Analyzing Binary Data: Is your Model Truly Pairwise?

Clélia de Mulatier[1], Paolo Mazza[2], and Matteo Marsili[3]

[1]University of Pennsylvania, Department of Physics and Astronomy, Philadelphia, USA
[2]Scuola Internazionale Superiore di Studi Avanzati (SISSA), Trieste, Italy
[3]The Abdus Salam International Centre for Theoretical Physics (ICTP), Trieste, Italy

Introduction

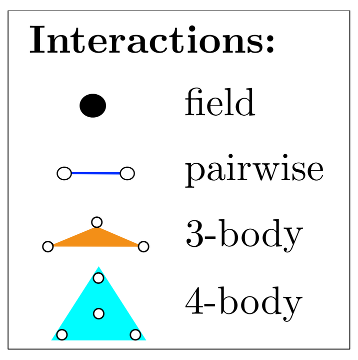
A fundamental issue in data analysis is to **find *the* model** that best captures the patterns hidden within the data, despite the random errors that effect them. This model should be **complex enough** to be able to fit the data, but **simple enough** to capture only the relevant patterns of the data.

In the context of **binary data**, pairwise spin models – Ising models – have been widely used to address this question. But, are pairwise models the best for your data? Are the relevant patterns of your data truly pairwise?

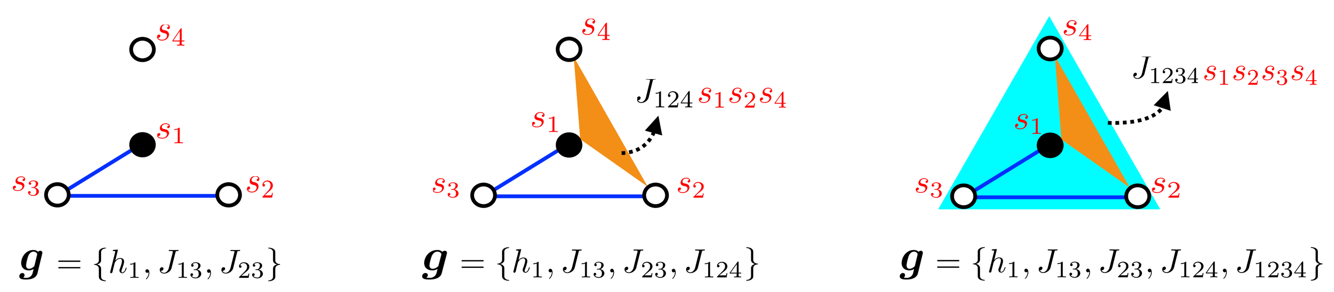
Pairwise Models are not necessarily *simpler*

Q? Which is *the best* spin model for your data?

$$P(\boldsymbol{s}\,|\,\boldsymbol{g},\mathcal{M}) = \frac{1}{Z_{\mathcal{M}}(\boldsymbol{g})}\exp\Big(\sum_{\mu\in\mathcal{M}} \overset{\text{parameter}}{g_\mu}\, \underset{\text{Spin operator}}{\phi_\mu(\boldsymbol{s})}\Big)$$



Ex.



How? Two approaches. Find *the* model $\mathcal{M}$:
- with the highest evidence, $P(\hat{\boldsymbol{s}}\,|\,\mathcal{M})$ $\longrightarrow$ Bayesian approch;
- which achieves the shortest description of the data, $L(\hat{\boldsymbol{s}}\,|\,\mathcal{M})$ $\longrightarrow$ MDL principle.

For large datasets,
the two criteria are identical [1,2]:
(assuming Jeffrey's prior)

$$\log P(\hat{\boldsymbol{s}}\,|\,\mathcal{M}) = \overbrace{\underbrace{\log P(\hat{\boldsymbol{s}}\,|\,\mathcal{M},\boldsymbol{g}^*)}_{\text{Maximum Log-Likelihood},\ \propto N}}^{\text{Quality of the Fit}} - \overbrace{\left[\underbrace{\frac{K}{2}}_{\text{Number of Parameters } K}\log\left(\frac{N}{2\pi}\right) + \underbrace{c_{\mathcal{M}}}_{\text{Due to Geometry}}\right]}^{\text{Model Complexity}} + O\left(\frac{1}{N}\right) = -L(\hat{\boldsymbol{s}}\,|\,\mathcal{M})$$

