Math 801, Fall 2011

A. Assadi

Review of Spectral Theory for Symmetric Matrices and Operators

Preliminary Discussion. The goal of this section of the lecture notes is to provide the background for the method of Proper Orthogonal Decomposition and various approaches to Model Reduction in dynamical systems and numerical solution of partial differential equations that arise in applications to biology. Numerical simulations of dynamics in biological phenomena require the models to have only a relatively small set of differential equations, because the modeler is constrained to integrate the equations on a spatial-temporal grid that could be computed in a reasonable time. On the other hand, as we have argued earlier, mathematical study of complex dynamical systems essentially belongs to the type of reasoning that is not constrained by pre-set limitations of a particular model, leading to the heuristic argument for their study in an infinite dimensional setting. Even when the final numerical simulation must take place on a finite and rather small discrete approximation, the modeler must ascertain the approximation errors would not exceed the given constraints for biological relevance of the outcomes. One way to handle these opposing demands is to seek a priori estimates on the error bounds when a mathematically rigorous "infinite model" is approximated by a "finite numerical model". In the case of infinite-dimensional evolution equations or partial differential equations, Galerkin Projection is a preferred method to turn them into a finite set of ordinary differential equations while maintaining a mathematically rigorous approach to error estimation. Classically, the Galerkin method projects the functions defining the original equation onto a finite-dimensional subspace of the original phase space that is spanned by eigen-functions of the (differential) operators at hand. However, the Galerkin projection argument is valid without any reference to the nature of the subspace being spanned by the eigen-functions of the original equation. In deriving low-dimensional models we shall ultimately wish to use subspaces spanned by (small) sets of empirically determined vectors in order to establish a firm estimate that relies on the observations and experimental data. Our aim in these notes is to outline a treatment of Galerkin projection that uses any suitable set of basis functions. The particular set of basis functions that we prefer to use are obtained from the method of Proper Orthogonal Decomposition (POD), or Karhunen-Loève Decomposition (KLD) that we shall discuss below in more details. The POD relies on two sets of fundamental mathematical concepts that we shall sketch in this background section. The first is a more sophisticated understanding of *averaging*, primarily insofar as it is needed to justify interchange of averaging and the inner product operations. The second set of concepts are usually part of the so-called *Hilbert-Schmidt theory* that provide conditions under which we are ensured of having a discrete collection of eigenvalues and a corresponding set of orthonormal eigen-functions in a Hilbert Space setting. The extension of eigenvalues in linear algebra to infinite dimensional Hilbert spaces and their operators is the subject of the Spectral Theory that we overview next.

The method of Proper Orthogonal Decomposition and various other approaches to Model Reduction in dynamical systems and numerical solution of PDEs have an overall intuitive reasoning similar to the change of variables familiar from calculus of several variables: We use change of variables when the given form of a function F or a system of equations $\mathcal{E}$ must be modified to a simpler and more computationally manageable form—as in calculation of surface and volume integrals where F(x , y , z) is the integrand, for example. The main step in change of variables is to come up with a new set of coordinates $x=x(\xi_1 ,\xi_2 ,\xi_3)$, $y=y(\xi_1 ,\xi_2 ,\xi_3)$, $z=z(\xi_1 ,\xi_2 ,\xi_3)$ to substitute for (x , y , z), such that the transformed function $\Phi(\xi_1 ,\xi_2 ,\xi_3)$ defined by $F(x(\xi_1 ,\xi_2 ,\xi_3) , y(\xi_1 ,\xi_2 ,\xi_3) , z(\xi_1 ,\xi_2 ,\xi_3)) = \Phi(\xi_1 ,\xi_2 ,\xi_3)$ is easier to work with.

In modeling biological (and numerous other dynamic phenomena), one encounters the "*inverse problem*", where observations of the system are recorded as a cloud of points in a vector space V, and the modeler is required to reconstruct compact and efficient representations of the events and objects in space-time using the data provided by the observation. This is indeed one major case in providing the mathematical structure of the model *ab initio* from the data. The task, therefore, would be to estimate (or "*to learn*") the potentially nonlinear and often stochastic rules that govern the course of events from data. Clearly, the simpler the symbolic (or rule-based) representation of the desired model, the easier and less-error-prone would be the mathematical extraction of the model from data. These considerations motivate a number of approximation methods that we shall see later.

