Principal dynamical components

MANUEL D. DE LA IGLESIA
Courant Institute

AND

ESTEBAN G. TABAK
Courant Institute

Abstract

A procedure is proposed for the dimensional reduction of time series. Similarly to principal components, the procedure seeks a low-dimensional manifold that minimizes information loss. Unlike principal components, however, the procedure involves dynamical considerations, through the proposal of a predictive dynamical model in the reduced manifold. Hence the minimization of the uncertainty is not only over the choice of a reduced manifold, as in principal components, but also over the parameters of the dynamical model, as in autoregressive analysis and principal interaction patterns. Further generalizations are provided to non-autonomous and non-Markovian scenarios, which are then applied to historical sea-surface temperature data.

1 Introduction

Complex systems typically involve a large number of degrees of freedom. Thus to elucidate the fundamental mechanisms underlying one such system's behavior, one may consider its projection onto smaller-dimensional manifolds, selected so as to capture as much of the dynamics as possible. A tool frequently used for this purpose is principal components [12], whereby a linear subspace of prescribed dimensionality of the phase-space of observations is sought, so as to maximize the amount of the variability that is preserved when the data are projected onto it.

Given a dataset z_j, $j \in [1,\dots,N]$, where each observation z_j consists of n real numbers, its first m ($m \leq n$) principal components are given by $x_j = Q_x'(z_j - \bar{z})$, where $\bar{z}$ is the mean value of z, and Q_x is an $n \times m$ matrix with orthonormal columns, chosen so that $\sum_{j=1}^{N} \left\| (z_j - \bar{z}) - Q_x x_j \right\|^2$ is as small as possible. From a statistical perspective, among all m-dimensional subspaces, x is the one whose knowledge minimizes the uncertainty of z. The matrix Q_x consists of the first m columns of U in the singular value decomposition

$$Z' = USV',$$

where the elements Z_j^i of the matrix $Z' \in R^{n \times N}$ contain the ith component of the jth observation minus its mean value $\bar{z}^i$ over all observations, $U \in R^{n \times n}$ and $V \in R^{N \times N}$

Communications on Pure and Applied Mathematics, Vol. 000, 0001–0034 (2000)

are orthogonal matrices, and $S \in R^{n\times N}$ is the diagonal matrix of singular values of Z', the eigenvalues of the empirical covariance matrix $C = Z'Z$ sorted in decreasing order.

In the probabilistic scenario underlying this procedure, the z_j's are independent samples of a Gaussian distribution $\mathcal{N}(\mu,\Sigma)$, $\bar{z}$ is an estimate for its mean μ, and the principal components estimate the principal axes of the covariance matrix Σ, sorted in decreasing order by the fraction of the total variance that they explain. Yet principal components are often sought for data that do not quite fit this scenario. Of particular concern to us here is the situation where the z_j's form a time series, representing snapshots of the vector z at equidistant times t_j. In this context, the dimensional reduction by principal components, oriented toward data compression, lacks any concept of *dynamics*: the various snapshots z_j are treated as independent observations, which renders immaterial even the order in which they are sorted. If there is an underlying dynamics, this is neither unveiled nor exploited by the analysis.

An example is provided by the Empirical Orthogonal Functions (EOFs) [8, 16, 22, 23] –the name given to principal components in climate studies–, which take a time series of atmospheric or oceanic data, subtract its time average or climatology, and find those modes that explain the largest share of its variability. These modes may then be assigned a dynamical interpretation, yet no dynamics ever entered into their calculation: just the static variability of the data, treated as a series of independent, unsorted observations.

In this paper, we develop an alternative methodology, highly reminiscent of the principal-component framework, but with a dynamical core. We seek, as in principal components, a hierarchy of manifolds, that we name "principal dynamical components". Attached to these manifolds is a model of predictive dynamics. The cost function to minimize has, as in principal components, the variability in the unrepresented variables, but also the fraction of the variability in the preserved variables that is not explained by the dynamics. Thus the dynamical components are characterized not by capturing most of the system's variability, but by explaining dynamically its largest possible share. Hence this methodology can be thought of as a blend of autoregression analysis [1, 27], which is used as a reduced dynamical model, and principal components, though the criterium for selecting a reduced manifold differs from the latter's. This proposal has much in common with that of principal interaction patterns, proposed by Hasselman and further developed by Kwasniok [9, 13, 14, 15], since both address simultaneously dynamics and dimensional reduction.

Various other approaches have been pursued to build low-dimensional dynamical models from time series, such as the Box-Jenkins methodology [1] in econometry, singular spectrum analysis [2, 6], reduced stochastic models [11, 17], continuous Markov chains [4], regime identification techniques [5, 10], balanced truncation methods [7, 19] and nonlinear principal component analysis [3, 18, 20, 21].

There is no way we can do justice in this introduction to such a rich and fast-growing field. The methodology proposed in this paper has, in our view, the virtue of conceptual simplicity: a low-dimensional manifold and associated dynamical model are sought, so as to minimize the 2-norm of the predictive uncertainty over the given time-series.

We first present the new methodology as a natural extension of principal components, in a linear, autonomous framework, with a dynamic manifold given by $x = Q_x' z$, where Q_x is a fixed $n \times m$ orthogonal matrix, and the dynamics by $x_{j+1} = Ax_j$, where A is another fixed, $m \times m$ matrix. The definition of principal dynamical components results in a minimization problem over both Q_x and A, considered in detail in the appendix. Section 2 presents this problem and provides an efficient methodology to solve it. Yet many real problems are not autonomous: climate dynamics, for instance, is season-dependent. In Section 3 we extend the methodology to non-autonomous situations and, more generally, to accommodate for the presence of exogenous variables and external controls, that appear in many engineering applications. Here Q_x and A depend on time and on those external variables. Section 4 extends the procedure further to handle non-Markovian processes, where the dynamics involves more than the immediate past. We illustrate the procedure throughout with synthetic data and, in Section 5, we concern ourselves with a real application to time series of sea-surface temperature over the ocean. Section 6 gives a probabilistic interpretation of the principal dynamical component procedure, which provides a conceptual extension to general nonlinear, non-Gaussian settings. The development of effective algorithms for the numerical implementation of this broad generalization will be pursued elsewhere.

2 The linear, autonomous framework

The probabilistic set-up for principal component analysis consists of independent observations drawn from a Gaussian distribution. The natural extension to time series has observations z_j , $j \in [1,\ldots,N]$ drawn from the linear Markovian dynamics

$$z_{j+1} = \mathcal{N}(A^z z_j, \Sigma^z).$$

Here the matrix A^z models autocorrelation, and $\mathcal{N}$ represents a Gaussian process with mean $A^z z_j$ and covariance matrix Σ^z. Neither A^z nor Σ^z are known to us; instead, we seek an m-dimensional manifold $x = Q_x' z$ and reduced dynamics

$$\tilde{x}_{j+1} = Ax_j,$$

such that the *predictive* uncertainty or cost

$$c = \sum_{j=1}^{N-1} \left\| z_{j+1} - Q_x \tilde{x}_{j+1} \right\|^2 = \sum_{j=1}^{N-1} \left\| z_{j+1} - Q_x A Q_x' z_j \right\|^2$$

is minimal. This is the conceptual basis of what we shall denote *linear autonomous principal dynamical component* analysis.

It is convenient to introduce y for the orthogonal complement of x, so that

$$z = [Q_x Q_y] \begin{pmatrix} x \\ y \end{pmatrix},$$

where $Q = [Q_x Q_y]$ is an orthogonal matrix. Since the dynamics of y is not explained by the model, we have $\tilde{y}_{j+1} = 0$.

2.1 Two-dimensional case

The simplest scenario, appropriate for a first view of the proposed algorithm, has the observations z_j in a two-dimensional space, $n = 2$, and seeks a reduced manifold x of dimension $m = 1$. We introduce the following notation:

$$z = \begin{pmatrix} A \\ P \end{pmatrix},$$

where, mimicking an application to climate dynamics, A stands for Atlantic and P for Pacific spatially-averaged sea-surface temperatures,

$$x = A\cos(\theta) + P\sin(\theta),$$
$$y = -A\sin(\theta) + P\cos(\theta),$$

where the angle θ defines the direction of the dynamic component x in (A, P) space, and

$$\tilde{x}_{j+1} = a x_j,$$

with the stretching factor a describing the deterministic component of the reduced dynamics.

The cost function adopts the form

$$\begin{aligned} c(\theta, a) &= \sum_{j=1}^{N-1} \left\| \begin{pmatrix} A_{j+1} - \tilde{A}_{j+1} \\ P_{j+1} - \tilde{P}_{j+1} \end{pmatrix} \right\|^2 = \sum_{j=1}^{N-1} \left\| \begin{pmatrix} x_{j+1} - \tilde{x}_{j+1} \\ y_{j+1} - \tilde{y}_{j+1} \end{pmatrix} \right\|^2 \\ &= \sum_{j=1}^{N-1} \left\| \begin{pmatrix} x_{j+1} - a x_j \\ y_{j+1} \end{pmatrix} \right\|^2 = \sum_{j=1}^{N-1} (y_{j+1})^2 + (x_{j+1} - a x_j)^2 . \end{aligned}$$

By contrast, the corresponding cost function for regular principal components in this 2-dimensional scenario is

$$c_{pc}(\theta) = \sum_{j=1}^{N} {y_j}^2 :$$

the amount of variability in the unrepresented variable y.

The minimization of c can be solved iteratively: in each step, we first update the parameter θ by second order descent of c with a fixed, and then a by minimizing c with θ fixed. This minimization procedure is described in detail in appendix A.1.

As an illustration, we created data from the dynamical model

$$\begin{aligned} x_{j+1} &= a x_j + r_x \eta_j^x, \\ y_{j+1} &= r_y \eta_j^y, \end{aligned}$$

for $j = 1, \ldots, N-1$ (the initial values x_1 and y_1 are picked at random), where $N = 1000$, the $\eta_j^{x,y}$ are independent samples from a normal distribution, and we adopted the values $a = 0.6$ for the dynamics[1], and $r_x = 0.3$ and $r_y = 0.6$ for the amplitudes of the noise in x and y. Then we rotated the data through

$$\begin{aligned} A_j &= x_j \cos(\theta) - y_j \sin(\theta), \\ P_j &= x_j \sin(\theta) + y_j \cos(\theta), \end{aligned}$$

with $\theta = \frac{\pi}{3}$, and provided the A_j and P_j as data for the principal dynamical component routine. The results are displayed in Figure 2.1. The first plot shows the "observations" in the plane (A, P). These are treated as independent samples in a regular principal component analysis; we keep instead track of their sequential order, represented by the dotted lines in the plot. For this data, the first regular principal component, drawn in black, is in fact orthogonal to the principal dynamical component, drawn in green. The reason is that the total variability has a larger y-component, due to the bigger amplitude of the noise in y, while all the variability that is explainable dynamically is in x. This is an extreme example where regular principal components yield a leading mode that is absolutely irrelevant from a dynamical viewpoint. The other three plots in the figure display the evolution of the estimates for a and θ and the cost function c, as functions of the step-number. The dotted lines, drawn for reference, have the exact values of a and θ in the data, as well as the unexplainable part of the cost, $c = \frac{1}{N-1} \sum_{j=1}^{N-1} \left(r_x \eta_j^x \right)^2 + \left(r_y \eta_j^y \right)^2 \approx r_x^2 + r_y^2 = 0.45$. Notice the fast convergence to the exact solution, that in this example took 14 steps.

