Beyond Empirical Orthogonal Functions: explaining Climate Variability through Optimal Transport

Esteban G. Tabak
Courant Institute of Mathematical Science
New York University

Workshop on Big Data and the Environment

Buenos Aires, November 2015.

Explanation of Variability: Empirical Orthogonal Functions

1. **Data:** variables of interest (temperature, pressure, etc.) on a regular grid, at regular time intervals, with mean (climatology) substracted.
2. **Data arrangement:** re-arrange the spatial grid into a column vector x, form matrix $X = \left\{X_i^j\right\}$ with one column x^j per time.
3. **Singular Value Decomposition:**

$$X = U\Sigma V' = \sum_k \sigma_k u_k v_k',$$

$$\sigma_k \geq 0, \quad (u_k, u_l) = (v_k, v_l) = \delta_k^l.$$

The u_k are the EOFs (principal components of X, modes of variability of the system), each "explaining" a fraction $\frac{{\sigma_k}^2}{\sum_l {\sigma_l}^2}$ of the total variability.

Issues

1. **Interpretability:** Are we talking about internal modes of variability or responses to external factors?
2. Why should the modes of variability be orthogonal, why linear, why time independent.
3. **Data:** Data is typically not gathered on regular grids at regular intervals, hence the need to interpolate or re-analyze.
4. In order not to mix pears and apples, one restricts data to averages over a day-month-season-year: data waste, poorly resolved dynamics.

Proposal: an alternative, natural methodology for the explanation of variability, with interpretable factors and none of the issues above.

A starting point: consolidation of databases (addressing explanation through filtering)

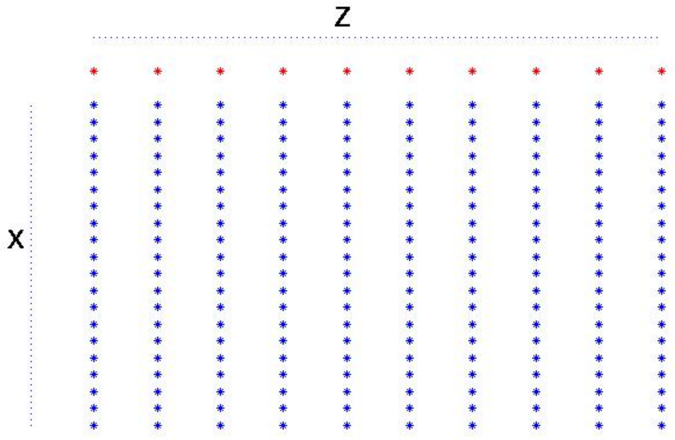
Distinct collections of observations of a single variable or sets of variables: different groups and protocols, different apparatuses, alternative methodologies, improved technology.

Problem: how to put these datasets together.

If the various sets are just amalgamated without further ado, a large fraction of the variability can be attributed to the different data sources.

One would like to clean the data (x) from any indication of its source (z), removing the source-idiosyncratic component of each measurement.

Anti-supervised learning

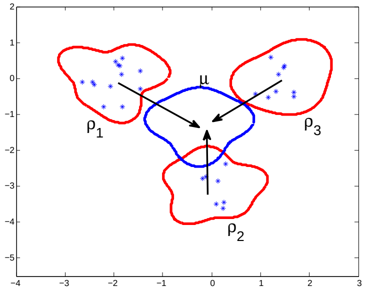


Relation to optimal transport

Filtering all information relating to a factor z (source) from a set of samples x_i is transforming the data

$$x \to y, \quad y = y(x; z)$$

so that one cannot infer from y_i the the corresponding label z_i: the distribution $\mu(y)$ underlying the y_i must be independent of the label.



