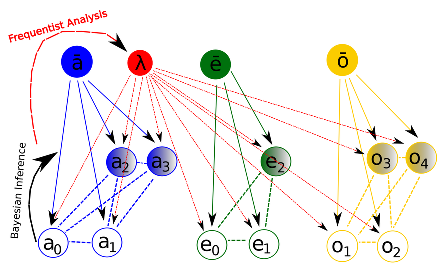


DATA MINING: BAYESIAN NETWORKS IN QUANTUM CHROMODYNAMICS

EMANUELE BAGNASCHI, MATTEO CACCIARI,
ALBERTO GUFFANTI, LAURA JENNICHES

DATA MINING

Data mining is the computational process of analyzing a large set of data.

Examples:

- **Bayesian networks [1, 2]:** probabilistic graphical model, set of random variables (nodes: observable quantities, hidden/ unknown parameters) with conditional dependences, priors on random variables.
- **Neural networks:** inspired by human learning, system of interconnected *neurons* (nodes) which compute values from input, parameters tuned by learning algorithm.
- **Fuzzy logic:** Truth values ranging from 0 to 1, concept of fuzzy set membership (how much is a variable in a set vs. how probable is it that it is in the set in probability theory).
- **Decision tree learning:** input variables (interior nodes), target variables (leaves), train tree by splitting up the data set

Here, we present an example of a Bayesian network.

QUANTUM CHROMODYNAMICS

Quantum Chromodynamics (QCD) is the theory of strong interactions (between quarks and gluons) in the Standard Model of particle physics.

- **Confinement:** The force between quarks grows larger with distance ⇒ they are only observed in bound states (protons, neutrons, ...).
- **Asymptotic Freedom:** at high energies (short distances), the strong coupling becomes small.
- **Perturbative expansions:** For small coupling, observables O can be calculated as an expansion in the strong coupling

$$O = \sum_{n=0}^{\infty} \alpha_s^n c_n$$

- In **high energy particle physics**, such observables can be decay rates, event shapes, production cross sections, sum rules .... These observables receive QCD corrections and can be written as a series expansion because of strongly interacting particles coupling to the particles involved in the decay or production process.

PRIORS [3]

In Bayesian networks, we encode all information which we have before starting the analysis in priors on the random variables. For the Bayesian network on the right, we need three priors (if λ is determined in a frequentist analysis).

- **Likelihood**

$$f(c_n|\bar{c}) = \frac{1}{2\bar{c}} \begin{cases} 1 & \text{if} \quad |c_n| \le \bar{c} \\ 0 & \text{if} \quad |c_n| > \bar{c} \end{cases}$$

- **Shared information and independence**

$$f(\{c_i, i \in I\}|\bar{c}) = \prod_{i \in I} f(c_i|\bar{c})$$

- **Hidden parameter** (non-affirmative)

$$f_\varepsilon(\bar{c}) = \frac{1}{2|\log \varepsilon|} \frac{1}{\bar{c}} \begin{cases} 1 & \text{if} \quad \varepsilon \le \bar{c} \le 1/\varepsilon \\ 0 & \text{else} \end{cases}$$

