# Towards fully Bayesian analyses in Lattice QCD

arXiv:2302.06550

Julien Frison

Lattice Field Theory Seminar HU/DESY,

08.04.2023

# Overview

Introduction

Fundamentals

State-of-the-art from a Bayesian perspective

Implementation

A few models

(Truly) Bayesian Model Averaging

Information Criteria

Conclusion

# Objectives

We want:

- efficient “learning” of physical parameters
- well-defined probabilistic interpretation
- unified and consistent framework
- combine strengths of current methods
- flexible model building with arbitrary assumptions
- metrics to test any assumption

# Current methods

## Bootstrap/Resampling

Poor support of auto-correlations

## $\Gamma$ method

Gaussian approximations and linearisation

## $\chi^2$ fit

- Gaussian likelihood
- Covariance needs to be known in advance and precisely
- Often unstable. No theoretical convergence toward smthing meaningful with finite data.

## Akaike IC

- Requires a reliable knowledge of correlated $\chi^2$
- Even more: needs data **parametrisable** by a **regular** model
- Nb of models to explore quickly explode $\Rightarrow$ computing time ($\times$bootstrap)

# My solution

## Make it bayesian from the start to the end!

- We directly get distributions and confidence intervals
- Every assumption is packed into the model, which can be made arbitrarily complicated
- Distance from model to truth can always be evaluated, with a robust criterion
- The HMC (a second one) makes it doable in practice

# Fundamentals

# The Bayes formula

$$P(a|y,M) = \frac{P(y|a,M)P(a|M)}{P(y|M)} \tag{1}$$

| Bayesian vocabulary | Interpretation |
|---|---|
| parameter | The results we want |
| posterior distribution | Uncertainties |
| likelihood | The statistical model we are fitting |
| prior | Arbitrary to some extent, incomplete prior knowledge |
| marginal distribution | Often "just a normalisation" |

Part of the family of *generative* machine-learning models thanks to the PPD:

$$P(y'|y,M) = \int P(y'|a,M)P(a|y,M)da \tag{2}$$

# Difference with frequentist approach

In a pure frequentist approach, arguments are based only on:

1. The likelihood
2. A **choice** of model (including values of its params) $=$ *Null hypothesis*
3. An arbitrary cutoff for hypothesis testing (e.g. $p < 0.05$ or $\chi^2/\mathrm{dof} < 1.42$)

We **assume** that the model is true, then estimate how much a weird coincidence our data is.
**By itself it does not allow to say *anything* about the uncertainties on models and parameters.**

# Maximum likelihood and maximum a posteriori

Frequentist analyses often use the maximum likelihood estimator (MLE)

$$a_{MLE} = \underset{a}{\operatorname{argmax}}\, P(y|a, M) \tag{3}$$

The maximum a posteriori (MAP) has a more direct interpretation

$$a_{MAP} = \underset{a}{\operatorname{argmax}}\, [P(y|a, M)P(a|M)] \tag{4}$$

Agree when prior flat (just as arbitrary as any prior)

**Dangerous**: peaks/singularities/funnels
Infinitesimal volumes are irrelevant, we want confidence intervals

# Conjugate priors

## Definition

**If the posterior distribution $P(a|y)$ belongs to the same family* of probability distributions than the prior distribution $P(a)$ then the prior is called a conjugate prior for the likelihood $P(y|a)$**

* Not the same distribution (i.e. not a fixed point): values of parameters can vary

## Uses

- Analytical calculations are simplified
- Many are well-known and tabulated
- Uninformative priors can be interpreted as limits of any of these conjugate priors
- Parameters are usually easy to interpret and priors easiers to choose
- Adding more data can be seen as pushing the parameters along a flow