2.2 The multidimensional case

For dimensions n bigger than two, we write

$$z = [Q_x Q_y] \begin{pmatrix} x \\ y \end{pmatrix},$$

and

$$x_{j+1} = A x_j .$$

The minimization problem that defines $Q = [Q_x Q_y]$ and A is

$$\min_{Q,A} c = \sum_{j=1}^{N-1} \left\| z_{j+1} - Q \begin{pmatrix} A Q_x{}' z_j \\ 0 \end{pmatrix} \right\|^2 = \sum_{j=1}^{N-1} \left\| \begin{pmatrix} x_{j+1} - A x_j \\ y_{j+1} \end{pmatrix} \right\|^2 .$$

Notice that Q and A are not uniquely defined: any pair of orthogonal bases for the optimal subspaces represented by x and y will give rise to different Q's and A's representing the same dynamics. The algorithm proposed below walks nicely around this degeneracy, avoiding unnecessary re-parameterizations of the two subspaces[2].

[1] The value of $|a|$ needs to be smaller than one for the time series not to blow up.

[2] After performing the optimization, one can, if desired, resolve the degeneracy in the description of the dynamical manifold by choosing a natural basis for x, such as the one made out of the

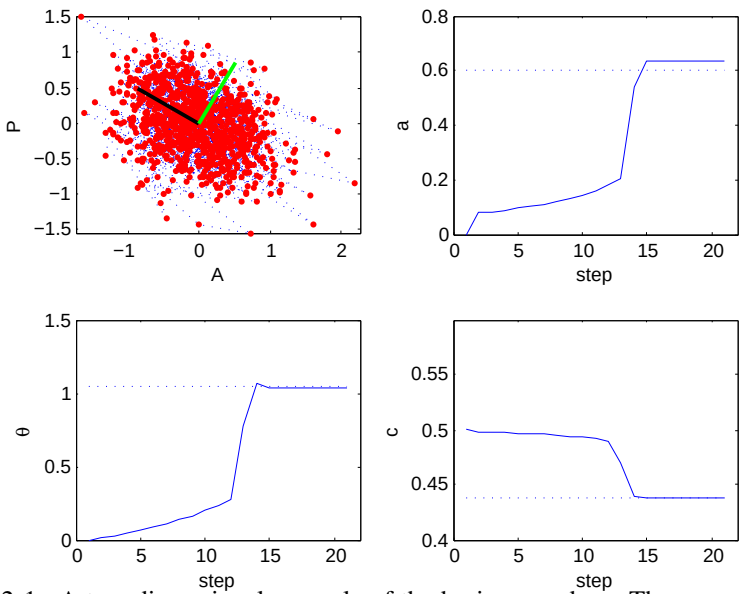

FIGURE 2.1. A two-dimensional example of the basic procedure. The first plot displays the data points, with dotted lines joining successive observations, and the directions for the first regular principal component –in black– and the principal dynamical component –in green–, which in this case are orthogonal to each other. The other three plots show the evolution of the estimates for the parameters a and θ for the dynamics and reduced manifold, and of and the cost function c –normalized by $(N-1)$– as functions of the step-number, with their exact values as dotted lines.

As in the two-dimensional scenario, the most straightforward methodology, described in appendix A.2, decouples the descent steps for A and Q. For the descent steps in Q, we propose two different approaches. In the first, described in appendix A.2, we perform an elementary rotation at a time, using the Givens rotation matrix for a plane picked at random. Such approach can be costly in high dimensions: seeking the optimal manifold through two-dimensional rotations requires the use of many random planes. It would be far more effective if one could perform high-dimensional orthogonal transformations at once, differentiating not with respect to the angle θ in an arbitrary plane, but with respect to a general orthogonal matrix Q. This is what we call a Lie algebra approach, described in detail in appendix A.2.

To create a simple synthetic example for the multidimensional case using a Givens rotation approach, we chose $n = 5$ and $m = 2$, and created data from the

principal components of A. For non-normal A's, there are two such bases: the eigenvectors of $A'A$ and those of AA'. Both are significant and sorted by sensitivity to perturbations: the former gives the directions where initial perturbations yield the highest effect; the latter, the directions where these effects manifest themselves after a time-step.

dynamical model

$$x_{j+1} = Ax_j + r_x \eta_j^x,$$
$$y_{j+1} = r_y \eta_j^y,$$

for $j = 1,\ldots,N-1$, where $N = 1000$, the $\eta_j^{x,y}$'s are independent samples from a normal distribution –two and three dimensional vectors respectively– and we adopted arbitrarily the values

$$A = \begin{pmatrix} 0.4569 & 0.3237 \\ -1.0374 & 1.0378 \end{pmatrix}$$

for the dynamics, and $r_x = 0.3$ and $r_y = 0.6$ for the amplitudes of the noise in x and y. Then we rotated the data through an arbitrary orthogonal matrix,

$$z = [Q_x Q_y] \begin{pmatrix} x \\ y \end{pmatrix},$$

with

$$Q_x = \begin{pmatrix} -0.7044 & -0.3823 & -0.3407 & -0.1985 & -0.4497 \\ 0.5754 & -0.1555 & -0.1798 & 0.2477 & -0.7423 \end{pmatrix}',$$

and generated the data displayed in the first panel of Figure 2.2. Running our algorithm on these data yields estimates A^* and Q_x^* for A and Q_x that, as remarked before, are not univocally defined. Indeed, the algorithm found

$$A^* = \begin{pmatrix} 0.6505 & 0.2401 \\ -1.1591 & 0.8685 \end{pmatrix}$$

and

$$Q_x^* = \begin{pmatrix} -0.8143 & -0.3258 & -0.2896 & -0.2545 & -0.2865 \\ 0.4089 & -0.2108 & -0.2570 & 0.1873 & -0.8290 \end{pmatrix}',$$

quite different in appearance from their exact values above.

To verify that Q_x and Q_x^* span the same plane and that A and A^* represent the same transformation in the corresponding coordinates, we project the two columns of Q_x^* onto the space spanned by those of Q_x, through the projection $P(Q_x^*)' = BQ_x'$, with $B = (Q_x^*)' Q_x$, and define the relative errors

$$e_Q = \frac{\| (Q_x^*)' - BQ_x' \|}{\|Q_x\|}, \quad e_A = \frac{\|A - B^{-1}A^*B\|}{\|A^*\|},$$

which vanish only if the two pairs of matrices represent exactly the same reduced manifold and dynamics.

The results are displayed in Figure 2.2. The first plot shows the first three components of the data points z_j. The second plot displays the evolution of the normalized cost function c as a function of the step-number, with the dotted line displaying the exact value of the unexplainable part of the cost, $c_* = \frac{1}{N-1} \sum_{j=1}^{N-1} \left(r_x \|\eta_j^x\| \right)^2 +$

$\left(r_y \|\eta_j^y\|\right)^2 \approx 2r_x^2 + 3r_y^2 = 1.26$. The third and fourth plots display the evolution of the errors e_Q and e_A defined above. Notice again the fast convergence of the algorithm to the exact solution.

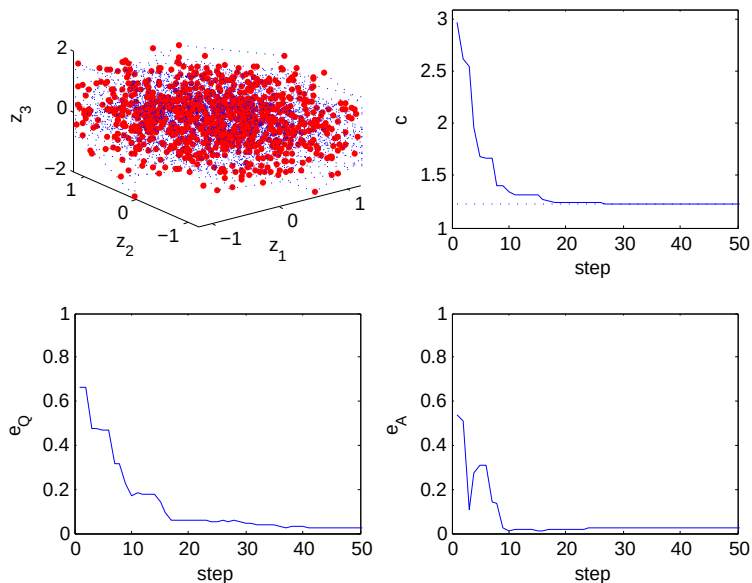

FIGURE 2.2. A multidimensional, autonomous example, with $n = 5$ and $m = 2$. The first plot displays the first three coordinates of the data points, with dotted lines joining successive observations, the second plot shows the evolution of the cost function, with its exact value as a dotted line, and the third and fourth plots display the evolution of the errors e_Q and e_A.

As an example of the descent procedure through general orthogonal transformations, we display in Figure 2.3 the results of a run similar to the one in Figure 2.2, but with much higher dimensionality: 5000 snap shots of an 80-dimensional vector z, of which a reduced 17-dimensional dynamical submanifold x is sought. We have adopted isotropic noise, $r_x = r_y = 0.3$. A few things are worth noticing in this run. On the one hand, the fast convergence: in around 50 steps one reaches a manifold which captures essentially all the dynamical information. This number of steps is to a large degree dimension independent: it follows from the choice of a learning rate, $\varepsilon = 0.1$ for this run, and a typical angle to rotate, of order $\pi/2$ (this accounts for the first 10-20 steps; the learning rate decreases exponentially as one approaches the optimal solution). Also worth noticing are three related phenomena: that the routine finds a total cost slightly below the exact one underlying the data, that the manifold it converges to is close to but not exactly the one the data have been built from, and that the matrix representing the dynamics, though not

exact either, is closer to the one proposed than the manifold where it acts. All three facts share the same explanation: since we have a relative small number of observations, $N = 5000$, for a space of dimension $n = 80$, random autocorrelation among small dimensional submanifolds of the data may overcome the smallest eigenvalues of the underlying matrix A. Hence more variability can be explained by this random autocorrelation than by those smallest eigenvalues, thus accounting for the smaller value of the cost function found and the discrepancy among the manifolds. The fact that the discrepancy in A is smaller follows from the fact that the directions associated with those smallest eigenvalues have little weight in the norm of A.

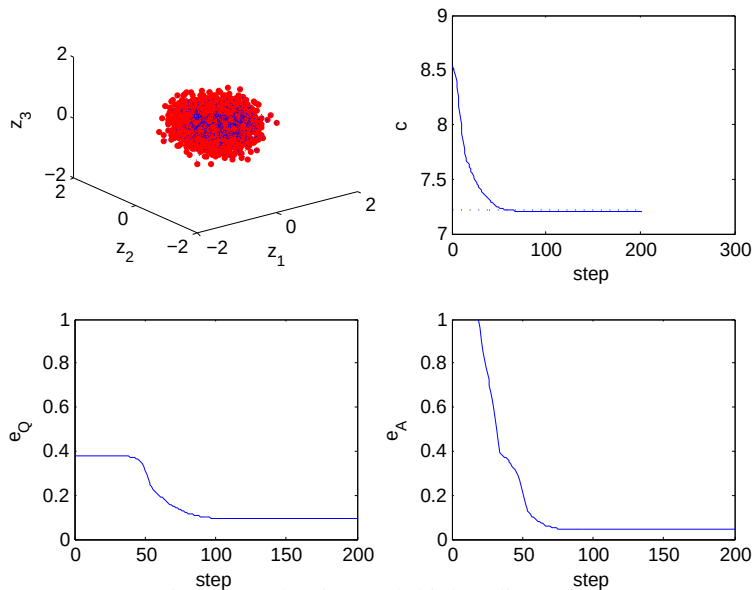

FIGURE 2.3. Same as Figure 2.2, but in much higher dimensions: $n = 80$ and $m = 17$, run through gradient descent in Q. The number of steps to convergence is roughly independent of the dimensionality of the space. For a relatively small number of observations (5000 in this run), one should not look for dynamical manifolds of too-large dimensionality, else one would be capturing just random autocorrelation among the data.