Formulation

$$x \to y, \quad y = y_k(x)$$

$$\forall A \quad \int_{y_k^{-1}(A)} \rho_k(x)\ dx = \int_A \mu(y)\ dy$$

$$\min_{\mu, y_k} D = \sum_{k=1}^{K} P_k \int c\left(x, y_k(x)\right) \rho_k(x)\ dx.$$

Canonical cost:

$$c(x, y) = \|y - x\|^2.$$

Formulation in terms of samples

1) Original formulation (*alla* Monge): maps $y = y_k(x)$,

$$y_k : \rho_k(x) \to \mu(y),$$

$$\min_{\mu, y_k} D_M = \sum_{k=1}^{K} P_k \int c\left(x, y_k(x)\right) \rho_k(x)\ dx.$$

2) Relaxation (*alla* Kantorovich): couplings $\pi_k(x, y)$,

$$\int \pi_k(x, y)\ dy = \rho_k(x), \quad \int \pi_k(x, y)\ dx = \mu(y),$$

$$\min_{\mu, \pi_k} D_K = \sum_{k=1}^{K} P_k \int c\left(x, y\right) \pi_k(x, y)\ dxdy.$$

Formulation in terms of samples (continuation)

3) Dual problem: Lagrange multipliers:

$$\int \pi_k(x,y)\, dy = \rho_k(x) \quad \rightarrow \quad \phi_k(x)$$

$$\int \pi_k(x,y)\, dx = \mu(y) \quad \rightarrow \quad \psi_k(y).$$

$$\max_{\phi_k,\psi_k} \sum_{k=1}^{K} \int \phi_k(x) \rho_k(x)\, dx,$$

$$\phi_k(x) + \psi_k(y) \leq P_k c(x,y), \; \sum_{k=1}^{K} \psi_k(y) \geq 0.$$

4) In terms of the data x_i, k_i,

$$\max_{\phi_k,\psi_k} \sum_{k=1}^{K} \frac{1}{m_k} \sum_{k_i=k} \phi_k(x_i), \quad \phi_k \in F.$$

Two considerations

- Effect of the selection of a space F for ϕ_k on the restrictions of the primal problem:

$$\int \pi_k(x,y)\ dy = \rho_k(x) \rightarrow$$

$$\forall u(x) \in F, \quad \int \left[\pi_k(x,y)\ dy - \rho_k(x)\right] u(x)\ dx = 0.$$

Example: if F is the space of quadratic functions, $\left[\int \pi_k(x,y)\ dy\right]$ must agree in expected value and variance with ρ_k.

- Connection between Kantorovich's dual and Monge's primal for the canonical cost:

$$y_k(x) = x - \nabla\phi_k(x).$$

Consequence: the example above yields linear maps.

A poor man's solution: linear maps in one dimension

For each component of x, propose $y_k = \alpha_k x + \beta_k$.

Procedure:

1. group the $\{x_i\}$ per class k,
2. estimate $\bar{x}_k$, σ_k (empirical mean and standard deviation),
3. optimal transport + canonical cost $\rightarrow$

$$\bar{y} = \sum_k P_k \bar{x}_k, \quad \sigma_y = \sum_k P_k \sigma_k,$$

4. compute

$$\alpha_k = \frac{\sigma_y}{\sigma_k}, \quad \beta_k = \bar{y} - \alpha_k \bar{x}_k$$

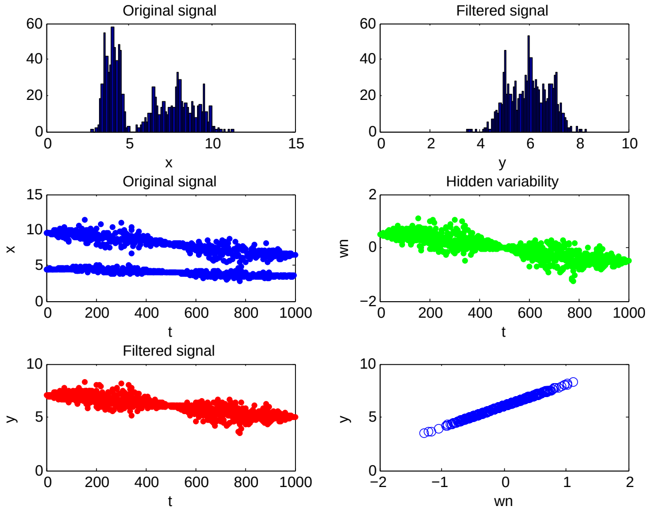
5. and filter

$$y_i = \alpha_{k_i} x_i + \beta_{k_i}.$$

Example

We create synthetic data, consisting of a signal $x_j = F(z_j, w_j)$, where w_j, the "hidden signal", is white noise with time-dependent parameters, $z_j \in \{0, 1\}$, the "source", is chosen randomly for each j, and F is a linear function of w_j, with parameters depending on z_j.

