the entire space is covered.

RBF approximation [5] provides an adaptive basis given that the centers $\{c_i\}$ may be modified to suit the data set. An RBF approximation is written

$$y = w_0 + \sum_{j=1}^{N_c} w_j \phi(\|x - c_j\|), \tag{5.3}$$

where $\phi(\cdot)$ is a fixed function centered at the point c_j, N_c is the number of basis functions employed in the expansion, and the w_j are r-dimensional "weight vectors." In our investigation we restricted our attention primarily to the thin plate spline RBF

$$\phi(r) = r^2 \ln r \tag{5.4}$$

and the Gaussian RBF

$$\phi(r) = \exp(-r^2/\alpha). \tag{5.5}$$

The thin plate spline RBF provides a global set of basis functions while the Gaussian RBF is effectively local when α is small enough. In the case of the Gaussian RBF, the size of the domain of each function is effectively determined by α. It can be shown that every continuous function on a compact domain may be approximated uniformly by a linear combination of RBFs with centers in the domain. General criteria for such RBFs have been characterized [25].

A system of linear equations may be obtained by evaluating (5.3) at each of the P data points. In particular, the domain values $x^{(i)}$ consist of the image of the original data under the projection, i.e., $x^{(i)} \in \pi\mathcal{A}$, while the range values of g consist of the associated points in the orthogonal complement $y^{(i)} \in (I - \pi)\mathcal{A}$. At the ith data point we require

$$y^{(i)} = w_0 + \sum_{j=1}^{N_c} w_j \phi(\|x^{(i)} - c_j\|). \tag{5.6}$$

To determine the associated linear system we assemble the output data points in the data matrix $Y \equiv [y^{(1)}|y^{(2)}|\cdots|y^{(P)}]$. If we also define $\phi_j^{(i)} \equiv \phi(\|x^{(i)} - c_j\|)$, then the interpolation matrix Φ may be defined as

$$\Phi = \begin{pmatrix} 1 & \phi_1^{(1)} & \phi_2^{(1)} & \dots & \phi_{N_c}^{(1)} \\ 1 & \phi_1^{(2)} & \phi_2^{(2)} & \dots & \phi_{N_c}^{(2)} \\ 1 & \phi_1^{(3)} & \phi_2^{(3)} & \dots & \phi_{N_c}^{(3)} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & \phi_1^{(P)} & \phi_2^{(P)} & \dots & \phi_{N_c}^{(P)} \end{pmatrix}.$$

Thus we must solve the linear system

$$Y = W\Phi^T \tag{5.7}$$

for the unknown $q \times (N_c + 1)$ weight matrix

$$W = [w_0|w_1|w_2|\dots|w_{N_c}].$$

Formally, a solution may be obtained by computing the pseudoinverse of Φ giving

$$W^T = \Phi^\dagger Y^T. \tag{5.8}$$

This pseudoinverse $\Phi^\dagger$ is found by first computing the SVD of Φ

$$\Phi = U\Delta V^T$$

from which we obtain

$$\Phi^\dagger = V\Delta^\dagger U^T,$$

where $\Delta^\dagger$ is a diagonal matrix which has nonzero elements equal to the inverse of the significant singular values of Φ. To demonstrate the reconstructions obtained using the RBF approach we used projected data from both the KS equation and the digital faces. Our main objective was to determine the impact of the achieved tolerance κ_π on the quality of the reconstruction for a given number of centers. No attempt was made to optimize the reconstructions; in fact the RBF centers were simply chosen randomly from the data set. A particularly poor choice of centers may result in a spike in the reconstruction error. This effect can be ignored in that the same set of centers was used in all the expansions.

In our numerical experiments we do indeed observe a striking improvement in the data reconstructions as the minimum tolerance κ_π achieved is increased. For example, as shown in Figure 5.1, a data set of 100 points sampled from the limit cycle solution of the KS equation with $\alpha = 84$ was accurately reconstructed from the 2-dimensional secant basis but poorly reconstructed from the 2-dimensional data SVD basis. Again, the data SVD basis achieved a tolerance of only $\kappa_\pi = 0.03$ for 2 modes while the secant SVD basis achieved a very favorable tolerance of over $\kappa_\pi = 0.9$. This small κ_π is a significant impediment to the RBF's ability to approximate the inverse to any accuracy. In addition, the maximum reconstruction error over all the data points was significantly worse for the data SVD basis than for the adapted secant basis.