2.3 Non-zero means

We have worked so far under the assumption that all means have been removed from the problem: the plane x goes through the origin, and the transformation given by the matrix A is linear, not affine. If the observations z have a well-defined mean (that is, if there is not a trend over time that makes the local mean of z evolve), these assumptions are fine: it is enough to remove from z its mean –the "climatology" of atmosphere-ocean science– ad initio, and add it back at the end. However, for the

non-autonomous scenario to be described below, it will be necessary to consider nontrivial means. In order to have our methodology prepared for this more general case, we consider the means in our present autonomous situation too, even though they have no practical consequence. Then we write

$$z - \bar{z} = Q \begin{pmatrix} x \\ y \end{pmatrix},$$

and

$$x_{j+1} = Ax_j + b.$$

It is convenient to partition $\bar{z}$ into its x and y components,

$$\bar{x} = Q_x' \bar{z}, \quad \bar{y} = Q_y' \bar{z}.$$

The addition of the mean $\bar{x}$, however, is unnecessary, for its effects can be absorbed into the drift b. In this case we have to include in the minimization problem the drift b and the mean $\bar{y}$. The detail of the gradients with respect to these new two variables can be found in the Appendix A.2.

3 Non-autonomous problems

We have considered up to now only autonomous problems, where the manifold x and the corresponding dynamical model are assumed to be time-independent. Yet there are many examples of practical importance where this assumption does not hold. Consider, for instance, climate-related data, such as monthly averages of sea-surface temperatures at various locations, recorded over many years. One should expect much of the dynamics to depend on seasonal changes in insolation. We should, accordingly, have a time-dependent dynamical model, with a period of one year. Similarly, in long series of economic or financial data, we should expect a change in the dynamics as populations or affluence levels change, new markets arise, new tools are developed. The corresponding dynamical model should not longer be constant, nor periodic as in the seasonal case, but rather evolve slowly, with scale separation between the time-scale of the dynamics and that of the evolution of the model itself (without the hypothesis of scale separation, little can be inferred statistically from the data, since the dynamical model can be adjusted instantly to account for each individual observation).

To incorporate this into our framework, it is enough to add a qualifying subindex "t" (or more precisely "j", since our time-series are discrete) to the various functions involved: Q, A, b and $\bar{y}$, plus the requirements of periodicity or scale separation. For instance, Q_t should satisfy either $Q_{t+T} = Q_t$ in the periodic case, or $\|Q_{t+1} - Q_t\| \ll 1$ for slowly varying trends. In this section, we discuss how to modify the methodology of Section 2.2 so as to make it applicable to the non-autonomous linear case.

The idea is simple: in the notation of the previous section, we are seeking a time-dependent orthogonal transformation Q_t and mean $\bar{y}_t$, and a time-dependent dynamical model parameterized by A_t and b_t. To this end, in each descent step,

we pick at random a time t_0 and propose, in order to update Q, a time-dependent rotation angle θ_t in the k-l plane, centered at $t = t_0$ [3] . Similarly, we propose time-dependent variations for $\bar{y}$, A and b:

$$\theta_t = \alpha F(t), \quad \bar{y}_t = \bar{y} + v F(t), \quad A_t = A + B F(t), \quad b_t = b + d F(t),$$

where $F(t)$ is a given scalar function, centered at t_0, and satisfying the corresponding restrictions: periodicity, slow variation, etc., and the parameters α, a scalar, v and d, vectors, and B, a matrix, are computed by descent of the cost function as before. The details of the minimization problem can be found in appendix A.3.

The time t of the non-autonomous scenario discussed above is just one example of an *exogenous variable*: one whose state is known independently at all times, and that may affect the dynamics of the z's. Other examples are state variables of a bigger system of which the z's are only a small part; and external controls.

One can collectively denote these exogenous variables s, and apply a straightforward generalization of the procedure above, where F is now a function of s rather than the single variable t. In appendix B we describe a few choices for F that we have found practical.

For clarity, we illustrate the non-autonomous procedure through a simple example where $n = 2, m = 1$. We created data from the dynamical model

$$\begin{aligned} x_{j+1} &= a_j x_j + b_j + r_x \eta_j^x, \\ y_{j+1} &= \bar{y}_{j+1} + r_y \eta_j^y, \end{aligned}$$

for $j = 1, \ldots, N-1$, where $N = 1000$, the $\eta_j^{x,y}$'s are independent samples from a normal distribution, $r_x = 0.3$ and $r_y = 0.6$, and we adopted the values $a_j = \frac{6}{5}\cos^2\left(\frac{2\pi t_j}{T}\right)$ for the dynamics, $b_j = \frac{1}{2}\sin\left(\frac{2\pi t_j}{T}\right)$ for the drift, and $\bar{y}_j = \frac{2}{5}\cos\left(\frac{2\pi t_j}{T}\right)$ for the non-zero mean of y, where $t_j = j$ and $T = 12$, mimicking the twelve months of the year that we will find again in our application to the sea-surface temperature field in Section 5. Then we introduce, as before, "Atlantic" and "Pacific" temperatures

$$\begin{aligned} A_j &= x_j \cos(\theta_j) - y_j \sin(\theta_j), \\ P_j &= x_j \sin(\theta_j) + y_j \cos(\theta_j), \end{aligned}$$

with $\theta_j = \frac{\pi}{6}\sin\left(\frac{2\pi t_j}{T}\right)$, and provide the A_j and P_j as data for the principal dynamical component routine.

For this example, we have adopted the trial function F from (B.1). The results are displayed in Figure 3.1. The first plot shows the "observations" in the plane (A, P), with the first regular principal component drawn in black, and the 12 first

[3] For simplicity, we describe all the extensions from here on in terms of the procedure using Givens rotations. Adapting them to the far more efficient descent through the Lie algebra of the orthogonal transformations is straightforward.

principal dynamical components, one for each month, drawn in green. The other plots in the figure display the evolution of the normalized cost function c, and the estimated results for $a(t)$, $b(t)$, $\bar{y}(t)$ and $\theta(t)$ at convergence (we only show the first two periods). The dotted lines, drawn for reference, have the exact values of $a(t)$, $b(t)$, $\bar{y}(t)$ and $\theta(t)$ in the data, as well as the unexplainable part of the cost, $c_* = \frac{1}{N-1}\sum_{j=1}^{N-1}\left(r_x\eta_j^x\right)^2 + \left(r_y\eta_j^y\right)^2 \approx r_x^2 + r_y^2 = 0.45$. Again, the algorithm detects essentially the exact solution to the problem; the number of required steps, about 60, is bigger than before, because various different trial functions $F(t)$ are involved, requiring at least one step for each.

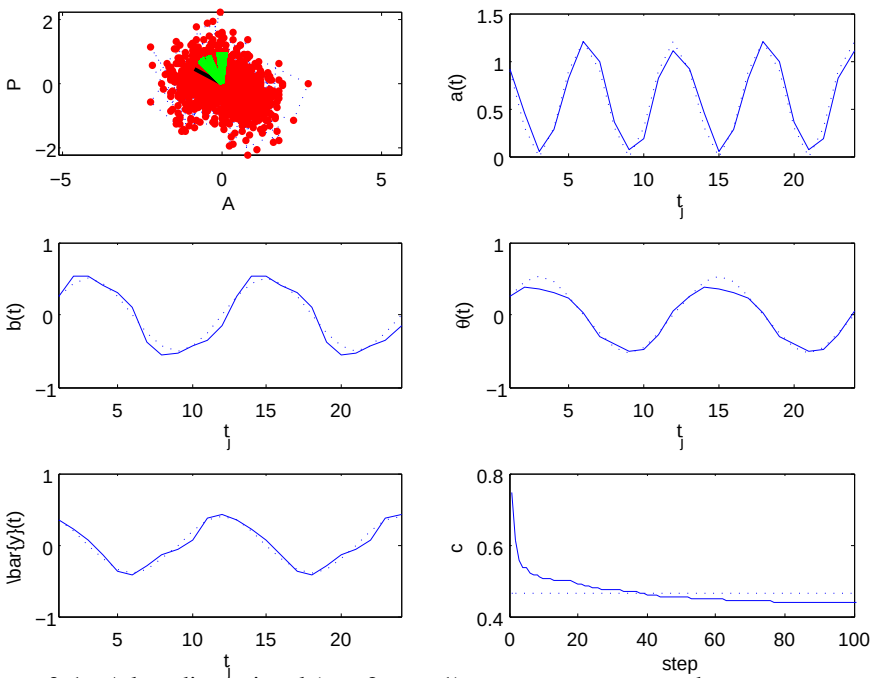

FIGURE 3.1. A low dimensional ($n = 2, m = 1$), non-autonomous problem. The first plot displays the data points, with dotted lines joining successive observations, the first principal component in black, and the twelve monthly first principal dynamic components in green. The other plots show the final estimates for $a(t)$, $b(t)$, $\bar{y}(t)$ and $\theta(t)$, for two periods of twelve snap-shots each, and the evolution of the cost function, with the exact answers in dotted lines.

4 Higher order processes

We have considered so far dynamical models without memory, where the current state of the system determines its future evolution through the matrix A. Yet many real processes are not well-described by such models. For instance, if the observations consist only of positions x_j in a system with non-negligible inertia, one

would expect a better prediction by using, in lieu of the unavailable velocity field, a second order model, $x_{j+1} = D(x_j, x_{j-1})$. Studying systems like this involves no significant change in our procedure: either we extend the phase-space from the line of x_j's to the plane of pairs (x_j, x_{j-1}) or, equivalently, consider matrices A that are rectangular, with twice as many columns as rows. Entirely similar considerations apply to higher order processes with longer memory.

We describe here the non-Markovian, non-autonomous case of order r, since the autonomous scenario is just a special case of the non-autonomous one, and the case with more general exogenous variables s is entirely similar. Our reduced dynamical model now adopts the form

$$x_{j+1} = D = b + \sum_{i=1}^{r} A_i x_{j-i+1}, \tag{4.1}$$

where the drift b and the matrices $A_i, i = 1, \ldots, r$, as well as the orthogonal matrix Q_x defining the x's, may in general be time-dependent. Each algorithmic step, we update these matrices through

$$(A_i)_t = A_i + B_i F(t), \quad b_t = b + d F(t), \quad \bar{y}_t = \bar{y} + v F(t), \quad \theta_t = \alpha F(t),$$

where $F(t)$ is a given trial function as described in appendix B.

The cost function adopts the form

$$c = \sum_{j=r}^{N-1} \left\| y_{j+1} \right\|^2 + \left\| x_{j+1} - D \right\|^2,$$

since the first $x_1, \ldots, x_r$ are not specified by the dynamics. The details of the minimization problem in this situation can be found in appendix A.4.