Example (data and results)



Interpretation and extensions

Filtering data source is not different from explaining away discrete variability factors: season, day vs. night, etc.

Natural extension: continuous factors z, such as time of the day or year:

$$y_k(x) \to y(x|z), \quad \rho_k(x) \to \rho(x|z).$$

In our poor man's solution, $\rho(x|z) \to \bar{x}(z),\ \sigma(z)$,

$$\bar{y} = \frac{1}{m}\sum_j \bar{x}(z), \quad \sigma_y = \frac{1}{m}\sum_j \sigma(z_j),$$

$$y_i = \alpha\left(z_i\right) x_i + \beta\left(z_i\right),$$

$$\alpha(z) = \frac{\sigma_y}{\sigma(z)}, \quad \beta(z) = \bar{y} - \alpha(z)\bar{x}(z).$$

Continuous factor (continued)

To avoid granularity, we may propose for instance

$$\bar{x}(z) = A + Bz, \quad \sigma(z) = e^{C+Dz},$$

and fit (A, B, C, D) to the data through maximal likelihood:

$$(A, B, C, D) = \arg\max L = \sum_j \log\left[N\left(x_j | \mu(z_j), \sigma(z_j)\right)\right].$$

For multifactors (z vectorial),

$$\bar{x}(z) = A + B \cdot z, \quad \sigma(z) = e^{C+D \cdot z}.$$

Furthermore, z can include any number of nonlinear features.

Time series

Consider for instance a time series generated by a Markov process,

$$x_{n+1} = F(x_n, w_n),$$

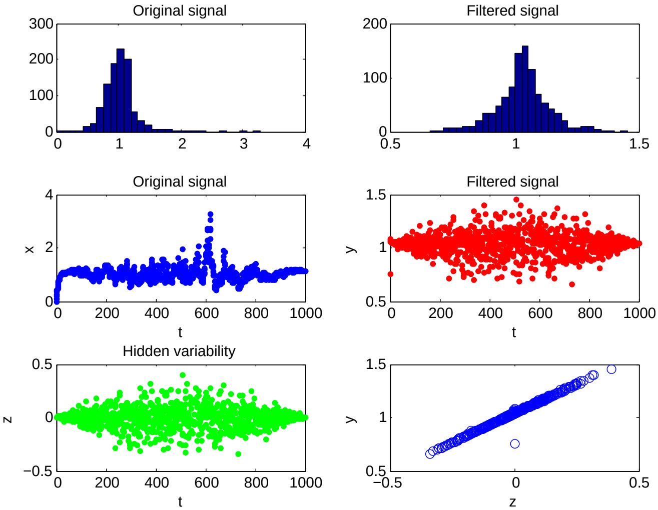
where only the x are observed, and F and w_n are unknown. Then the factor that plays the role of "z" is the prior element x_n in the time series.

In the example below, we generated a time series from the model

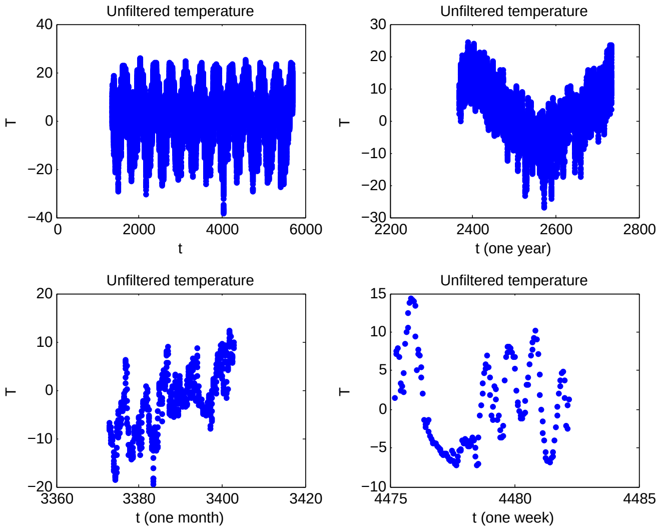
$$x_{n+1} = 0.1 + \frac{10}{11} x_n + e^{x_n - 1} w_{n+1},$$