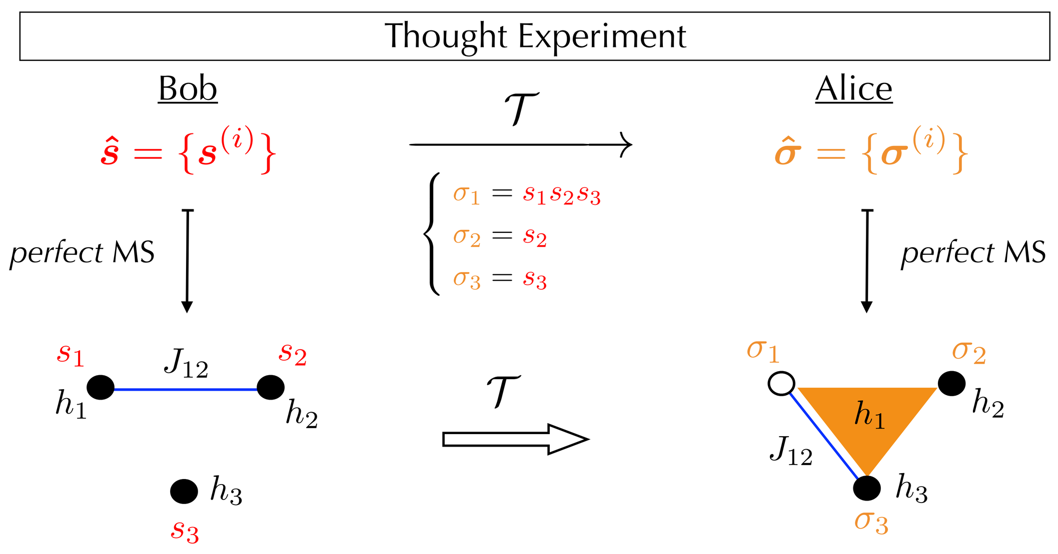
Due to

Problem? difficult task due to the huge number of models and the difficulty to evaluate $P(\hat{\boldsymbol{s}}\,|\,\mathcal{M})$ or $L(\hat{\boldsymbol{s}}\,|\,\mathcal{M})$, even with the expansion.

Common simplification, reduce the space of models to models with only field and pairwise interactions, which are simpler to infer and to interprete.

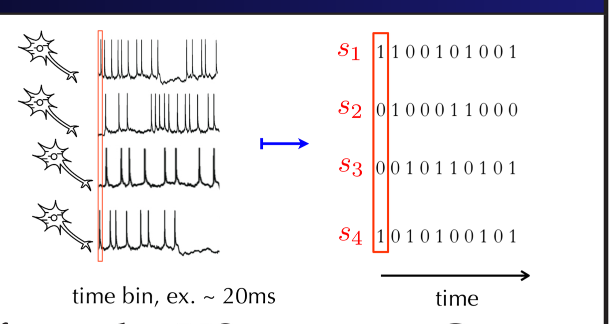
Is it a good idea? Are these models really *simpler*?

Thought experiment:
$\Longrightarrow$ Pairwise interactions don't play any special role.
$\Longrightarrow$ Model selection shouldn't be basis dependent...

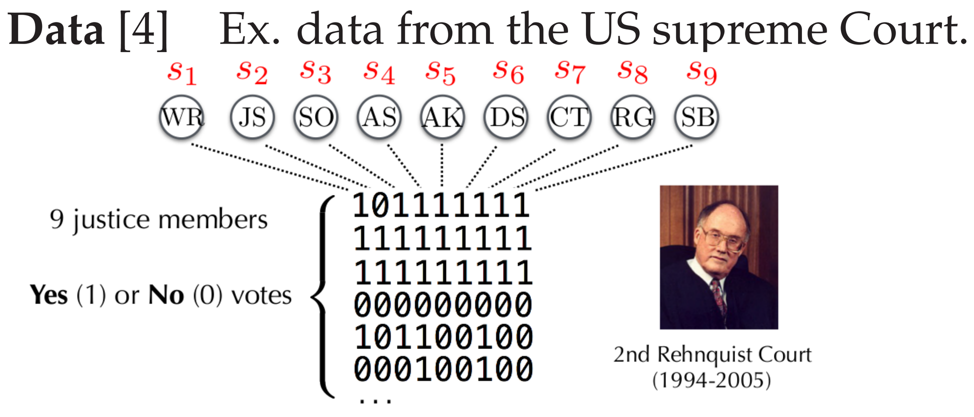


Examples of Binary Datasets

Neuronal Data [?,3,5]