REFERENCES

[1] G. D'Agostini, Improved iterative Bayesian unfolding

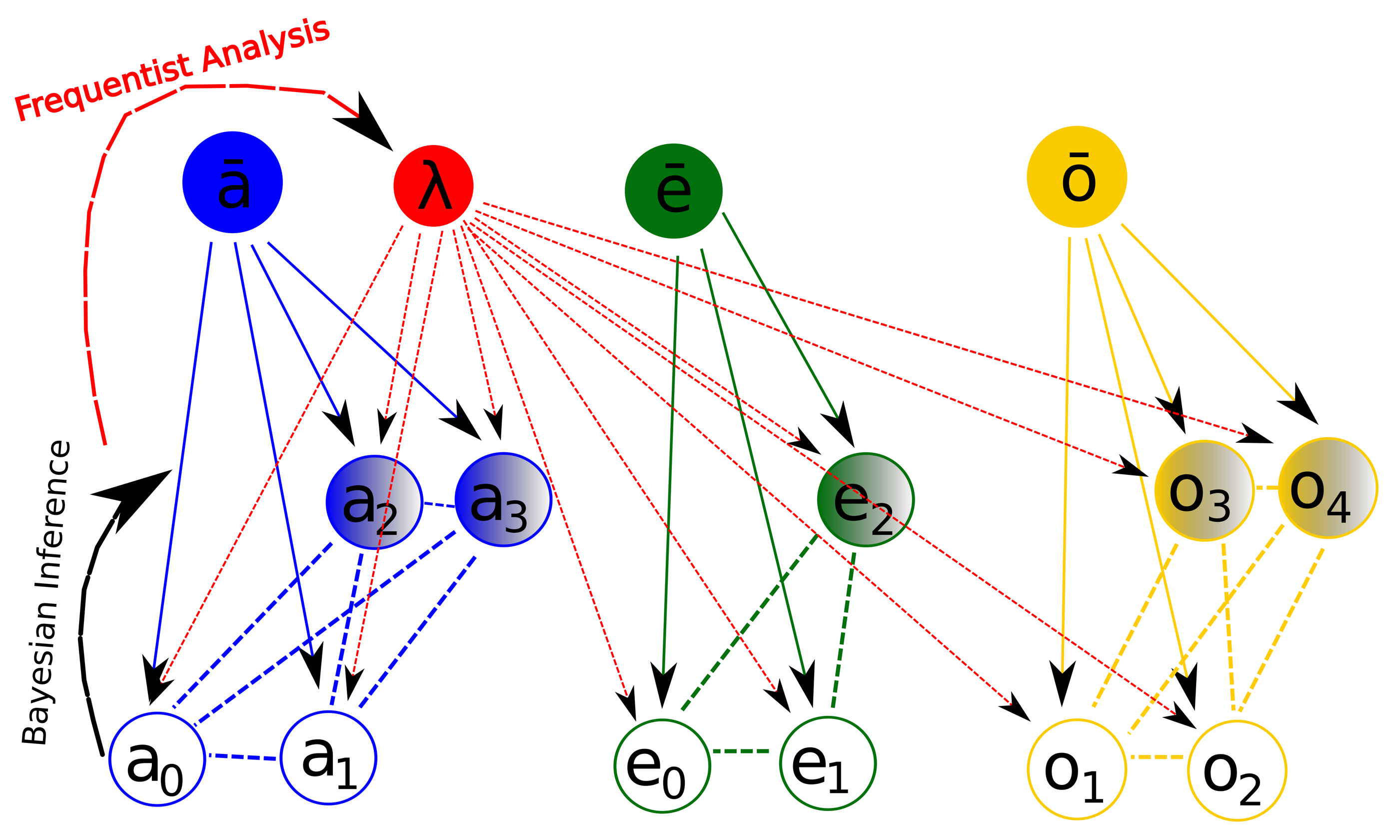
[2] G. D'Agostini, Probability and Measurement Uncertainty in Physics - a Bayesian Primer

[3] M. Cacciari and N. Houdeau, PoS EPS -**HEP2011** (2011) 275.

[4] E. Bagnaschi, M. Cacciari, A. Guffanti and L. Jenniches, An extensive survey of perturbative theoretical uncertainty estimates

BAYESIAN NETWORK

Central question: How large are the uncertainties introduced by neglecting higher orders in perturbative QCD? Find an estimator for the remainder $\Delta_k = \sum_{n=k+1}^{\infty} \alpha_s^n c_n$ (→ see introduction to QCD in the left column).



Random variables for three observables

$$O_k = \sum_{n=0}^{k} \left(\frac{\alpha_s}{\lambda}\right)^n \tilde{c}_n \quad \text{for} \quad \tilde{c}_n = c_n \lambda^n,\ c = a, e, o$$

- **Observable quantities:** Known coefficients $a_0, a_1, e_0, e_1, o_1, o_2$.
- **Hidden parameters:** $\bar{a}, \bar{e}, \bar{o}$.
- **Unknown parameters:** $\lambda, a_2, a_3, e_2, o_3, o_4$.