Given the similar minimum projected secant norms for the case of $\alpha = 87$ with either the secant basis or the adapted secant basis, we employed the former in our

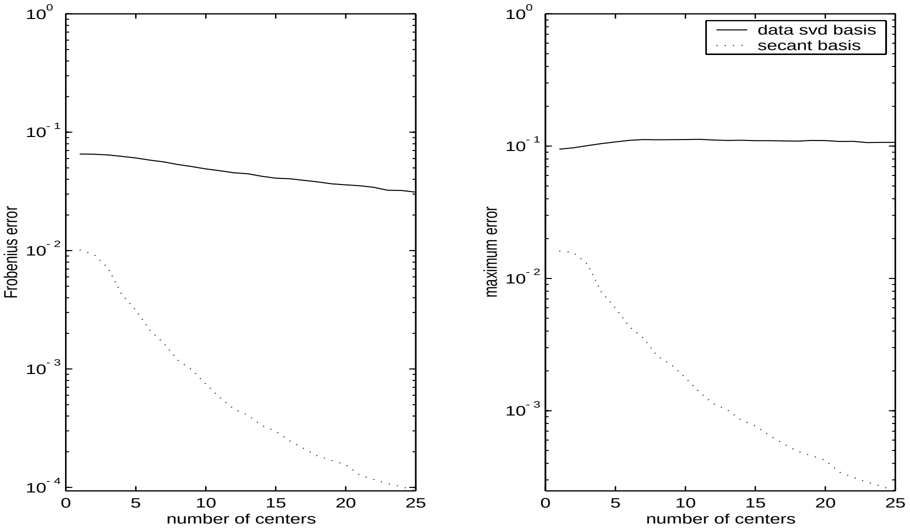

FIG. 5.1. $\alpha = 84$. *A comparison of reconstruction errors as a function of the number of centers for the* 2*-dimensional secant basis and* 2*-dimensional data SVD basis. Displayed on the left is the relative error based on the Frobenius norm of the reconstructed data matrix. On the right the maximum error is plotted. The radial basis function centers are selected randomly. All errors are averaged over* 25 *different sets of centers. Thin plate splines were used as the radial basis functions.*

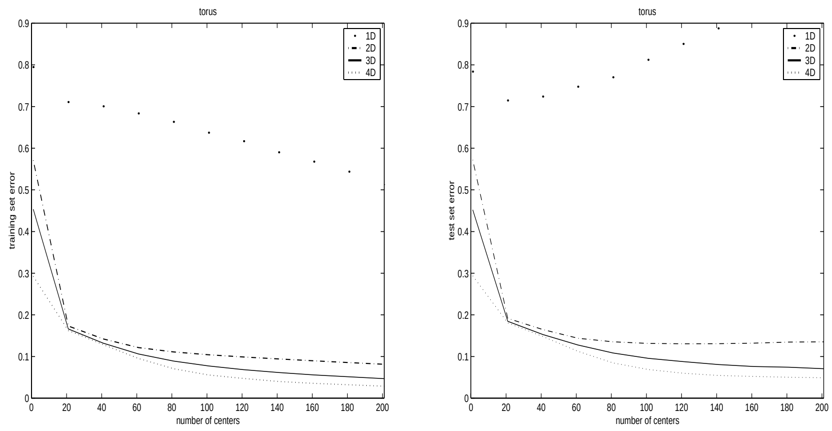

FIG. 5.2. $\alpha = 87$. *Left: Relative Frobenius error for reconstructed training data matrix. Right: Relative Frobenius error for reconstructed test data matrix. The top dotted curve shows the large errors associated with reconstructing a* 1*-dimensional projection. The reconstruction errors decrease for projections onto two, three, and four dimensions. The generalization error is approximately* 5% *for the test data set using* 200 *centers and a* 4*-dimensional domain.*

reconstructions. The Fourier basis achieved a tolerance of 0.25 using four dimensions. (Note that Whitney's theorem states that generically no more than five dimensions are required to embed a torus.) A data set of 500 points was collected for this experiment. The first 400 points were used for approximating the inverse and the last 100 reserved for testing the reconstruction error, i.e., they were used as a validation set. The results of the RBF constructions using 1–200 centers for projection dimensions of one, two, three, and four are shown in Figure 5.2.