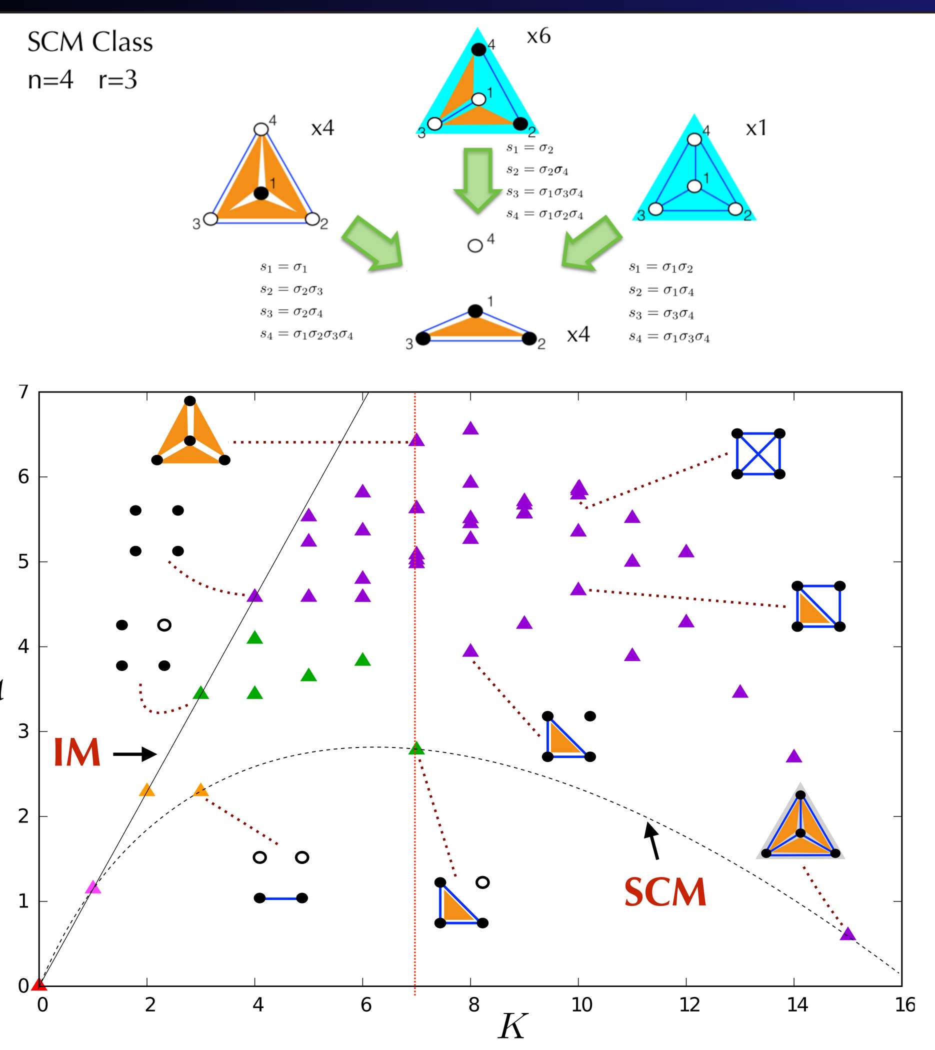


Voting Data [4] Ex. data from the US supreme Court.