- **Bayesian inference** to determine posterior distributions $f(c_{k+1}|c_0, \ldots, c_k)$. For small $\alpha_s \sim 0.1 \ldots 0.2$ (as in pert. QCD), we can approximate $\Delta_k \approx \alpha_s^{k+1} c_{k+1}$.
- **Frequentist methods** to determine optimal expansion parameter α_s/λ (not uniquely defined in QCD).

BAYESIAN INFERENCE [3]

- Goal: Obtain **posterior conditional density distribution** for the first unknown coefficient c_{k+1} given the coefficients $c_0, \ldots, c_k$

$$f(c_{k+1}|c_0, \ldots, c_k) = \frac{f(c_0, \ldots, c_k, c_{k+1})}{f(c_0, \ldots, c_k)}.$$

- The **density functions** $f(c_0, \ldots, c_k)$ and $f(c_0, \ldots, c_k, c_{k+1})$ can be obtained by marginalizing $f(c_0, \ldots, c_k, \bar{c})$

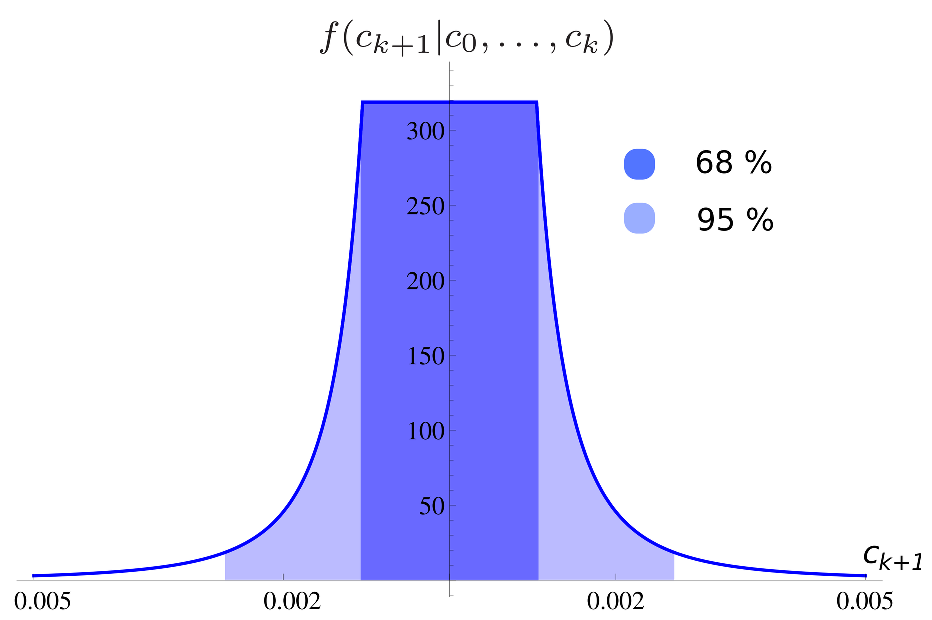
$$f_\varepsilon(c_0, \ldots, c_k) = \int d\bar{c}\ f_\varepsilon(c_0, \ldots, c_k, \bar{c}) = \int d\bar{c}\ f_\varepsilon(c_0, \ldots, c_k|\bar{c}) f_\varepsilon(\bar{c}).$$

- Here, we need the explicit dependence on the small parameter ε, which we can send to zero later on. Because of the **independence** of the likelihood functions, we can write

$$f_\varepsilon(c_0, \ldots, c_k) = \int d\bar{c}\ f(c_0, \cdots, c_k|\bar{c}) f_\varepsilon(\bar{c}) = \int d\bar{c} \prod_{i=0}^{k} f(c_i|\bar{c}) f_\varepsilon(\bar{c})$$

- Plugging in the **priors** (defined on the left), we obtain the conditional density function (which provides us with the information we need to calculate uncertainty bands).