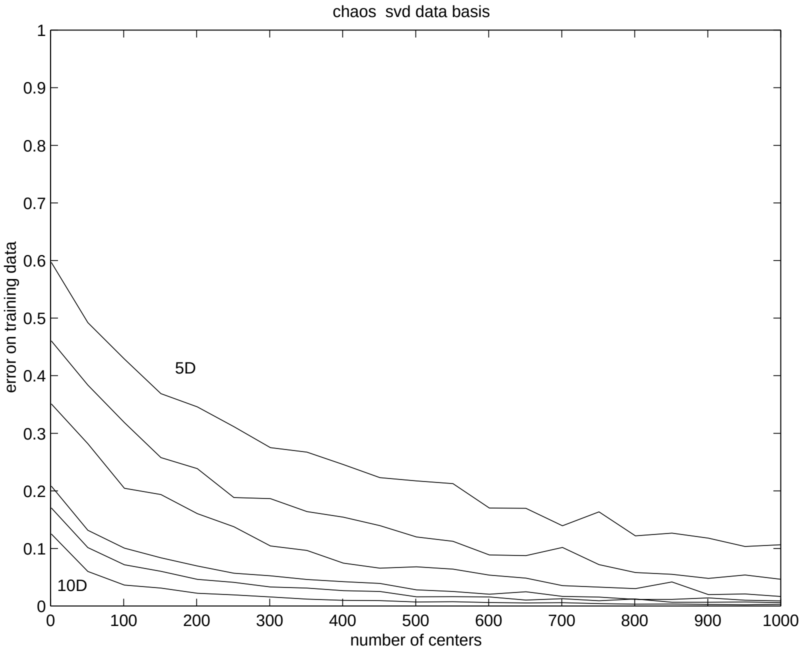

FIG. 5.3. *Reconstruction errors of the training data as a function of dimension for Fourier (sinusoidal) basis for KS data with $\alpha = 91$.*

The last reconstruction results we present correspond to the case $\alpha = 91$. Now 4000 data points were used for constructing the radial basis function inverse as well as a data SVD basis. In this example we contrast the results of using a basic linear reconstruction for a given dimension with our method which fits the residual of the linear reconstruction with radial basis functions. Thus, when the number of centers in the reconstruction is zero, the error corresponds to that produced by the optimal linear subspace only. For example, as shown in Figure 5.3, the reconstruction of the data matrix after a projection onto five dimensions results in a relative Frobenius error of 60%. Using 1000 centers to approximate the residual reduces the error to about 10%. As the projection dimension is increased the error due to the truncating in the best linear subspace is reduced, but the use of the radial basis functions always significantly improves the errors.

5.2. The Lipschitz condition.

An inverse is said to be well-conditioned if small perturbations in its domain produce changes in the range which are, in some sense, small. This issue arises here because we need to fit π^{-1} using discrete data. In particular, we are interested in fitting an inverse projection which interpolates sensibly between the known data points to give a good approximation to $\mathcal{U}$. We observe that, for this, it is not sufficient to require the mean square error or the maximum error—computed over the data—to be attained to within a desired tolerance. For good interpolation, or generalization, it is also necessary for the inverse to be well-conditioned.