Again we choose, for the sake of clarity, to illustrate the procedure in its simplest possible setting, which is autonomous, with $n = 2$, $m = 1$, and $r = 3$, the order of the Non-Markovian process. We created data from the dynamical model

$$\begin{aligned} x_{j+1} &= a_1 x_j + a_2 x_{j-1} + a_3 x_{j-2} + r_x \eta_j^x, \\ y_{j+1} &= r_y \eta_j^y, \end{aligned}$$

for $j = 3, \ldots, N-1$, where $N = 1000$, the $\eta_j^{x,y}$'s are independent samples from a normal distribution, and we adopted the values $a_1 = 0.4979, a_2 = -0.2846, a_3 = 0.1569$ for the dynamics and $r_x = 0.3$ and $r_y = 0.6$ for the amplitudes of the noise in x and y. Then, as before, we define

$$\begin{aligned} A_j &= x_j \cos(\theta) - y_j \sin(\theta), \\ P_j &= x_j \sin(\theta) + y_j \cos(\theta), \end{aligned}$$

with $\theta = \frac{\pi}{3}$, and provide the A_j and P_j as data for the principal dynamical component routine. The results are displayed in Figure 4.1. Again the procedure converges to the exact answer, this time for all elements of the multi-step dynamics.

As in the first example, the first regular principal component is orthogonal to the principal dynamical component, thus capturing none of the system's dynamics.

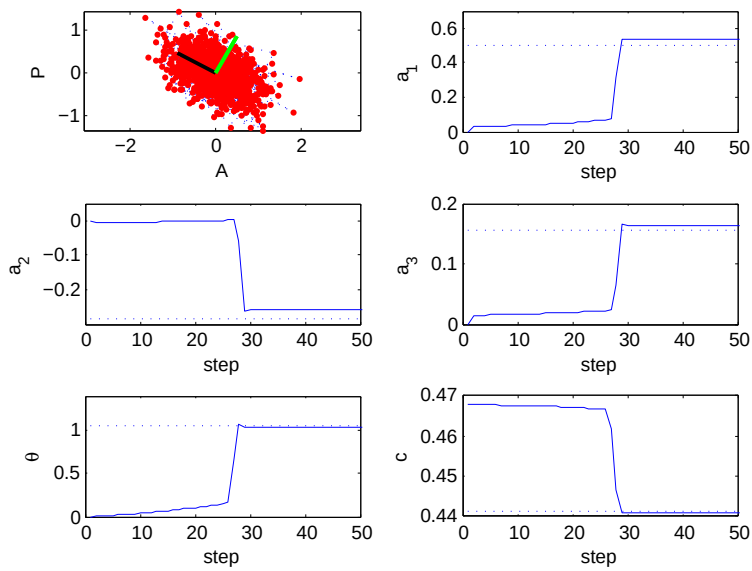

FIGURE 4.1. A multi-step process of order 3. The first plot displays the data points, with dotted lines joining successive observations, their regular first principal component in black and their first principal dynamical component in green. The other plots show the evolution of the estimates for a_1, a_2, a_3 and θ, as well as of the cost function, with their exact values and the exact unexplainable part of the cost, $c_* = \frac{1}{N-1}\sum_{j=1}^{N-1}\left(r_x\eta_j^x\right)^2 + \left(r_y\eta_j^y\right)^2 \approx r_x^2 + r_y^2 = 0.45$, displayed in dotted lines.

5 A real application: the global sea-surface temperature field

To see the workings of the new procedure on real data, we have chosen a topic of present concern: the estimation of climatic variations and trends. For this, we use a database of monthly averaged extended reconstructed global sea surface temperatures based on COADS data (see [25]) from January 1854 to October 2009, and ask whether we can extract from these a reduced low dimensional dynamical model. A few before-hand considerations are in order:

- Climate dynamics, a real pressing issue, is treated just as an illustration in this methodological paper. A far more in-depth treatment of how much principal dynamical components can help increase climate predictability and elucidate its causal relations will be pursued elsewhere.

- The ocean is not an isolated player in climate dynamics: it interacts with the atmosphere and the continents, and is also affected by external conditions, such as interannual variations in solar radiation and human-related release of CO_2 into the atmosphere. The latter are examples of slowly varying external trends that fit naturally into our non-autonomous setting –the seasonal variations giving its periodic component. As for the land and atmosphere, their dynamics is typically faster than that of the oceans, and can be conceptually divided into two components: a part that is slaved to the state of the ocean's surface temperature –and hence can in principle be included in its dynamical model–, and one that can be treated as external noise. Including explicitly land and atmospheric observations involves at least two further challenges, that will be pursued elsewhere: handling data with disparate units –such as atmospheric pressure, ice extent and ocean temperature–, and allowing for multiple time-scale dynamical models.
- Even within the ocean, the surface temperature does not evolve alone: it is carried by currents, and it interacts through mixing with lower layers of the ocean. As mentioned in Section 4, one way to account for unobserved variables is to make the model non-Markovian: discrete time derivatives of the sea-surface temperature provide indirect evidence on the state of those hidden variables.

We have adopted as our dataset the sea-surface temperature monthly means between January 1854 to October 2009 of the 50 points displayed on the map in Figure 5.1, covering much of the world oceans in a roughly homogeneous manner.

In order to apply our methodology to the data, we need to select a class of trial functions from Subsection B, the dimension m for the reduced manifold x, and the order r of the non-Markovian process. The trial functions for the periodic component that we have used for these runs are the monthly discrete δ-functions from (B.2), with $T = 12$. This takes to a new depth the idea behind the use of a "monthly climatology" in climate studies: not only the climatological mean is computed independently for each month, but also the dynamical model and manifold may change significantly from month to month. In the runs reported here, we have not modeled any inter-annual trend.

The following physical considerations suggest picking $r = 3$ for the order of the Markov process. A simplified conceptual model for the upper mixed layer of the ocean is that of a rotating shallow layer of water, forced by the atmosphere from above and the deep ocean from below. In such model, the active dynamical variables are the two horizontal components of the velocity and the layer's thickness. The surface temperature can be thought of as an emergent of the evolution of these three variables and the external forcing. Conservation of mass and horizontal momentum, the core dynamics of the layer, are three differential equations, each involving one time derivative. Hence reducing the system to a single variable –the temperature, the only one available in the data– yields a third order differential equation: two time derivatives relate to the evolution of gravity waves, the third to

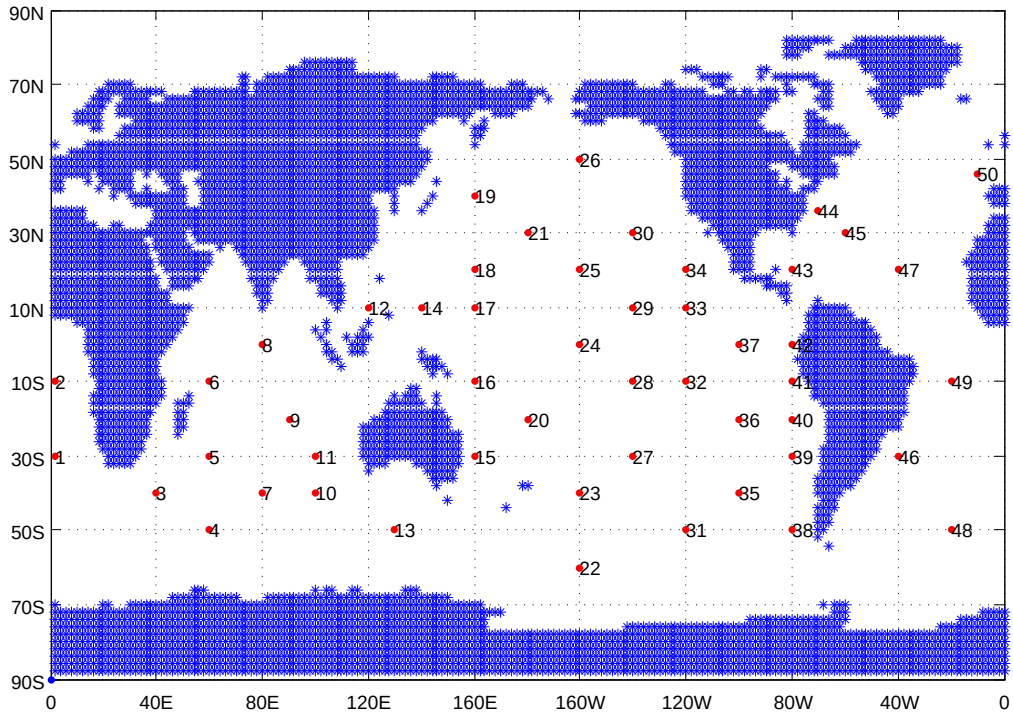

FIGURE 5.1. The 50 points on the ocean used for the procedure.

the potential vorticity. In our discrete setting, this corresponds to a Markov process of order three.

Figure 5.2 illustrates a line of reasoning for choosing the values of m and r. The figure on the left shows, for a fixed $m = 4$, the evolution of the final error when we move the order of the non Markovian process from $r = 1$ to $r = 6$[4]. For reference, the dotted line shows the value of $\frac{1}{N}\sum_{i=5}^{N} S_i^2$, where the S's are the singular values of the real dataset, with the monthly climatology subtracted. We find for $r = 3$ the steepest drop of the final error, consistent with our reasoning above. Therefore we pick $r = 3$ for the order of our non Markovian process. In the figure on the right, we observe, for this fixed $r = 3$, the evolution of the final error when $m = 1, \ldots, 6$. The isolated points correspond to the sum of squared singular values, $\frac{1}{N}\sum_{i=m+1}^{N} S_i^2$. We observe that for $m = 4$ this error matches almost exactly the one from the dynamical components. This can be interpreted in the following way: for smaller values of m,

[4] This final error is calculated for a number of steps such that the difference between the final error and the error 1000 steps before is less than 0.01; therefore the number of steps used for each value of r may be different. Notice that we are using the descent procedure through Givens rotations, hence the large number of steps involved. When a far more thorough exploration of the application of principal dynamical components to the climate is pursued elsewhere, a basic ingredient for efficiency will be the descent through general orthogonal transformations Q described at the end of Section 3.

accounting for the dynamics allows us to reduce the information loss even beyond the theoretical maximum –for autonomous settings– provided by the singular value decomposition. Beyond $m = 4$, on the other hand, the biggest share in the further reduction of information loss is probably due to the increased bare dimensionality of the model, more than to a further refinement of the dynamics. Hence we pick $m = 4$ for the dimension of our reduced dynamical manifold.

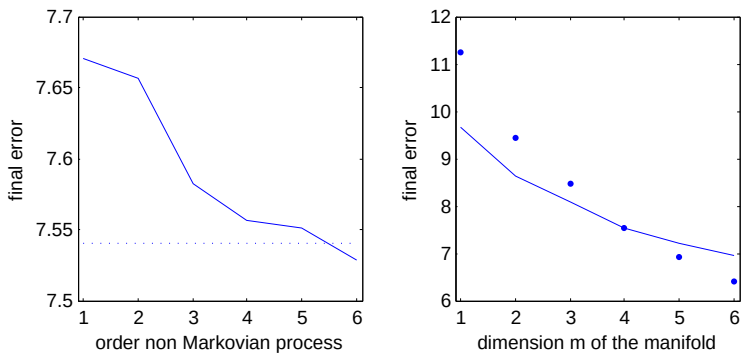

FIGURE 5.2. Predictive uncertainty as a function of the dimension m of the reduced dynamical manifold and the order r of the process. The figure on the left shows the final error for a fixed $m = 4$ and $r = 1,\ldots,6$. The dotted line is the sum $\frac{1}{N}\sum_{i=5}^{N} S_i^2$ of the squared singular values of the data, with the monthly climatology removed. The figure on the left shows the final error for a fixed $r = 3$ and $m = 1,\ldots,6$, with the isolated points displaying the corresponding uncertainty in the standard principal component procedure, $\frac{1}{N}\sum_{i=m+1}^{N} S_i^2$.