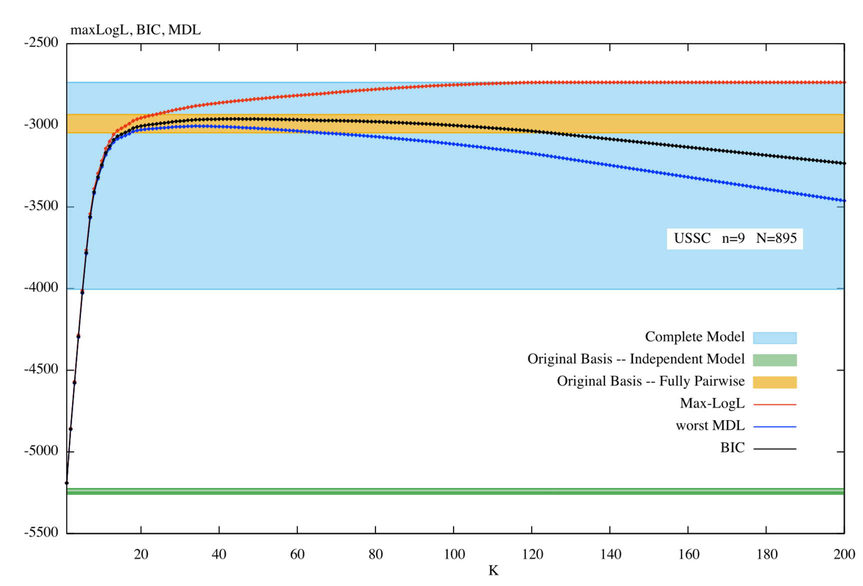
Financial Data [6] Time series, at each time step each stock i takes the value:
- $s_i = +1$, if the stock has gone up (profit);
- $s_i = -1$, if the stock has gone down (loss).

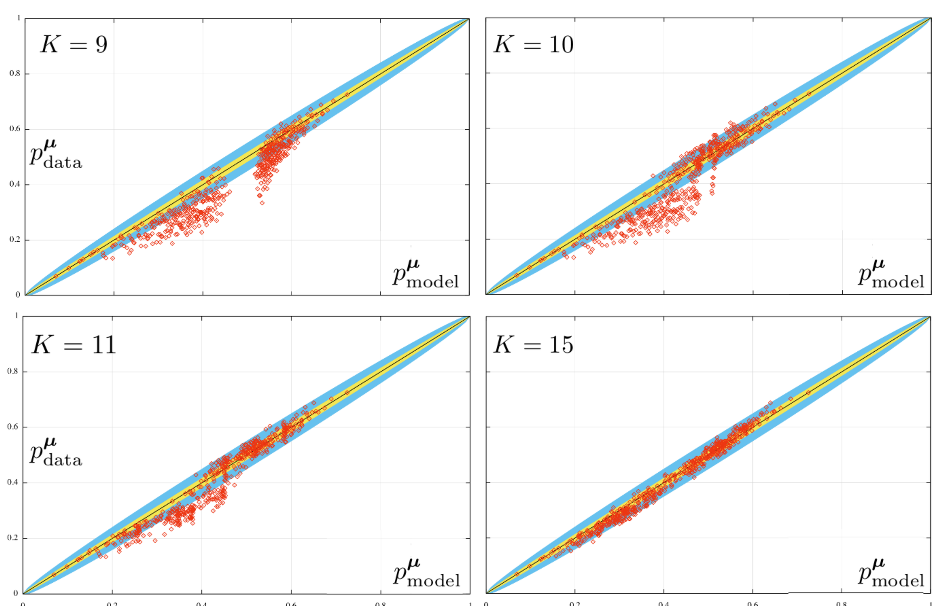
Equivalence classes



Approach 1: Shortest Path in the Space of Models

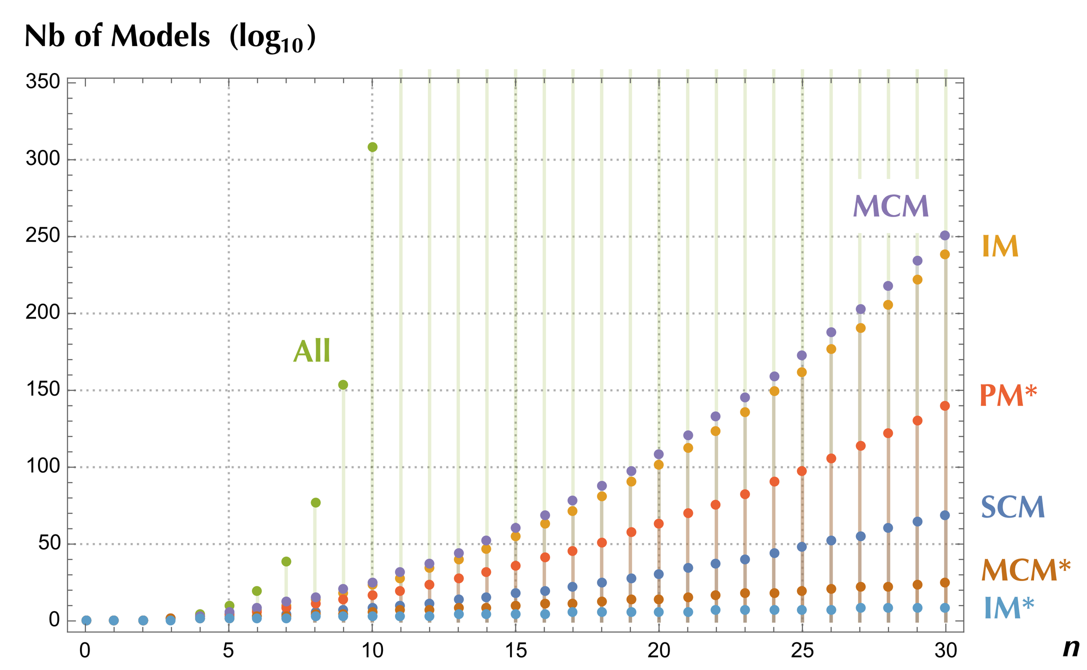
How? Start $\mathcal{M}_0 = \emptyset$. At each step K, add to the model the operator ϕ_K whose statistics is the less captured by the current model: $\mathcal{M}_{K+1} = \mathcal{M}_K \cup \{\phi_K\}$.





