Noise in measurements and data analysis: State estimation, models from data, and forecasts

Holger Kantz

Max Planck Institute for the Physics of Complex Systems, Dresden

1. Assessing and reducing uncertainty of recorded data
2. Model equations from observed data
3. Forecasting: dealing with uncertainty of the "initial condition"

Sorry, no technology, no informatics

Errors on time series data

Abstract setting

signal $x(t)$, recorded data s_n.

In between:
Measurement uncertainty and digitization, noise in transmission channel, uncertainty of clock, systematic measurement errors.

We like to assume:
$s_n = x(t_n) + \xi_n$, where ξ_n is i.i.d. (additive measurement noise).

Sometimes very bad approximation!

Dynamical measurements

Measurement device is a dynamical input-output system: measurement output $y(t)$, system input $x(t)$:

$$\dot{y}(t) = f(y, x)$$

such that: if $x(t) = x_0$: $y(t) \to y_\infty(x_0)$ for $t \to \infty$
(dissipative system, globally attracting fixed point).

If $x(t)$ a true function of time: non autonomous dynamics of $y(t)$.

Problem 1: $y(t_n) \neq y_\infty(x(t_n))$ (no instantaneus relaxation).

Dynamical measurements: $y(t) \neq y_\infty(x(t))$

Instead: the output $y(t_n)$ for given $x(t_n)$ may depend on the past $x(t),\ t < t_n$.

Skew dynamical system (measurement device plus object of investigation)
Difficulties to define error bars and standards.

Introduces temporal correlations into the measurement errors!
Introduces correlations between signal (its time derivative) and errors.

Technological solution: relaxation time much shorter than the sampling interval.
Partial fix: considering the linear effects as low-pass filter ([Badii et al. (1988)]: can increase the dimension.)

Noise reduction

Removing additive noise

Assume noisy data $s_n = x(t_n) + \xi_n$ with some noise.

Goal of noise reduction: estimate ξ_n and form $x_n = s_n - \xi_n$.
(individually for every n)

Criteria to distinguish between signal and noise

Standard technique: linear filters in the Fourier domain

Example: low-pass filter:
Assume: $x(t)$ is smooth function of time, noise ξ is i.i.d.
Then:
$x(t_{n+k}) = x(t_n) + k\Delta t\dot{x}(t_n) + O(k^2\Delta t^2)$
$\frac{1}{2}(s_{n-1} + s_{n+1}) = x(t_n) + \frac{1}{2}(\xi_{n+1} + \xi_{n-1}) + O(\Delta t^2)$
Variance of "new" noise $\frac{1}{2}(\xi_{n+1} + \xi_{n-1})$ is reduced by $1/2$.

Wiener filter

Noise model: moving average (MA) process with known power spectrum

data model: nontrivial power spectrum (preferably with peaks, e.g., AR).

Optimal linear filter: Wiener filter:

$$\hat{S}_k = \sqrt{\frac{S_k^{(s)} - S_k^{(n)}}{S_k^{(s)}}} S_k^{(s)}$$

$S_k^{(y)}$ power of signal y in frequency bin k.

Unsuccessful if too much overlap between noise power spectrum and signal power spectrum, non-Gaussian noise.

Other criteria than in the frequency domain?

Noise reduction exploiting redundancy

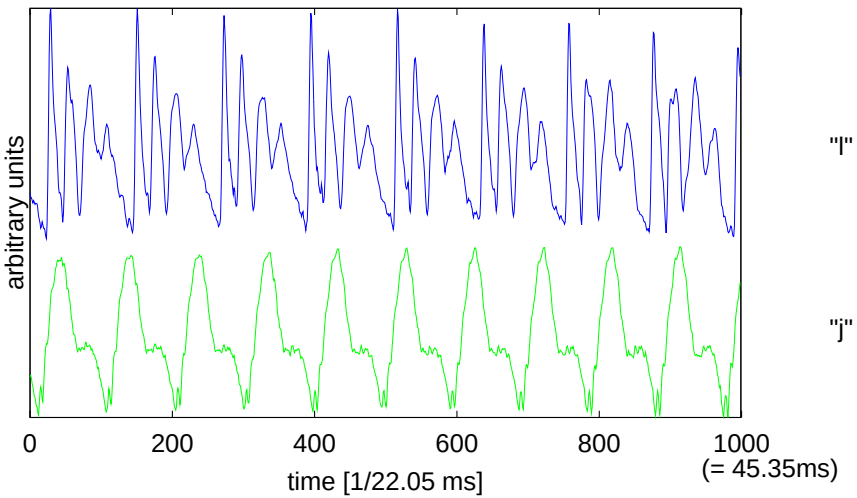
If signal satisfies E.N. Lorenz' concept of analogy:
similar data segments $(x_k, x_{k+1}, \ldots)$ imply a similar continuation
(justified for deterministic origin, but also for a "strong" grammar, i.e., for high redundancy in stochastic data, e.g., human language):