Next we show various results for the chosen parameters, $m = 4$ and $r = 3$, displayed for the time window from January 1991 to January 1999, which includes three El Niño years, represented by vertical lines; one of them, in 1998, the strongest ever recorded. Figure 5.3 shows the evolution of the four components of the manifold x in solid lines and, in dotted lines, the same components predicted from the prior three months.

One question one may ask is whether the reduced dynamical manifold x is dominated by a small set of locations on the ocean. This would be manifest in having the columns of Q_x dominated by a few significant rows. Yet the columns of Q_x do not have a meaning per se: x is a four dimensional manifold, but each component x^i lacks individual meaning. To fix a reference frame in the manifold x, we resort to the matrix A_1: its four left principal components U in $A_1 = USV'$ provide a natural set of coordinates in x-space. Figure 5.4 displays the first four columns of $Q_x(t)U(t), t = 1,\ldots,12$. We observe between four and six dominant

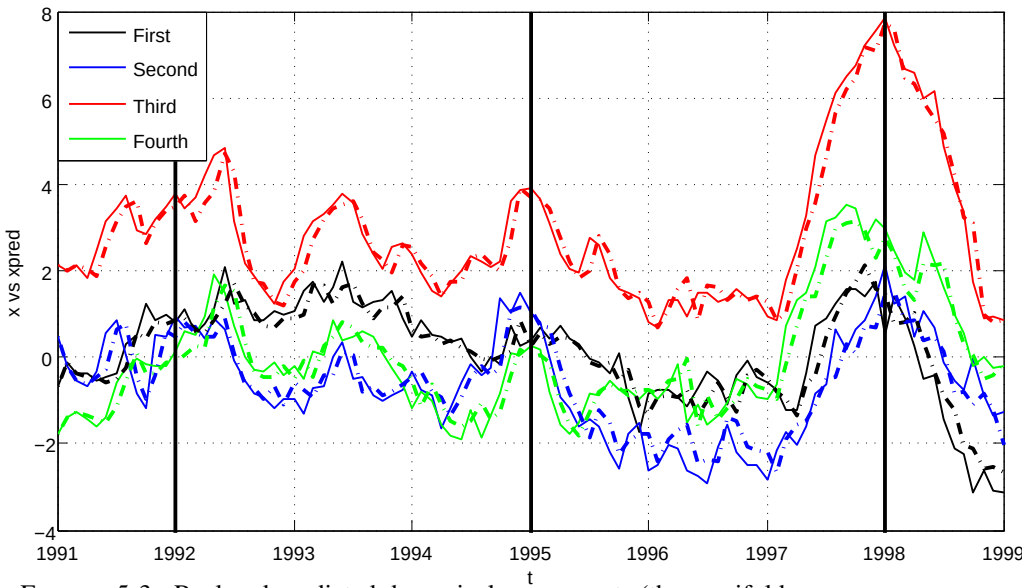

FIGURE 5.3. Real and predicted dynamical components (the manifold x). The vertical lines mark El Niño years.

peaks; the four clearest ones corresponding to the points 19, 24, 37 and 41 in Figure 5.1. This suggests that a reduced dynamical model for the ocean could be built from four to six selected locations. Notice that these four points are on the Pacific ocean, in locations that one would naturally associate with the strongest El Niño signals.

Figure 5.5 shows the observed ocean surface temperature for these four points, comparing them with the ones predicted by the algorithm, in dotted lines. We see that the approximation is quite sharp, particularly near El Niño years, where changes of temperatures are most significant. Even though we have chosen to plot only these four temperatures, all of the 50 points used are well-predicted by the procedure [5].

Finally, we monitor the evolution of a measure of the global anomalies associated with El Niño. To this end, we compute a discrete analogue of the running 3-month mean SST anomaly in the El Niño regions [26]. In particular, we average the temperatures on the points 16, 17, 24, 28, 29, 32, 33, 37, 41 and 42 on the map in Figure 5.1, for a time window from February 1964 to October 2009. These 10 points are not all strictly included in what are known as El Niño regions (there are four of them, 1+2, 3, 4 and 3.4), but they are the closest on our discrete map to the union of all of them. We observe in Figure 5.6 the warm (positive) peaks, coinciding with El Niño years, the cold (negative) peaks corresponding to La Niña

[5] By "prediction" here we mean prediction based on the three previous values of x; the possibility of a longer term prediction using principal dynamical components will be explored elsewhere.

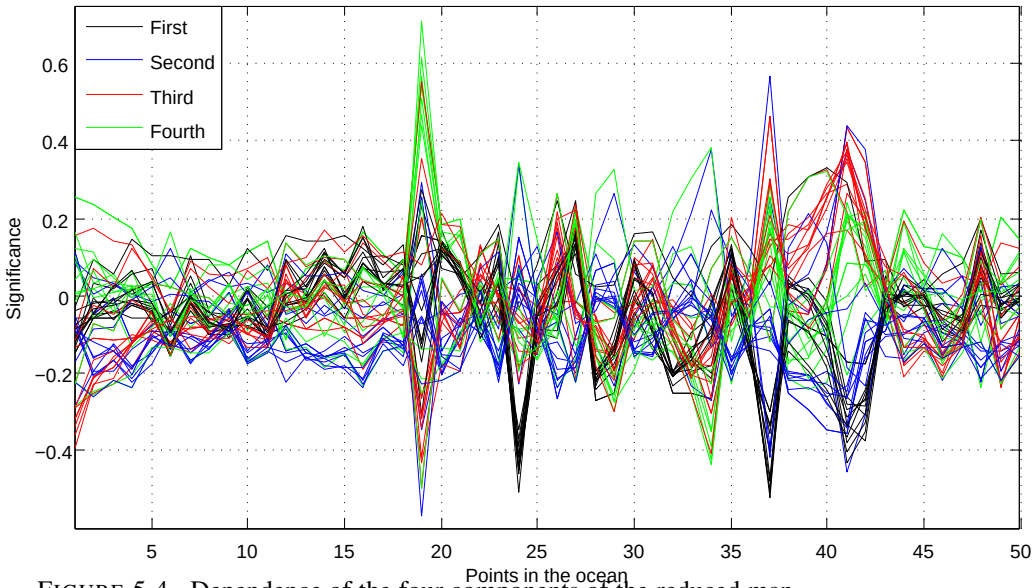

FIGURE 5.4. Dependence of the four components of the reduced manifold x on the individual locations on the ocean for the twelve months of the year. A natural coordinate system in x is the one provided by the principal components of the first dynamical matrix A_1. Notice that four to six points on the ocean dominate the dynamics.

years and, in dotted lines, the predicted values of these SST anomalies generated by the principal dynamical component procedure.

6 Probabilistic perspective and extension to nonlinear dynamics

Throughout this article, we have defined and developed the principal dynamical component procedure in terms of the minimization of a specific cost function: the sum of squares of the prediction errors. In this section, we assign a meaning to this cost in terms of the log-likelihood function of a probabilistic model. Framing the principal dynamical component procedure in a probabilistic setting has two main advantages: to permit a more thorough interpretation, and to extend its applicability beyond the linear models developed in this article. We sketch such generalization in this section; its algorithmic implementation, under current development, will be presented elsewhere.

Generally, a probabilistic model for a time series $z_j \in R^n$ involves the transition probability density

$$T\left(z_{j+1}|z_j\right).$$

(This corresponds to the Markovian, autonomous scenario, the only one that we address in this section. The extension to non-autonomous and non-Markovian cases, involving a transition probability density of the form $T(z_{j+1}|z_j, z_{j-1}, \ldots, z_{j-r}, t, s)$,

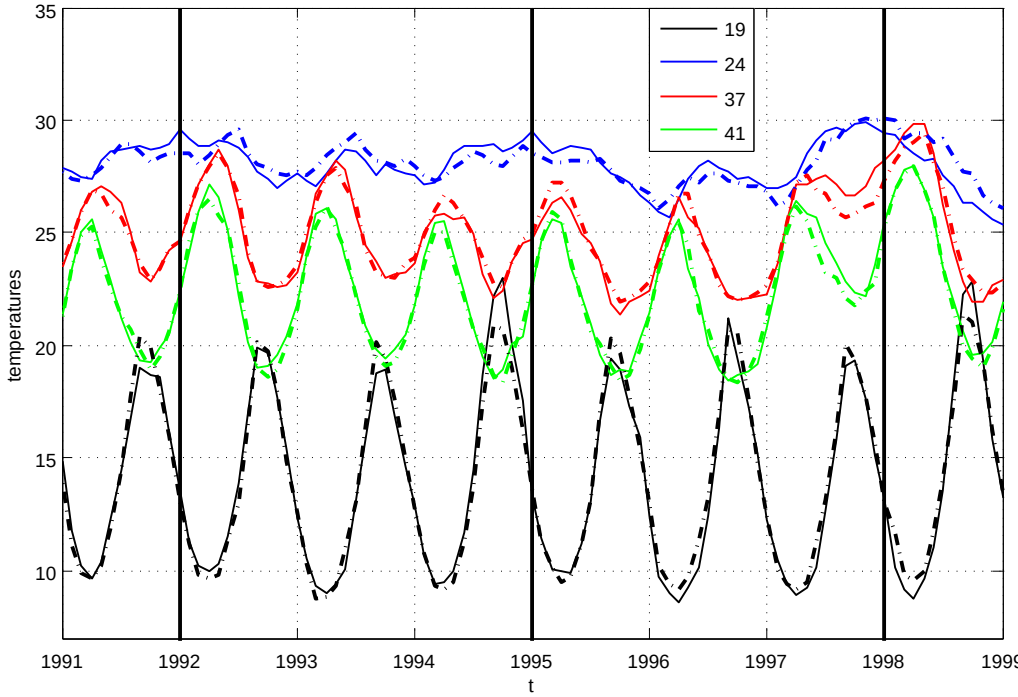

FIGURE 5.5. Observed and predicted ocean surface temperature for four selected points. The vertical lines mark El Niño years.

is straightforward). The principal dynamical component proposal considers a dimensional reduction of such transition probability density, using the following elements:

- A coordinate system $z = z(x,y)$, $x \in R^m$, $y \in R^{n-m}$, with corresponding projection operators P_x and P_y:

$$x = P_x(z(x,y)), \quad y = P_y(z(x,y)).$$

- A reduced dynamical model given by a transition probability density in R^m:

$$d\left(x_{j+1}|x_j\right).$$

- A probabilistic *embedding*

$$e(y|x).$$

The transition probability density for z is then given by

$$T\left(z_{j+1}|z_j\right) = J\left(z_{j+1}\right) e\left(y_{j+1}|x_{j+1}\right) d\left(x_{j+1}|x_j\right),$$

where $x = P_x(z)$, $y = P_y(z)$, and $J(z)$ is the Jacobian determinant of the coordinate map $z \to (x,y)$.

