An overview of statistical methods

How does it fit all together?

* Background Review: fundamental statistical measures (e.g. PDF, moments of a PDF, JPDF, random variable and function of random variables)

$\Downarrow$

the goal is to provide us basic statistical descriptions of the system we want to study.

* Combining MODELS and OBSERVATION:

Typically we have a set of variable

$$\underline{y} = \begin{bmatrix} \\ \\ \\ \end{bmatrix} \quad \text{and} \quad \underline{x} = \begin{bmatrix} \\ \\ \\ \\ \end{bmatrix}$$

and a model that provides a relationship between $\underline{x}$ and $\underline{y}$ so that:

$$\underline{y} = \underset{\text{model}}{\underline{\underline{E}}}\, \underline{x} + \underset{\text{noise}}{\underline{n}}$$

* maybe non-linear $\underline{y}(t) = \underline{\underline{E}}(\underline{x}_t) + \underline{n}$

Examples

$\underline{y}$ = $\underline{\underline{E}}$ $\underline{x}$ + $\underline{n}$

$\underline{y}$ data / variable we want to model	= $\underline{\underline{E}}$ model	$\underline{x}$ model input parameters + $\underline{n}$
1 2 concentration of a pollutant in water/air $\underline{y} = \underline{c}_t$	Advection/diffusion equation ① $\frac{dc}{dt} = \underset{\text{velocity}}{\underline{u}} \cdot \nabla c + \underset{\text{source}}{Q}$ discretize and integrate ② $c_{t+1} = c_t + \int rhs\, dt$ ③ $\underline{c}_{t+1} = \underline{\underline{R}}(t, t+1)\, \underline{c}_t$ $\underbrace{\underline{c}_t}_{\underline{y}} = \underbrace{\underline{\underline{R}}(t_0, t)}_{\underline{\underline{E}}}\, \underbrace{\underline{c}_{t_0}}_{\underline{x}}$	$\underline{x} = \underline{c}_{t_0}$ initial condition or $\underline{x} = \begin{bmatrix} \underline{c}_{t_0} \\ \underline{u} \\ \underline{Q} \end{bmatrix}$ initial cond. + model parameters
2 1 global temperature trends $\underline{y} = \begin{bmatrix} T(t_1) \\ T(t_2) \\ \vdots \\ T(t_N) \end{bmatrix}$	$T(t) = \underset{\text{trend component}}{at} + \underset{\text{exponential component}}{bt^2}$ $\underbrace{\begin{bmatrix} T_1 \\ \vdots \\ T_N \end{bmatrix}}_{\underline{y}} = \underbrace{\begin{bmatrix} t_1 & t_1^2 \\ t_2 & t_2^2 \\ \vdots & \vdots \\ t_N & t_N^2 \end{bmatrix}}_{\underline{\underline{E}}} \cdot \underbrace{\begin{bmatrix} a \\ b \end{bmatrix}}_{\underline{x}}$	$\underline{x} = \begin{bmatrix} a \\ b \end{bmatrix}$ the model parameters

$\underline{y}$	$\underline{\underline{E}}$	$\underline{x}$ $+ \underline{n}$
3 Precipitation in the Tropics $\underline{y} = \text{precip}(x, y)$	Multilinear regression of precip with SST e.g. $\text{precip}_1 = a_{11} SST_1 + a_{12} SST_2 \ldots$ spatial location $\underline{\underline{E}} = \begin{bmatrix} a_{11} & a_{12} & \cdots \\ a_{21} & \ddots & \\ & & a_{NM} \end{bmatrix}$	$\underline{x} = \begin{bmatrix} SST_1 \\ STT_2 \\ \vdots \\ \vdots \end{bmatrix}$
4 Your own research example.		

A

Forward ('linear') modelling

if you know $\underline{x}$ but not $\underline{y}$

$$\underline{y} = \underline{\underline{E}}\,\underline{x} + \underline{n}$$

B Inverse or "adjoint" modeling

if you know $\underline{y}$ but not $\underline{x}$

$$\underline{y} = \underline{\underline{E}}\,\underline{x} + \underline{n}$$

$$\underline{\underline{E}}^T \underline{y} = \underline{\underline{E}}^T \underline{\underline{E}}\,\underline{x} + \underline{\underline{E}}^T \underline{n}$$

$$\underline{\underline{E}}^T (\underline{y} + \underline{n}) = \underline{\underline{E}}^T \underline{\underline{E}}\,\underline{x}$$

$$(\underline{\underline{E}}^T \underline{\underline{E}})^{-1} \underline{\underline{E}}^T (\underline{y} + \underline{n}) = \underline{x}$$

it involve an inverse matrix of the model $(\underline{\underline{E}}^T\underline{\underline{E}})^{-1}$

it involves the adjoint of the model $\underline{\underline{E}}^T$

• combining models and observations

1) "EMPIRICAL" (e.g. system of linear equations) with parameters: $y = a x$ or $y = a \cos(\omega t)$ ← you choose the model

least square typically used to fit model to observations and computer aposteriori errors.