One of our primary goals in requiring that points in the high-dimensional space be mapped such that

$$\|\pi x - \pi x'\| \geq \kappa_\pi \|x - x'\|$$

is to improve the conditioning of the inverse mapping. We may draw the connection as follows: let $y = \pi x$, $y' = \pi x'$. It follows that

$$\|\pi^{-1} y - \pi^{-1} y'\| \leq \kappa_\pi^{-1} \|y - y'\|,$$

that is, the inverse projection is Lipschitz with constant κ_π^{-1}. This condition captures the idea of conditioning of the inverse projection. In particular, we shall treat the Lipschitz constant, κ_π^{-1}, as a condition number, L, for π^{-1}:

$$L = \frac{1}{\kappa_\pi}. \tag{5.9}$$

There is, of course, a gap here between theory and practice. We have only an estimate of the Lipschitz constant of π^{-1} based on the discrete data set at our disposal. The expectation is that—given our assumption that $\mathcal{U}$ is smooth—a dense enough sampling of data will lead to a reliable estimate for π^{-1}—and hence L—extended to the whole of $\pi\mathcal{U}$.

Our proposal to use RBF models for estimating π^{-1} allows a comparison to be made between models and the data. RBF models are generally Lipschitz—indeed, for the choices of ϕ considered here, they are even smooth—so we can estimate the Lipschitz constant for an RBF model. We write $\pi^{-1}x = (x, g(x))$ where

$$g(x) = W \circ \tilde{\phi}(x) \quad \text{and} \quad \tilde{\phi} : \mathbb{R}^p \to \mathbb{R}^{N_c+1},$$

$\tilde{\phi}(x) = (1, \phi(\|x - c_1\|, \ldots, \phi(\|x - c_{N_c}\|))^T$ and $W : \mathbb{R}^{N_c+1} \to \mathbb{R}^r$ is linear. It follows that

$$\begin{aligned} \|\pi^{-1}x - \pi^{-1}x'\|^2 &= \|(x, g(x)) - (x', g(x'))\|^2 \\ &= \|x - x'\|^2 + \|W \circ (\tilde{\phi}(x) - \tilde{\phi}(x'))\|^2 \\ &\leq \|x - x'\|^2 + \|W\|^2 \|\tilde{\phi}(x) - \tilde{\phi}(x')\|^2 \\ &\leq \|x - x'\|^2 + \|W\|^2 L_\phi^2 N_c \|x - x'\|^2; \end{aligned}$$

then

$$\|\pi^{-1}x - \pi^{-1}x'\| \leq \|x - x'\| \sqrt{1 + L_\phi^2 \|W\|^2 N_c}.$$

Here N_c is the number of centers and L_ϕ is the Lipschitz constant for the selected RBF ϕ. (Since it is usual to use smooth ϕ, e.g., thin plate spline, cubic, Gaussian, multiquadric, we can estimate L_ϕ using the greatest magnitude of $D\phi$. In the case of the RBF fucntions with global support we must restrict our interest to a bounded domain.)

Because of our direct estimate from data of κ_π^{-1} we can expect that there is a pair x, x' such that $\|\pi^{-1}x - \pi^{-1}x'\| = \kappa_\pi^{-1}\|x - x'\|$. It follows that a necessary condition for the RBF to be able to model π^{-1} is that

$$\kappa_\pi^{-1} < \nu \equiv \sqrt{1 + L_\phi^2 \|W\|^2 N_c} \approx L_\phi \|W\| \sqrt{N_c}. \tag{5.10}$$

The quantity ν is a characteristic of the RBF model. In particular, the above inequality can be interpreted as a lower bound on the norm of the weight matrix. Note that