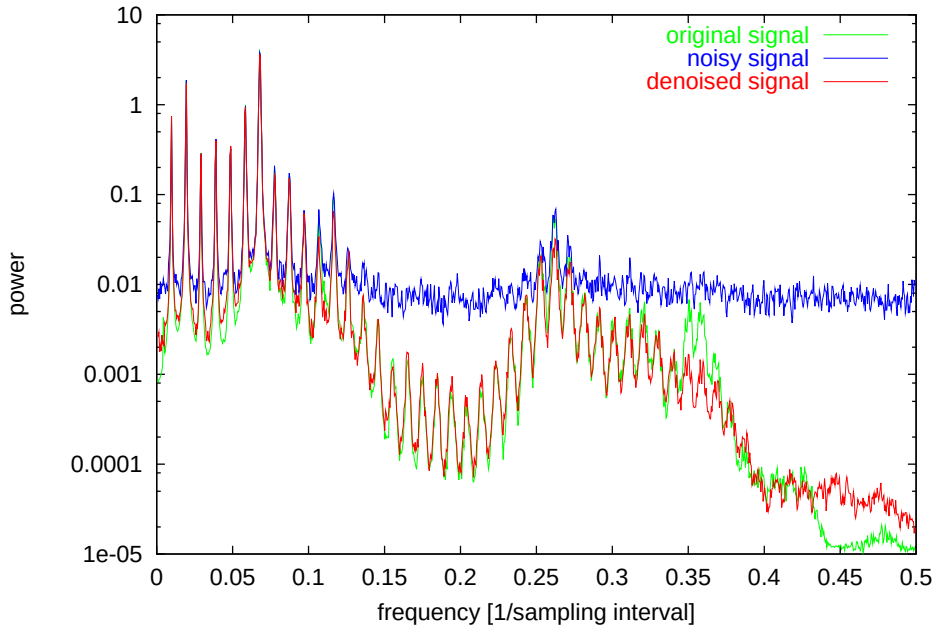
Synchronised average over similar time series segments (plus higher order corrections: [Grassberger et al. 1993]) (non-causal!)

$$\hat{s_n} = \frac{1}{||\mathcal{U}_\epsilon(\mathbf{s_n})||} \sum_{k \in \mathcal{U}_\epsilon(\mathbf{s_n})} s_k$$

where $\mathbf{s}_n = (s_{n-m}, s_{n-m+1}, \ldots, s_n, s_{n+1}, \ldots, s_{n+m})$

Example: human articulated voice:





Redundancy concept useful for image processing or video streams?

Problem 2:
Find other concepts of distinction between signal and noise

Constructing model equations from data

Concept of deterministic models

Takens time delay embedding with embedding dimension m, deterministic map:

$$x_{n+1} = f(x_n, x_{n-1}, \ldots, x_{n-m+1})$$

Or alternatively: vector valued observations $\vec{x}$,

$$\vec{x}_{n+1} = \vec{f}(\vec{x}_n)$$

Standard approach: Ansatz for f, least squares problem for parameter fitting :