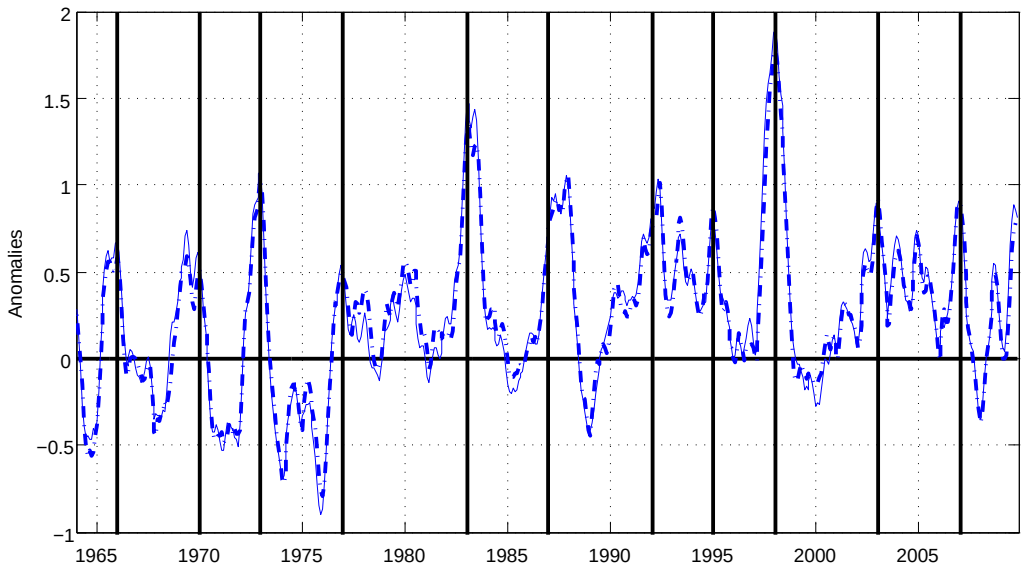

FIGURE 5.6. Observed and predicted 3-month mean SST anomalies from February 1964 to October 2009, quantifying El Niño and La Niña intensities. Only El Niño years are marked with vertical lines; La Niña years correspond to strong negative anomalies.

A natural measure of the goodness of the model is the log-likelihood function

$$L = \sum_{j=1}^{N-1} \log\left[T\left(z_{j+1}|z_j\right)\right].$$

In particular, in the setting of Section 2, we have the projections

$$P_x(z) = Q_x' z, \quad P_y(z) = Q_y' z, \tag{6.1}$$

where $Q = [Q_x Q_y]$ is orthogonal, so $J(z) = 1$. The embedding and reduced dynamics are given by the isotropic Gaussians

$$e(y|x) = \mathcal{N}(0, \sigma^2 I_{N-m}) \tag{6.2}$$

and

$$d\left(x_{j+1}|x_j\right) = \mathcal{N}(Ax_j, \sigma^2 I_m), \tag{6.3}$$

where I_k stands for the $k \times k$ identity matrix. Consequently, the log-likelihood function is given by

$$L = \sum_{j=1}^{N-1} - \left[\frac{n}{2}\log(2\pi) + n\log(\sigma) + \frac{1}{2\sigma^2}\left(\|x_{j+1} - Ax_j\|^2 + \|y_{j+1}\|^2\right)\right].$$

Thus maximizing the log-likelihood L over Q and A is equivalent to minimizing the cost function

$$c = \frac{1}{N-1} \sum_{j=1}^{N-1} \left(\|x_{j+1} - Ax_j\|^2 + \|y_{j+1}\|^2 \right) ,$$

that we have used throughout the paper; the corresponding optimal value of σ is given by

$$\sigma = \left(\frac{c}{n} \right)^{\frac{1}{2}} .$$

This interpretation immediately suggests the following generalization, which remains within the realm of Gaussian distributions and linear maps: keep the orthogonal projections in (6.1), but replace the embedding (6.2) and dynamical model (6.3) by the more general

$$\begin{aligned} e(y|x) &= \mathcal{N}(0, \Sigma_y), \\ d(x_{j+1}|x_j) &= \mathcal{N}(Ax_j, \Sigma_x), \end{aligned}$$

where Σ_x and Σ_y are general covariance matrices. The resulting log-likelihood function is

$$L = \sum_{j=1}^{N-1} -\frac{1}{2} \left[\log((2\pi)^n |\Sigma_x||\Sigma_y|) + \left(x_{j+1} - Ax_j, \Sigma_x^{-1}(x_{j+1} - Ax_j) \right) + \left(y_{j+1}, \Sigma_y^{-1} y_{j+1} \right) \right] .$$

This formulation has the advantage of providing a natural ranking of the coordinates x and y, through the principal components of the corresponding covariance matrices.

More generally, one can propose different, typically nonlinear, families of distributions, projections and dynamical models, and maximize the corresponding log-likelihood function. The proposed distributions can be given parametrically, in which case the maximization of the log-likelihood is over their parameters, or non-parametrically, for instance as an extension of the methodology proposed in [24]. Thus the principal dynamical component methodology extends naturally to very general scenarios, with nonlinear reduced dynamical manifolds, stochastic, nonlinear dynamical models, and non-Gaussian embeddings. This extension, however, goes beyond the scope of this paper, and will be pursued elsewhere.

7 Conclusions

A new methodology has been developed for the dimensional reduction of time series. The procedure seeks a low dimensional manifold x and a dynamical model $x_{j+1} = D(x_j, x_{j-1}, \ldots, t)$ that minimize the predictive uncertainty of the series. The procedure has been successfully tested on synthetic data, and illustrated with a real application to time series of sea-surface temperature over the ocean. Finally, a probabilistic interpretation of the principal dynamical component procedure was proposed, providing a conceptual extension to general nonlinear, non-Gaussian settings.

One item not addressed in this article is the selection of a time-scale: it is assumed throughout that the time-intervals Δt between the times t_n for which data is available correspond to the natural time-scale for the dynamics. Yet this is not necessarily the case; the observations may over or under-resolve the underlying dynamical processes. In the former case, one might apply the procedure developed here not to the individual snap-shots provided by the observations but to their averages over longer time-windows compatible with the dynamics. A more sophisticated option, beyond the scope of this article, is to consider a multi-scale version of the algorithm, combining short and long-term dynamical components.

Appendix: The minimization problem

In this appendix we give full details about the minimization problem for the parameters A and Q described in Sections 2, 3 and 4.

A.1 Two dimensional linear, autonomous case

The descent steps in a are given by

$$\frac{\partial c}{\partial a} = -2\sum_{j=1}^{N-1}(x_{j+1} - ax_j)\,x_j.$$

Equating $\frac{\partial c}{\partial a}$ to zero yields the standard regression formula

$$a = \frac{\sum_{j=1}^{N-1} x_j x_{j+1}}{\sum_{j=1}^{N-1} x_j^2}.$$

If now we update x and y through a further rotation

$$x \leftarrow x\cos(\theta) + y\sin(\theta)\,,$$

$$y \leftarrow -x\sin(\theta) + y\cos(\theta)\,,$$

we have

$$\frac{\partial c}{\partial \theta} = 2a\sum_{j=1}^{N-1}\Big[(ax_jy_j - (x_{j+1}y_j + x_jy_{j+1}))\cos(2\theta) + \Big(x_{j+1}x_j - y_{j+1}y_j + \frac{a}{2}\left(y_j^2 - x_j^2\right)\Big)\sin(2\theta)\Big].$$

Rather than seeking a closed expression for θ that would make this derivative vanish –notice that θ is implicitly included in the definition of the x and y's–, it is preferable to descend the gradient

$$\frac{\partial c}{\partial \theta}\Big|_{\theta=0} = 2a\sum_{j=1}^{N-1}[ax_jy_j - (x_{j+1}y_j + x_jy_{j+1})]$$

or, more efficiently, to involve also the second derivative

$$\left.\frac{\partial^2 c}{\partial \theta^2}\right|_{\theta=0} = 2a \sum_{j=1}^{N-1} \left[2x_{j+1}x_j - 2y_{j+1}y_j + a\left(y_j^2 - x_j^2\right)\right],$$

and compute the θ that minimizes the quadratic local approximation to c:

$$\theta = \theta_q = -\frac{\left.\frac{\partial c}{\partial \theta}\right|_{\theta=0}}{\left.\frac{\partial^2 c}{\partial \theta^2}\right|_{\theta=0}}.$$

A little extra care is required when applying the quadratic approximation far from the optimal θ: if $\left.\frac{\partial^2 c}{\partial \theta^2}\right|_{\theta=0} \leq 0$, then we must do descent instead:

$$\theta = -\varepsilon_l \left.\frac{\partial c}{\partial \theta}\right|_{\theta=0}, \tag{A.1}$$

where $\varepsilon_l > 0$ is a chosen learning rate. Also, if θ_q is too big, we must limit our step size:

$$|\theta| = \max(|\theta_q|, \varepsilon),$$

where ε is the maximum allowable step in θ. It is sensible to relate the values of the two ε's through

$$\varepsilon_l = \frac{\varepsilon}{\sqrt{\varepsilon^2 + \left(\left.\frac{\partial c}{\partial \theta}\right|_{\theta=0}\right)^2}},$$

which, when applied to (A.1), yields descent steps of size bounded by ε, and much smaller near the optimal θ.

A.2 Multidimensional linear, autonomous case

The descent steps in A are given by

$$\frac{\partial c}{\partial A} = -2 \sum_{j=1}^{N-1} (x_{j+1} - Ax_j) x_j'. \tag{A.2}$$

Instead of descending the gradient we can, as in the two-dimensional case, directly solve $\frac{\partial c}{\partial A} = 0$, yielding

$$A = X_1 X_0' \left(X_0 X_0'\right)^{-1},$$

where

$$X_0 = [x_1, \ldots, x_{N-1}] \quad \text{and} \quad X_1 = [x_2, \ldots, x_N].$$

For the descent steps in Q, we propose two different approaches. In the first, we use representations of any orthogonal matrix Q as a product of Givens rotation, in which case we only have to differentiate with respect to the angle θ. In the second, we use a Lie algebra approach, where we differentiate with respect to a general orthogonal matrix Q.

A Givens rotation approach for Q

We note that any orthogonal matrix can be factorized as a product of Givens rotations of the form

$$R_{kl}(\theta) = \begin{pmatrix} 1 & \dots & 0 & \dots & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \dots & \cos(\theta) & \dots & \sin(\theta) & \dots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \dots & -\sin(\theta) & \dots & \cos(\theta) & \dots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & \dots & 0 & \dots & 1 \end{pmatrix},$$

which act in the plane of the two coordinates k and l, rotating them an angle θ. Then we can, in each descent step, pick at random the two indices k and l, and perform a rotation following the derivative of c with respect to θ at $\theta = 0$. From the observation above about degeneracy, however, we note that picking both k and l from either the dynamical coordinates x or their orthogonal complement y alone, serves no purpose other than re-parametrization. Then we always pick k at random in $[1,\dots,m]$, and adopt $l = m+h$, with h picked at random in $[1,\dots,n-m]$. In order to consider arbitrary directions in these two manifolds though, we first perform a random orthogonal transformation to each:

$$x \to Q^r_x x, \quad y \to Q^r_y y,$$

where $Q^r_{x,y}$ are random orthogonal matrices.

