Bayesian Methods

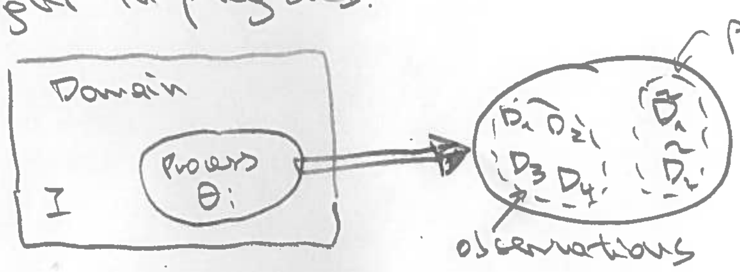
alexander.rothkopf@uis.no
@rothkopfAK

Motivation: - Combining theory and experimental results to explain and predict measurable processes (scattering, transport..)

- Extracting properties of the quark-gluon-plasma from heavy-ion collision data conditioned on a pheno model
  J. Bernhard Nature Physics 15 113 (2019)
- Extract from lattice QCD real-time dynamics of quarks & gluons -> spectral functions, transport coeff., PDF's
  A.R. Front. Phys. 10 28995 (2022)
- Determine Wilson coeff. and their uncertainties from a combined theory & data analysis.
  S. Wesolowski J.Phys G. 43 074001 (2015)

Books: Bayesian Rethinking, R. McElrath
Information Theory, Inference & learning Algorithms, D MacKay

Insight in physics:

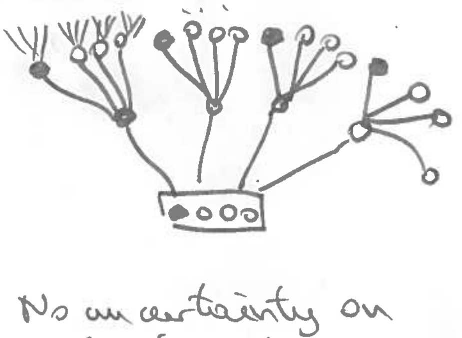


[Inverse problem]

Domain knowledge
↳ set of potential models

Task: Infer most plausible model and parameters, then make predictions.

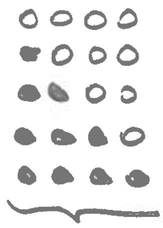
Probability Theory: Efficient formalism for the forward problem: known process, estimate outcomes. [frequentist: probabilities as outcome of repeated experiments]



How probable to draw ●○○○ 2 $\frac{1}{4}\frac{3}{4}\frac{1}{4} = \frac{3}{64}$

No uncertainty on state of system

Bayesian statistics "Plausibility theory"



Efficient formalism for the inverse problem: unknown process (both type & parameters) inferred from observations.

[plausibility: generalized concept of (un)certainty assigned NOT only to data but parameters & models]
{still between [0,1] etc}

What is the most plausible composition of an urn, given observed draws? (prior knowledge: 4 balls): uncertainty on state of system, need to survey all potential cases.

Intuitively: The more observations the more certain

Bayesian Inference (or Bayesian learning) is a stringent application of conditional probabilities. Forces us to make explicit ALL uncertainty by assigning plausibility distributions to data & models.

Starting point: joint probability

$$P(D,\Theta,I) = P(D|\Theta,I)\,P(\Theta,I) = P(\Theta|D,I)\,P(D,I) \quad \Big\} \quad P(\Theta|D,I) = \frac{\overbrace{P(D|\Theta,I)}^{\text{likelihood}}\,\overbrace{P(\Theta,I)}^{\text{prior}}}{\underbrace{P(D,I)}_{\text{normalization}}}$$

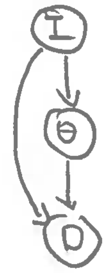
D: ●○● ; Θ: how many ● ; I: 4 balls

Likelihood: How is data generated (conditioned on model)

Prior: Plausibility of model & parameters from domain knowledge

Normalization: Plausibility relative to all potential models

Example of a simple hierarchical / multilevel model (see also Bayesian graph networks)



$$P(●●○○\,|\,●○●,4) = \frac{P(●○●\,|\,●●○○,4)\;P(●●○○,4)}{\text{all possible compositions}}$$