where w_n is white noise modulated by a sinusoidal amplitude.

Time series (data and results)



Real data: hourly temperature in Boulder, CO

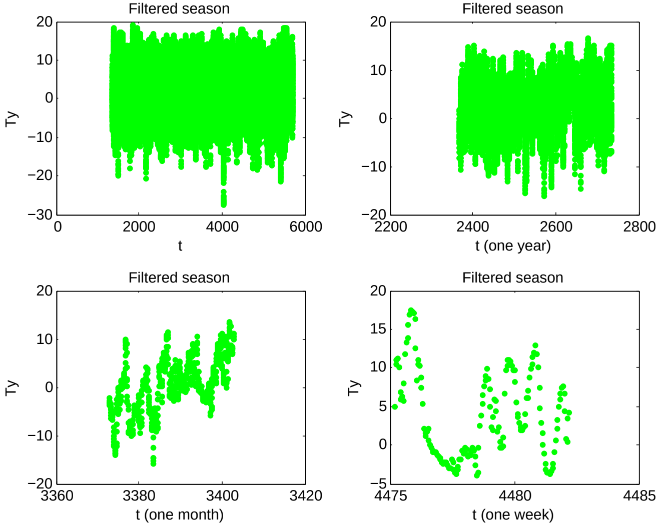


Filtering the season

$$t_y = \frac{2\pi t}{365.25}$$

$$z = \begin{pmatrix} \cos\left(t_y\right) \\ \sin\left(t_y\right) \\ \cos\left(2t_y\right) \\ \sin\left(2t_y\right) \\ \cos\left(3t_y\right) \\ \sin\left(3t_y\right) \\ \cos\left(4t_y\right) \\ \sin\left(4t_y\right) \end{pmatrix}$$