2) "DYNAMICAL" (e.g. 1) an atmospheric general circulation model
2) eq. describing the ~~seismic~~ evolution of seismic waves) ← you choose the dynamics

with parameters: initial and boundary conditions, convective parametrization, coefficients of the stress tensor, ... in general anything that you think is uncertain in the model that needs to be "constrained" by the available information.

again least square tecnique or typically used to fit model to observations —

fiting the model requires varying the parameters and understanding how the model responds to these changes "in a linear way" → 1) concept of model sensitivity
2) variational data assimilation
3) inverse modeling / adjoint modeling

3) "STATISTICAL MODEL" : mo the core of the model is based on statistics computed from the data.

⇓

most simple statistic is based on linear relationships and involves a covariance or multiple covariances. → regressions or multivariate regressions

⇓

The linear relationship is typicaly used in interpolation → space → extrapolation
↘ time → forecasting

How to build a covariance?

Models build on covariance often lead to the so called "statistical optimal estimators"

CASE 1: ↔ CASE 2:

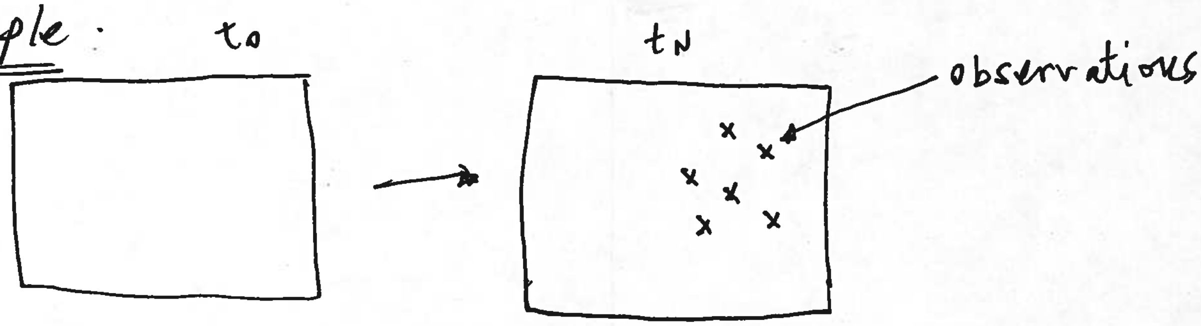
you can compute the covariance from the data (derive the statistical optimal estimator now see page 7 of notes)!

you ~~have~~ have to assume the shape of the covariance using other information or assumptions

"EMPIRICAL" and "DYNAMICAL" models provide such covariance information!

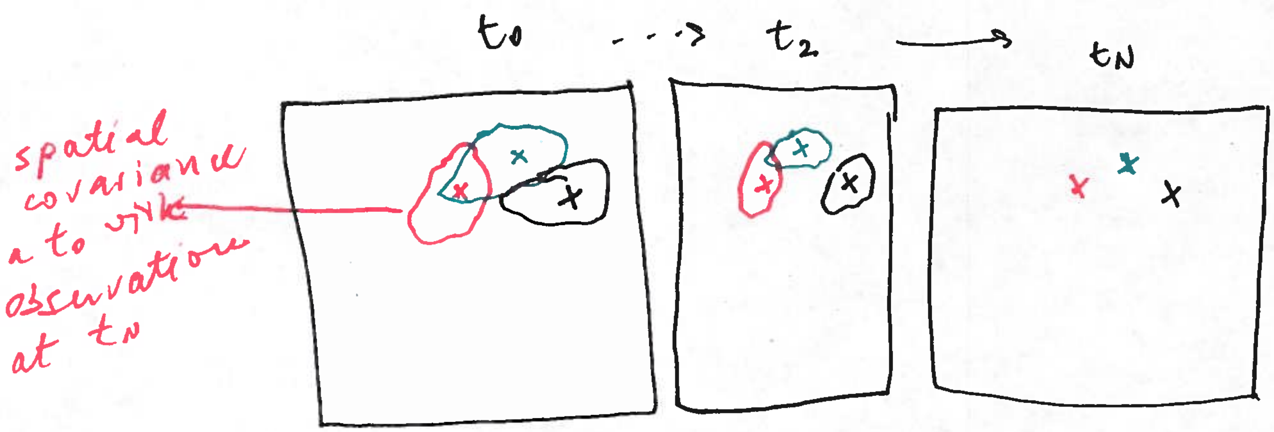
<u>NOTE:</u> when using "EMPIRICAL" and "DYNAMICAL" models statistics are still needed to constraint the unobserved degrees of freedom.

<u>example:</u>

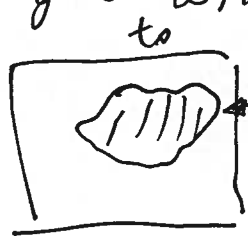


1) Assume you know the dynamics that go from $t_0 \to t_N$
2) you have a set of TEMPERATURE OBSERVATIONS at t_N and would like to know the temperature field at t_0.

3) Assume that using the model "dynamics" you can retrive the optimal covariance between the observations and the temperature field at t_0



Do the same for each point observation. At the end you will obtain a map at t_0



where is the area for which observations at t_N provide us information for.