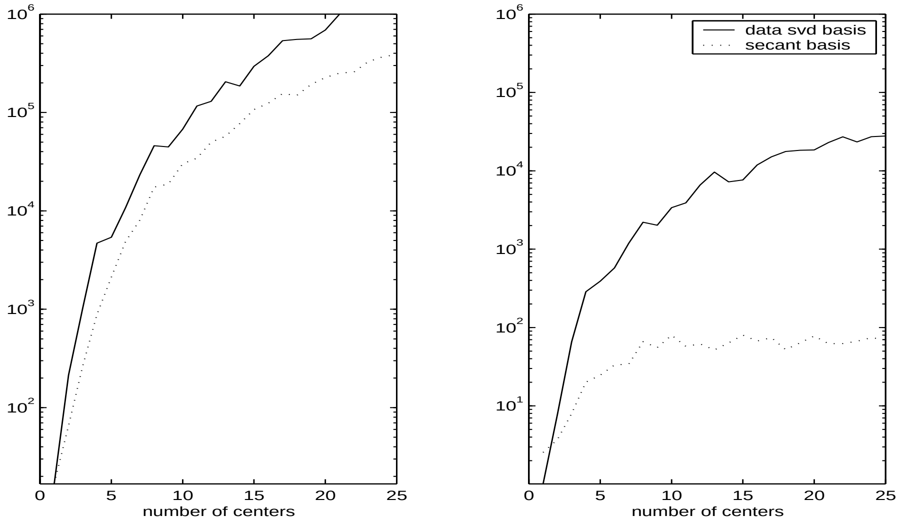

FIG. 5.4. $\alpha = 84$. *Left: Condition number of the interpolation matrix as a function of the number of centers in the radial basis function approximation. Right: Complexity of the RBF approximation as computed by* (5.10).

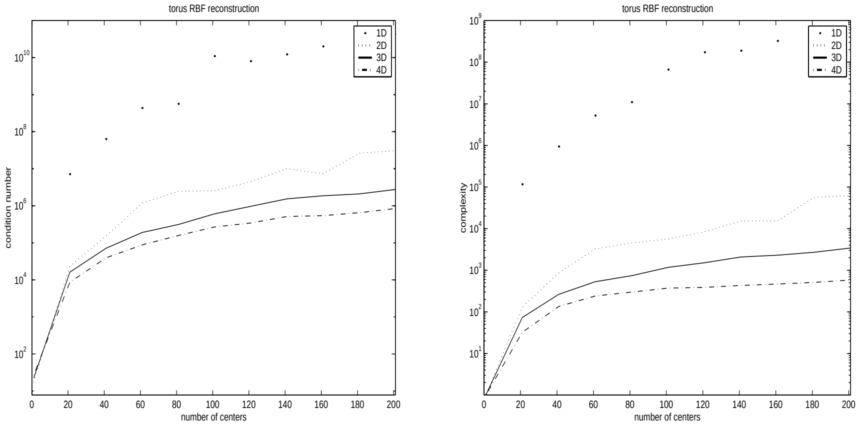

FIG. 5.5. $\alpha = 87$. *Left: Condition number of the interpolation matrix as a function of the number of centers in the radial basis function approximation. Right: Complexity of the RBF approximation as computed by* (5.10).

if the inverse we hope to estimate is ill-conditioned in the sense that L is large, this relationship shows that we need a weight matrix with large norm. Referring to (5.8), we see that it may be necessary to achieve this by including small singular values in the pseudoinverse of Φ, that is, by making the norm of $\Delta^\dagger$ large. If this is the case, the resulting model is unlikely to be robust. Note that this analysis applies to both the local and global representations. We anticipate its bound will be sharper in estimating the complexity of RBFs with local receptive fields, such as the Gaussian.

Given that solving the least squares problem (5.7) for the weight matrix W is generally ill-conditioned, the SVD approach is recommended. We observed in our calculations that the condition numbers of the interpolation matrices generally increased with the number of centers. However, the condition number appears to be inversely proportional to the minimum projected secant norm of the RBF input data. See, for example, Figure 5.4 corresponding to the RBF reconstructions calculated in Figure 5.1, i.e., the KS limit cycle data. Note that in this example the 2-dimensional projection onto the secant basis has a much larger minimum projected secant norm than the projection on the data SVD basis. Another interesting side effect of increasing the value of κ_π achieved for a basis is the reduction of complexity in the reconstruction. The measure of complexity ν, as defined by (5.10), is plotted for the RBF reconstruction of the limit cycle data on the right in Figure 5.4. In this example the Lipschitz constant given in (5.9) is approximately $L = 1.1$ for the inverse of the adapted secant basis projection and $L = 33$ for the data SVD projection. Thus, the bound ν is not very sharp for this problem. It is interesting to observe that our measure of complexity appears to level off with the secant basis but continue to increase as a function of the number of centers for the ill-conditioned reconstruction.