$$\langle (x_{n+1} - f(x_n, \ldots, x_{n-m+1}))^2 \rangle = \text{ min}$$

errors in variables

$$\langle (x_{n+1} - f(x_n, \ldots, x_{n-m+1}))^2 \rangle = \min$$

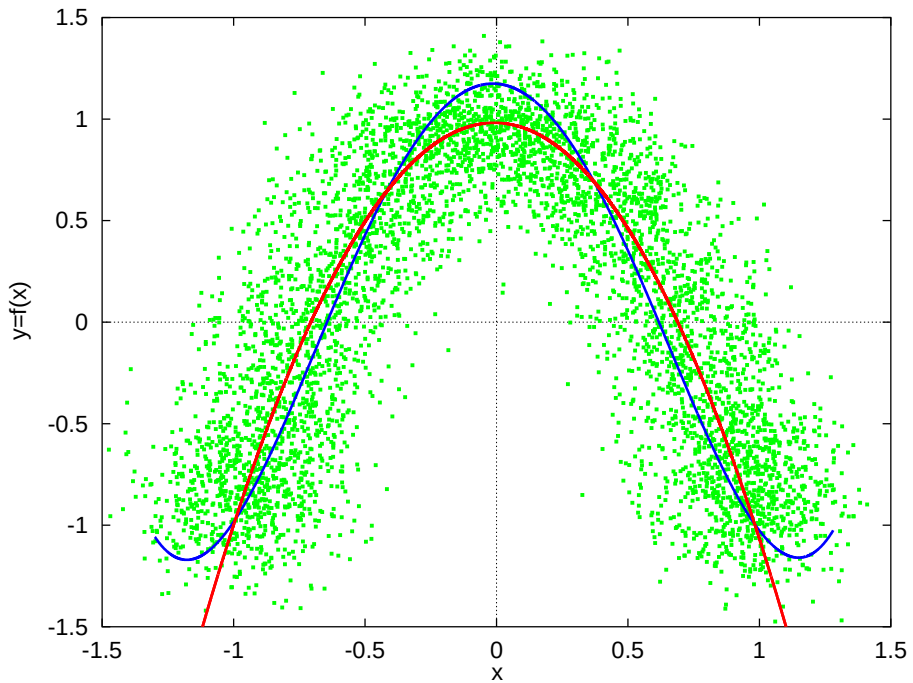
Noise on x_{n+1}: no problem, maximum likelihood model for uncorrelated data sample (not true for time series data).

However: if x_{n+1} is noisy, then also x_n, x_{n-1}, ...

Problem 3:

errors in independent variables.

formal solution:
total least squares
partial solutions:
shadowing (e.g., multiple shooting [Bock & Plitt (1984)]),
tractable for chaotic systems?



Stable models

How to guarantee a stable model?

Linear (ARMA, ARIMA) models: Known criteria for stability, known algorithms to construct stable models from data (e.g., [Box & Jenkins (1970)]).

Nonlinear (stochastic) models: no general constraints of model parameters which guarantee stability
No algorithms known which generate only stable models.

Consequence: maximum likelihood estimates of model parameters are not guaranteed to yield a stable model.
Problem 4: Ensuring stability of fitted (nonlinear) models or finding the closest stable model to an unstable fitted model

Lack of structural stability

Generic dynamical systems are not structurally stable.
I.e., even if solutions remain in a bounded regime (=stability of the model), their asymptotic behaviour may sensitively depend on parameters.

Problem 5: robust fitting results for non-hyperbolic systems?
How to ensure that the asymptotics of the model is similar to the asymptotics of the data?
partial solution (purely deterministic): shadowing.
or: stochstic models (Langevin equations)
[Friedrich & Peinke (1997)]
[Compare also talks by A. Duggento and A.A. Dubkov (yesterday)]
Non Gaussian noises? Correlated noises?

time discrete stochastic models

estimate transition probabilities (e.g., by kernel estimators):

$$p(x_{n+1}|x_n, x_{n-1}, \ldots, x_{n-m})$$

- guaranteed to be stable
- no choice of basis functions: flexible

essential: Implies a Markov approximation

Problem 6: Order of the Markov approximation?
Criteria for truncation of the memory needed.
[Paparella et al. (1997), Kantz et al. (2004)]

State estimation

identifying a system's state

Scenario (as in weather forecasting):
Run the dynamical model and simultaneously observe reality.
When model state deviates from real state: update model state (= data assimilation).

How to estimate the current state of the real world?

Construct a blend of the model state vector and the noisy observations as the most probable state of reality.

Simple setting: All system variables observed (noisy).
More difficult setting: Incomplete observations (noisy).

Linear system: Kalman filter.
Extended Kalman filter: linearization of a nonlinear system.
other extensions for nonlinear systems are available.

Data assimilation

Problem 7: Data assimilation remains a widely open field

essential: efficiency of the algorithms

(e.g., numerical weather prediction ECMWF global model: 50% of all computation time for a 10 day (medium range) weather forecast just for data assimilation and ensemble breeding).