For each Givens rotation, we have

$$\frac{\partial c}{\partial \theta}\Big|_{\theta=0} = -2\sum_{j=1}^{N-1}\left[y^h_{j+1}A_k x_j + y^h_j \sum_{p=1}^{m} A^k_p \left(x^p_{j+1} - A_p x_j\right)\right] \tag{A.3}$$

and

$$\frac{\partial^2 c}{\partial \theta^2}\Big|_{\theta=0} = 2\sum_{j=1}^{N-1}\left[x^k_{j+1}A_k x_j - 2y^h_{j+1}A^k_k y^h_j + \sum_{p=1}^{m}\left[x^k_j A^k_p (x^p_{j+1} - A_p x_j) + (y^h_j A^k_p)^2\right]\right]. \tag{A.4}$$

As before, θ can be computed so as to minimize the quadratic local approximation to c:

$$\theta = \theta_q = -\frac{\frac{\partial c}{\partial \theta}\Big|_{\theta=0}}{\frac{\partial^2 c}{\partial \theta^2}\Big|_{\theta=0}},$$

with the same caveats on big steps as in the one-dimensional case.

A Lie algebra approach for Q

A general orthogonal transformation depends on $n(n-1)/2$ parameters which, for Q near the identity, are best represented as the independent entries of a skew-symmetric matrix S (the Lie algebra of the orthogonal group):

$$Q = e^S, \quad \text{with} \quad S_{ij} = -S_{ji}.$$

When S is small, each entry S_{ij} defines a two-dimensional differential rotation in the plane ij. In our case, however, we are not interested in rotations within the x or y manifolds, since these would act only as re-parameterizations of the manifolds. Then we need only to consider the $m(n-m)$ entries where, borrowing the notation from the first procedure, $i = k \leq m$ and $j = m+h$, and their skew-symmetric counterparts. For these, we can compute the gradient of the cost c,

$$G_{ij} = \frac{\partial c}{\partial S_{ij}},$$

and write, by descent,

$$Q = e^S, \quad S_{ij} = -\frac{\varepsilon}{\sqrt{\varepsilon^2 + |G|^2}} G_{ij},$$

where

$$|G|^2 = \sum_{i,j} G_{ij}^2.$$

This makes steps of size ε far from the optimal Q, and smaller near it, to avoid oscillations. The computation of $Q = e^S$ is far less expensive that it would be for a general $n \times n$ matrix, since S has only $m(n-m)$ independent entries, and a simple block structure, that can be exploited to reduce the cost of the calculation significantly.

Alternatively to the gradient descent above, we can compute also the Hessian

$$H_{ij}^{kl} = \frac{\partial^2 c}{\partial S_{ij} \partial S_{kl}},$$

and can write the second order step

$$Q = e^S, \quad \sum_{ij} H_{ij}^{kl} S_{ij} = -G_{kl}.$$

This is not very efficient in this crude form though, since it requires solving a system of equations in $m \times n$ unknowns. There are efficient ways of carrying the descent in Q to second order, but these will be presented elsewhere.

The first derivatives of c with respect to S_{ij} are the same as those with respect to θ in (A.3), with $k = i$ and $h = j - m$. The diagonal elements of the Hessian, H_{ij}^{ij}, also agree with the second derivatives in (A.4). For the off diagonal terms, we have

$$\frac{\partial^2 c}{\partial S_{i_1 j_1} \partial S_{i_2 j_2}} = -2 \sum_{j=1}^{N-1} \left[y_{j+1}^{h_1} A_{k_1}^{k_2} y_j^{h_2} + y_{j+1}^{h_2} A_{k_2}^{k_1} y_j^{h_1} - y_j^{h_1} y_j^{h_2} {A^{k_1}}' A^{k_2} \right] \quad \text{for} \quad j_1 \neq j_2$$

and

$$\frac{\partial^2 c}{\partial S_{i_1 j_1} \partial S_{i_2 j_2}} = -\sum_{j=1}^{N-1} \Big[-\left(x_{j+1}^{k_2} A_{k_1} + x_{j+1}^{k_1} A_{k_2}\right) x_j - \left(x_j^{k_2} A_{k_1} + x_j^{k_1} A_{k_2}\right) x_{j+1}$$

$$+ \left(x_j^{k_2} A^{k_1\prime} + x_j^{k_1} A^{k_2\prime}\right) A x_j + \left(A_{k_1}^{k_2} + A_{k_2}^{k_1}\right) y_j^h y_{j+1}^h - 2\left(y_j^h\right)^2 A^{k_1\prime} A^{k_2} \Big] \quad \text{for} \quad j_1 = j_2$$

with $k_{1,2} = i_{1,2}$ and $h_{1,2} = j_{1,2} - m$ (differential rotations do not commute to second order, hence the complex look of this last expression).

Non-zero means case

For the non-zero means we have the gradients

$$\frac{\partial c}{\partial b} = -2 \sum_{j=1}^{N-1} \left(x_{j+1} - (A x_j + b)\right) \tag{A.5}$$

and

$$\frac{\partial c}{\partial \bar{y}} = -2 \sum_{j=1}^{N-1} \left(y_{j+1} - \bar{y}\right), \tag{A.6}$$

that can be used either for descent or for the direct calculation of the optimal b and $\bar{y}$.

A.3 Non-autonomous case

The descent steps in this situation are similar to the formulas given in Section A.2, but introducing non-autonomous considerations. Then equations (A.3) and (A.4) generalize into

$$\frac{\partial c}{\partial \alpha}\Big|_{\alpha=0} = -2 \sum_{j=1}^{N-1} \left[w^{j+1} y_{j+1}^h \left(A_k x_j + b_k\right) + w^j y_j^h \sum_{p=1}^{m} A_p^k \left(x_{j+1}^p - A_p x_j - b_p\right) \right]$$

and

$$\frac{\partial^2 c}{\partial \alpha^2}\Big|_{\alpha=0} = 2 \sum_{j=1}^{N-1} \Big[(w^{j+1})^2 x_{j+1}^k \left(A_k x_j + b_k\right) - 2 w^{j+1} w^j y_{j+1}^h A_k^k y_j^h +$$

$$\sum_{p=1}^{m} \left[(w^j)^2 x_j^k A_p^k (x_{j+1}^p - A_p x_j - b_p) + (w^j y_j^h A_p^k)^2 \right] \Big],$$

and equations (A.2), (A.5) and (A.6) into

$$\frac{\partial c}{\partial B} = -2 \sum_{j=1}^{N-1} w^j \left(x_{j+1} - (A x_j + b)\right) x_j',$$

$$\frac{\partial c}{\partial d} = -2 \sum_{j=1}^{N-1} w^j \left(x_{j+1} - (A x_j + b)\right)$$

and

$$\frac{\partial c}{\partial v} = -2 \sum_{j=1}^{N-1} w^j \left(y_{j+1} - \bar{y}\right),$$

where the weights w^j are given by

$$w^j = F(t_j).$$

As before, equating the derivatives with respect to B, d and v to zero provides simple closed forms for B, d and v, while α can be found through the minimization of a local quadratic approximation to c. The appropriate choice of F is discussed in the Appendix B.

A.4 Higher order case

In this case, we have

$$\frac{\partial c}{\partial B_h} = -2 \sum_{j=r}^{N-1} w^j \left(x_{j+1} - D\right) x'_{j-h+1},$$

for every $h = 1, \ldots, r$, where D is given by (4.1),

$$\frac{\partial c}{\partial d} = -2 \sum_{j=r}^{N-1} w^j \left(x_{j+1} - D\right),$$

and

$$\frac{\partial c}{\partial v} = -2 \sum_{j=r}^{N-1} w^j \left(y_{j+1} - \bar{y}\right),$$

where $w^j = F(t_j)$.

It is possible to get explicit expressions for $B_1, \ldots, B_r$ by equating all $\frac{\partial c}{\partial B_h}$ to zero. We introduce the $m \times m$ matrices

$$X_0^{h,k} = \sum_{j=r}^{N-1} \left(w^j\right)^2 x_{j-h+1} x'_{j-k+1}, \quad X_1^h = \sum_{j=r}^{N-1} w^j \left(x_{j+1} - b\right) x'_{j-h+1},$$

where $X_0^{h,k} = \left(X_0^{k,h}\right)'$. In terms of these, we get a block system of linear equations:

$$\begin{aligned}
B_1 X_0^{1,1} + B_2 X_0^{2,1} + \cdots + B_r X_0^{r,1} &= X_1^1, \\
B_1 X_0^{1,2} + B_2 X_0^{2,2} + \cdots + B_r X_0^{r,2} &= X_1^2, \\
\vdots \qquad\quad \vdots \qquad\qquad \vdots \qquad\quad & \vdots \\
B_1 X_0^{1,r} + B_2 X_0^{2,r} + \cdots + B_r X_0^{r,r} &= X_1^r,
\end{aligned}$$

which determines the matrices $B_1, \ldots, B_r$.

For the angle α we proceed as in the previous sections, though a quadratic approximation to c, using

$$\frac{\partial c}{\partial \alpha}\Big|_{\alpha=0} = -2\sum_{j=r}^{N-1}\left[w^{j+1}y^h_{j+1}(D)_k + \sum_{p=1}^{m}(D\alpha)_p\left(x^p_{j+1}-(D)_p\right)\right],$$

and

$$\frac{\partial^2 c}{\partial \alpha^2}\Big|_{\alpha=0} = -2\sum_{j=r}^{N-1}\left[-\left(w^{j+1}\right)^2 x^k_{j+1}(D)_k + 2w^{j+1}y^h_{j+1}(D\alpha)_k \right.$$
$$\left. + \sum_{p=1}^{m}\left[(D\alpha\alpha)_p\left(x^p_{j+1}-(D)_p\right) - [(D\alpha)_p]^2\right]\right],$$

where

$$(D\alpha) = \sum_{i=1}^{r}(A_i)^k w^{j-i+1}y^h_{j-i+1}, \quad (D\alpha\alpha) = -\sum_{i=1}^{r}(A_i)^k\left(w^{j-i+1}\right)^2 x^k_{j-i+1}.$$

Appendix: Trial functions

In this appendix we consider the issue of how to pick the functions $F(s)$ and corresponding weights $w^j = F(s^j)$ (here we use s to denote either time or other exogenous variables).We describe a few choices that we have found practical. First of all, for the autonomous case, we have the trivial

$$F = 1\,.$$

This should still be used in the more general case, to capture the s-independent components of Q and A, but must be alternated with other functions $F(s)$ with non-trivial s-dependence.

B.1 One-dimensional functions

When s represents time, the domain of $F(s)$ must be either the real line –for the trend– or a parametrization of the unit circle –for periodic factors such as the seasons. A sensible choice for the trend is

$$F(s) = \frac{S}{\sqrt{S^2+L^2}} - \bar{F}, \quad S = s - s_0,$$

displayed on Figure B.1, depending on the choice of a center s_0, picked at random at each step, and a mollification parameter, the length-scale L (see below for more details). As $L \to 0$, $F(s)$ becomes piecewise constant, with a discontinuity at $s = s_0$. For larger values of L, the transition between the two constant states is smoothed over an interval of order L. The subtraction of the mean $\bar{F}$ over the observations is intended to decouple the effect of these steps from the ones using $F = 1$, a function concerned only with the mean. At the initial stages of the algorithm, L should be large, providing a global perspective; then it should decrease gradually, to tune the finer, more local details.

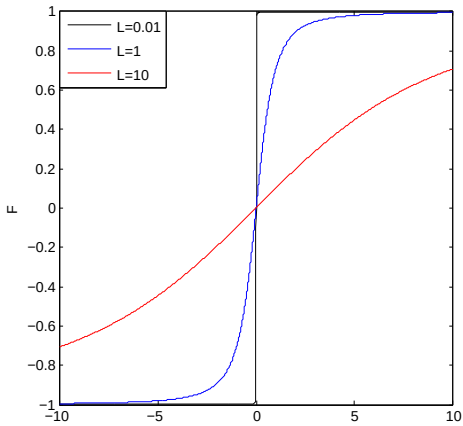

FIGURE B.1. Plot of the function $F(S) = \frac{S}{\sqrt{S^2+L^2}}$ over the interval $[-10, 10]$ for different values of L.