The KS torus example shows the most dramatic improvement in condition number going from a 1-dimensional projection to a 2-dimensional projection; see Figure 5.5. This is not surprising given that the minimum projected secant norm is zero for 1-dimensional projections to either the Fourier basis or adapted secant basis and improves to approximately 0.05 (Fourier basis) and 0.1 (adapted secant basis) for the corresponding 2-dimensional projections. Further increases in κ_π as a function of dimension are correlated with a reduction in the complexity bound ν as well as the condition number of the interpolation matrix; see Figure 5.5.

6. A comparison of reduction techniques.

Whitney's embedding theorem permits a relative classification of reduction procedures according to their generally attainable reduction dimensions

$$q > q' \geq d \geq m' \geq m, \tag{6.1}$$

where m is the topological dimension of the manifold $\mathcal{U}$. The minimum reduction dimension that can be attained generically, according to Whitney's embedding theorem, for the procedure outlined in this paper is $d = 2m + 1$. If both G and H are linear, the attainable reduction dimension will be designated q' while if both G and H are nonlinear then we will denote the reduction dimension by m'.

6.1. Global methods.

We deem a reduction method to be global if the support of the basis functions is the entire domain.

Global Case I. G linear, H linear. For global linear reduction and reconstruction mappings it is generally not possible to attain the Whitney limit $2m+1$. In fact, depending on the distribution of the data set in the ambient space, the lowest attainable reduction dimension q' may be far greater than the Whitney limit.

Several well-known related techniques, including principal component analysis (PCA), the Karhunen–Loève (KL) procedure, and the SVD, determine optimal mappings G and H over all orthogonal, or unitary, transformations. These linear methods use bases which serve to encapsulate, or span, the data.

Global Case II. G linear, H nonlinear. This case is the main subject of this paper. As described in section 3 Whitney's theorem guarantees that, generically, a projection π restricted to $\mathcal{U}$ is an embedding of $\mathcal{U}$ in $\mathbb{R}^{2m+1}$, so we know there is an inverse $\pi^{-1} : \pi\mathcal{U} \subset \mathbb{R}^{2m+1} \to \mathbb{R}^q$. This nonlinear inverse affords reduction dimensions at least as small as $2m+1$ where m is the topological dimension, a significant improvement on using an optimal linear inverse. We may further distinguish this approach from

Case I above, by observing that since the inverse is nonlinear the data are being parameterized rather than spanned.

This approach is based on the concept of a "good projection." Random projections, or even projections based on optimal bases, such as PCA, do not ensure in any sense that the inverse map will be well-conditioned. Note also the reduction mapping, i.e., projection π, is determined independently of the inverse π^{-1}. In practice this leads to improved scalability over the case where both maps are nonlinear (see below).

Global Case III. G nonlinear, H nonlinear. When both G and H are nonlinear it may be possible to obtain a reduction dimension of $m' < 2m + 1$ thus improving upon the Whitney limit. For example, it may not be possible to embed a given circle in the plane via a projection. (Whitney's theorem gives an upper bound of dimension $d = 3$.) However, every circle is homeomorphic to the unit circle and thus there always exist nonlinear G and H which will achieve this embedding.

Nonlinear reduction may be implemented as a nonlinear autoassociative, or bottleneck, neural network as proposed in [22, 24]. This network has also been applied to the modeling and analysis of dynamical systems; see, e.g., [14, 18, 20]. The mappings G and H are represented by sigmoidal feedforward neural networks and are trained such that the composition $H \circ G$ approximates the identity mapping by minimizing the standard mean square error. Note that the range, or output, of G is not specifically known during training and that the functions G and H are being adapted simultaneously during the training phase (a nonlinear optimization problem). This approach, while powerful in theory, appears to be ill-conditioned [16]. We note also the inherent nonuniqueness of the solution, given that if $H \circ G$ minimizes the mean square error, so does $H \circ F \circ F^{-1} \circ G$. Also, this method, often referred to as nonlinear principal component analysis (NLPCA), typically only optimizes the mean square error, although it is possible to incorporate user-defined constraints in NLPCA [20].

6.2. Local methods. Local methods are based on constructing an array of reduction and reconstruction mappings defined over local regions. For example, the local regions may consist of Voronoi polyhedra which are determined via a variety of clustering algorithms. Local methods are especially attractive given that the computations associated with creating models for each local region may be done in parallel.