Temperature with season filtered



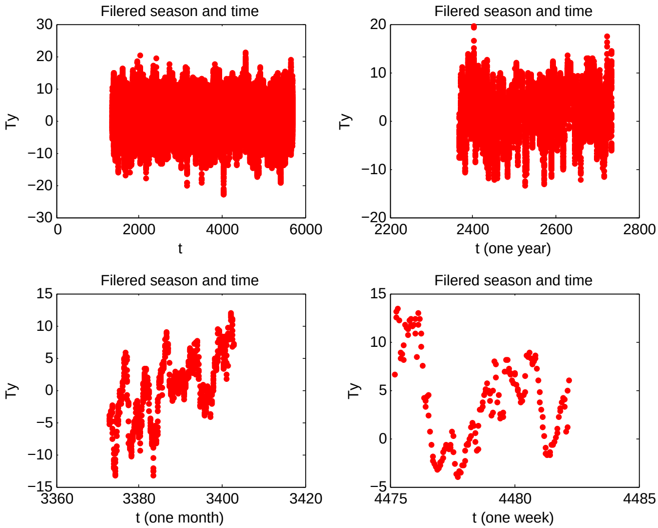
Filtering time and season

$$t_d = 2\pi t$$

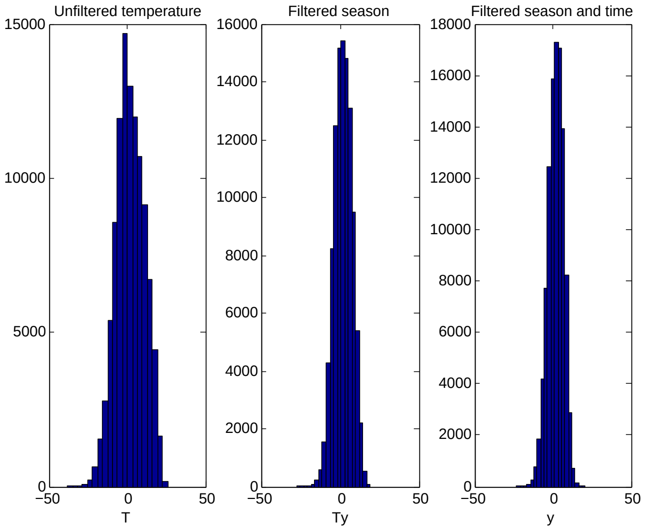
$$t_y = \frac{2\pi t}{365.25}$$

$$z = \begin{pmatrix} \cos(t_y) \\ \sin(t_y) \\ \cos(t_d) \\ \sin(t_d) \\ \cos(t_y)\cos((t_d) \\ \dots \\ \sin(4t_y)\sin(4t_d) \end{pmatrix}$$

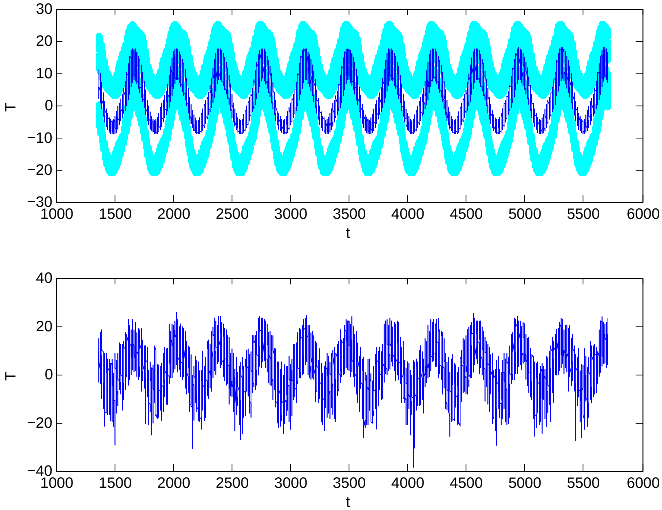
Temperature with time and season filtered



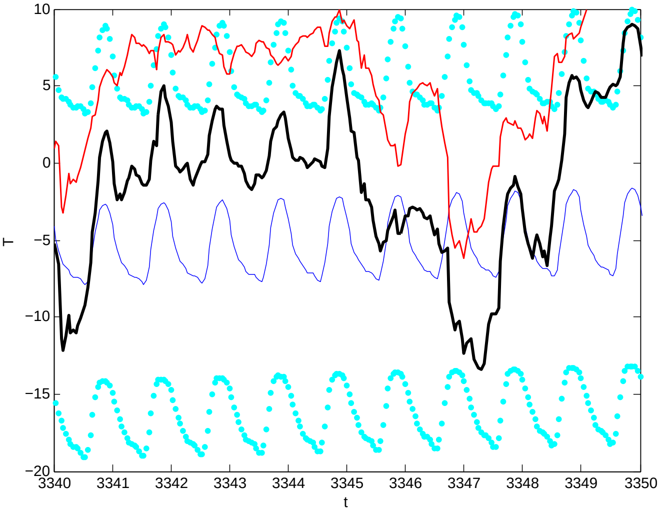
Histograms


Unfiltered temperature
Filtered season
Filtered season and time
15000
10000
5000
0
−50
0
50
T
16000
14000
12000
10000
8000
6000
4000
2000
0
−50
0
50
Ty
18000
16000
14000
12000
10000
8000
6000
4000
2000
0
−50
0
50
y