# State-of-the-art from a Bayesian perspective

# The bootstrap

- Way to get an estimate of uncertainty from a method which does not contain such a concept
- In today's ML: related to bagging (e.g. random forests)
- Can be interpreted as a Bayesian model by itself, essentially mixture of Diracs
- As a model, it is incapable of any generalisation:
  - **Non-parametric** is good because it applies to any data (no underfitting),
  - but you have to **break** that at some point if you want **physical parameters**
- Block bootstrapping is a notoriously inefficient way to deal with long-range **correlations**
  Becomes a serious issue nowadays with topology freezing

# The $\Gamma$ method

Consists in two parts:

1. Markov's version of the Central Limit Theorem for "primary quantities"
2. Taylor expansion (in practice linearisation) to propagate errors to "derived quantities"

> More efficient at dealing with (moderate) auto-correlations

> We would like to keep in our Bayesian models this possibility of explicitly describing auto-correlations

> Unfortunately it makes strong Gaussian assumptions everywhere

- Often contradicted by skewness in the bootstraps for noisy signals
- Does not go well with pseudo-Bayesian Model Averaging

1. amounts to inferring means from a **Gaussian likelihood***

* more precisely an AR(p) with Gaussian innovation, see later

2. corresponds to local expansions of the posterior **around the MLE**

# The $\chi^2$ fit (GLS)

Minimise a norm:

$$\bar{\chi}^2_C(a) = [\bar{y} - f(a)]^\dagger \, C^{-1} \, [\bar{y} - f(a)] \tag{5}$$

Corresponds to computing the MLE for the gaussian likelihood

$$P(y|a, M_C) \propto e^{-\bar{\chi}^2_C(a)/2} \tag{6}$$

Fixed $C$ (empirical covariance or its diagonal) is part of the model, $a$ is free and minimised. This **approximation** can be seen as model

$$P(y|a, C, M) \propto \delta\left[C_{ij} - \langle (y_i - \bar{y})^\dagger (y_j - \bar{y}) \rangle\right] e^{-\bar{\chi}^2_{C/\sqrt{n}}(a)/2}, \tag{7}$$

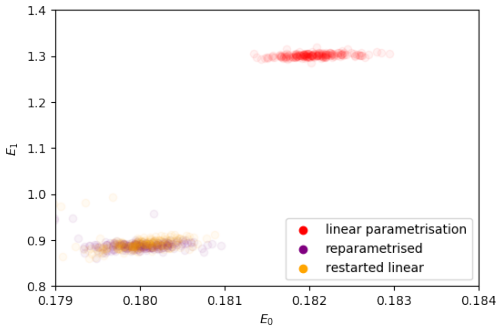
$$\chi^2_C(a) = \sum_{i=1}^{n} [y_i - f(a)]^\dagger \, C^{-1} \, [y_i - f(a)] \, . \tag{8}$$

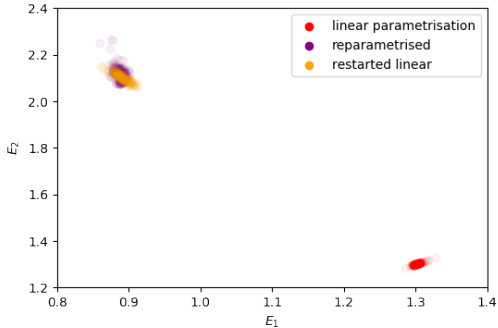
Will see later how it becomes asymptotically justified, but for **finite data** might be wrong. Empirical covariance might not even be invertible (should have probability zero)

# An example of $\chi^2$ failure

Bootstrapped multi-exponential $\chi^2$ stuck at local minima, sensitive to reparametrisation and fine-tuning of initial conditions

Bayes with MCMC does not have this issue, even with naïve parametrisation and flat prior (explores the whole space, not trying to converge to a single point)



# A comment on Gaussianity

## Central Limit Theorem

Counter-intuitively, assuming gaussianity configuration-by-configuration is **not a stronger assumption** than "only" on averages: one obtains the **same posterior**:

$$P(a|y, M_C) \propto e^{-\chi^2_C(a)/2} \propto e^{-\bar{\chi}^2_{C/\sqrt{n}}(a)/2} \tag{9}$$