Local Case I. G linear, H linear. In this case the SVD may be applied to the data contained in each Voronoi region. The primary advantage is that the reduction to the local topological dimension m may be achieved in theory. If the data lie on a smooth manifold, the tangent space (and its dimension) may be determined by assessing which of the singular values scale linearly with the radius of the ball as proposed in [6].

The primary advantage of the local linear-linear method is that it permits a very fast reduction scheme once the Voronoi regions have been determined. The efficiency of the method is due to the fact that the (linear) inverse mapping is immediately available via the SVD. The main disadvantage of the method is that to actually achieve the theoretical reduction to the topological dimension m, the linear reconstruction requires a very dense partitioning of the ambient space to approach the accuracy of reconstruction which is possible when a (local) nonlinear method is employed. Examples of applications of local principal component analysis, or local SVD, include [7, 6, 4, 12].

Local Case II. G linear, H nonlinear. Again, for *locally* defined mappings G and H it is possible to obtain a reduction of the data to the topological dimension [8]. The primary benefit of employing a nonlinear reconstruction mapping H is the significant reduction of the number of local regions required to reconstruct $\mathcal{U}$ to a specified accuracy. The additional expense incurred by this procedure results from the fact that the nonlinear mapping must be approximated, e.g., using RBFs. We note that the methodology proposed in this paper for global reductions may also be employed locally where G is linear and H is nonlinear.

An example of this class of local method, referred to as *neural charts*, was developed in [10]. It proposes an initial (tree-based) clustering of the data set. The reduction space may be determined using the scaling laws proposed in [6]. The nonlinear inverse may be determined using multilayer perceptrons or using RBFs on the locally defined regions.

7. Summary and future work. We have presented a new approach for dimensionality reduction of data sets which exploits Whitney's theorem from differential topology. This theorem indicates the theoretically attainable reduction dimension to be $2m+1$ for the case that the data are sampled from an m-dimensional submanifold. Motivated by the proof of this theorem, we propose an adaptive secant algorithm which produces increasingly good projections. A consequence of this construction is that the nonlinear inverse or reconstruction mapping is Lipschitz and the value of the Lipschitz constant is inversely proportional to the quality, or tolerance, achieved by the projection.

To put this work in context, a comparison of basic reduction procedures was made based on the nature of the mappings G and H; the theoretically attainable dimensions were indicated in each case. Given the differences in computational expenses among the methods it is appropriate to combine the procedures to form hybrid methods. For example, a global linear reduction may be used as a preprocessing stage to obtain an initial reduction to dimension q'. Following this by the linear/nonlinear global reduction, the subject of this paper, will result in a further reduction to dimension $2m+1$. To achieve even further reduction this could be followed either by a global nonlinear reduction procedure, such as a bottleneck neural network to attain the dimensionality m', or a local reduction procedure, such as neural charts, to obtain a parameterization in terms of the local intrinsic dimension m.

The reduction procedure presented here was developed as a global method. However, the basic approach may be applied locally to produce a reduction of the data. In particular, locally defined good projections may be calculated such that the inverse is especially easy to reconstruct. Furthermore, the number of points associated with each locally defined region should be small enough to greatly speed up the search for good projections. This work will be the subject of future investigations.

REFERENCES

[1] M. Barnsley, *Fractals Everywhere*, Academic Press, Boston, 1988.

[2] A. R. Barron, *Universal approximation bounds for superpositions of a sigmoidal function*, IEEE Trans. Neural Networks, 39 (1993), pp. 930–945.

[3] A. R. Barron, *Approximation and estimation bounds for artificial neural networks*, Mach. Learning, 14 (1994), pp. 115–133.

[4] D. Broomhead, R. Indik, A. Newell, and D. Rand, *Local adaptive Galerkin bases for large-dimensional dynamical systems*, Nonlinearity, 4 (1991), pp. 159–197.

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

[6] D. S. Broomhead, R. Jones, and G. P. King, *Topological dimension and local coordinates from time series data*, J. Phys. A, 20 (1987), pp. L563–L569.

