Are pairwise models really simpler?

Alberto Beretta[1], Claudia Battistin[2], Clélia de Mulatier[1], Iacopo Mastromatteo[3], Matteo Marsili[1]

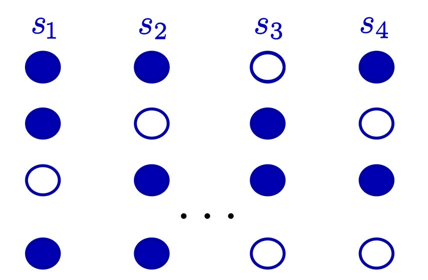
(1) The Abdus Salam International Centre for Theoretical Physics (ICTP), Trieste, Italy
(2) Kavli Institute for Systems Neuroscience, NTNU, Norway
(3) Capital Fund Management, Paris, France

Introduction – Bayesian Model Selection on spin models

A fundamental issue in data analysis is to find the model that best captures the patterns hidden within the data, despite the random errors that effect them. The model should be complex enough to be able to fit the data, but simple enough to capture its main patterns.

Consider a system of n spins that take random values in $\{-1,+1\}$:

Ex. $n=4$

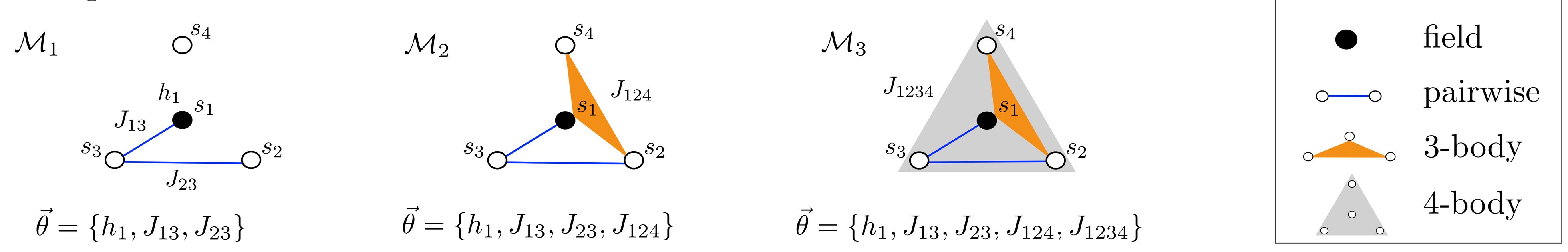


$\vec{s}^{(1)} = \{1,1,-1,1\}$
$\vec{s}^{(2)} = \{1,-1,1,-1\}$
$\dots$
$\vec{s}^{(N)} = \{1,1,-1,-1\}$

Data set: N samples $\hat{s} = \{\vec{s}^{(i)}\}$

Q? What would be the best *model* for the system, that could explain what we observe/reproduce similar data?