Sending $\varepsilon \to 0$, the **conditional density function** is:



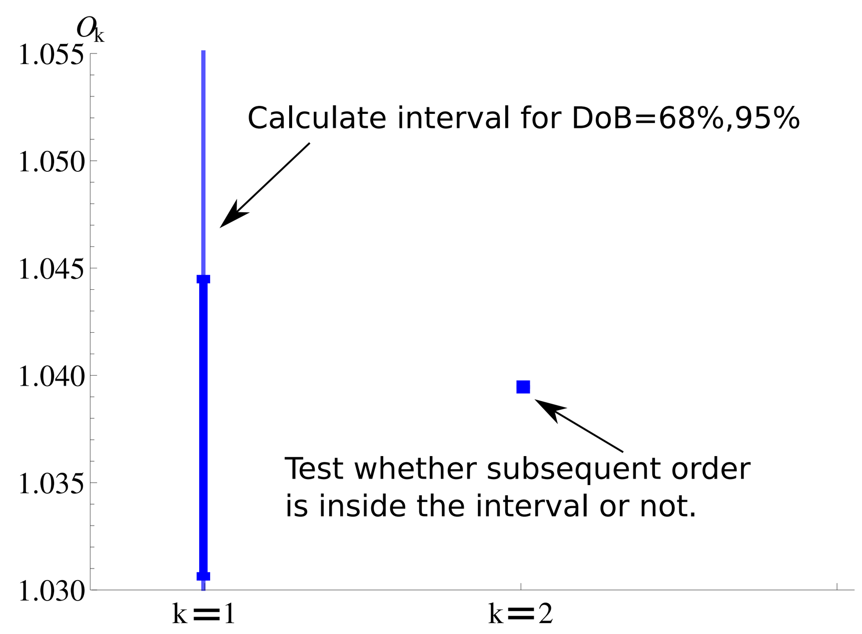
A FUTURE DIRECTION

- Use results for uncertainty bars when comparing experimental results to theory.
- Use Bayesian inference (instead of frequentist approach) to determine optimal λ.

FREQUENTIST ANALYSIS [4]

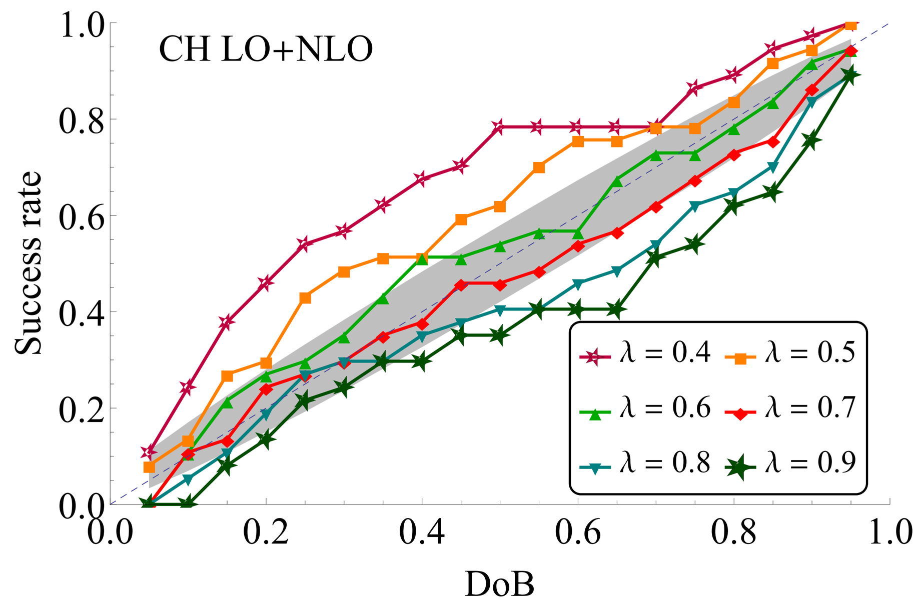
Use set of 20 observables to find *optimal* expansion parameter α_s/λ.

1. Calculate uncertainty interval for one observable at order k for given degree of belief (DoB).
2. Test whether known higher order is within DoB interval or not.



3. Repeat this for all observables and at all orders, define success rate (how many higher order contributions are within).

Success rate vs. input DoB. Optimal λ value: success rate close to input DoB (blue, dashed line):



Scan through DoB values (0.05 to 0.95, steps of 0.01), histogram optimal λ values (success rate close to input DoB).