$n$ observations fitted with a covariance $C \Leftrightarrow 1$ observation fitted with a covariance $C/n$
General property of Gaussians, does *not* depend on the true distribution of $y_i$

## Gaussian likelihood vs Gaussian posterior

One should note mistake one for the other.
Gaussian likelihoods depend on data being summed (cfg/vol/hits/...; avoid ratio/log/meff/...)
Gaussian posteriors require additionally a linear model

# Implementation

# Applying the HMC

We do not need a closed formula for $P(a|y, M)$, we can just draw $a_1, a_2, a_3, \ldots$.
Exactly what our good old HMC does!

$$P(a|y, M) = \frac{P(y|a, M)P(a|M)}{P(y|M)} \tag{10}$$

| Bayesian vocabulary | LQCD analogue |
|---|---|
| parameter | configuration |
| posterior distribution | |
| likelihood | $e^{-S}$ |
| negative log-likelihood | action |
| prior | |
| marginal distribution | partition function |

# Software

- We show tests with PyMC
  In Python and simple to use, but several alternatives exist
- Vectorisation and Automatic Differentiation (HMC forces) handled by Theano
  Made for somewhat complex ML methods & Deep Learning
- Writing a model is then very simple, and it can be anything:
  - Just write as many terms as you want in an expansion
  - With extra human effort you can marginalise irrelevant RV
  - Does not have to be parametric:
    Bayesian Bootstrap, Gaussian Processes, Bayesian Neural Networks, ...
  - Does not need to be the *true* model:
    Models are always an approximation, to be checked a posteriori on data (IC)
- Runs on a laptop but scales with cluster/GPU

# Alternatives to HMC

Once a Bayesian model is defined, one can use, roughly from the "simplest" to the most complicated:

- Maximum A Posteriori (through Scipy)
- Taylor Expansion around MAP (not fully implemented)
- Variational Inference
- Normalising Flows
- **Hamiltonian Monte Carlo** (NUTS)
- **Sequential Monte Carlo**
- Langevin Dynamics (through Jax)
- Normalising Flows w/ Neural Networks + HMC (NeuTraHMC in Pyro)

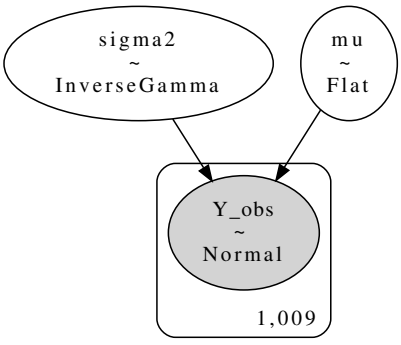
# Data

We show results on a pion correlator for the CLS H101 symmetric ensemble
We stick to this example but this would be valid for any kind of fit or statistical analysis:

- Combined fits for form factors
- Continuum and chiral fits
- $z$ expansion with unitarity constraints [Flynn:2303.11285]
- Phase-shift fits
- Spectral function reconstruction [Rothkopf:2208.13590]
- Topological susceptibility
- ...

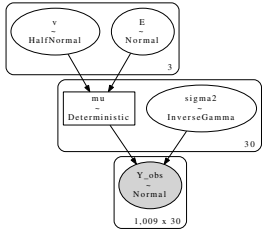
# A few models

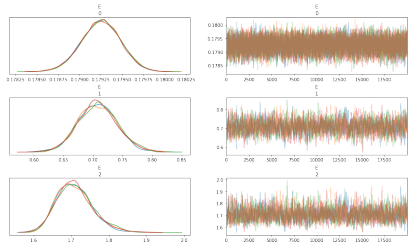
# A trivial model




- PyMC automatically makes such graphs: parameters/priors upwards, observable below can have many layers
- Data: one number $Y_{obs}$ $\times$ 1009 configurations
- **We want to infer $\mu$**
- $\sigma^2$ is just a nuisance parameter
- Adding more nuisance parameters is trivial
- The MLE gives us the usual point estimates: empirical mean and variance
- With $\sigma^2$ frozen this would be a $dof = 0$ fit

# Uncorrelated model for a two-point function

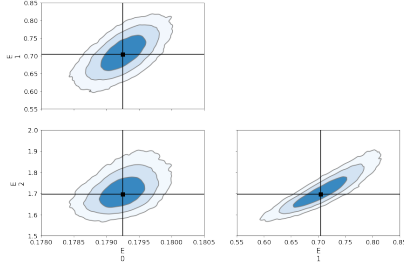


## Model description

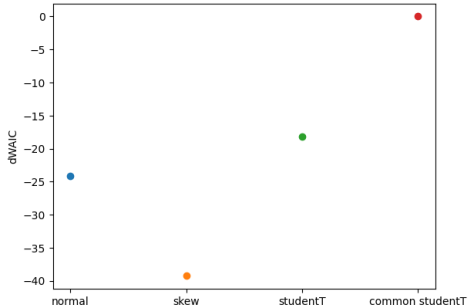
$$y \sim \mathcal{N}(\mu, \sigma), \quad \mu = \sum_{i=1}^{3} v_k \exp(-E_k t) \qquad (11)$$

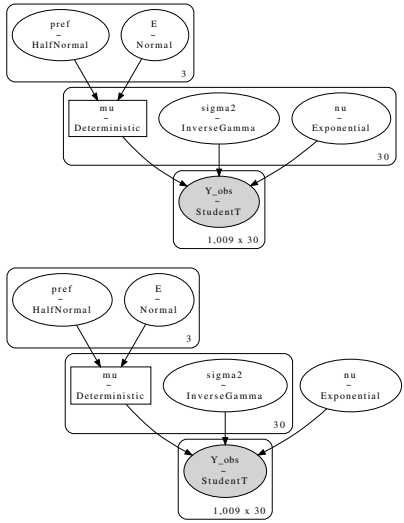
Inverse-Gamma is the conjugate prior for Gaussians
Very uninformative priors enough for good stability



# Probing non-gaussianity

- Building non-Gaussian models is trivial
- Sampling can become inefficient
- Easy to check on a single slice
- Here we build full models and check IC
- Not very interesting with *this* data



# Correlating Euclidian times(1/2)

- **Formally trivial**: replace Normal by MvNormal, $\sigma$ by $C$, Gamma by Wishart
- Poorly conditioned $C$ will be suppressed, never singular even with low stat
- However, **in practice** sampling Matrices can be difficult
- Using the conjugate prior allows a simple **marginalisation** to bypass this issue

$$\mathcal{W}(C^{-1}|V,\nu) = \frac{|C^{-1}|^{(\nu-p-1)/2} e^{-\operatorname{Tr}(V^{-1}C^{-1})/2}}{2^{\frac{\nu p}{2}}|V|^{\nu/2}\Gamma_p(\frac{\nu}{2})} \tag{12}$$

Uninformative when $\nu \ll n$.
We saw the $p = 1$ case earlier, which is called the $\Gamma$ distribution:

$$\Gamma(x|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}. \tag{13}$$

It is a generalisation of the $\chi^2$ distribution for non-integer dof (role played by $\alpha$ or $\nu$).