Ex. of spin models $\mathcal{M}_i$

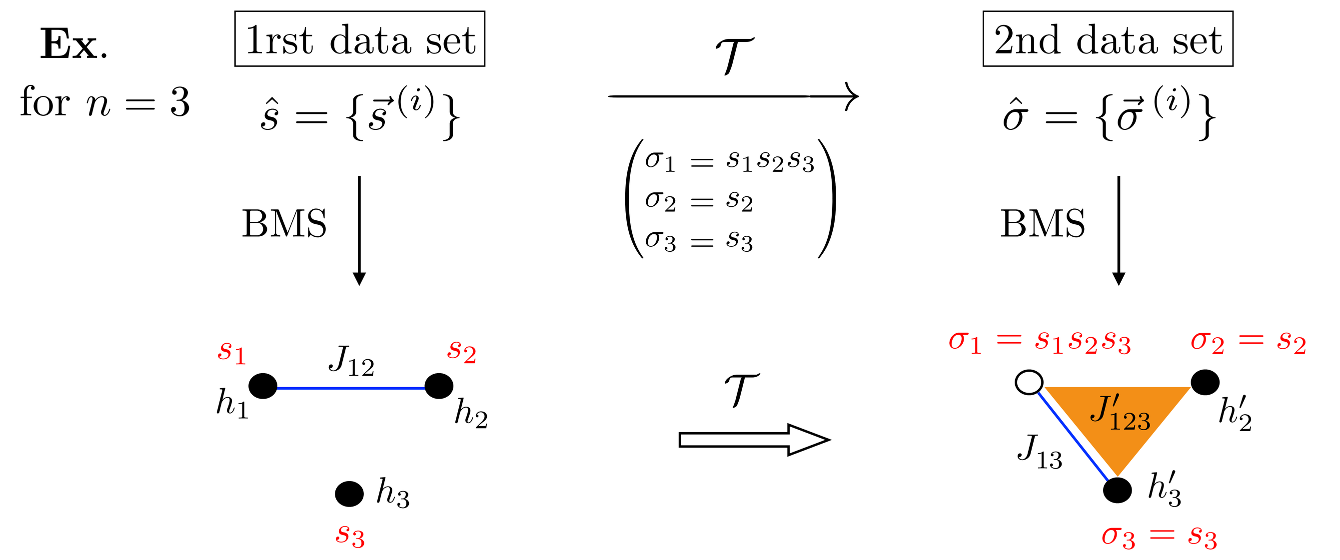


How? Classifying all the possible models $\mathcal{M}_i$ to find the one with the highest *posterior probability* $P(\mathcal{M}_i\,|\,\hat{s})$. However, this is a very difficult task, due to the huge number of models.

To simplify, it is common to reduce the space of the possible models to models with only field and pairwise interactions.

Is it a good idea? Are these models simpler?

$\Longrightarrow$ Pairwise interactions don't seem to play any special role.



How many models?

For the spin models:

- $2^n - 1$ possible interactions
- 2^{2^n-1} possible spin models

Ex. $n=2:$ 8 models; $n=3:$ 128 models; $n=4:$ 32768 models; $n=5:$ 2147483648 models

Models with only fields and pairwise interactions:

- $n\,(n+1)/2$ possible interactions
- $\sim 2^{n^2}$ possible models

Ex. $n=2:$ 8 models; $n=3:$ 64 models; $n=4:$ 1024 models; $n=5:$ 32768 models

Space of PDFs

∘ For a given a model $\mathcal{M}$, the family of probability distributions $\{P(\vec{s}\,|\,\vec{\theta},\mathcal{M})\}_{\vec{\theta}}$ forms a Riemannian manifold with natural coordinates $\vec{\theta}$ [1].

∘ Each point of this space is a probability distribution $P(\vec{s}\,|\,\vec{\theta},\mathcal{M})$.

∘ The natural metric on this manifold is given by the *Fisher Information Matrix* [1]:

$$J_{qk}(\vec{\theta}) = \partial_{\theta_q}\partial_{\theta_q} \log Z_{\mathcal{M}}(\vec{\theta}) = \langle f_q f_k\rangle - \langle f_q\rangle\langle f_k\rangle$$

∘ Varying the parameters $\vec{\theta}$ from $\mathrm{d}^K\vec{\theta}$ gives rise to distributions similar to $P(\vec{s}\,|\,\vec{\theta},\mathcal{M})$ that correspond to nearby points in the manifold, contained in the small volume:

$$\mathrm{d}V_{\mathcal{M}} = \sqrt{\det J(\vec{\theta})}\,\mathrm{d}^K\vec{\theta}$$

Complexity of spin models

Using Bayes' theorem: $P(\mathcal{M}\,|\,\hat{s}) = \dfrac{P(\hat{s}\,|\,\mathcal{M})\,P_0(\mathcal{M})}{P(\hat{s})}$, where

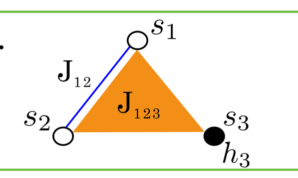
$$P(\hat{s}\,|\,\mathcal{M}) = \int \mathrm{d}^K\vec{\theta}\; \underbrace{P(\hat{s}\,|\,\vec{\theta},\mathcal{M})}\; P_0(\vec{\theta}\,|\,\mathcal{M})$$

In absence of information, the prior $P_0(\mathcal{M})$ can be taken uniform, and models may be ranked directly with $P(\hat{s}\,|\,\mathcal{M})$.