Example. A good example to have in mind is the following method to simplify representation of the surface of an ellipsoid. Let F(x , y , z)=c define a surface for each value of c.

$$F(x , y , z) = 1009x^2 + 1176xy - 542x + 522y^2 - 8700y - 600xz + 544z^2 + 128*25z + 216yz + 83.$$

You can verify that a change of variable formula

$$x = 4u + 3v \ ;$$
$$y = 3u - 4w \ ;$$
$$z = 3v + 4w \ ;$$

and its inverse transform back and forth between F(x , y, z)=c and $\Phi(\xi_1 , \xi_2 , \xi_3)$ =c , where

$$\Phi(\xi_1 , \xi_2 , \xi_3) = 9(\xi_1 - 1)^2 + 25(\xi_2 - 1)^2 + 9(\xi_3 - 1)^2 \ .$$

Of course, a further substitution { $\zeta_1 = \xi_1 - 1$, $\zeta_2 = \xi_2 - 1$, $\zeta_3 = \xi_3 - 1$ } transforms the last expression into the much simpler form to represent the family of surfaces via the equation $\Psi(\zeta_1 , \zeta_2 , \zeta_3)$ = c, where

$$\Psi(\zeta_1 , \zeta_2 , \zeta_3) = 9\zeta_1{}^2 + 25\zeta^2 + 9\zeta_3{}^2 \ .$$

In applications, it could happen tat we are given a data set S whose description would require estimating the structure of surfaces that approximate the data, for example, as in having a time-dependent family of surfaces given by F(x , y , z) = c , where c is proportional to the time parameter. In this situation, it is desirable to *transform the data* (which would be used to define estimation formulas for the surfaces) into a new form that could lend itself more easily to such simpler approximation formulas.

We continue our discussion with the following simple version of the Spectral Theorem for symmetric matrices that is also quite easy to prove. The simple yet basic observation that we need to prove this theorem is “the eigenvectors corresponding to different eigenvalues are orthogonal.” The intuition from statement and proof of this theorem extends to more general circumstances, including the case of Hilbert spaces that is of particular significance in our setting.

Theorem I. Let A be a real symmetric n x n matrix. Suppose A has *n distinct eigen-values* $\lambda_1, \ldots, \lambda_n$ and corresponding unit eigenvectors $\{u_1, \ldots , u_n\}$ (i.e with normalized length $\|u_k\| = 1$, k = 1,..., n.) Then, $\{u_1, \ldots , u_n\}$ is an orthonormal basis of eigenvectors. Let Q = (q_{ij}) be the orthogonal matrix with the columns $(q_{1k} , \ldots, q_{nk})$ are the coordinates of the eigenvectors u_k (k = 1,..., n.) with respect to the standard basis. Then A = $Q^{-1}DQ$, where D is a diagonal matrix with the eigen-values λ_k along the diagonal, and $Q^{-1} = Q^t$. Recall Q^t denotes the transpose, and for complex-valued matrices, Q^* denotes the Hermitian (or conjugate) transpose, i.e. Q^* = complex conjugate (Q^t).

Let us illustrate how this result follows from showing that the eigenvectors corresponding to different eigenvalues are orthogonal. In this simplified case, the assumption that there are *n distinct eigenvalues* $\lambda_1,..., \lambda_n$ allows us to check that the corresponding eigenvectors $\{u_1,... , u_n\}$ are pairwise orthogonal. The latter assertion, in turn, implies that $\{u_1 ,... , u_n\}$ forms a basis for R^n and the claims would follow. We complete the proof by observing that for a pair λ_j and λ_k that correspond to two different (non-zero) eigenvalues, the symmetric matrix A satisfies the following:

$$\begin{aligned}\langle Au_j, u_k\rangle &= \langle u_j, Au_k\rangle\ ,\\ \langle Au_j, u_k\rangle &= \langle \lambda_j u_j, u_k\rangle = \lambda_j\langle u_j, u_k\rangle\ ,\\ \langle u_j, Au_k\rangle &= \langle u_j, \lambda_k u_k\rangle = \lambda_k\langle u_j, u_k\rangle.\end{aligned}$$

Thus $(\lambda_j - \lambda_k)\langle u_j, u_k\rangle = 0$, or $\langle u_j, u_k\rangle = 0$ as claimed.