[7] K. Fukunaga and D. Olsen, *An algorithm for finding intrinsic dimensionality of data*, IEEE Trans. Comput., C-20 (1971), pp. 176–183.

[8] V. Guillemin and A. Pollack, *Differential Topology*, Prentice–Hall, Englewood Cliffs, NJ, 1974.

[9] M. W. Hirsch, *Differential Topology*, Grad. Texts in Math. 33, Springer, New York, 1976.

[10] D. Hundley, M. Kirby, and R. Miranda, *Empirical dynamical system reduction* II*: Neural charts*, in Semi-Analytic Methods for the Navier-Stokes Equations, K. Coughlin, ed., CRM Proc. Lecture Notes 20, AMS, Providence, RI, 1999, pp. 65–83.

[11] J. M. Hyman, B. Nicolaenko, and S. Zaleski, *Order and complexity in the Kuramoto-Sivashinsky model of weakly turbulent interfaces*, Phys. D, 23 (1986), pp. 265–292.

[12] N. Kambhatla and T. K. Leen, *Dimension reduction by local principal component analysis*, Neural Comput., 9 (1997), pp. 1493–1516.

[13] I. G. Kevrekidis, B. Nicolaenko, and J. C. Scovel, *Back in the saddle again: A computer assisted study of the Kuramoto-Sivashinsky equation*, SIAM J. Appl. Math., 50 (1990), pp. 760–790.

[14] I. Kevrekidis, R. Rico-Martinez, R. Ecke, R. Farber, and A. Lapedes, *Global bifurcations in Rayleigh-Bénard convection. Experiments, empirical maps and numerical bifurcation analysis*, Phys. D, 71 (1994), pp. 342–362.

[15] M. Kirby, *Minimal dynamical systems from partial differential equations using Sobolev eigenfunctions*, Phys. D, 57 (1992), pp. 466–475.

[16] M. Kirby, *Ill-conditioning and gradient based optimization of multi-layer perceptrons*, in Mathematics in Signal Processing IV, J. McWhirter and I. Proudler, eds., Inst. Math. Appl. Conf. Ser. 67, Oxford University Press, London, 1998, pp. 223–237.

[17] M. Kirby and D. Armbruster, *Reconstructing phase-space from PDE simulations*, Z. Angew. Math. Phys., 43 (1992), pp. 999–1022.

[18] M. Kirby and R. Miranda, *Nonlinear reduction of high-dimensional dynamical systems via neural networks*, Phys. Rev. Lett., 72 (1994), pp. 1822–1825.

[19] M. Kirby and R. Miranda, *The remodeling of chaotic dynamical systems*, in Intelligent Engineering Through Artificial Neural Networks 4, S. Dagli, B. Fernandez, J. Ghosh, and R. S. Kumara, eds., The American Society of Mechanical Engineers, New York, 1994, pp. 831–836.

[20] M. Kirby and R. Miranda, *Empirical dynamical system reduction* I*: Global nonlinear transformations*, in Semi-Analytic Methods for the Navier-Stokes Equations, K. Coughlin, ed., CRM Proc. Lecture Notes 20, AMS, Providence, RI, 1999, pp. 41–64.

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

[22] M. A. Kramer, *Nonlinear principal component analysis using autoassociative neural networks*, AIChE J., 37 (1991), pp. 233–243.

[23] P. Lancaster and M. Tismenetsky, *The Theory of Matrices*, Academic Press, New York, 1985.

[24] E. Oja, *Data compression, feature extraction, and autoassociation in feedforward neural networks*, in Artificial Neural Networks, T. Kohonen, K. Mäkisara, O. Simula, and J. Kangas, eds., Elsevier Science Publishers, New York, 1991, pp. 737–745.

[25] M. Powell, *The theory of radial basis functions in* 1990, in Advances in Numerical Analysis II: Wavelets Subdivision and Radial Basis Functions, W. Light, ed., Oxford University Press, London, 1992, pp. 105–210.

[26] L. Sirovich and M. Kirby, *A low-dimensional procedure for the characterization of human faces.*, J. Opt. Soc. Amer. A, 4 (1987), pp. 519–524.