We can be more specific: calling L_0 the largest length scale in s, we need L_0/L steps to cover it with transitions of length L. Then the amount dt of algorithmic time spent using a length L should satisfy

$$\frac{dL}{dt} \propto L,$$

leading to the expression

$$L = L_0 \left(\frac{L_f}{L_0}\right)^{\frac{k}{k_{tot}}},$$

where k is the step number, k_{tot} the total number of steps, and L_f the smallest length scale to be used, not to over-resolve the dynamics.

In the periodic case, we can make an entirely analogous proposal:

$$F(s) = \frac{\sin(S)}{\sqrt{4\sin^2(S/2)+L^2}}, \quad S = \frac{2\pi(s-s_0)}{T}, \tag{B.1}$$

where T is the period; see Figure B.2.

Sometimes s can adopt only a discrete set of values: the months of the year, an on-off control, etc. In that case, it may be useful to consider signature functions F that are one on each of these values at a time, and zero on the others:

$$F_i(j) = \delta_{mod(j,T),i}, \tag{B.2}$$

where T is the integer period.

An alternative to the $F(s)$'s above, which have local derivatives but global effects, are the more localized bumps given by

$$F(s) = \frac{L^3}{(S^2+L^2)^{3/2}} - \bar{F},$$

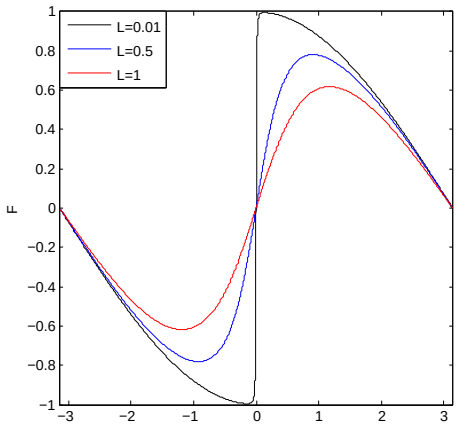

FIGURE B.2. Plot of the function $\bar{F}(S) = \frac{\sin(S)}{\sqrt{4\sin^2(S/2)+L^2}}$ over the interval $[-\pi,\pi]$ for different values of L.

and

$$F(s) = \frac{L^3}{(4\sin^2(S/2)+L^2)^{3/2}} - \bar{F},$$

displayed in Figures B.3 and B.4. We can also alternate between the two, or among more proposals satisfying different needs.

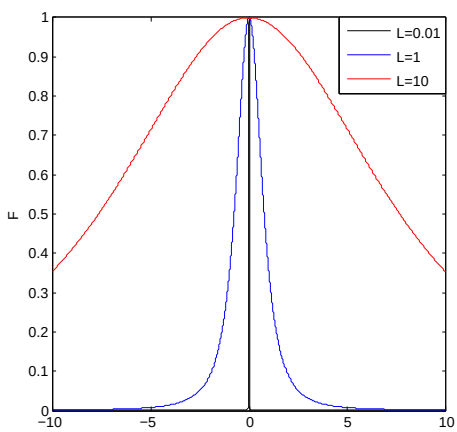

FIGURE B.3. Plot of the function $\bar{F}(S) = \frac{L^3}{(S^2+L^2)^{3/2}}$ over the interval $[-10,10]$ for different values of L.

Still another natural alternative in the periodic case is to use Fourier components

$$F_k^c = \cos(kS), \quad F_k^s = \sin(kS), \quad S = \frac{2\pi s}{T}.$$

There is no need for a center s_0 here, since the use of both sines and cosines renders the F spatially homogeneous. Each step one must use either F^c or F^s with probability $1/2$ each (discounting the steps with $F = 1$), and an integer value for

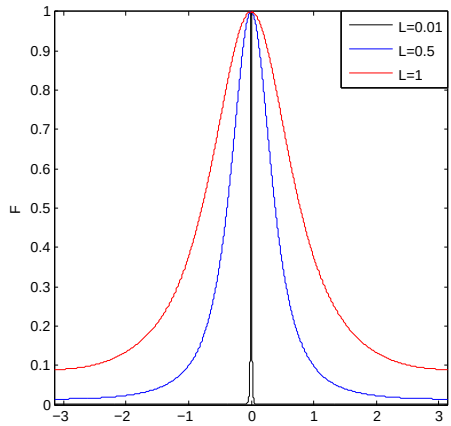

FIGURE B.4. Plot of the function $\hat{F}(S) = \frac{L^3}{(4\sin^2(S/2)+L^2)^{3/2}}$ over the interval $[-\pi,\pi]$ for different values of L.

the wave number k. The latter should be sampled from a distribution that decays rapidly with k, so as to result into smooth composite functions.

An advantage of the use of Fourier modes, particularly when only a finite number K of modes is allowed, is that one can store the accumulated amplitude added to each mode at the various steps, and thus end up with explicit expressions for the non-autonomous dynamical matrix $A(s)$ and shift $b(s)$ as finite Fourier series. To obtain a similar bonus for the non-periodic case (i.e., for representations of the trend), one would need to replace the functions above by others that do not involve a variable length-scale L and random point s_0. A simple choice is that of monomials

$$F_k(s) = s^k,$$

up to a power K. Then the dynamics is represented by a matrix A and a vector b that depend explicitly on s through polynomials of degree K.

When the problem has more that one kind of variable –some periodic and some trendy, for instance–, we can alternate the various types of function $F(s)$ among steps.

B.2 Multidimensional choices

When s lives in a multidimensional space, we can still use the one-dimensional proposals involving s_0 and L above, but picking the direction of space in which they apply each step at random. Yet this is not a very effective procedure when the dimensionality of s is large. An alternative is to use radial functions centered at s_0, such as

$$F(s) = e^{-r^2}, \quad r = \frac{\|s - s_0\|}{L}.$$

Acknowledgment.

The work of M. D. de la Iglesia is partially supported by D.G.E.S, ref. BFM2006-13000-C03-01 and ref. MTM2009-12740-C03-02, Junta de Andalucía, grants FQM-262, FQM-481, P06-FQM-01738, P09-FQM-4643 and Subprograma de estancias de movilidad posdoctoral en el extranjero, MICINN, ref. -2008-0207, and that of E. G. Tabak is partially supported by the National Science Foundation under grant number DMS 0908077.

Bibliography

[1] Box, G. and Jenkins, G., "Time series analysis: Forecasting and control", *San Francisco: Holden-Day*, 1970.

[2] Broomhead, D. S., and King, G. P., "Extracting qualitative dynamics from experimental data", *Physica D*, **20** (1986), 217–236.

[3] Christiansen, B. "The Shortcomings of Nonlinear Principal Component Analysis in Identifying Circulation Regimes", *J. Climate*, **22** (2005), no. 22, 4814–4823.

[4] Crommelin, D. and Vanden-Eijnden, E., "Fitting timeseries by continuous-time Markov chains: a quadratic programming approach", *J. Comp. Phys.*, **217** (2006), 782–805.

[5] Franzke, C., Horenko, I., Majda, A., and Klein, R., "Systematic metastable atmospheric regime identification in an AGCM", *J. Atmos. Sci.*, **66** (2009), 1997–2012.

[6] Ghil, M. et al., "Advanced spectral methods for climatic time series", *Rev. Geophys.*, **40** (2002), 3.1–3.41.

[7] Gugercin, S. and Antoulas, A., "A survey of model reduction by balanced truncation and some new results", *Int. J. Control*, **77** (2004), no. 8, 748–766.

[8] Hannachi, A., Jolliffe, I. T. and Stephenson, D. B., "Empirical orthogonal functions and related techniques in atmospheric science: A review", *Int. J. Climat.*, **27** (2007), no. 9, 1119–1152.

[9] Hasselmann, K., "PIPs and POPs: The Reduction of Complex Dynamical Systems Using Principal Interaction and Oscillation Patterns", *J. Geophys. Res.*, **93** (1988), no. D9, 11015-11021.

[10] Horenko, I., "On identification of nonstationary factor models and its application to atmospherical data analysis", *J. Atm. Sci.*, **67** (2010), pp. 1559-1574.

[11] Horenko, I., Klein, R., Dolaptchiev, S., and Schuette, C., "Automated generation of reduced stochastic weather models i: simultaneous dimension and model reduction for time series analysis", *SIAM MMS*, **6** (2008), 1125–1145.

[12] Jolliffe, I. T., "Principal component analysis", *Springer series in Statistics, Springer-Verlag*, 1986.

[13] Kwasniok, F., "The reduction of complex dynamical systems using principal interaction patterns", *Phys. D*, **92** (1996), no. 1-2, 28–60.

[14] Kwasniok, F., "Optimal Galerkin approximations of partial differential equations using principal interaction patterns", *Phys. Rev. E*, **55** (1997), no. 5, 5365–5375.

[15] Kwasniok, F., "Low-dimensional models of complex systems using principal interaction patterns", Proceedings of the Second World Congress of Nonlinear Analysts, Part 1 (Athens, 1996) *Nonlinear Anal.*, **30** (1997), no. 1, 489–494.

[16] Lorenz, E. N., "Empirical orthogonal functions and statistical weather prediction", *Statistical forecast project report 1, Dept. of Meteor., MIT*, 1956.

[17] Majda, A., Timofeyev, I., and Vanden-Eijnden, E., "Models for stochastic climate prediction", *PNAS*, **96** (1999), 14687–14691.

[18] Malthouse, E.C., "Limitations of nonlinear PCA as performed with generic neural networks", *IEEE Trans. Neural Networks*, **9** (1998), no. 1, 165–173.

[19] Merhmann, V. and Stykel, T., "Balanced truncation model reduction for large-scale systems in descriptor form", *Technical report 157-2004*, TU Berlin, 2004.

[20] Monahan, A.H., "Nonlinear Principal Component Analysis by Neural Networks: Theory and Application to the Lorenz System", *J. Climate*, **13** (2000), no. 4, 821–835.

[21] Monahan, A.H. and Fyfe, J.C. "Comments on "The Shortcomings of Nonlinear Principal Component Analysis in Identifying Circulation Regimes"", *J. Climate*, **20** (2007), no. 2, 375–377.

[22] Monahan, A.H., Fyfe, J.C., Ambaum, M.H.P., Stephenson, D.B., North, G.R., "Empirical Orthogonal Functions: The Medium is the Message", *J. Climate*, **22**, no. 24 (2009), 6501–6514.

[23] Preisendorfer, R. W., "Principal Component Analysis in Meteorology and Oceanography", *Elsevier, New York*, 1988.

[24] Tabak, E. and Vanden-Eijnden, E., "Density estimation by dual ascent of the log-likelihood", *Comm. Math. Sci.*, **8** (2010), 217-233.

[25] The International Research Institute for Climate and Society, http://iridl.ldeo.columbia.edu/SOURCES/.NOAA/.NCDC/.ERSST/

[26] Trenberth, K. E., and Stepaniak, D. P., "Indices of El Niño evolution", *J. Climate*, **14** (2001), 1697-1701.

[27] Wei, W. W., "Time series. Univariate and multivariate methods", *Addison-Wesley Publishing Company, Advanced Book Program, Redwood City, CA*, 1990.

Received Month 200X.