Proof of the more general form of the theorem above requires the following observation. In the simplified form above, the $n \times n$ matrix A has the additional property that its characteristic equation $f(\lambda) = \det(A — \lambda I) = 0$ has *n* distinct roots. It thus remains to remove this redundant assumption and generalize the argument to the case that $f(\lambda)$ has multiple roots, or where $f(\lambda)$ has less than *n* different roots. We leave the details of this as an exercise, and demonstrate an application.

Consider the initial value problem that requires finding the solution $f : [0,1] \to R^n$ for the system

$$\frac{df(t)}{dt} = Af(t)\ , f(0) = x_0\ ,\ 0 \le t \le 1.$$

Assume further that the $n \times n$ matrix $A = (a_{jk})$ is real-valued, symmetric with constant coefficients. Systems of this form arise in numerous applications and their behavior could be quite complicated. We proceed by selecting an orthonormal basis of eigenvectors of A, say $\{u_1,... , u_n\}$ and form the square matrix *Q* with columns comprised of the coordinates of the eigenvectors of A with respect to the standard basis. Then $B = Q^{-1}AQ$, where *B* is the diagonal matrix with the eigenvalues $\lambda_1,..., \lambda_n$ along the diagonal. Introduce the new variable $g(t) = Q^{-1}f(t)$, or $f(t) = Qg(t)$, where $g : [0,1] \to R^n$. Then, the equation $\frac{df(t)}{dt} = Af(t)$ takes the form

$$\frac{dQg(t)}{dt} = AQg(t),$$
$$\frac{dg(t)}{dt} = Q^{-1}AQg(t) \ ,$$
$$\frac{dg(t)}{dt} = Bg(t),$$

Here, we use the fact that Q is independent of time. As a result of this substitution, we get a system in which the matrix of coefficients is in diagonal form in the new variables. The initial values also change accordingly by substitution. The solution of the new system is readily obtained from a similar one-dimensional ODE for each one of the diagonal elements, because the individual equations are now decoupled, and $g(t)$ could be obtained from the following matrix exponential formula:

$$\begin{pmatrix} \exp(\lambda_1 t) & 0 & 0 & .... & 0 \\ 0 & \exp(\lambda_2 t) & 0 & .... & 0 \\ . & . & . & . & . \\ 0 & 0 & 0 & .... & \exp(\lambda_n) \end{pmatrix}$$

In the new coordinate system, the dynamics of this system is easy to grasp, as the reader is invited to verify directly.

The General Spectral Theorem for Symmetric Matrices

Above, we saw that eigenvalues of a matrix A are roots of the characteristic equation det(A - λI)= 0. In principle, we can find the eigenvalues and eigenvectors of given matrix by first solving the characteristic equation to find all the eigenvalues, and then for each eigenvalue A find corresponding eigenvectors by solving the linear system of equations (A — λI)x = 0. When the matrix A has a very large size, or if we need to consider an operator A defined on a Hilbert space, this approach is no longer suitable, and we present an alternative way of finding the eigenvectors (or constructing them) and eigenvalues of a symmetric matrix A. It turns out that this method also proves the Spectral Theorem for a symmetric $n \times n$ matrix A in the general case with possibly multiple roots. The proof constructs an orthonormal basis of eigenvectors $\{u_1,\dots, u_n\}$ of A by constructing the eigenvectors one by one starting with u_1. When A is a self-adjoint (i.e. symmetric in the finite-dimensional case) operator on a Hilbert space $\mathcal{H}$, then we seek to define a complete basis for the entire $\mathcal{H}$ (or at least for a suitable subspace of interest.)