[http://www.ecmwf.int/products/forecasts/guide/user_guide.pdf]

Nonlinear system: invariant density (invariant measure) is (close to) singular, true state vectors located on "attractor".

Unstable directions due to chaos.

Ensemble breeding

Ensembles of initial conditions

Nonlinear systems with uncertain initial conditions:
Create an ensemble of "plausible" initial conditions,
run the dynamics on **all** of them to explore the variablity of the future.

Problem 8: Construct plausible ensembles

a) ensemble should respect the error covariance statistics
b) ensemble members should be on the attractor
c) few ensemble members should explore a high dimensional phase space

established solutions: bred vectors, Lyapunov vectors,
more recent: covariant Lyapunov vectors gurantee b), help for c), but numerically expensive.
[Ginelli et al. (2007), Lopez, Guiterez (2007)].

propagating PDFs

The ensemble represents a probability density function (PDF) for the unknown state of reality.

Problem 9: Estimate propagated PDFs from propagated ensembles.

Failure of predictions

wrong prediction:
- perfect initial conditions, bad model (data driven or first principles)
- perfect model, bad initial conditions

Problem 10:
identify model errors despite errors in initial conditions

Conclusions

Problems in state estimation and model construction

1. Dynamical measurement errors depend on the signal
2. criteria for the distinction of signal and noise
3. Fitting model parameters: errors in independent variables
4. Selecting model classes which are stable
5. Guaranteeing correct asymptotic model dynamics despite lack of structural stability
6. Markov order of discrete time stochastic models
7. fast algorithms for data assimilation: blending model state with measurements
8. construct ensembles of initial conditions
9. derive time evolution of pdfs from time evolution of ensembles
10. model errors versus errors in initial conditions

MPIPKS, Dresden, Germany
IFISC, Palma de Mallorca, Spain

Trends in Complex Systems

International Workshop on Extreme Events:

Theory, Observations, Modeling, and Prediction

Palma de Mallorca, November 10 - 14, 2008

Scientific coordination:

Holger Kantz, MPIPKS **Manuel A. Matías**, IFISC

This workshop launches the program **Trends in Complex Systems** [http://www.pks.mpg.de/~tcs or http://ifisc.uib.es/tcs]. We will address the current state of research of extreme events to stimulate new work in this field. The talks will cover a wide range of issues related to extreme events in complex systems, such as listing systems and model classes where evidence for extreme events exists, presenting current statistical tools for the characterization of extreme events including temporal and spatial correlations, and highlighting examples where a detailed understanding of underlying mechanisms has been achieved. This will be completed by the issue of prediction and predictability, of scoring and evaluating predictions, and control of extreme events, as well as considerations about the discrepancy between physical impact of some event and societal impact.
In addition to the invited speakers, the workshop facilities enable us to target an audience of about 60 additional participants, who will also present their work in contributed talks and posters. A special emphasis of this workshop lies on discussion, for which ample time is available.

Invited speakers include: (* to be confirmed)

Michael Ghil, Paris	**Bruce Malamud*, London**	**Lenny Smith*, London**
Peter Grassberger, Calgary	**Edward Ott, Maryland**	**Sorin Solomon, Jerusalem**
Jürgen Kurths, Potsdam	**Jean-François Pinton, Lyon**	**Didier Sornette, Zurich**
Juan M. López, Santander	**Sidney Redner, Boston**	**Raúl Toral, Mallorca**

Applications for participation and contributions are welcome and should be made by using the application form on the workshop web page (please see URL below). The number of attendees is limited. The applicants' registration fee for the workshop is 100 Euro. Costs for accommodation will be covered during the workshop. Limited funding is available on request to partially cover travel expenses.

Deadline for applications is: September 10, 2008.

CONTACT:
Ms. Claudia Pönisch - Workshop Secretary
exev08@pks.mpg.de
http://www.pks.mpg.de/~exev08
Tel.: +49-351-871-2198 / Fax: +49-351-871-2199
Local IFISC Contact - Ms. Marta Ozonas - exev08@ifisc.uib.es
Tel.: +34-971-173290 / Fax: +34-971-173248

MAX-PLANCK-GESELLSCHAFT

Consejo Superior de Investigaciones Científicas
Universitat de les Illes Balears