Time Series Analysis via Machine Learning
Revealing Decadal Variability and Intermittency in the North Pacific Sector of a Coupled Climate Model

Dimitris Giannakis, Andrew Majda
Courant Institute of Mathematical Sciences, NYU

2011 NASA Conference on Intelligent Data Understanding
Mountain View, CA — October 21, 2011
The data set lies on a hypersurface $S$ embedded in a high-dimensional Hilbert space $H$.

- E.g., $H = \mathbb{R}^n$, where $n$ is the number of gridpoints.
- Each point in $S$ corresponds to an observation “snapshot.”
Singular Spectrum Analysis (SSA)*

Data matrix of \( s \) observations:

\[
X = \begin{pmatrix}
\uparrow & \uparrow & \uparrow & \cdots & \uparrow \\
x_0 & x_{\delta t} & x_{2\delta t} & \cdots & x_{(s-1)\delta t} \\
\downarrow & \downarrow & \downarrow & \cdots & \downarrow \\
\end{pmatrix}, \quad x_t \in \mathbb{R}^n.
\]

Singular value decomposition (SVD):

\[
X = U \Sigma V^T,
\]

\[
U = \begin{pmatrix}
\uparrow & \cdots & \uparrow \\
u_1 & \cdots & u_n \\
\downarrow & \cdots & \downarrow \\
\end{pmatrix}, \quad \Sigma = \begin{pmatrix}
\sigma_1 & 0 \\
0 & \ddots \\
0 & \sigma_{\min\{n,s\}} \\
\end{pmatrix}, \quad V = \begin{pmatrix}
\uparrow & \cdots & \uparrow \\
v_1 & \cdots & v_s \\
\downarrow & \cdots & \downarrow \\
\end{pmatrix}.
\]

**spatial patterns** \( u_i \in \mathbb{R}^n \)  \hspace{1cm} **singular values** \( \sigma_i > 0 \)  \hspace{1cm} **temporal patterns** \( v_i \in \mathbb{R}^s \)

- \( u_i \) are the principal axes of the ellipsoid associated with \( XX^T \).
- \( \sigma_i v_i \) are linear projections of the data onto those axes.
- Procedure might not work well if the geometry of the data set is nonlinear.

---

Seek an operator which is

1. Linear,
2. Defined in a basis-independent manner,
3. Compatible with the nonlinear geometry of $S$.

**temporal patterns:**

square-integrable scalar functions on $S$

$$f \in L^2(S, \nu)$$

**linear map** $A$

**spatial patterns:**

vectors in $H$

$$y \in H$$

**Strategy**

Perform data analysis via low-rank approximations of $A$.

- Well-defined notion in linear algebra.

$$y = A(f) = \int_S d\mu(x_t) f(x_t)x_t$$

- $\mu$: Riemannian measure of $S$. 

Geometrical perspective
Generically, for *incomplete observations* the system trajectory is self-intersecting.

Nearby points in $S$ can actually be far apart in the phase space of the underlying dynamical system.

Embedding “unravels” the data manifold, so that trajectories of $x_t$ are non-intersecting.

$x_t \mapsto X_t = (x_t, x_{t-\delta t}, x_{t-2\delta t}, \ldots, x_{t-\Delta t}) \in M$.

Local neighborhoods of $M$ encode similar dynamical features.
Laplacian eigenfunctions

\[ \Delta \phi_i = \lambda_i \phi_i, \]
\[ \Delta = -\mu^{-1} \frac{\partial}{\partial \xi^i} \mu g^{ij} \frac{\partial}{\partial \xi^j}, \]
\[ 0 = \lambda_0 < \lambda_1 \leq \lambda_2 \leq \cdots \]

The eigenfunctions of the Laplace-Beltrami operator \( \Delta \) form a natural orthonormal basis of \( L^2(M, \mu) \):

\[ \int d\mu(x) \phi_i(x) \phi_j(x) = \delta_{ij}. \]

- \( g \) is the metric tensor.
- \( \mu = (\det g)^{1/2} \) is the Riemannian measure of \( M \).

Eigenfunctions “pick out” different regions of the data manifold.

Good candidates for capturing distinct patterns of variability.

Computed numerically via sparse graph-theoretic algorithms.*

Let \( \{\phi_0, \ldots \phi_l\} \) be the leading \( l \) Laplacian eigenfunctions, and \( \{e_1, e_2, \ldots\} \) an orthonormal basis of \( H \).

The matrix elements of \( A \) with this choice of bases are

\[
A_{ij} = \langle A(\phi_j), e_i \rangle = \int d\mu(x_t) \phi_j(x_t)x^i_t,
\]

where \( x^i_t = \langle x_t, e_i \rangle \).

Spectral decomposition of \( A \):

\[
A_{ij} = \sum_{k=1}^{r} u_{ik} \sigma_k v_{jk}, \quad r = \text{rank} A \leq \min\{l, \text{dim} H\},
\]

Spatio-temporal analysis of the signal:

\[
x_t = \sum_{k=1}^{r} x^k_t, \quad \text{with} \quad x^k_t = u_k \sigma_k v_k(t),
\]

\[
u_k = \sum_{i=1}^{\text{dim} H} e_i u_{ik}, \quad v_k(t) = \sum_{j=1}^{l} v_{jk} \phi_j(x_t).
\]
The North Pacific sector of CCSM3

- Input data is the depth-averaged upper 300-m sea temperature field of the Community Climate System Model 3 (CCSM3).
- Monthly samples from a 700-year equilibrated run at 1° resolution.
- Two-year lag window.
- Time-lagged embedding is crucial to capture distinct patterns of variability.

<table>
<thead>
<tr>
<th>Number of samples</th>
<th>8400</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physical space dimension</td>
<td>534 (# grid points)</td>
</tr>
<tr>
<td>Embedding space dimension</td>
<td>$534 \times 24 = 12,816$</td>
</tr>
</tbody>
</table>
Families of temporal modes

- **Periodic**: Annual and semiannual cycles.
- **Low-frequency**: Spatial patterns resembling the Pacific Decadal Oscillation.
- **Intermittent**: Processes associated with the Kuroshio current and variations in the subpolar and subtropical gyres.
Temperature field reconstructions
Conclusions & outlook

1. Nonlinear Laplacian spectral analysis captures intermittent processes, which carry low variance and are not accessible to classical SSA.

2. Spatial patterns lead to high-quality orthogonal bases for dimensionality reduction of dynamical systems.

3. Ongoing & future work:
   - Analysis of DNS data (2D Rayleigh-Benard convection at $Ra \sim 10^6$).
   - Can this framework be used to build predictive models in the GCM and/or DNS context?