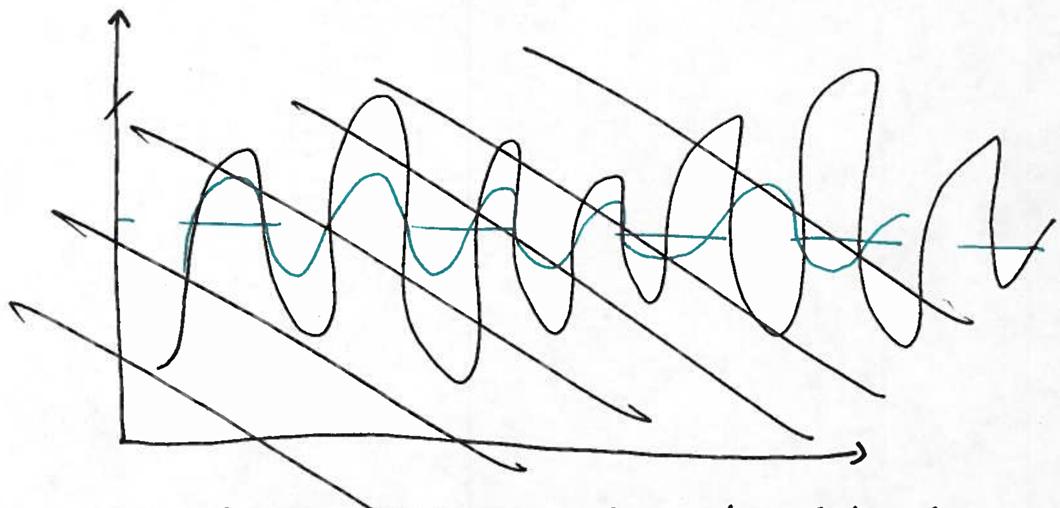
This means that the observations at t_N provide $\emptyset$ information (= carry $\emptyset$ relationship) with the TEMPERATURE field at t_0 outside this area.

All the TEMP. points at t_0 are the degrees of freedom $\rightarrow$ this implies that we have a large number of unobserved degree of freedom for the system (= observations do not provide sufficient information to constrain them).

So what do you do?

Let us look at "EMPIRICAL MODELS" again:

FIGURE 1



consider a time series of TEMPERATURE and assume you want to remove the seasonal cycle. One can do this by fitting the following function.

$$y = a_0 + b\sin(\omega t) + c\cos(\omega t)$$

Once the fit is done the variance of the seasonal cycle is

$$\frac{b^2 + c^2}{2} = \text{variance of seasonal cycle}$$

In principal one could do the same exercise for each frequency and study the variance explained by each of those components $\longrightarrow$ SPECTRA

$\Downarrow$

spectral analysis.

SPECTRAL analysis is an example of decomposing a signal on a different basis, in this case the sin / cos functions $\rightarrow$ FOURIER SERIES.

$\Downarrow$

SIGNAL DECOMPOSITION

other types of decompositions exist depending on the choice of "basis set" or "basis function" (eg. sherical harmonics, etc...)
~~[illegible]~~ ... $\rightarrow$ wavelet analysis.

Let us now consider a "statistical model"

$$y = \underset{\text{param.}}{\alpha} x + \underset{\text{error}}{n} \qquad \text{model}$$

take the mean $\langle\ \rangle$

$$\langle y \rangle = \alpha \langle x \rangle + \langle n \rangle$$

define $y' = y - \langle y \rangle$

$$y' = \alpha x' + n'$$

now compute the covariance between $\langle x' y' \rangle$

$$\langle x' y' \rangle = \alpha \langle x' x' \rangle + \underbrace{\cancel{\langle n' n' \rangle}}_{= \emptyset} \Rightarrow \alpha = \frac{\langle x' y' \rangle}{\langle x' x' \rangle}$$

If you define y = as the model output

x = as the model input

$\underline{\underline{C}}_{oo}$ = covariance of the outputs × outputs

$\underline{\underline{C}}_{oi}$ = covariance of the outputs × inputs

vector of output variable

$$\underline{o} = \underbrace{\underline{\underline{C}}_{oi}\, \underline{\underline{C}}_{ii}^{-1}}\, \underline{i}$$

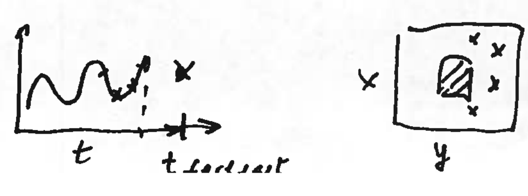
vector of input data

statistical model based on the covariance

↳ FUNDAMENTAL RESULT which is at the root of most statistical
① interpolation
② extrapolation
③ hindcast = time interpolation
④ forecast

The flavor of the various methods in ①, ②, ③, ④ depends on how one specifies the "covariance"

example:



∴ One can decompose the covariance in eigenvalues and eigenmodes → EOFs

another example of signal decomposition, in which the basis set is made up of ~~modes~~ orthogonal modes explaining a certain fraction of the variance.