What about $P(●●○○,4)$? Apriori no reason to favor one composition over other 1/5. (BEWARE: choosing a good representation of ignorance may be non-trivial, see e.g. Jeffrey's prior)

	L[●○●]	prior	post.
○○○○	0	1/5	0
●○○○	3/64	1/5	0,15
●●○○	8/64	1/5	0,4
●●●○	9/64	1/5	0,45
●●●●	0	1/5	0

- Distribution over parameters
- simple to count L for 3 draws

What if new draw comes in [●○●●]

Efficient <u>Bayesian updating</u>: Posterior → new prior
likelihood → prob. of SINGLE new draw cond. on model

L[●]	prior	post.
0	0	0
1/4	0,15	0,07
2/4	0,4	0,35
3/4	0,45	0,59
1	0	0

New observation brings [●●●○]

⇒ iterated <u>learning process</u>

Interpretation of prior plausibility: model (parameter) uncertainty informed by domain prior knowledge.
(in the past prior often chosen for computational convenience "conjugate prior" or on generic information theory arguments)

BEWARE: Bayes is explicit in modeling choices. Non-Bayesian approaches often contain hidden prior assumptions, which are difficult to probe.

Example Machine learning: Training with data, choice of learning cost functional often includes regulators, structure of NN.

<u>Likelihood</u>: Basis for χ^2-fitting ≡ maximum likelihood fit in Bayesian view constant prior $P(\Theta, I) = 1$

For a process with two outcomes Binomial $P_B(n_a, n_{tot}, p) = \frac{n_{tot}!}{n_a(n_{tot}-n_a)!} p^{n_a}(1-p)^{n_{tot}-n_a}$
(relevant for <u>classification</u> tasks)

For continuous variables with observations fulfilling the central limit theorem $P(D_n|\Theta, I) \sim N(\mu(D_n, I, \Theta), \sigma)$
(example body height & weight)

<u>linear regression</u> $\mu = \Theta_1 + \Theta_2 \cdot D_n + \ldots$ predictor variable (Model has "=" not "~")

In Bayesian setting: expressivity of model is uncertain (functional dependence on predictor & # of params) and parameters Θ carry prior distribution.

(Jargon: linear models with Gaussian prior on slope parameter called ridge regression.)

Evaluating the posterior

I Quick and cheap: Maximum aposteriori (MAP) + Fisher curvature

1) Compute point estimate $\left.\frac{\delta P(\Theta|D,I)}{\delta\Theta}\right|_{\Theta=\Theta_{MAP}} = 0$ "most probable Θ"

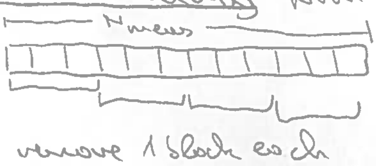
2) Quadratic approx to posterior $\delta^2 P/\delta\Theta^2$ "how shallow is maximum?"

BEWARE: Gaussian approx of tailed distribution (or even multimodal)

$\Rightarrow$ Exploit explicit dependence of $P(\Theta|D I)$ on data uncertainty & prior

statistical error / systematic error

Data uncertainty from (Jackknife) resampling:



N_B: $\to \bar{D}_1$, $\to \bar{D}_2$, $\to \bar{D}_3$, $\to \bar{D}_4$

$\left.\begin{matrix} f(\bar{D}_1) \\ f(\bar{D}_2) \\ f(\bar{D}_3) \\ f(\bar{D}_4) \end{matrix}\right\} \bar{\bar{f}} \qquad \sigma_{\bar{\bar{f}}}^2 = \frac{N_B - 1}{N_B} \sum_n \left(f(\bar{D}_n) - \bar{\bar{f}} \right)^2$

If one needs $\sigma_{\bar{D}}$ for $f(\bar{D}, \sigma_{\bar{D}})$ autocorrelations can be lessened by

Estimation of systematic error here "unsystematic". Variation of prior probability from judicious choice.

II Full Markov Chain Monte Carlo (from a prob. perspective 1701.02434) hands-on in tutorials

Powerful tool for inference and prediction since access to full posterior. This includes BOTH stat. & systematic uncertainties.