##1:K	2:maxlogL	3: BIC	5:Op
##			
1	-5188.26	-5190.74	000100100
2	-4855.53	-4860.49	000001010
3	-4568.7	-4576.14	000000011
4	-4283.84	-4293.76	100010000
5	-4010.52	-4022.92	100000100
6	-3764.56	-3779.43	101000000
7	-3542.3	-3559.66	010000010
8	-3388.53	-3408.37	001000001
9	-3294.46	-3316.77	000000100
10	-3217.8	-3242.59	010000000
11	-3141.58	-3168.86	101111001
12	-3098.4	-3128.15	000011000
13	-3056.17	-3088.4	010001011
14	-3031.83	-3066.55	101110000
15	-3018.23	-3055.43	001111000
16	-3004.28	-3043.95	001011010
17	-2989.99	-3032.14	100000000
18	-2972.42	-3017.05	000001000
19	-2962.6	-3009.71	111101111
20	-2954.65	-3004.24	001101100

How many... ?



Approach 2: Selection Within and Between Classes

Independent Models (IM) $\longrightarrow \mathcal{C}_{ind}(n,r)$

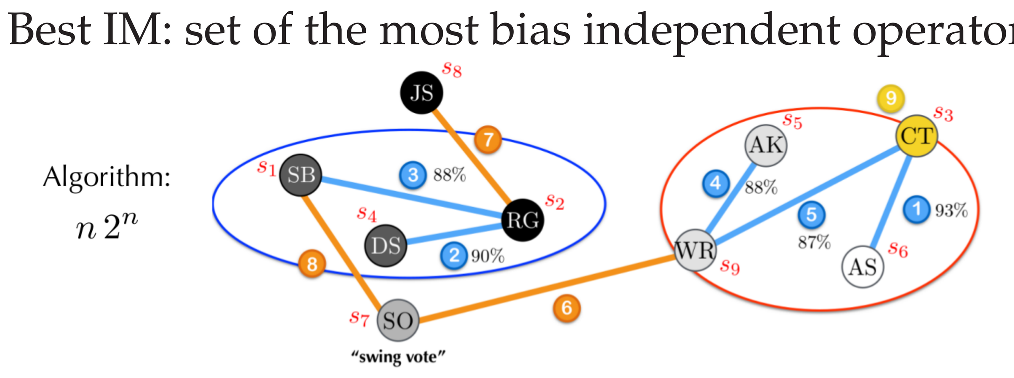
$(n+1)$ classes $\qquad \sim \dfrac{2^{nr}}{r!}$ models/class $\qquad K = r$

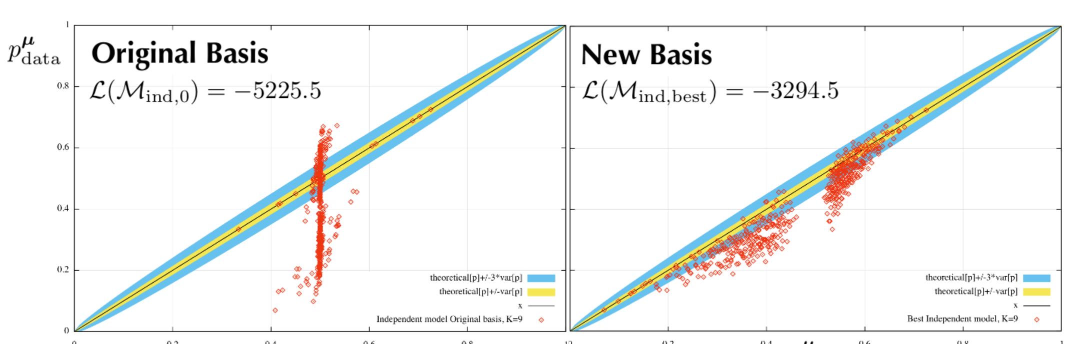
$|\mathcal{C}_{ind}(9,r)| = \{1, 511, 1.10^5, 2.10^7, 3.10^9, 3.10^{11}, 2.10^{13}, 1.10^{15}, 7.10^{16}, 2.10^{18}\}$

$$\mathcal{L}^{\text{ind}}_{max} = -\sum_{i=1}^{r} S[p_i] - (n-r)N\log(2)$$

$S[p] = -[p\log p + (1-p)\log(1-p)] \qquad p_i = \mathbb{P}_{\text{data}}[\phi_i(\boldsymbol{s}) = -1]$

$\Longrightarrow$ Best IM: set of the most bias independent operators





— Decreasing order of relevance
— Possibility, reduce the dimension of the model
— These Interactions must be in the model

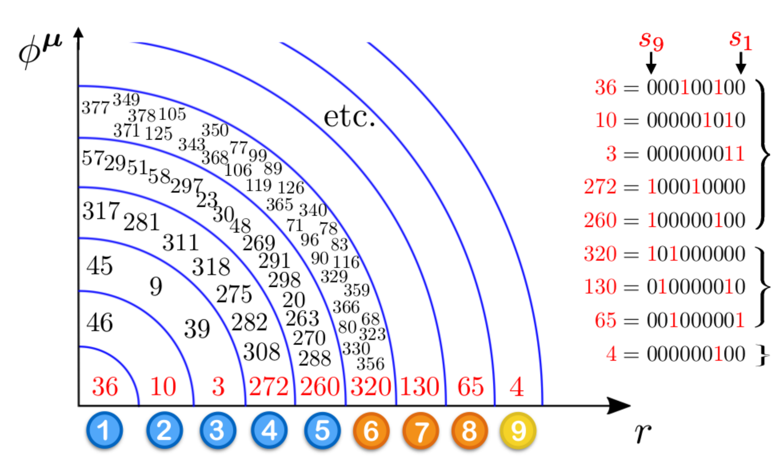
Sub-Complete Models (SCM) $\longrightarrow \mathcal{C}_{sc}(n,r)$

$(n+1)$ classes $\qquad \sim 2^{r(n-r)}$ models/class $\qquad K = 2^r - 1$

$|\mathcal{C}_{sc}(9,r)| = \{1, 511, 4.10^4, 8.10^5, 3.10^6, 8.10^5, 4.10^4, 511, 1\}$

$$\mathcal{L}^{sc}_{max} = \sum_{\sigma_{|r}} k_{\sigma_{|r}} \log\left(\frac{k_{\sigma_{|r}}}{N}\right) - (n-r)N\log(2)$$

$\longrightarrow$ Monte Carlo simulation in each class:



$\longrightarrow$ Evidence exact and easy to compute (no need to infer):

$$P(\hat{\boldsymbol{s}}\,|\,\mathcal{M}_{sc}) = \frac{1}{2^{N(n-r)}}\,\frac{\Gamma(2^{r-1})}{\pi^{2^{r-1}}\,\Gamma(N+2^{r-1})}\prod_{\sigma_r}\Gamma\left(k_{\sigma_r}+\frac{1}{2}\right)$$

Minimally Complex Models (MCM) $\longrightarrow \mathcal{C}_{mc}(n,\{r_i\})$

$\sum_{r=0}^{n}\text{part}(r)$ classes $\qquad K = \sum_{\{r_i\}}(2^{r_i}-1)$

$|\mathcal{C}_{mc}(9,\{1,2,3\})| = 1.5\,10^{13}$

$n = 9 \qquad r_1 = 3 \qquad r_2 = 1 \qquad r_3 = 2 \qquad r = 6 \qquad K = 11$

$\longrightarrow$ Contains all IM and SCM.

All rank ordered: 26 443 models