Constructing the First Eigenvector *u_1* proceeds by considering a variational problem asking for vectors $v \in R^n$ that minimize the Rayleigh quotient

$$F(x) = \frac{\langle Av, v\rangle}{\langle v, v\rangle}.$$

The function $F(x)$ is homogenous of degree zero, that is, $F(\lambda x) = F(x)$, $\lambda \neq 0$. This allows us to use unit vectors in calculation of the minimum, because $F(v) = F(\frac{v}{\| v \|})$, $v \neq 0$. Since the unit sphere in the Euclidean space is compact (bounded and closed), the minimum of the function *F* exists, so we can find a solution: $F(\hat{v}) = \min\{F(v) : v \neq 0, \ \| v \| = 1\}$. It is worth noticing that the function *F* is Lipschitz, so we could also argue for existence of a solution using the contraction mapping theorem without appealing to the compactness of the restricted domain of F. This remark is particularly useful when we must consider linear operators on Hilbert (or Banach) spaces of infinite dimension. Next, we set $u_1 = \hat{v}$ and proceed to check that indeed we have an eigenvector for the operator $A : H \to H$, where H is either a Hilbert space or the n-dimensional Euclidean space. First, notice that the function $F(v)$ is differentiable on the sphere $\{v : \ \| v \| = 1\}$, and that it attains its minimum at $u_1 = \hat{v}$. Therefore, the gradient of $F(v)$ is defined and must vanish at $u_1 = \hat{v}$. Following this argument, and having in mind that A is symmetric (self-adjoint) lead us through the computations below:

$$\begin{aligned}\nabla F(v) &= \nabla\left(\frac{\langle Av, v\rangle}{\langle v, v\rangle}\right)\\ &= \frac{2(\langle v, v\rangle) Av - \langle Av, v\rangle .2v}{\langle v, v\rangle^2}.\end{aligned}$$

Since $\nabla F(\hat{v}) = 0, \ \langle \hat{v}, \hat{v}\rangle A\hat{v} - \langle A\hat{v}, \hat{v}\rangle \hat{v} = 0.$ Therefore, $A\hat{v} = \frac{\langle A\hat{v}, \hat{v}\rangle}{\langle \hat{v}, \hat{v}\rangle}\hat{v}$, that shows $u_1 = \hat{v}$ is an eigenvector with eigen-value $\langle A\hat{v}, \hat{v}\rangle$ for the unit vector $\hat{v}$ (recall $\langle \hat{v}, \hat{v}\rangle = 1$).

Having constructed the first eigen-vector, the remaining argument proceeds by induction once we realize that the orthogonal complement to $\hat{v}$ is also invariant under A, and the entire process could be repeated for the restriction of domain of A to this smaller subspace. As a result, the second eigenvalue will be the minimum of $F(x)=\dfrac{\langle Av,v\rangle}{\langle v,v\rangle}$ for those vectors that are already perpendicular to $\hat{v}$, which implies that the second eigenvalue λ_2 will be either smaller than λ_1 or equal to it, which handles the case of having multiplicity for the roots of the characteristic function of A. Another remark is that the gradient vanishes at maximum value of $F(x)=\dfrac{\langle Av,v\rangle}{\langle v,v\rangle}$ as well. For the case where the symmetric matrix under study arises as the correlation matrix of an rectangular matrix, we could use maximum value of this functional and give a proof of the Singular Value Decomposition or the Principal Components Analysis.

Theorem II. (The Spectral Theorem). Let A be a real symmetric n x n matrix or a self-adjoint operator on a real (or complex) Hilbert space $\mathcal{H}$. Then A has real eigenvalues and there exists a basis for $\mathcal{H}$ corresponding to unit eigenvectors of A. When $\mathcal{H}$ is finite dimensional and $\{u_1,\dots, u_n\}$ is an orthonormal basis of eigenvectors of A, we could form the orthogonal transformation of $\mathcal{H}$ via the matrix Q = (q_{ij}) with the columns $(q_{1k},\dots, q_{nk})$ that are the coordinates of the eigenvectors u_k (k = 1,..., n.) with respect to the standard basis. In this basis, the operator A is represented by a diagonal matrix D = $Q^{-1}AQ$ with the eigen-values of A along the diagonal.