N independent set of data $\searrow$ $= \prod_{j=1}^{N} P(\vec{s}^{\,j}\,|\,\vec{\theta},\mathcal{M})$

Probability that the spin system is in the configuration $\vec{s}^{(i)}$:

Ex.



$$\mathcal{M} = \begin{cases} \vec{f}(\vec{s}) = (s_3, s_1 s_2, s_1 s_2 s_3) \\ \vec{\theta} = (h_3, J_{12}, J_{123}) \end{cases}$$

$$P(\vec{s}^{\,j}\,|\,\vec{\theta},\mathcal{M}) = \frac{\mathrm{e}^{\sum_{k=1}^{K} \boxed{f_k(\vec{s}^{\,j})}\,\theta_k}}{\boxed{Z_{\mathcal{M}}(\vec{\theta})}}$$

Partition function

Spin operator: product of the spins involve in the interaction k

Expanding for a large size N of the data set, finally leads, in the framework of Bayesian Model Selection, to [2,4]:

$$\log P(\hat{s}\,|\,\mathcal{M}_i) = \underbrace{\log P(\hat{s}\,|\,\mathcal{M}_i,\theta^*)}_{\propto N} - \underbrace{\frac{K}{2}\log\left(\frac{N}{2}\right)}_{\propto \log N} - \boxed{c_{\mathcal{M}}} + O\left(\frac{1}{N}\right)$$

Maximum log-Likelihood

Penalty term number of parameter K

Penalty term geometrical complexity

Geometrical Complexity

$$c_{\mathcal{M}} = \log \int_{\mathbb{R}^K} \sqrt{\det J(\vec{\theta})}\,\mathrm{d}^K\vec{\theta}$$

Prior on the values of $\vec{\theta}$

Best choice in absence of any information [2]:

Jeffreys' prior: $\quad P_0(\vec{\theta}\,|\,\mathcal{M}) = \dfrac{\sqrt{\det J(\vec{\theta})}}{\int \sqrt{\det J(\vec{\theta})}\,\mathrm{d}^K\vec{\theta}}$

$\longrightarrow$ invariant under re-parametrisation [5];
$\longrightarrow$ uniform in the space of observables.

Complexity – interpretation

$$c_{\mathcal{M}} = \log V_{\mathcal{M}}$$

$V_{\mathcal{M}}$ is the total volume of the manifold defined by $\mathcal{M}$:

∘ Complexity represents how broad the model is in term of describing various probability distributions [2].

∘ A model is complex if it can fit a wide range of data.

Gauge transformations

$c_{\mathcal{M}}$ is expected to stay invariant under the transformation $\mathcal{T}$ introduced in **Ex a**. This invariance emerges explicitly when expressing $Z_{\mathcal{M}}(\vec{\theta})$ in the form:

$$Z_{\mathcal{M}}(\vec{\theta}) = 2^n \prod_{\mu\in\mathcal{M}} \cosh(\theta^\mu) \left[1 + \sum_{\ell\in\mathcal{L}} \prod_{\mu\in\ell} \tanh(\theta^\mu)\right]$$

- *Loop* ℓ: subset $\ell \subseteq \mathcal{M}$ such that $\prod_{\mu\in\ell} f^\mu(\vec{s}) = 1$;
- Set $\mathcal{L}$: set of all the loops of $\mathcal{M}$.

$Z_{\mathcal{M}}(\vec{\theta})$ depends on few characteristics of $\mathcal{M}$:

– number $|\mathcal{M}|$ of operators;
– the structure of the loops $\mathcal{L}$ of the $\mathcal{M}$;