# Correlating Euclidian times(2/2)

PPD for gaussian likelihood with known mean and a Wishart prior: multivariate Student-t distribution

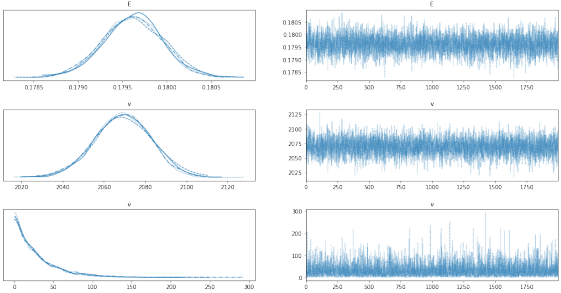
$$t_{\nu+n-p+1}\left(y \mid \mu, \frac{\left(V^{-1}+\sum_{i=1}^{n}(y_i-\mu)(y_i-\mu)^T\right)^{-1}}{\nu+n-p+1}\right). \tag{14}$$

## Convergence to Gaussian

- $n \to \infty$ gives some justification to the $\chi^2$ model
- Main difference is wider tail:
  be more tolerant with outliers since we do not perfectly know $C$
- However, scale matrix is computed from $\mu$ rather than $\langle y \rangle$

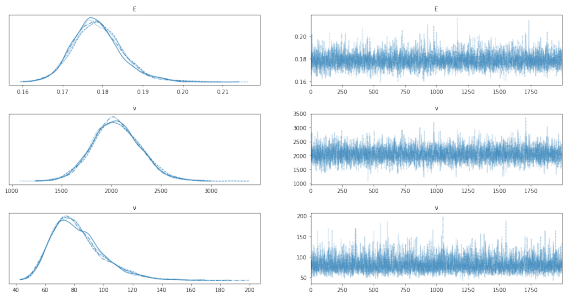
# Results for marginalised Wishart model

Full statistics:




$\nu$ small means uninformative prior is preferred

Low stat (n=9, singular empirical covariance):



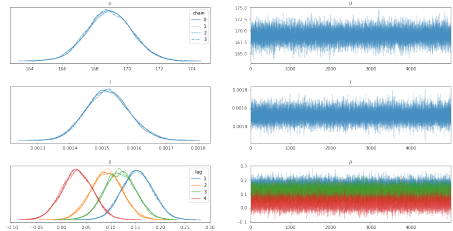

$\nu$ increases to avoid singularities

# Auto-regressive model

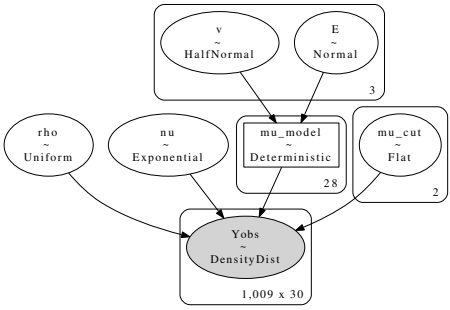
- Until now we neglected auto-correlations (should have binned)
- Now we treat our data as time series
  Auto-correlations are explicitly described
- AR model encodes **long-range correlations** with a few parameters (modes)

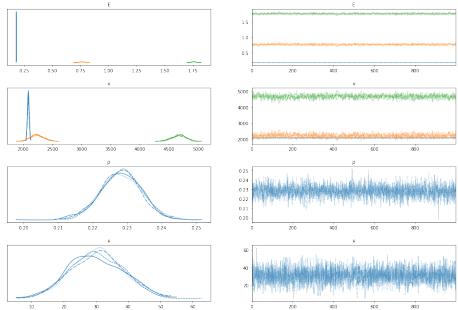
$$y_i = \rho_0 + \sum_{j=1}^{r} \rho_j y_{i-j} + \xi_i, \quad \xi_i \sim \mathcal{N}(0, \tau) \qquad (15)$$

- $\rho_1$ is related to the popular $\tau_{exp}$, cutting in $r$ is similar to choosing window in $\Gamma$ method
- Larger $r$ stable but cut non-significant terms
- Here apply on a single time-slice as a first illustration



# Multi-exponential model with full correlations



- Mixing correlated model (in Euclidian time) with marginalised Wishart prior...
- ... and auto-regressive model (auto-correlation between configurations)

$$y_\tau(t) = \rho_0 + \sum_{i=1}^{r} \rho_i y_{\tau-i}(t) + \xi_\tau(t), \qquad \langle \xi_\tau(t)\xi_\tau(t') \rangle \neq 0 \tag{16}$$

# (Truly) Bayesian Model Averaging

# Data cuts

- Regardless of the method you use for model averaging (or model selection), comparing models only makes sense if they are applied to the **same data**
- This is still compatible with performing cuts: here cuts mean applying **trivial submodels** to some areas of the data.
- In the case of 2-pt functions for instance we cut $t_{min}$:

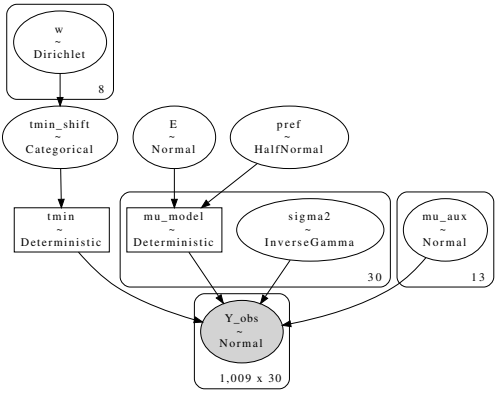
$$y(t < t_{min}) \sim \mathcal{N}(\mu_{aux,t}, \sigma_{aux,t}) \quad (17)$$

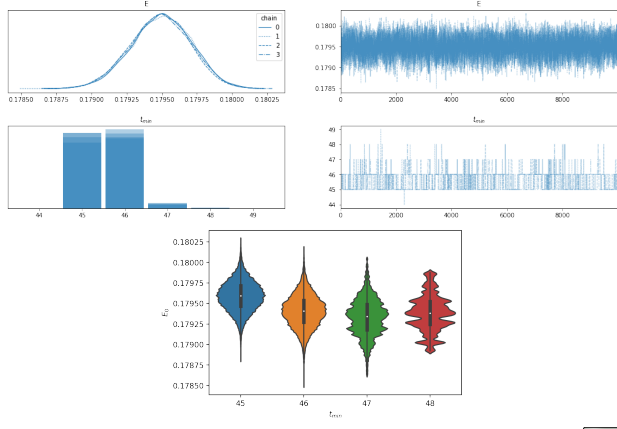
$$y(t \geq t_{min}) \sim \mathcal{N}(f(a,t), \sigma_t) \quad (18)$$

- In principle $\mu_{aux,t}, \sigma_{aux,t}$ are easy to marginalise, but they are also easy to sample

# Mixture model

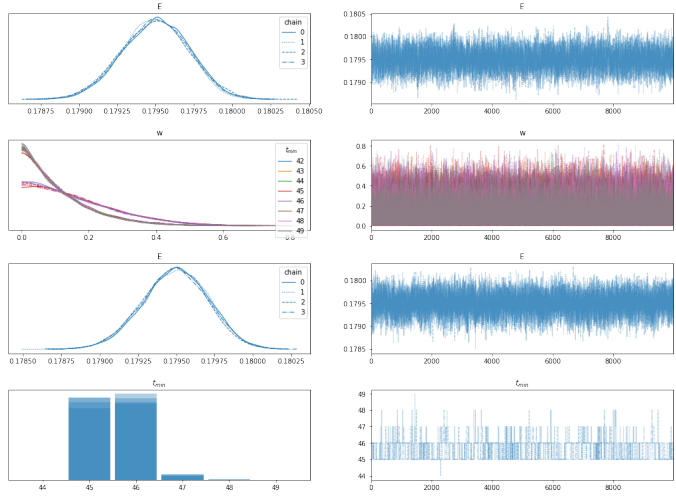
- There are now **many ways** to do model averaging (i.e. include "*systematics*")
- One is to simply put everything into a **single model**:



# $t_{min}$-marginalised or not

- Sampling **discrete variables** can be dangerous
- Marginalising can be a practical solution
- Here both work and agree
- Non-marginalised allow to extract more information



# Information Criteria

# Widely Applicable Information Criteria

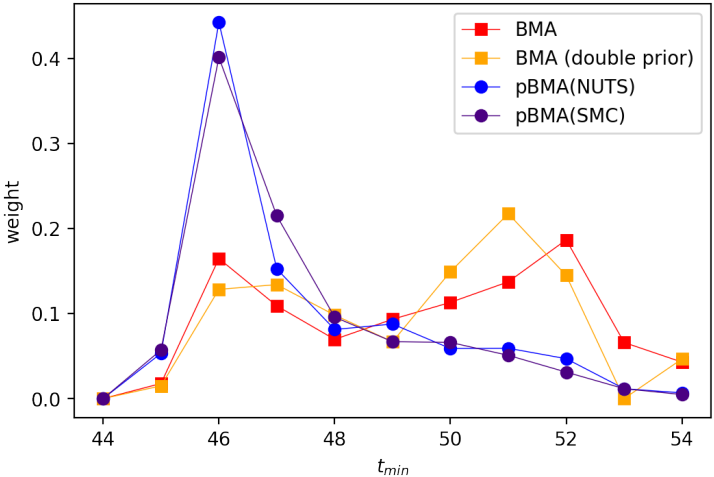
- Generalisation of Akaike IC
- Built for a **generic ML model/method**
- Computing it is just calling a function, regardless of the model
- **Generalisation error of our PPD** from this training data to hypothetical future test data
- Related to **Kullback-Leibler divergence** from model to truth
- Also related to (Leave-One-Out) Cross Validation
- One term against under-fitting and one correction against over-fitting

$$WAIC = \sum_i \log\left[P(y_i \mid y)\right] - k_{WAIC} \tag{19}$$

The *effective* nb of parameters $k_{WAIC}$ depends on fluctuations inside $P(y_i \mid y)$

# BMA vs pBMA

!!! WAIC and Bayes Factors (WBIC) are not exactly the same thing:



# Takeuchi Information Criteria

- Applies to MLE, such as traditional $\chi^2$ fits
- Very similar to WAIC, more general than AIC:
  **the model does not need to parametrise the truth**
- Can be computed analytically for an uncorrelated $\chi^2$:

$$k_{TIC} \simeq \mathrm{Tr}\left[(G^\dagger C_W G)(G^\dagger G)^{-1}\right] = \mathrm{Tr}\left[\mathcal{P} C_W\right], \tag{20}$$

$$TIC = \chi^2_{MLE} - 2E(\chi^2|M^*), \tag{21}$$

where $C_W = WCW^\dagger$ ($C$ is the true covariance for the *true* model $M^*$)
$F_{i\alpha} = \partial f_i / \partial a_\alpha$ ($i \leq p, \alpha \leq k$) is the $n \times k$ Jacobian matrix of the fitting function
$G = WF$ and $\mathcal{P} = G(G^\dagger G)^{-1} G^\dagger$ is a projector.

# Conclusion

# Conclusion

- Fully bayesian framework **well-defined theoretically and put in practice**
- On a simple problem it already tends to outperform standard techniques
  stable 3-state fits, correlations and auto-correlations, low stat, non-gaussian posterior, ...
- Benefits likely to me more obvious on more complicated problems
- More work needed to build good models
  Case-by-case problem, and to some extent it is a good thing
- All **assumptions can be checked/compared with IC**

If you do not like HMC sampling:

- Once a bayesian model is properly defined you can always go back to simpler approximations ($\chi^2$ model, MLE, variational inference, ...)
- Helps to understand things such as the TIC
  Now you can use that for your old-style uncorrelated $\chi^2$

# Thanks for your attention!