Conditional Expectation as the Basis for Bayesian Updating

Hermann G. Matthies

Bojana V. Rosić, Elmar Zander, Alexander Litvinenko, Oliver Pajonk

Institute of Scientific Computing, TU Braunschweig
Brunswick, Germany

wire@tu-bs.de http://www.wire.tu-bs.de

15_Cond-Exp.tex,v 2.9 2017/07/18 00:35:08 hgm Exp

Overview

1. BIG DATA

2. Parameter identification

3. Stochastic identification — Bayes's theorem

4. Conditional probability and conditional expectation

5. Updating — filtering.

Representation of knowledge

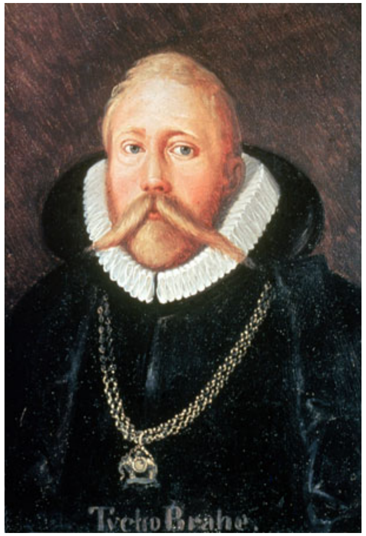
Data from measurements, sensors, observations $\Rightarrow$
one form of knowledge about a system.
'Big Data' considers only data — looking for patterns, interpolating, etc.

Mathematical / computational models of a system represent another form of knowledge — 'structural' knowledge — about a system.
These models are often generated based on general physical laws (e.g. conservation laws), a very compressed form of knowledge.

These two views on systems are not in competition,
they are complementary.
The challenge is to combine these forms of knowledge
— in form of a synthesis.

Knowledge may be uncertain.

Big Data 16th century

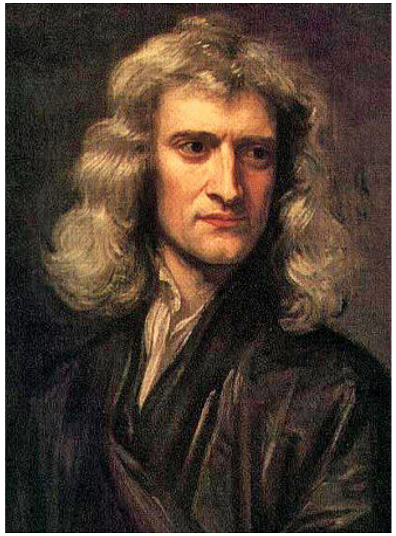


Tycho Brahe
(1546 – 1601)

Data

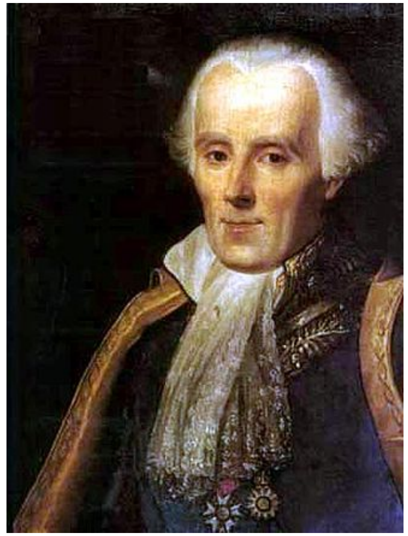
Johannes Kepler
(1571 – 1630)

Description



Isaac Newton
(1643 – 1727)

Understanding

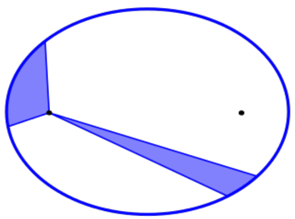


Pierre-Simon Laplace
(1749 – 1827)

Perfection

I. Newton: *The latest authors, like the most ancient, strove to subordinate the phenomena of nature to the laws of mathematics.*

Kepler's 2nd law:



(adapted from M. Ortiz)

BIG DATA

Mathematically speaking, big data algorithms
(feature / pattern recognition) are
regression (generalised interpolation) methods.

Often based on deep artificial neural networks (deep ANNs),
combining many inputs (= high-dimensional data).

Deep networks are connected to
sparse tensor decompositions
(buzzword: deep-learning).

Although often spectacularly successful,
as knowledge representation, it is difficult to extract insight.

But there is a connection of such regression to Bayesian updating.

Inference

Our uncertain knowledge about some situation is described by probabilities. Now we obtain new information.
How does it change our knowledge — the probabilistic description?
Answered by T. Bayes and P.-S. Laplace more than 250 years ago.

Thomas Bayes
(1701 – 1761)

Pierre-Simon Laplace
(1749 – 1827)

Synopsis of Bayesian inference

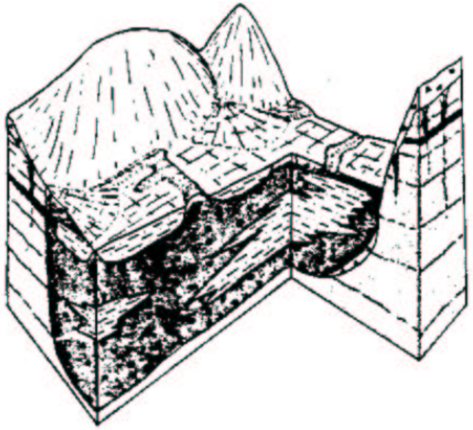
We have a some knowledge about an event $\mathcal{A}$,
but it can not be observed directly.

After some new information $\mathcal{B}$ (an observation, a measurement),
our knowledge has to be made consistent with the new information,
i.e. we are looking for conditional probabilities $\mathbb{P}(\mathcal{A}|\mathcal{B})$.

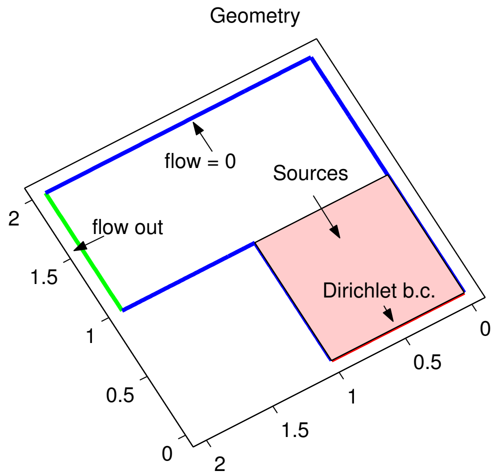
The idea is to change our present model by just so much
— as little as possible — so that it becomes consistent.

For this we have to predict — with our present knowledge / model —
the probability of all possible observations and
compare with the actual observation.

Model inverse problem



Aquifer



2D Model

Governing model equations:

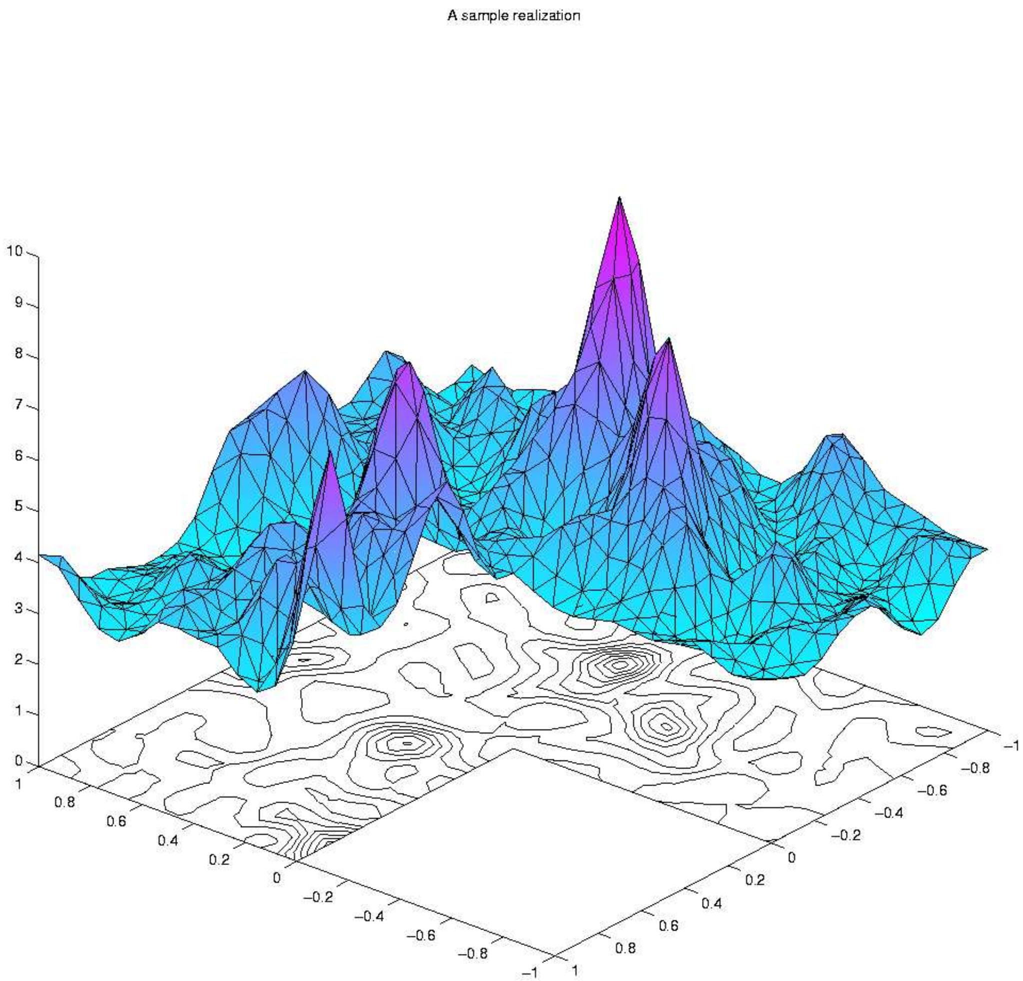
$$\varrho \frac{\partial u}{\partial t} - \nabla \cdot (\kappa \cdot \nabla u) = f \quad \in \mathcal{G} \subset \mathbb{R}^d.$$

Parameter $q = \log \kappa$.

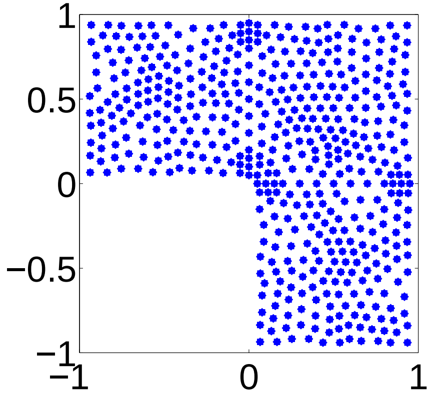
Conductivity field κ, initial condition u_0, and state $u(t)$ may be unknown.

They have to be determined from observations $Y(q; u)$.

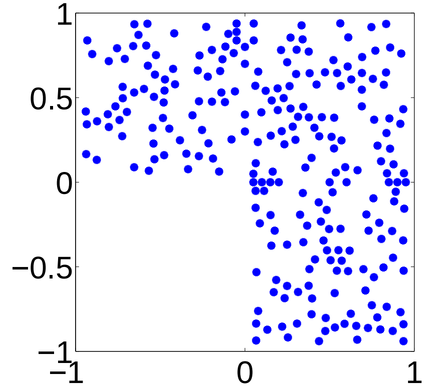
A possible realisation of $\kappa(x, \omega)$



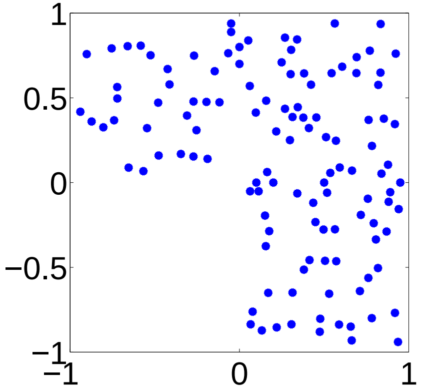
Measurement patches



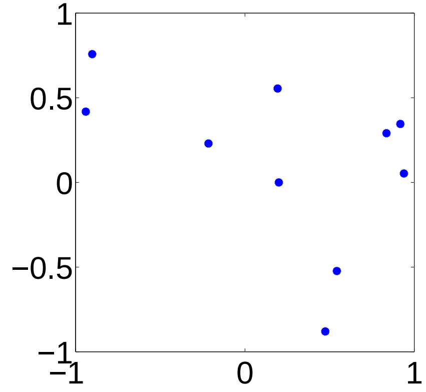
447 measurement patches



239 measurement patches

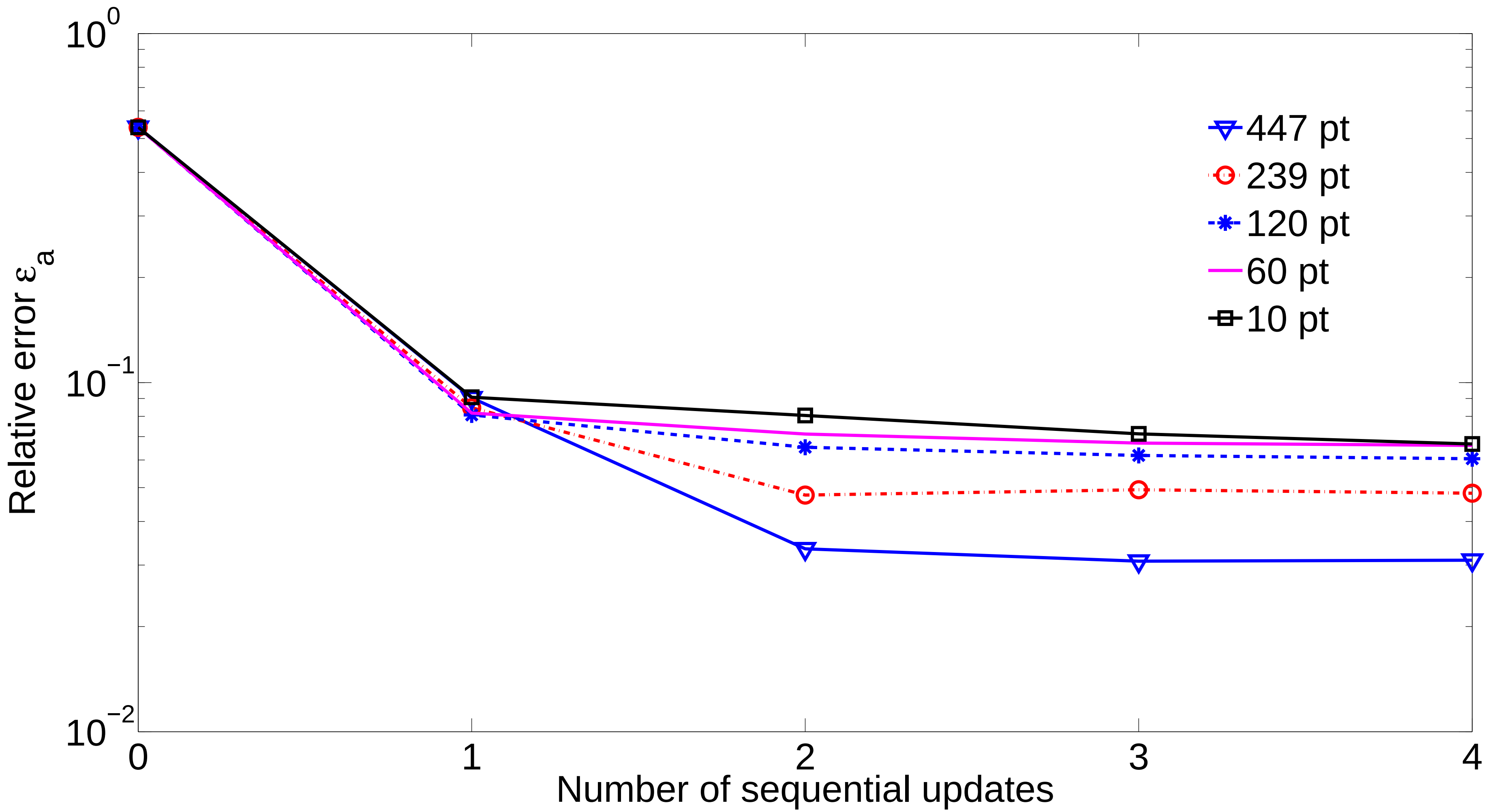


120 measurement patches

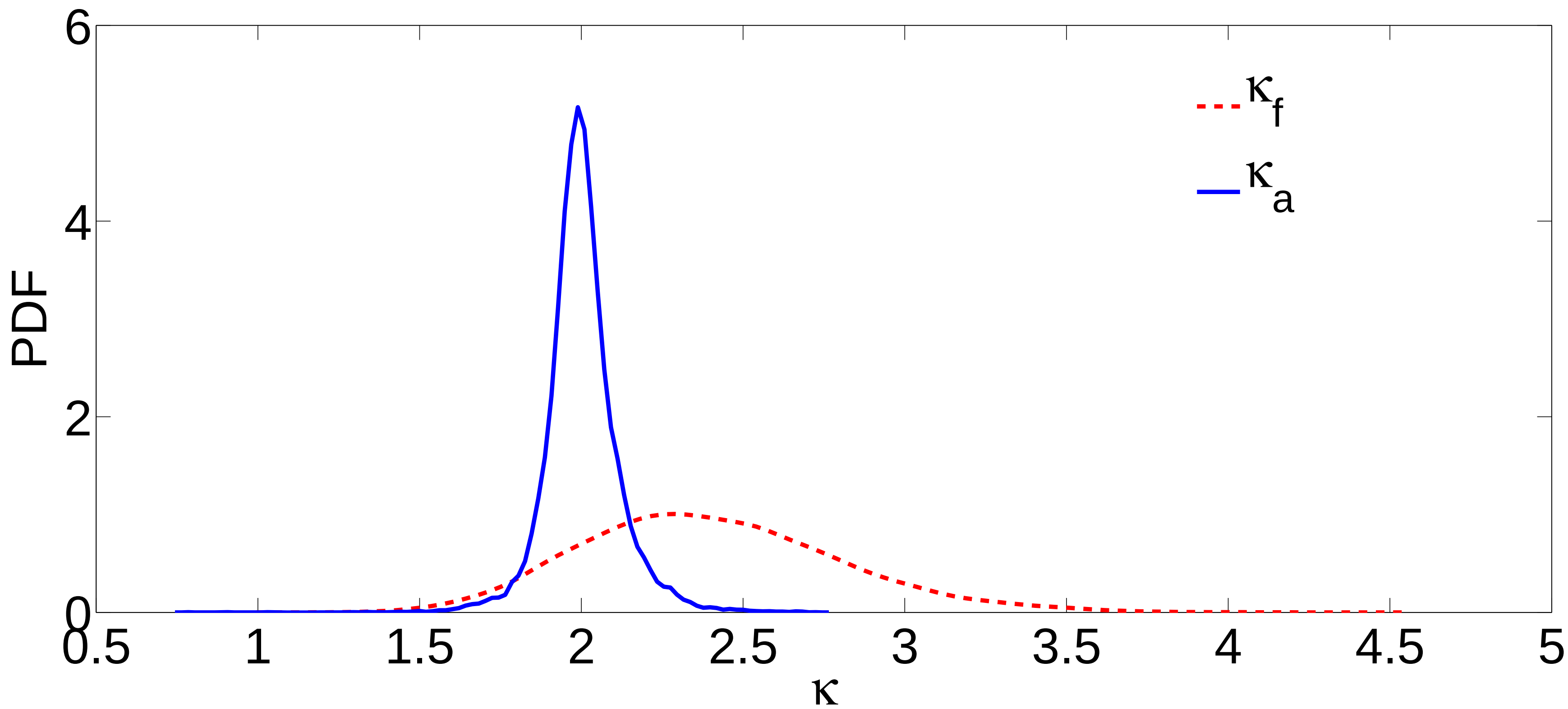


10 measurement patches

Convergence plot of updates



Forecast and assimilated pdfs



Forecast and assimilated probability density functions (pdfs) for κ at a point where $\kappa_t = 2$.

Setting for identification

General idea:

We observe / measure a system, whose structure we know in principle.
The system behaviour depends on some quantities (parameters),
which we do not know $\Rightarrow$ uncertainty.

We model (uncertainty in) our knowledge in a Bayesian setting:
as a probability distribution on the parameters.

We start with what we know a priori, then perform a measurement.
This gives new information, to update our knowledge (identification).

Update in probabilistic setting works with conditional probabilities
$\Rightarrow$ Bayes's theorem.

Repeated measurements lead to better identification.

Mathematical formulation of model

Consider operator equation, physical system modelled by A:

$$\mathrm{d}u + A(u;q)\,\mathrm{d}t = g\,\mathrm{d}t + B(u;q)\,\mathrm{d}W \qquad u \in \mathcal{U},$$

$\mathcal{U}$ — space of states, g a forcing, W noise, $q \in \mathcal{Q}$ unknown parameters.

Well-posed problem: for q, g and initial cond. $u(t_0) = u_0$
a unique solution $u(t)$, given by the flow or solution operator,

$$S : (u_0, t_0, q, g, W, t) \mapsto u(t;q) = S(u_0, t_0, q, g, W, t).$$

Set extended state $\xi = (u, q) \in \mathcal{X} = \mathcal{U} \times \mathcal{Q}$,
advance from $\xi_{n-1} = (u_{n-1}, q_{n-1})$ at time t_{n-1} to $\xi_n = (u_n, q_n)$ at t_n,

$$\xi_n = (u_n, q_n) = (S(u_{n-1}, t_{n-1}, q_n, g, W, t_n), q_n) =: f(\xi_{n-1}, w_{n-1}).$$

This is the model for the system observed at times t_n.
Applies also to stationary case $A(u;q) = g$.

Mathematical formulation of observation

Measurement operator Y with values in $\mathcal{Y}$:

$$\eta_n = Y(u_n; q) = Y(S(u_{n-1}, t_{n-1}, q, g, W, t_n); q).$$

But observed at time t_n, it is noisy y_n with noise ϵ_n

$$y_n = H(\eta_n, \epsilon_n) = H(Y(u_n; q), \epsilon_n) =: h(\xi_n, \epsilon_n) = h(f(\xi_{n-1}, w_{n-1}), \epsilon_n).$$

For given g, w, measurement $\eta = Y(u(q); q)$ is just a function of q.
This function is usually not invertible $\Rightarrow$ ill-posed problem,
measurement η does not contain enough information.

Parameters q and initial state u_0 uncertain, modelled as RVs
$q \in \mathscr{Q} = \mathcal{Q} \otimes \mathcal{S} \Rightarrow u \in \mathscr{U} = \mathcal{U} \otimes \mathcal{S}$, with e.g. $\mathcal{S} = \mathrm{L}_2(\Omega, \mathbb{P})$ a RV-space.

Bayesian setting allows updating of information about $\xi = (u, q)$.
The problem of updating becomes well-posed.

Mathematical formulation of filtering

We want to track the extended state ξ_n
by a tracking equation for a RV x_n through observations $\hat{y}_n$.

- Prediction / forecast state is a RV $x_{n,f} = f(x_{n-1}, w_{n-1})$;
- Forecast observation is a RV $y_n = h(x_{n,f}, \epsilon_n)$, actual observation $\hat{y}_n$,
- Updated / assimilated $x_n = x_{n,f} + \Xi(x_{n,f}, y_n, \hat{y}_n)$,
- Hopefully $x_n \approx \xi_n$, and the update map Ξ has to be determined.
  $x_{n,i} := \Xi(x_{n,f}, y_n, \hat{y}_n)$ is called the innovation.

We concentrate on one step from forecast to assimilated variables.

- Forecast state $x_f := x_{n,f}$, forecast observation $y_f := y_n$,
- Actual observation $\hat{y}$ and assimilated variable

$$x_a := x_f + \Xi(x_f, y_f, \hat{y}) = x_n = x_{n,f} + \Xi(x_{n,f}, y_{n,f}, \hat{y}_n).$$

This is the filtering or update equation.

Setting for updating

Knowledge prior to new observation is also called forecast:

the state $u_f \in \mathscr{U} = \mathcal{U} \otimes \mathcal{S}$ and parameters $q_f \in \mathscr{Q} = \mathcal{Q} \otimes \mathcal{S}$
modelled as random variables (RVs),
also the extended state $x_f = (u_f, q_f) \in \mathscr{X} = \mathcal{X} \otimes \mathcal{S}$ and
the measurement $y(x_f, \varepsilon) \in \mathscr{Y} = \mathcal{Y} \otimes \mathcal{S}$.

Then an observation $\hat{y}$ is performed,
and is compared to predicted measurement $y(x_f, \varepsilon)$.

Bayes's theorem gives only probability distribution of
posterior or assimilated extended state x_a.

Here we want more: a filter $x_a := x_f + \Xi(x_f, y_f, \hat{y})$.

Using Bayes's theorem

Classically, Bayes's theorem gives conditional probability

$$\mathbb{P}(\mathcal{I}_x|\mathcal{M}_y) = \frac{\mathbb{P}(\mathcal{M}_y|\mathcal{I}_x)}{\mathbb{P}(\mathcal{M}_y)}\mathbb{P}(\mathcal{I}_x) \quad \text{for} \quad \mathbb{P}(\mathcal{M}_y) > 0.$$

Well-known special form with densities of RVs x, y (w.r.t. some background measure μ):

$$\pi_{(x|y)}(x|y) = \frac{\pi_{xy}(x,y)}{\pi_y(y)} = \frac{\pi_{(y|x)}(y|x)}{Z_y}\pi_x(x);$$

with marginal density $Z_y := \pi_y(y) = \int_{\mathcal{X}} \pi_{xy}(x,y)\ \mu(\mathrm{d}x)$
(from German Zustandssumme) — only valid when $\pi_{xy}(x,y)$ exists.

Problems / paradoxa appear when $\mathbb{P}(\mathcal{M}_y) = 0$ (and $\mathbb{P}(\mathcal{M}_y|\mathcal{I}_x) = 0$)
e.g. Borel-Kolmogorov paradox. Problem is limit $\mathbb{P}(\mathcal{M}_y) \to 0$,
or when no joint density $\pi_{xy}(x,y)$ exists.

Conditional probability

"Many quite futile arguments have raged—between otherwise competent probabilists—over which of these results is 'correct'." E.T. Jaynes

"The concept of a conditional probability with regard to an isolated hypothesis whose probability equals zero is inadmissible." A. Kolmogorov

$\Rightarrow$ How to use conditioning in these typical singular cases, where Bayes's formula is **not** applicable? $\Leftarrow$

With posterior / conditional measure $\mathbb{P}(\cdot|\mathcal{M}_y)$ one may compute the conditional expectation $\mathbb{E}\left(\psi|\mathcal{M}_y\right) = \int_{\Omega} \psi(\omega)\, \mathbb{P}(\mathrm{d}\omega|\mathcal{M}_y)$.

Kolmogorov turns it around and starts from conditional expectation operator $\mathbb{E}\left(\cdot|\mathcal{M}_y\right)$, from this conditional probability via

$\mathbb{P}(\mathcal{I}_x|\mathcal{M}_y) := \mathbb{E}\left(\mathbf{1}_{\mathcal{I}_x}|\mathcal{M}_y\right), \quad \mathbf{1}_{\mathcal{I}_x}(\xi) = 1$ for $\xi \in \mathcal{I}_x$, 0 otherwise.

Conditional expectation and probability

Expectation of a RV ψ: $\mathbb{E}\,(\psi) = \int_\Omega \psi(\omega)\, \mathbb{P}(\mathrm{d}\omega)$.

$\mathbb{E}\,(\cdot)$ as a functional $\mathrm{L}_2(\Omega, \mathfrak{A}) = \mathcal{S} \to \mathbb{R}$, but also orthogonal projection

$$\mathbb{E} : \mathcal{S} = \mathrm{span}\{\mathbf{1}_\Omega\} \oplus \{\phi \in \mathcal{S} \mid \mathbb{E}\,(\phi) = 0\} \to \mathrm{span}\{\mathbf{1}_\Omega\}, \quad (\mathbf{1}_\Omega \equiv 1).$$

Conditional expectation is an orthogonal projection onto subspaces $\mathrm{L}_2(\Omega, \mathfrak{B}, \mathbb{P}) =: \mathcal{S}_\infty$ defined by sub-σ-algebras $\mathfrak{B} \subseteq \mathfrak{A}$:
Here $\mathfrak{B} = \sigma(y)$ — generated by measurement y, and
the subspace $\mathcal{S}_\infty$ is the space of all (measurable) functions of y.

$$\mathbb{E}(\cdot|\sigma(y)) := \mathbb{E}(\cdot|\mathfrak{B}) : \mathrm{L}_2(\Omega, \mathfrak{A}) = \mathcal{S} = \mathcal{S}_\infty \oplus \mathcal{S}_\infty^\perp \to \mathcal{S}_\infty$$

Call $\mathbb{E}(\cdot|y) := \mathbb{E}(\cdot|\sigma(y)) =: P_\infty$ the pre-conditional expectation.
$\mathbb{E}(\psi|y) \in \mathcal{S}_\infty$ is a RV, because y is. After observing $\hat{y}$ one has post-conditional expectation $\mathbb{E}(\psi|\hat{y}) \in \mathbb{R}$—new expectation after new $\hat{y}$.

The state of knowledge has changed, hence so has the expectation.

Conditional expectation

With orthogonal direct sum $\mathcal{S} = \mathcal{S}_\infty \oplus \mathcal{S}_\infty^\perp$ one has decomposition

$$\psi = P_\infty \psi + (\mathrm{I} - P_\infty)\psi = \mathbb{E}(\psi|y) + (\psi - \mathbb{E}(\psi|y)).$$

According to Pythagoras:

$$\|\psi\|_{\mathcal{S}}^2 = \|P_\infty \psi\|_{\mathcal{S}}^2 + \|(\mathrm{I} - P_\infty)\psi\|_{\mathcal{S}}^2 = \|\mathbb{E}(\psi|y)\|_{\mathcal{S}}^2 + \|(\psi - \mathbb{E}(\psi|y))\|_{\mathcal{S}}^2$$

Simple cases:

1. $\mathfrak{B} = \{\Omega, \emptyset\} \Rightarrow \mathbb{E}\left(\cdot|\mathfrak{B}\right) = \mathbb{E}\left(\cdot\right)$, the normal expectation.
2. $\mathfrak{B} = \mathfrak{A} \Rightarrow \mathbb{E}\left(\cdot|\mathfrak{B}\right) = \mathrm{I}_{\mathrm{L}_2}$, the identity on $\mathrm{L}_2(\Omega, \mathfrak{A}, \mathbb{P})$.
3. In our case $\mathfrak{B} = \sigma(y)$, the σ-algebra generated by measurement RV y (not so simple!).

Question: How to compute $P_\infty = \mathbb{E}\left(\cdot|y\right)$, and how to build filter Ξ to obtain $x_a := x_f + \Xi(x_f, y_f, \hat{y})$?

Representing and using the conditional expectation

As $P_\infty = \mathbb{E}\left(\cdot|y\right)$ is an orthogonal projection, for any ψ

$$\mathbb{E}(\psi(x)|y) := P_\infty(\psi(x)) = \arg\min_{p\in\mathcal{S}_\infty} \|\psi(x) - p\|^2_{\mathcal{S}}$$

The subspace $\mathcal{S}_\infty$ represents the available information,
conditional expectation $P_\infty\psi$ minimises $\Phi(\cdot) := \|\psi(x) - (\cdot)\|^2_{\mathcal{S}}$ over $\mathcal{S}_\infty$.

More general loss functions than minimising mean square error (MMSE) are possible, used in decision processes.

Taking $\psi_1(x) = x$, one obtains $P_\infty x = \mathbb{E}\left(x|y\right)$ and $\bar{x}^{|\hat{y}} := \mathbb{E}\left(x|\hat{y}\right)$.

Taking $\psi_2(x) = x \otimes x = x^{\otimes 2}$, one obtains $P_\infty(x \otimes x) = \mathbb{E}\left(x \otimes x|y\right)$,
from which one may compute the post-conditional covariance of x:

$$\mathrm{cov}_x^{|\hat{y}} = \mathbb{E}\left(x \otimes x|\hat{y}\right) - \bar{x}^{|\hat{y}} \otimes \bar{x}^{|\hat{y}}.$$

Update through conditional expectation

Reminder: want to find mapping / filter Ξ for assimilated x_a:

$$x_a := x_f + \Xi(x_f, y_f, \hat{y});$$

x_a with Bayesian posterior distribution resp. $\mathbb{E}\left(\psi(x_a)|\hat{y}\right)$ for all ψ.
As Bayesian update is costly, several approximations possible:

- The conditional expectation (CE-filter) update, with correct $\mathbb{E}\left(x_a|\hat{y}\right)$.
- Approximated by linearised version of the CE-update — the Gauss-Markov-Kalman filter (GMKF), where Ξ is linear in $\hat{y} - y$.
- The conditional expectation variance (CEV) update, both conditional expectation and covariance of x_a are correct.
- Approximated by linearised version of the CEV-update; (best linear Ξ).
- Computing an expansion (with truncation) of Ξ, resp. x_a.
- Better approximations using conditional expectation . . .

Possibility: CE-update / filter

The space $\mathcal{S}_\infty = \mathrm{L}_2(\Omega, \sigma(y), \mathbb{P})$ is the space of all functions of measurement / observation y. Taking first $\psi(x) = x$

$$\mathbb{E}\left(x|y\right) =: \phi_x(y) = \arg\min\{\|x - p\|^2_{\mathcal{S}} : p \in \mathcal{S}_\infty = \{p \in \mathcal{S} : p = \varphi(y)\}\}.$$

With this operator (conditional expectation) one may construct a new RV x_a with correct posterior.

First step: the "MMSE Bayesian update" x_a with correct conditional expectation $\bar{x}^{|\hat{y}}$ (CE-filter).

As $\mathbb{E}\left(x|y\right) =: P_\infty x$ is orthogonal projection onto $\mathcal{S}_\infty$, one has

$$\mathcal{S} = \mathcal{S}_\infty \oplus \mathcal{S}_\infty^\perp \;\Rightarrow\; x = P_\infty x + (\mathrm{I} - P_\infty)x = \phi_x(y) + (x - \phi_x(y)).$$

From this $x_a \approx \phi_x(\hat{y}) + (x_f - \phi_x(y_f)) = x_f + (\phi_x(\hat{y}) - \phi_x(y_f))$.

Obviously $\mathbb{E}\left(x_a|\hat{y}\right) = \mathbb{E}\left(x_f|\hat{y}\right) = \phi_x(\hat{y}) = \bar{x}^{|\hat{y}}$.

Further improvements by transforming $x_a - \bar{x}^{|\hat{y}} = x_f - \phi_x(y_f)$.

BIG DATA — Gauss-Markov-Kálmán filter

If one only wants $\mathbb{E}(x_f|\hat{y}) = \phi_x(\hat{y}) = \bar{x}^{|\hat{y}}$, then the function ϕ_x can be found through regression or machine learning / deep networks.

Estimation of $(x_f - \phi_x(y_f))$ is possible.

Further simplification / approximation:

if only linear (affine) functions $\varphi(y) = Ay + b$ are allowed:

$$K_x\, y + c = \arg\min\{\|x - p\|^2_{\mathcal{S}} : p \in \mathcal{S}_1 := \{p \in \mathcal{S} : p = Ay + b\}\},$$

$\phi_x(y) \approx K_x\, y + c =: P_1 x$ with Kálmán gain K_x. As $\mathcal{S}_1 \subseteq \mathcal{S}_\infty$,

$$\|x - \phi_x(y)\|^2_{\mathcal{S}} = \|x - P_\infty x\|^2_{\mathcal{S}} \leq \|x - P_1 x\|^2_{\mathcal{S}} = \|x - (K_x\, y + c)\|^2_{\mathcal{S}}.$$

From Kálmán gain K_x

$\Rightarrow$ Gauss-Markov-Kálmán filter (GMKF)

$$x_a \approx x_f + (K_x\, \hat{y} - K_x\, y_f) = x_f + K_x(\hat{y} - y_f).$$

Rudolf Kálmán
(1930 – 2016)

Numerical Remarks

- Parametric or stochastic problems — like stochastic PDEs — lead to solutions (states) in tensor product space.
- Stochastic forward solution allows identification
- "Curse of dimensionality" has to be controlled.
- Reduced order models can yield sparse (or low-rank) representations, with all work carried out on the low-rank approximation.
- After solution has been computed, is has to be processed further.
- If further processing is a tensor function, this might often be computed with little effort.

Computation of conditional expectation

Minimisation to compute conditional expectation for any RV $\psi(x)$:

$$\mathbb{E}\left(\psi|y\right) := P_\infty \psi = \phi_\psi(y) := \arg\min_{p \in \mathcal{S}_\infty} \|\psi(x) - p\|^2_{\mathcal{S}}.$$

Variational equation / Galerkin condition from minimisation:

$$\forall p \in \mathcal{S}_\infty : \quad \langle \psi(x) - \phi_\psi(y) \mid p \rangle_{\mathcal{S}} = \mathbb{E}\left((\psi(x) - \phi_\psi(y)) \cdot p\right) = 0.$$

GMKF was obtained by Galerkin approximation $\mathcal{S}_1 \subseteq \mathcal{S}_\infty$.

Minimisation may also be performed by Gauss-Newton methods.
Each iteration looks similar to Gauss-Markov-Kalman-filter (GMKF).
Various variations of iteration are possible,
e.g. BFGS-methods instead of Gauss-Newton.

In any case, it is in principle possible to compute $\mathbb{E}\left(\psi(x)|y\right)$ for any RV $\psi(x)$ to any desired accuracy, including a posteriori error control.

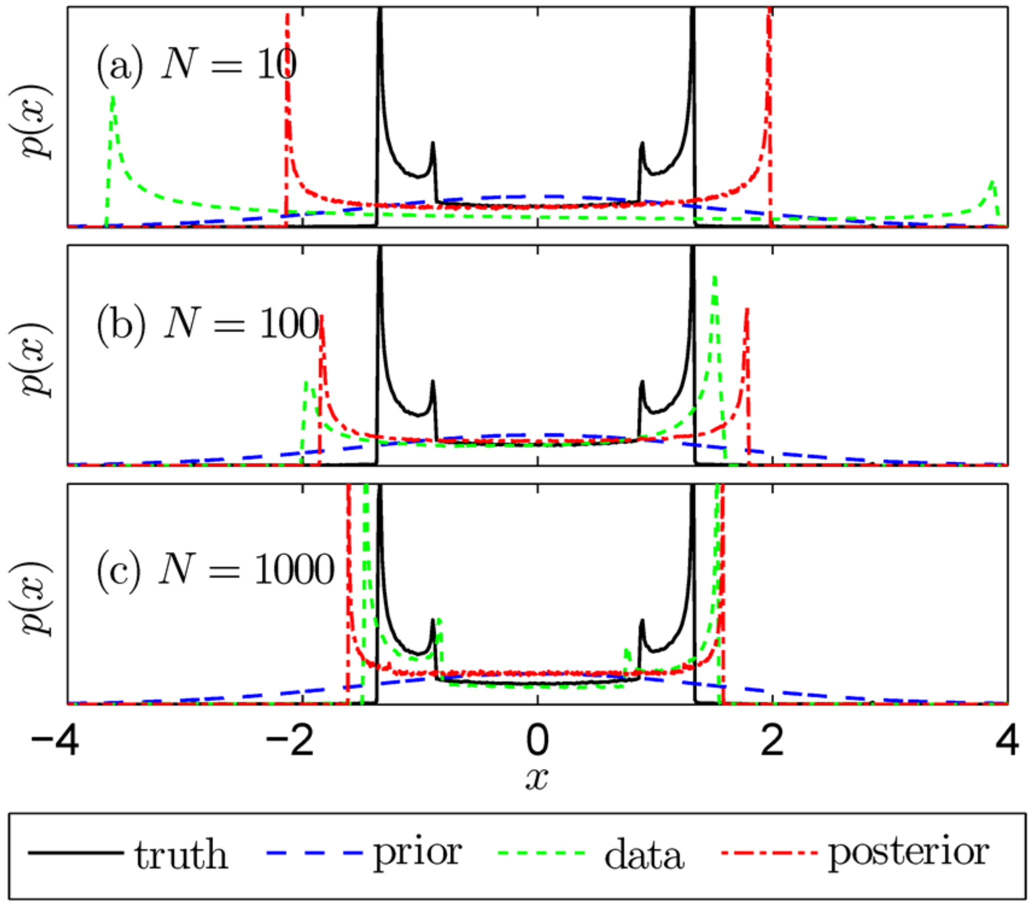
Example 1: Identification of multi-modal dist

Setup: Scalar RV x with non-Gaussian multi-modal "truth" $p(x)$; wide Gaussian prior; "large" Gaussian measurement errors.

Aim: Identification of $p(x)$.

10 updates of $N = 10, 100, 1000$ measurements.

Filter: GMK-filter — optimal linear filter — in PCE representation



Example 2: Lorenz-84 chaotic model

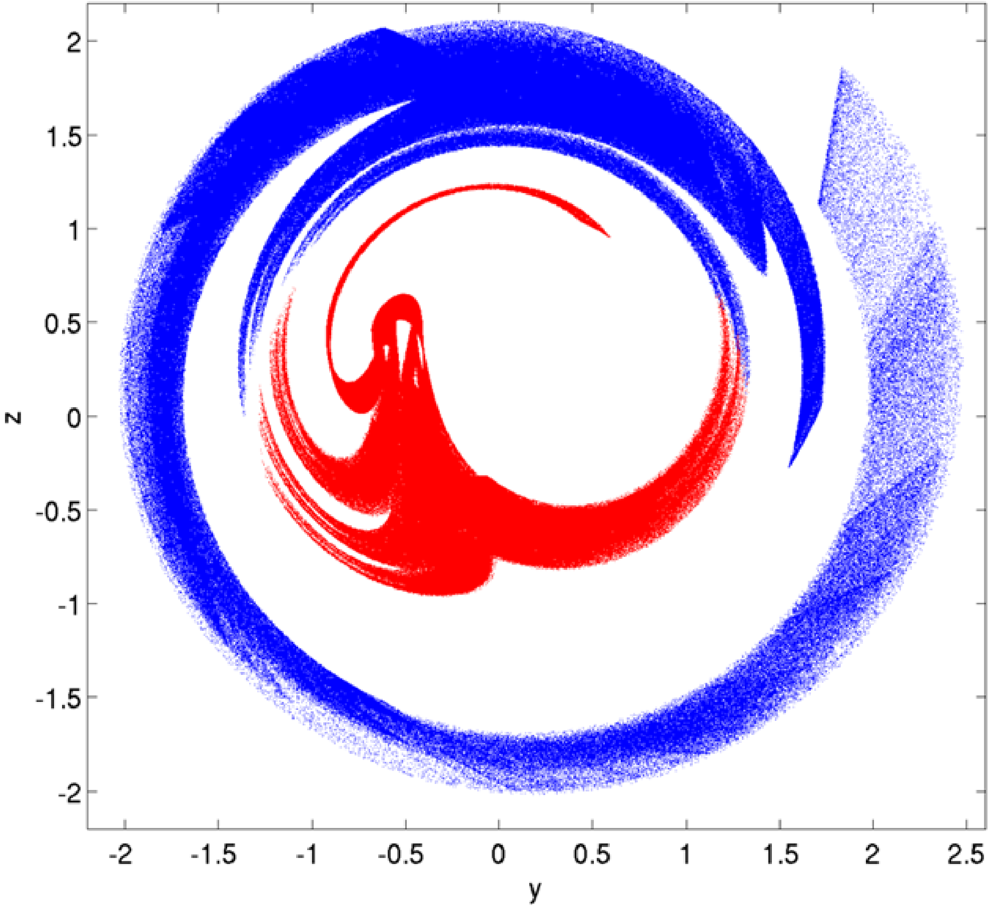
Setup: Non-linear, chaotic system

$$\dot{u} = f(u), \;\; u = [x, y, z]$$

Small uncertainties in initial conditions u_0 have large impact.

Aim: Sequentially identify state u_t.

Methods: GMK-filter in PCE representation and PCE updating

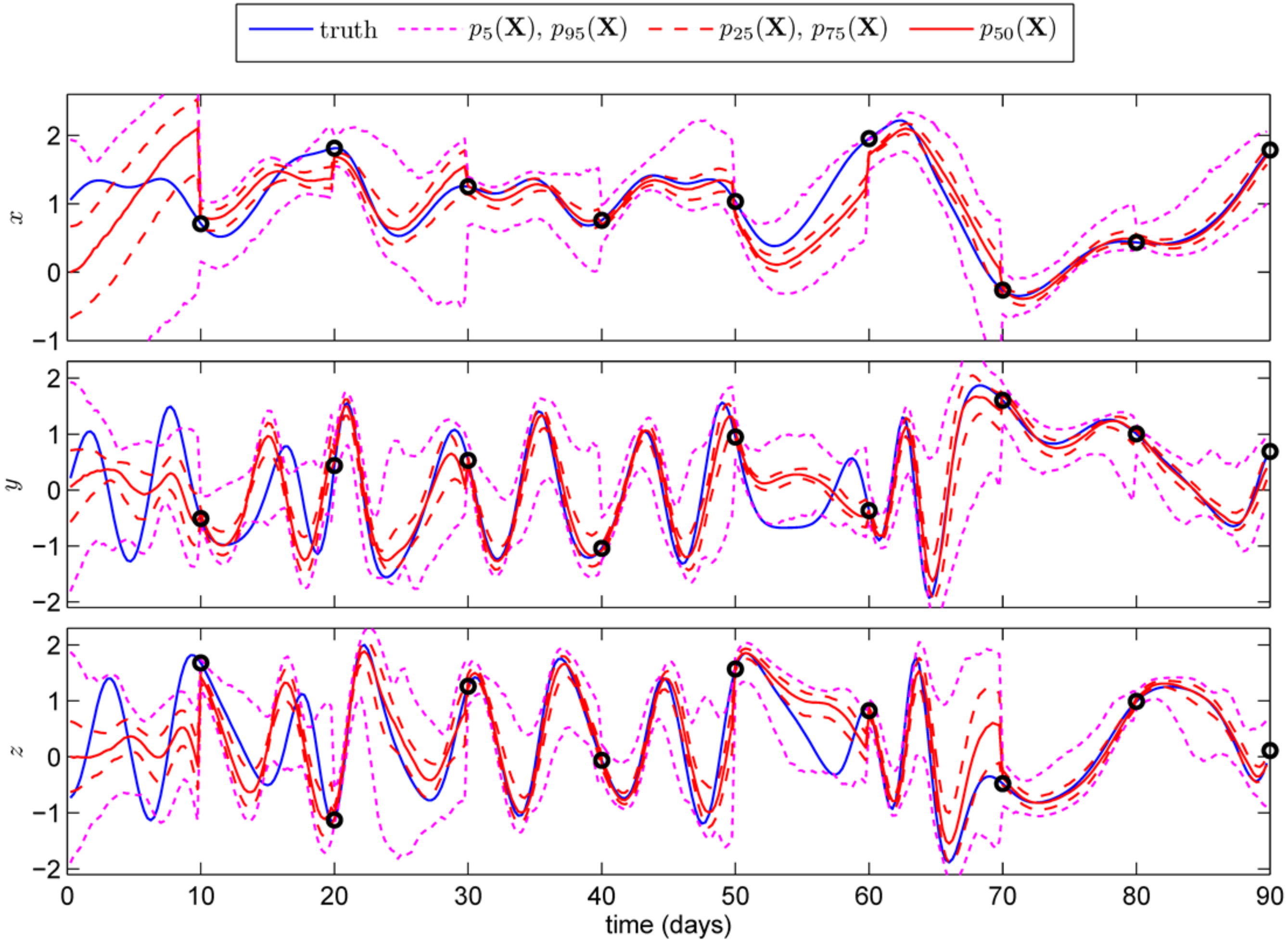


Poincaré cut for $x = 1$.

Example 2: Lorenz-84 PCE representation

PCE: Variance reduction and shift of mean at update points.

Skewed structure clearly visible, preserved by updates.



Summary

- UQ allows stochastic inverse identification as a well-posed problem, this Bayesian update is based on conditioning.
- Conditional probability is based on conditional expectation, starting point for numerics, connects to MMSE.
- Bayesian update may be presented as a filter, a simple approximation is GMKF, even simpler by machine learning.
- Works for
  - non-Gaussian distributions.
  - linear and nonlinear models and observation operator Y.
  - possible for ODEs, PDEs, processes, fields, etc.