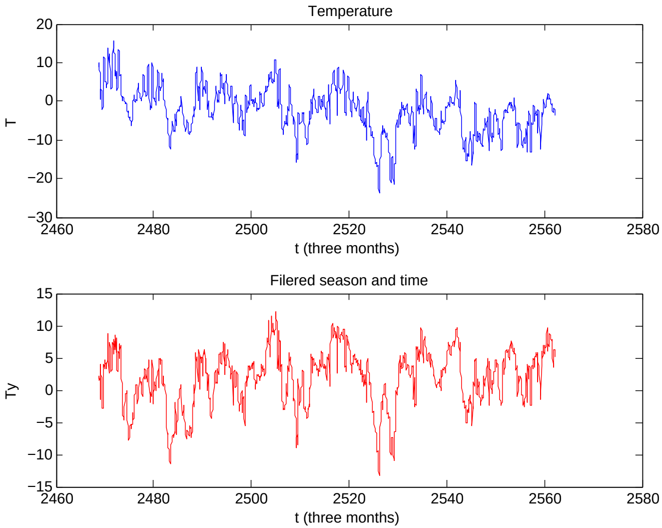
Predicted and realized variability



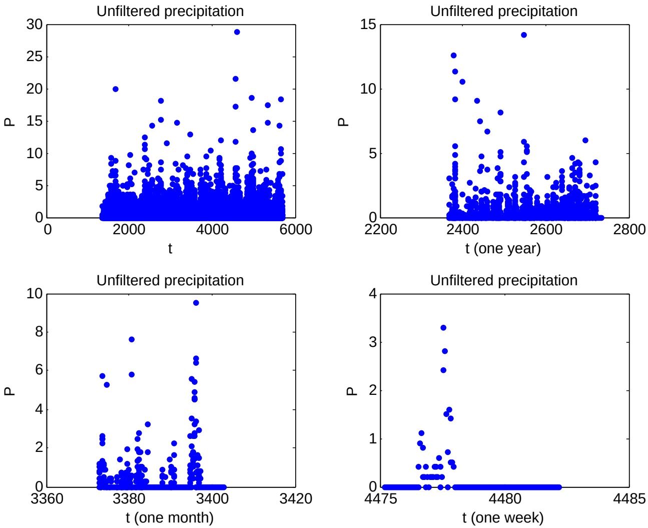
Predicted, realized and filtered variability (10 days)



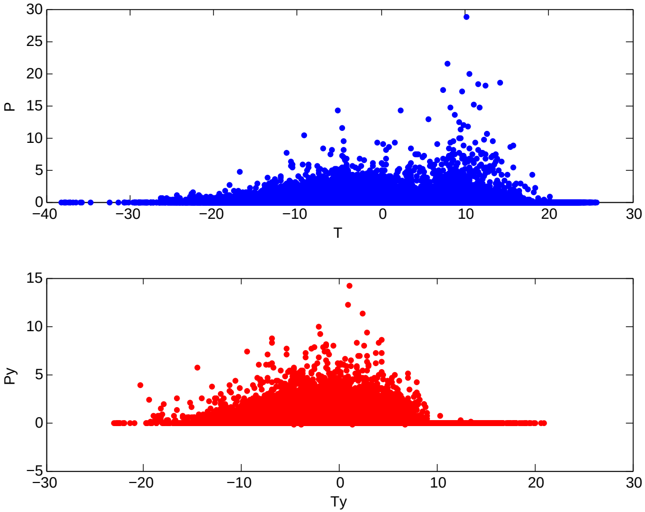
Realized and filtered variability (3 months)



Precipitation



Temperature vs. precipitation, realized and filtered



Further ado

Bring in more sites!

No current need to sample at regular intervals; with coordinates added as explanatory factors, no need for a spatial grid either.

Some of many other factors to add:

- Solar radiation
- Level of CO_2
- Altitude, distance from sea, nature of soil.

Factor discovery (clustering + explanation), including internal modes of variability

Dynamics, time series analysis

No need to stick to a poor man's tools.

Summary

- Generalizations of optimal transport provide a conceptual and computational framework for "anti-supervised learning": the removal from data of information that can be explained by external factors.
- A plethora of applications to climate, including consolidation of data sets, explanation of variability by external factors, discovery of internal modes of variability.
- Flexible and robust computational tools, ranging from state-of-the-art data-driven optimal transport to poor man solutions restricted to linear maps.
- Much to do: apply to much more data, theoretic and algorithmic developments, generalizations.