The Norm of a Symmetric Matrix

While an $n\times n$ matrix A naturally arises as an operator on a Hilbert space $\mathcal{H}$, the collection of all such operators has a versatile algebraic structure. In particular, the set of $n\times n$ matrices forms a vector space under addition and multiplication by scalars that are inherited from $\mathcal{H}$. The subset of $n\times n$ symmetric matrices forms a subspace of this vector space of operators. A useful norm (metric) could be defined for the set of all operators by the formula

$$\|A\|=\max\{\frac{\|Av\|}{\|v\|}: v\in R^n, v\neq 0\}.$$

From this definition, we see that always $\|Av\|\leq\|A\|.\|v\|$, so ||A|| is the smallest constant K such that $\|Av\|\leq K\|v\|$ over the entire space. As an application of the Spectral Theorem, we show that when A is symmetric, this constant happens to be computable in terms of eigenvalues of A. namely, $\|A\|=\max\{|\lambda_j|:\ \lambda_j$ is an eigenvalue of A$\}$. Write A=Q^tDQ with an orthogonal Q and a diagonal matrix D that has eigenvalues $\lambda_1,\ldots,\lambda_n$ of A along its diagonal. Since orthogonal transformations are isometries, $||A||=||Q^tDQ||=||D||$. From definition of the norm, we compute easily that $\| D \|= \max\{|\lambda_j|: \lambda_j \text{ is an eigenvalue}\}$ that completes the proof of the assertion.

Generalizations and Discussion

When the matrix A is not symmetric, its eigenvalues need not be real. If we extend the scalars from real numbers to complex numbers, then we can use the Fundamental Theorem of Algebra to conclude that any $n \times n$ matrix has precisely n eigenvalues (counting multiplicities). For non-symmetric matrices, we could focus our attention to the subspace spanned by eigenvectors corresponding to real eigenvalues of A. Since we can form a basis for this subspace comprised of eigenvectors of A, much of the discussion above applies to this situation.

Going back to the proof of the Spectral Theorem, we observe that the construction of the eigenvectors corresponding to the smallest or largest eigenvalues amounts to finding an absolute maximum or minimum of a function. While gradient was used in the proof, there is no reason to confine ourselves to the differential solution. For example, we shall see elsewhere in these lectures that a powerful method for finding maximum or minimum of a function is via evolutionary computation, such as genetic algorithms. These and similar considerations lead to design of efficient algorithms for construction of the first few eigen-vectors “on-line”, that is, without access to the entire data, especially when data is streaming through, as in the case of communication satellites and similar real-time data acquisition methods.

At this point, we must address the issue that non-mathematician readers will ask why we do not immediately resort to numerical simulation. Numerical methods are certainly a major tool in determining the behavior of non-linear systems and their importance and use will continue to grow with the wider availability of high performance computation. On the other hand, analytical methods, used hand in hand with careful simulations, are more useful now than ever. There are several arguments in favor of this assertion, and we shall mention a few:

(a) Resources such as computer time or human time to interpret and evaluate, a "complete" picture of the solutions to a system of reasonable dimension (say more than 4 is sufficient) that depends on several parameters (e.g. 3). It is crucial to locate important or interesting parameter ranges and regions of phase space, and preliminary analyses are typically required to achieve this. Numerical simulations often prompt and suggest such analyses.

(b) In the theory of nonlinear or chaotic dynamical systems, it is always an issue to account for the sensitivity of the system to choice of initial conditions. This difficulty renders numerical approximations subject to scrutiny and results in lack of certainty of the intuitive or heuristics resulting from them.

(c) Numerical analysis is the mathematical theory underlying numerical simulations. This theory typically deals with finite time integrations, while the framework of dynamical systems theory greatly benefits from infinite time properties, asymptotic and phenomena such as attractors. From this viewpoint, the numerical analysis and the analytic theories complement each other, and their combination provides a more reliable and satisfactory theory. See J. Guckenheimer. The role of geometry in computational dynamics, in Dynamics, Bifurcations and Symmetries, P. Chossat, Editor, pages 155-66, Kluwer, Dordrecht, 1994) for more detailed discussion and examples.