Inference $P(\Theta|D,I)$ Posterior predictive distribution

$$P(\tilde{D}|D) = \int d\Theta \; P(\tilde{D}|\Theta,I)\, P(\Theta|D,I)$$

Allows cross-validation: estimate $P(\Theta|D,I)$ on "training" subset of data and compare predictions $P(\tilde{D}|D)$ with "validation" data set.

How to summarize posterior? MAP = mode, mean, median?

Depends on context: If the loss incurred due to misprediction scales with certain power -> optimal choice

loss $|P_{est} - P_{true}| \rightarrow$ median $\qquad (P_{est} - P_{true})^2 \rightarrow$ mean

Model uncertainties: von Neumann's elephant

"with 4 parameters I can fit an elefant with 5 I can make his trunk wiggle."

Intuitive:
- Ockhams Razor: "models with fewer assumptions preferred"
- ignorance of cause: potential causes that produce data in more ways are more plausible
- adding more parameters improves fit (on training data)

Goal: model learns to generalize - predictive accuracy
(think of model fitting as data compression c.f. variational autoencoder)

Need measure of distance between distributions: via information entropy

Information: Reduction in uncertainty, derived from learning outcomes
uncertainty ≙ entropy $\quad H(p) = -\langle \log(p) \rangle = -\int dx\, p(x) \log(p(x))$

["Maximum Entropy" principle looks for "least surprising distribution"]
May be a good 1st guess BUT not always best choice when e.g. prior info avail.

Assume we know the correct target distribution $t(\theta)$

Concept of Divergence: Additional uncertainty induced by using a distribution p different from t.

cross-entropy

↳ Kullback-Leibler $\quad D_{KL}(t,p) = \int dx\, t(x) \Big(\log(t(x)) - \log(p(x)) \Big)$

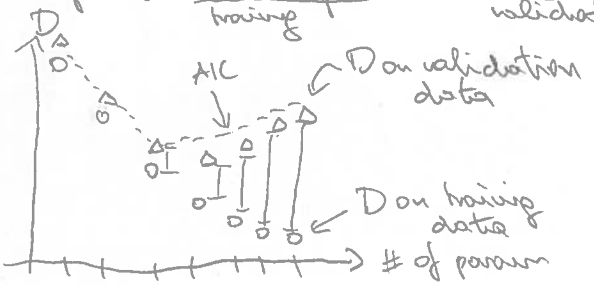
entropy of t

In practice t is unknown so instead often used deviance

$$D(p) = -2 \int dx \log(p(x))$$

Simplest model comp. based on $p = P(D|\theta, I)$
D is the log likelihood.

Compare in-sample (training) to out-sample (validation) deviance for model with diff. # of param.



Difference offers a measure for predictive accuracy.

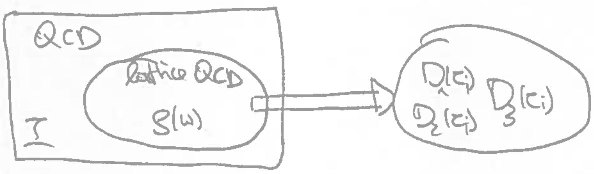
[BEWARE: at this point prior not considered]

<u>AKAIKE Information criterion</u>: attempts to estimate the validation deviance for models with flat prior $\underline{AIC = D_{train} + 2\#\text{param.}} \approx D_{validate}$

For a more robust measure of pred. accuracy when using non-flat priors:
WAIC (widely applicable IC) or LOO (leave-one-out cross validation)
for more details see 1507.04544

How to compare models $\Rightarrow$ average over models using info criteria
assign a model weight $w_m = \dfrac{\exp(-\frac{1}{2} IC_m)}{\sum_m \exp(-\frac{1}{2} IC_m)}$ works for IC's based on deviation scale, exponentiation gives prob.
sample from ensemble of models weighted by w_m.
Interpretation of w_m^{AIC}: "estimates the probability that the model will make the best prediction on new data conditional on set of models considered."