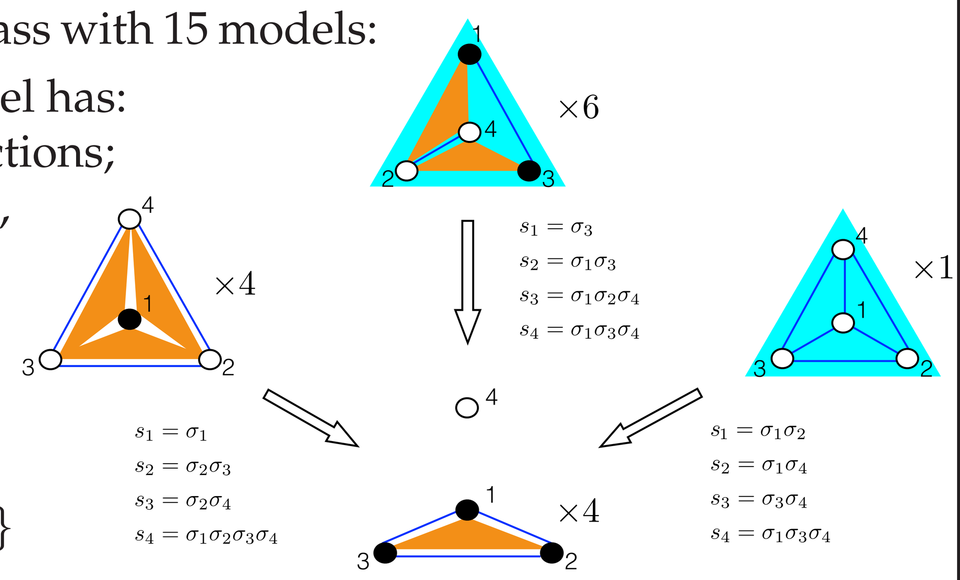
which are invariant under the transformation $\mathcal{T}$. We call (such a transformation) $\mathcal{T}$ a Gauge Transformation as it preserves the geometry of the model.

Thus $c_{\mathcal{M}}$ stays invariant under $\mathcal{T}$, which allows defining *complexity classes* of models (images through a GT and with the same value of $c_{\mathcal{M}}$).

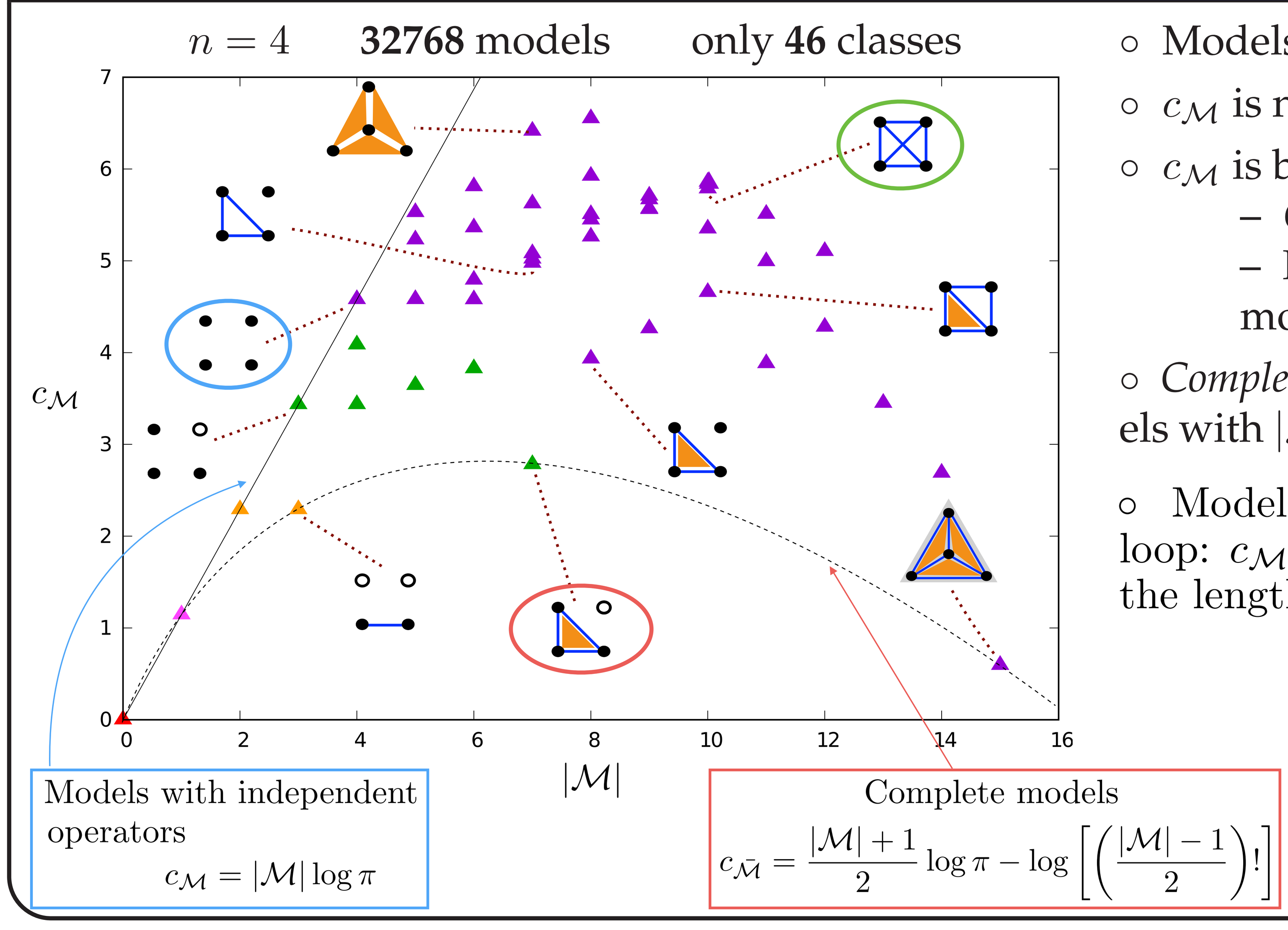
Ex. one class with 15 models:

Each model has:
– 7 interactions;
– 15 loops,

structure: $\{3^7, 4^7, 7^1\}$



Complexity classes



∘ Models with pairwise interactions are not simpler.

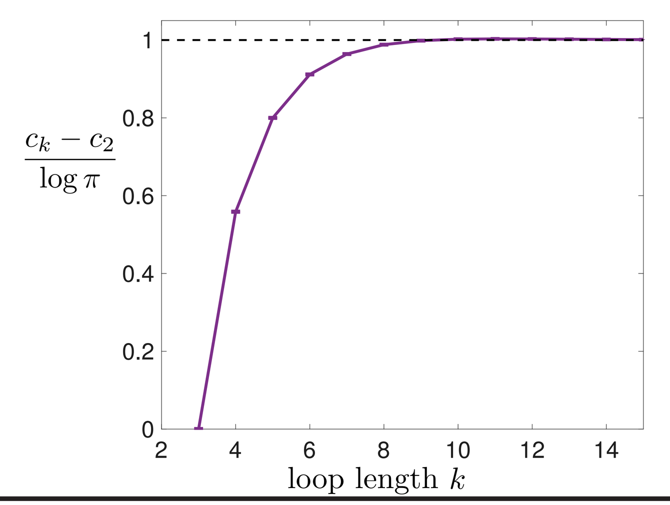
∘ $c_{\mathcal{M}}$ is not monotonic in the number of parameters.

∘ $c_{\mathcal{M}}$ is between two limit curves:

– *Complete models* are the simplest models;
– Models with only independent operators are the most complex.

∘ *Complementary classes*: same number of classes for models with $|\mathcal{M}|$ operators than for models with $2^n - 1 - |\mathcal{M}|$.

∘ Models with a single loop: $c_{\mathcal{M}}$ increases with the length of the loop.



Degenerate models

Degenerate models Operators share parameters:
$\longrightarrow$ $|\mathcal{M}|$ operators, but K parameters with $K < |\mathcal{M}|$;
$\longrightarrow$ α_k degeneracy of θ_k, for $k \in [1, K]$.

Result In general, degeneracy reduces the complexity.

Ex 1. models with $|\mathcal{M}| = \sum_{k=1}^{K} \alpha_k$ independent operators:

$$\exp c_{\mathcal{M}}^{non-deg} = \pi^{|\mathcal{M}|} \qquad \exp c_{\mathcal{M}}^{deg} = \pi^K \prod_{k=1}^{K} \sqrt{\alpha_k}$$

Ex 2. models with a single loop: