Bayesian Inference in Physics

U. von Toussaint
Introduction

Scientific Inference Cycle

I. Experiment \rightarrow II. Evaluation

IV. Hypothesis Verification \leftrightarrow III. Hypothesis Design
I. Experiment

II. Evaluation
Parameter Estimation

IV. Hypothesis Verification

III. Hypothesis Design
Model Selection

Prerequisite: Consistent reasoning...
# Outline

<table>
<thead>
<tr>
<th>I. Scientific Inference</th>
<th>Inference in Science Processing of Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>II. Model Comparison</td>
<td>Basic Concept Nuclear Reaction Analysis</td>
</tr>
<tr>
<td>III. Experimental Design</td>
<td>Basic Concept Nuclear Reaction Analysis</td>
</tr>
<tr>
<td>IV. Numerical Interlude</td>
<td>Nested Sampling</td>
</tr>
<tr>
<td>V. Conclusion</td>
<td>Summary Outlook</td>
</tr>
</tbody>
</table>
Inference in Science

Science: prior information + new data $\Rightarrow$ new knowledge

Prior information:
- old data
- calibration measurements, validation data
- theoretical considerations
- parameters
- model

New data:
- measurements
- calibration
- theories

But: prior information and data are uncertain
- new knowledge and derived hypotheses are uncertain
- inductive (instead of deductive) reasoning required
- Calculus for “uncertainties” needed (quantification)
Inference in Science

Science: prior information + new data ➔ new knowledge

a) deductive reasoning:

cause ➔ Effects or outcomes

Example: fair coin: \( p(7 \text{ heads} \mid 10 \text{ tosses}) = ? \)

b) inductive reasoning

possible causes ➔ Effects or observations

Example: 7 heads out of 10 tosses: \( p(\text{fair coin}) = ? \)
Inference in Science

Calculus:
Cox (1946): Basic requirements (i.e. transitivity, consistency) single out usual rules of probability theory to handle uncertainty

Please note: probability here not restricted to frequency interpretation:

Degree of belief about a proposition

Data:
- statistical (counting statistics)
- measurement uncertainty (ruler)
- systematic (e.g. misalignment)
- outliers

Hypotheses:
- parameter of interest
- nuisance parameters
- physical models
- future data

Notation: \(p(x|I)\)
\(p(x|I)\) describes how probability (plausibility) is distributed among possible choices for \(x\) for the case at hand (information \(I\))
Inference in Science

Calculus:
Cox (1946): Basic requirements (i.e. transitivity, consistency) single out *usual rules of probability theory* to handle uncertainty

Bayesian interpretation:
\[ p(x|I) \] describes how probability (plausibility) is distributed among possible choices for \( x \) for the case at hand (information \( I \))

Frequentist interpretation:
\[ p(x|I) \] describes how \( x \) is distributed throughout an infinite (hypothetical) ensemble
probability = frequency
Processing of information: Combination of (cond.) probability distributions

Conditional Probabilities:

\[ p(A|B) : \text{probability of proposition } A \text{ given truth of proposition } B: \]
\[ \text{quantification of uncertainty of } A \]

Probability Theory Axioms:

- sum rule (OR):
\[ p(H_1 + H_2 | I) = p(H_1 | I) + p(H_2 | I) - p(H_1, H_2 | I) \]

- product rule (AND):
\[ p(H_1, D | I) = p(D | H_1, I)p(H_1 | I) \]
\[ = p(H_1 | D, I)p(D | I) \]

\[ p(H_1 | D, I) = \frac{p(D | H_1, I)p(H_1 | I)}{p(D | I)} \]

Bayes Theorem
Bayes' theorem:

\[ p(H_1 \mid D, I) = \frac{p(D \mid H_1, I) p(H_1 \mid I)}{p(D \mid I)} \]

**Prior**: knowledge before experiment (logically)
**Likelihood**: Probability for data if the hypothesis was true
**Posterior**: Probability that the hypothesis is true given the data
**Evidence**: normalization; important for model comparison

*Maximum Likelihood* approach (parameters which maximise probability for data) *does not* yield most likely parameters!

Examples:

\[ p(\text{wet street} \mid \text{rain, I}) \neq p(\text{rain} \mid \text{wet street, I}) \]
\[ p(\text{female} \mid \text{pregnant, I}) \neq p(\text{pregnant} \mid \text{female, I}) \]

*) in general, except e.g. flat, unbounded priors
Processing of information

**Toy Example: Mass of elementary particle**

- Prior knowledge: $0 \leq m \leq m_{\text{upper}} = 0.2$
- Measurement uncertainty $\sigma = 0.15$
- Measured data:
  - $d_1 = 0.09$,
  - $d_2 = -0.2$,
  - $d_3 = 0.05$

Maximum likelihood: $m = -0.02 \ (< 0!)$

Bayes:

1) Assign prior: $p(m \mid I) = \begin{cases} 1/m_{\text{upper}}, & \text{if } 0 \leq m \leq m_{\text{upper}}; \\ 0, & \text{otherwise}, \end{cases}$

2) Assign likelihood: $p(d \mid m, \sigma, I) = \frac{1}{\sigma \sqrt{2\pi}} \exp \left( -\frac{1}{2} \frac{(d - m)^2}{\sigma^2} \right)$

3) Compute posterior using Bayes' theorem:
Processing of information

posterior: \[ p(m \mid d, \sigma, I) = \frac{p(m \mid I) \prod_{i=1}^{N} p(d_i \mid m, \sigma, I)}{Z} \]

contains complete information:

summarizing quantities:
- mode: \( m = 0 \)
- mean: \( \langle m_0 \rangle = 0.06 \)
- 95%-interval: \([0;0.145]\)

Non-gaussian likelihoods: 😁

E.g. \[ p(d \mid m, \beta, I) = \frac{\beta}{\pi \left( \beta^2 + (d - m)^2 \right)} \]

for \( d_3 \) with \( \beta = 0.05 \)
Processing of information

Reasoning about parameter $a$:
(uncertain) prior information $p(a|I)$
+ (uncertain) measured data $d = D \pm \varepsilon$
+ physical model $D = f(a)$

$\{ \begin{align*} p(d|a) \text{ likelihood distribution} \\ p(a|l) \text{ prior distribution} \end{align*} \}$

+ Bayes theorem

$$p(a \mid d) = \frac{p(d \mid a) \times p(a)}{p(d)}$$

posterior distribution

+ an additional (nuisance) parameter $b$:

$$p(a \mid d) = \int db \ p(a, b \mid d)$$
$$= \int db \ \frac{p(d \mid a, b) \ p(a, b)}{p(d)}$$

Marginalization: generalized error propagation

Clear recipe how to tackle a problem – possibly demanding mathematics/numerics

1) Quantify information at hand in probability distributions
2) Multiply probability distributions
3) Marginalize nuisance parameters
4) Analyze posterior distribution
Outline

II. Model Comparison

Basic Concept
Mass Spectroscopy
How many boxes are in the picture?"

Desired: **Occam's Razor** (Prefer simpler models (that fit the data))

*)MacKay, Information theory
Model Comparison

Bayesian Approach:

\[ I = (M_1 + M_2 + \ldots) \] — Specify a set of models.
\[ H_i = M_i \] — Hypothesis chooses a model.

*Posterior probability for a model:*

\[
p(M_i | D, I) = p(M_i | I) \frac{p(D | M_i, I)}{p(D | I)} \propto p(M_i) \mathcal{L}(M_i)
\]

But \[ \mathcal{L}(M_i) = p(D | M_i) = \int d\theta_i p(\theta_i | M_i) p(D | \theta_i, M_i). \]

Likelihood for model = Average likelihood for its parameters

\[ \mathcal{L}(M_i) = \langle \mathcal{L}(\theta_i) \rangle \]

• Bayes Model Comparison requires always alternative models
The Occam Factor:

\[ p(D|M_i) = \int d\theta_i \, p(\theta_i|M) \, L(\theta_i) \approx p(\hat{\theta}_i|M) \, L(\hat{\theta}_i) \, \delta\theta_i \]

\[ \approx L(\hat{\theta}_i) \frac{\delta\theta_i}{\Delta\theta_i} \]

\[ = \text{Maximum Likelihood } \times \text{Occam Factor} \]

Models with more parameters often make the data more probable—for the best fit.

Occam factor penalizes models for “wasted” volume of parameter space.

*) T. Loredo, Garching, 2003
Mass Spectrometry

Quadrupole mass spectrometry with electron impact ionization:

\[ A + e^- \Rightarrow A^+ + 2e^- \]

Electron impact ionization can lead to complex, instrument dependent fragmentation patterns:

Versatile tool for neutral gas analysis: fast, high dynamic range, flexible,...

\[ CH_4 + e^- \Rightarrow CH_3^+ + H + 2e^-, \quad CH_2^+ + 2H + 2e^- \]

\[ CH^+ + 3H + 2e^-, \quad C^+ + 4H + 2e^- \]

\[ H^+ + CH_3 + 2e^-, \quad ... (+^{13}C, D) ... \]
"a mass spectrometrist is someone, who figures out what something is, by smashing it with a hammer and looking at the pieces"

Quote from Th. Schwarz-Selinger
Mass Spectrometry

Model: \[ d_j = C x_j + \epsilon_j \] with cracking matrix \( C \) (\( C, x \) are uncertain/unknown)

Likelihood: \[
p(D|C, X, \{S\}, E, I) = \prod_j \frac{1}{\prod_i \sqrt{2\pi s_{ij}}} \exp \left( -\frac{1}{2} (d_j - C x_j)^T S_j^{-1} (d_j - C x_j) \right)
\]

\( S \) : measurement uncertainties
\( E \) : number of species

Prior terms:
- Concentrations \( x \) : depending on experiment (e.g. gas mixture)
- Cracking matrix \( C \) : exponential prior based on point estimates of Cornu&Massot (1979)

Probability for a set of species \( \{E\} \):
\[
p(E|D, \{S\}, I) = \frac{p(E|I) p(D|\{S\}, E, I)}{p(D|\{S\}, I)}
\]

with \[
p(D|\{S, E\}, I) = \int dC dX p(C|E, I) p(X|E, I) p(D|C, X, \{S\}, E, I)
\]

\[ \Rightarrow \text{High-dimensional integrals!} \]
\[ \text{MCMC-Integration} \]
Mass Spectrometry

Measured data:

![Graph showing mass/charge (amu/q) vs. relative intensity (a.u.) with a dynamic range of 4 o.m.](image)

**CH₄ plasma data**

U. von Toussaint, MPE, 25.01.2012
Mass Spectrometry

**input:**

**data:**
signal + error of 34 mass channels for 27 different plasmas conditions

calibration measurements + error for 11 species

**prior:**
cracking estimates for 14 species

**output:**
cracking pattern and concentrations (+ errors!) for 14 species (e.g. \( \text{C}_4\text{H}_2, \text{C}_3\text{H}_6 \))

\( \text{CH}_4 \) plasma data

'Shifted' cracking pattern for radical \( \text{CH}_3 \) very bad

How many radicals are in the plasma?

model comparison with „Occams Razor“

CH\textsubscript{3}, C\textsubscript{2}H\textsubscript{5}, H

fitting noise!
An extreme example of model comparison: Neural Networks without training data

U. von Toussaint et al, JAO, 2006

U. von Toussaint, MPE, 25.01.2012
<table>
<thead>
<tr>
<th>III. Experimental Design</th>
<th>Basic Concepts Nuclear Reaction Analysis</th>
</tr>
</thead>
</table>

U. von Toussaint, MPE, 25.01.2012
Experimental Design

Design of experiments:

Maximize information gain of planned experiments
Best performance for various physical scenarios,
    hardware constraints, parameter constraints, financial budget

How to quantify the strength (weakness) of an experiment?
How to quantitatively chose experimental design parameter(s) ?
    (spectral bands, number and position of line-of-sights, measurement times)
Experimental Design

Bayesian Decision Theory

Decisions depend on consequences
Might bet on improbable outcome if payoff is large

Utility functions
Compare consequences via utility quantifying the benefits of a decision
Choice of action: c (eg next accelerator energy $E_0$)
Utility = $U(c, o)$
Outcome: o (eg next yield $d(E_0)$)

Deciding amidst uncertainty
We are uncertain of what the outcome will be: average possible results

Expected Utility: $EU(c) = \sum_{\text{outcomes}} P(o|c) U(c, o)$

Best choice maximizes EU: $c^* = \arg \max_c EU(c)$
Experimental Design

Information as Utility:
Goal: minimize estimation uncertainty for parameter(s) \( a \), maximize information gain

Utility function: Kullback-Leibler divergence \( K \)
(\textit{generalizes} covariance-matrix)

\[
Utility(d,c) = K(d,c) = \int da \, p(a|d,c) \ln \frac{p(a|d,c)}{p(a)}
\]

(c=action, d=expected data)

Putting it all together:

\[
EU(c) = \int dd \, p(d|c) K(d,c) = \int da \, p(a) \int dd \, p(d|a,c) K(d,c)
\]

Best choice maximizes EU: \( c^* = \text{arg max} \) EU(c)

High-dimensional integration necessary: Markov Chain Monte Carlo-Methods
Experimental Design

- Hydrogen isotope depth profiling using NRA (not too many other ways)
- nuclear reaction: $d(^{3}\text{He},p)^{4}\text{He}$ : background free, range

$$Y_i \propto \int_0^\infty dx \sigma \left( E \left( \vec{c}(x), x, E_i0 \right) \right) \left( \frac{dE}{dx} \right) \cdot c_j(x) + \epsilon_i$$

V. Alimov, J. Roth, NIMB, 2006
Sequential Design: Which energy $E_0$ next?

- depth profile has to be parametrized: piecewise constant or analytically:
  
  \[ c(x) = \{a_0, a_1, a_2, a_3, \ldots \} \quad \text{or} \quad \; c(x) = a_0 \cdot \exp\left(-\frac{x}{a_1}\right) + a_2 \]
Experimental Design

Compute $EU(E)$, and select best energy

Integration by posterior sampling

$CPU: 1-4 \text{ min}$
Experimental Design

Optimize EU, using posterior sampling CPU: 1-4 min

Cycle: Prediction - Verification

U. von Toussaint, MPE, 25.01.2012
Optimize EU, using posterior sampling. 

CPU: 1-4 min

Cycle: Prediction - Verification
Optimize EU, using posterior sampling

CPU: 1-4 min

Cycle: Prediction - Verification
Experimental Design

• Quantification of information
  
  Value of measurements /diagnostics / experimental setups accessible
  Optimized measurements protocols (on-line sequential design)

• Linear and **nonlinear** problems can be tackled

• Necessary integrations computationally demanding:
  
  High-dimensional integrals in data- and parameter space
  Anything beyond greedy algorithms unexplored (n-step ahead designs)

• Considerable gain in efficiency and/or accuracy has been demonstrated

• **Result-driven** automated measurement strategies (robotics) conceivable
Outline

IV. Numerical Interlude

Nested Sampling
Nested Sampling*

Bayesian inference depends on (high-dimensional) integration...

\[
p(D \mid \theta, \mathcal{H}) \ p(\theta \mid \mathcal{H}) = p(\theta \mid D, \mathcal{H}) \ p(D \mid \mathcal{H})
\]

likelihood    prior    posterior    evidence

\[
p(\theta \mid D, \mathcal{H}) = \frac{p(D \mid \theta, \mathcal{H}) \ p(\theta \mid \mathcal{H})}{p(D \mid \mathcal{H})}
\]

posterior

evidence

\[
p(D \mid \mathcal{H}) = \int_{\theta} p(D \mid \theta, \mathcal{H}) \ p(\theta \mid \mathcal{H}) \ d\theta
\]

model comparison

\[
p(\mathcal{H}_1 \mid D) \propto p(D \mid \mathcal{H}_1) \ p(\mathcal{H}_1)
\]
\[
p(\mathcal{H}_2 \mid D) \propto p(D \mid \mathcal{H}_2) \ p(\mathcal{H}_2)
\]

*J. Skilling 2004; MacKay 2006
Nested Sampling

Relation to statistical physics: partition function

**Free energy**

Probability of macrostate $\mathcal{H}$ is proportional to:

$$Z(\beta, \mathcal{H}) = \int_\theta \exp(-\beta E(\theta, \mathcal{H})) \, d\theta$$

**Evidence**

How well a model $\mathcal{H}$ predicted the data:

$$p(D \mid \mathcal{H}) = \int_\theta p(D \mid \theta, \mathcal{H}) p(\theta \mid \mathcal{H}) \, d\theta$$

Common problem:

$$Z = \int_\theta L(\theta) P(\theta) \, d\theta$$

$L(\theta) \equiv \exp(-\beta E(\theta, \mathcal{H}))$  

Boltzmann factor

$L(\theta) \equiv p(D \mid \theta, \mathcal{H})$  

Likelihood function

Why is this integration difficult?
Nested Sampling

Typical situation for expectation values $\langle f \rangle$:

Typical situation for evidence calculation:
Nested Sampling

**Thermodynamic Integration:** Slowly introduce likelihood structure into integrand (similar: parallel tempering, perfect tempering):

\[ Z(\lambda) = \int d\lambda \Gamma^\lambda(x) \pi(x) \]

\( Z(0) = 1 \) and \( Z(1) = \text{Evidence} \)

\[ \ln(I) = \int_{\lambda=\cdot} d\lambda \frac{\partial \ln Z(\lambda)}{\partial \lambda} = \int_{\lambda=\cdot} d\lambda \int d\lambda \ln \Gamma(x) \rho_\lambda(x) \]
Nested Sampling

\[ Z = \int_{\theta} L(\theta)P(\theta) \, d\theta \]

Contour plot of \( L \)

\[ Z = \int_{\theta} L(\theta)P(\theta) \, d\theta = \int_{0}^{1} L(x) \, dx \]

\[ dx = P(\theta) \, d\theta \]

Key concept:

- Sort all points by \( L \)
Nested Sampling

\[ P(\theta^{(1)}) = P(\theta) \]

\[ P(\theta^{(i+1)}) \propto \begin{cases} 
P(\theta) & L(\theta) > L_i \\
0 & \text{otherwise} 
\end{cases} \]

\[ x_1 \sim \text{Uniform}(0, 1) \]

\[ \langle x_1 \rangle = \frac{1}{2} \]

\[ x_2 \sim \text{Uniform}(0, x_1) \]

\[ \langle x_2 \rangle = \frac{1}{2} \langle x_1 \rangle = \frac{1}{4} \]

\[ x_3 \sim \text{Uniform}(0, x_2) \]

\[ \langle x_3 \rangle = \frac{1}{8} \]
Nested Sampling

\[
P(\theta^{(i+1)}) \propto \begin{cases} 
  P(\theta) & L(\theta) > L_i \\
  0 & \text{otherwise}
\end{cases}
\]

- cf Annealing's intermediate distributions

\[
P(\theta | \beta) = \frac{1}{Z(\beta)} L(\theta)^\beta P(\theta)
\]

- Assuming uniform sampling subject to an energy constraint is possible, \( P(X) \) is simple

\[
\left\langle \log \frac{x_{i+1}}{x_i} \right\rangle = -1
\]

Order statistics for \( P(x) \)

\textit{independent of} \( L \)

- Nested sampling’s behaviour is invariant under monotonic transformations of \( L \)
Nested Sampling

**Pro:** Quite different approach, very general, easy to implement

**Con:** Uniform sampling under constraint?

The result:

\[ \hat{Z} = \sum_{i} \delta \hat{x}_i L_i \]

\[ \hat{x}_i \equiv \exp(-i/N) \]
Summary

- Bayesian probability theory provides consistent and transparent approach to the cycle of scientific inference

- Incorporation of available prior information is straightforward

- Drawback of numerical complexity mitigated by
  - New algorithms
  - Increasing computing power

- State of the Art: parameter estimation, model comparison

- Coming soon: probabilistic combination of diagnostics (IDA)

- Still (largely) unexplored:
  - potential of Experimental Design, e.g. robotics, self-adapting 'measurement'-strategies on computer simulations (e.g. automated MD potential generation)
  - novelty detection in large scale simulations / experiments
Outlook

• Promising and/or unexplored research directions

Bayesian Experimental Design

Large data sets and complex (simulation based models) require consistent response-surface estimates (O'Hagan)

Estimation of stochastic partial differential equations or functionals from data (e.g. for turbulence)

Data based model design and model estimation, e.g. with respect to possible causation instead of correlation only (Pearl)

• 32\textsuperscript{th} Workshop on Bayesian Inference and Maximum Entropy Methods at IPP Garching (15.-20. July 2012)
See: http://www.ipp.mpg.de/maxent2012
32\textsuperscript{th} Workshop on Bayesian Inference and Maximum Entropy Methods at IPP Garching (15.-20. July 2012)

See: http://www.ipp.mpg.de/maxent2012

Thank you for your attention