<u>QCD Example</u>: Bayesian inference of spectral functions



- wish to infer $\rho(\omega)$ from $D(\tau)$
- ill-posed since data sparse & approximate $C(\tau_i) + \varepsilon_i$

① infinitely many $\rho(\omega)$'s reproduce $C(\tau)$ within errors. ② Naive inversion leads to exponential dependence of resulting ρ on error in D

$$P(\rho|D,I) = \frac{P(D|\rho,I)\, P(\rho,I)}{P(D,I)}$$

Model for <u>likelihood</u> given by QCD
$D \sim \mathcal{N}(\mu,\sigma)$ $\quad \mu(\tau) = \int d\omega\, K(\omega,\tau)\, \rho(\omega)$

No uncertainty in $K(\omega,\tau)$ since ab-initio input.

$K(\omega,\tau) = \exp(-\omega\tau)$ $T=0$ spectroscopy, NRQCD at $T=0$ & $T>0$, PDFs from Hadronic tensor

$K(\omega,\tau) = \dfrac{\cosh(\omega(\tau-\beta/2))}{\sinh(\omega\beta/2)}$ $T>0$ hadrons

$K(x,\nu) = \cos(x\nu)$ $T=0$ pseudo PDFs

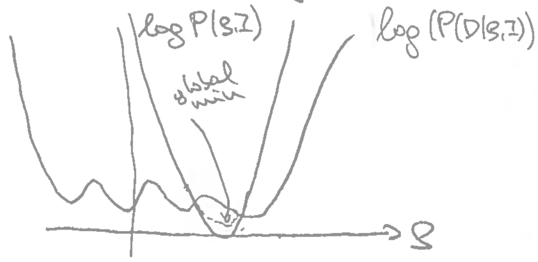
Note that correlations not only along MC time (autocorrelation) but also between the means of correlator at different τ's, i.e. <u>covariance matrix</u> w.r.t. mean

$$Cov_{ij} = \underbrace{\frac{1}{N(N-1)}}_{\text{assuming NO autocorr}} \sum_{\ell=1}^{N_{conf}} \left(D_\ell(\tau_i) - \bar{D}(\tau_i)\right)\left(D_\ell(\tau_j) - \bar{D}(\tau_j)\right)$$ is not diagonal

Instead of modeling cross-correlations → decorrelate data

$\mathrm{Cov} = R\, \mathrm{diag}(\sigma_i)\, R^{+}$ transform data in basis where Cov is diagonal → independent gaussians.

<u>Prior distribution</u> Acts as a regulator for the ill-posed inverse problem. Prior knowledge picks a most probable ρ among the degenerate extrema of the likelihood



In the literature: <u>Exploit positivity of ρ</u>

$$P_T(\rho, I) \propto \exp\left(-\alpha \int d\omega\, (\rho(\omega) - m(\omega))^2\right)$$

Tikhonov Gaussian prior (does not exploit positivity)

Jargon: $m(\omega)$ "default model" parametrizes maximum of prior. α width parameter

$$P_{TV}(\rho, I) \propto \exp\left(-\alpha \int d\omega\, \left|\frac{d\rho}{d\omega}\right|\right)$$

L1 / total variance denoising (also does not use positivity)

see e.g. Phys. Rep. 269 (1996) 133

$$P_{MEM}(\rho, I) \propto \exp\left(\alpha \underbrace{\int d\omega \left[\rho(\omega) - m(\omega) - \rho(\omega) \log\left(\frac{\rho(\omega)}{m(\omega)}\right)\right]}_{\text{Shannon-Jaynes entropy}}\right)$$

justified via axioms: Do not introduce correlations where none are present in the data. Challenge: originally from 2d image recon and axioms refer to that setting & assumes ρ is probability distribution w/o units.

$$P_{BR}(\rho, I) \propto N \exp\left(\int d\omega\, \alpha(\omega) \left[1 - \frac{\rho(\omega)}{m(\omega)} + \log\left(\frac{\rho(\omega)}{m(\omega)}\right)\right]\right)$$

see Y. Burnier, A.R. 1307.6106

justified via axioms: Scale invariance - units of ρ may not matter and smoothness of ρ is assumed. Weakest known regulator: let's the data speak but tendency to show more ringing than other priors.

In Bayesian continuum limit ($N_\tau \to \infty$, $\Delta D \to 0$) Bayes theorem guarantees correct inference of ρ. For finite # of datapoints N_τ and finite error ΔD choice of prior determines how efficiently limit is approached and what artifacts encountered on the way (oversmoothing vs. ringing e.g.)