Detecting Degeneracies for Simulation-based Inference

Chaipat (Tata) Tirapongprasert

Rev. 04.26.2026

The General Idea

- **Posterior:** $\mathcal{P}(\boldsymbol{\theta} \mid \mathbf{d}) \sim \mathcal{P}(\mathbf{d} \mid \boldsymbol{\theta})\mathcal{P}(\boldsymbol{\theta})$

- **Degeneracies:** Most posterior mass lies within an ε-neighborhood of a lower-dimensional manifold

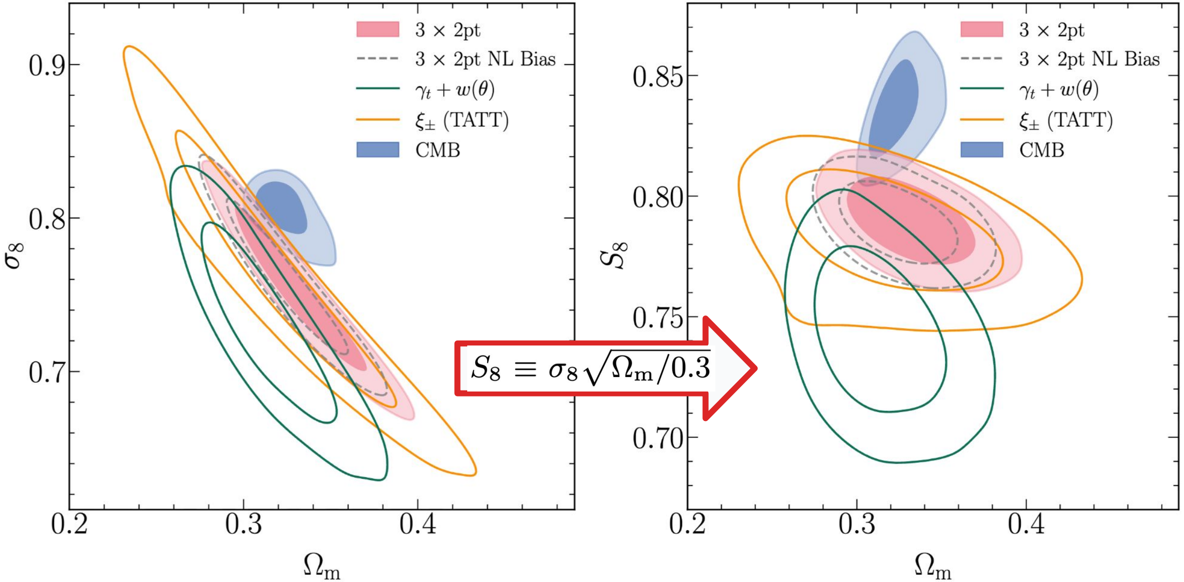
$$\mathcal{M}_c = \left\{ \boldsymbol{\theta} \in \mathbb{R}^M : F(\theta_{j_1}, \ldots, \theta_{j_k}) = c \right\}$$

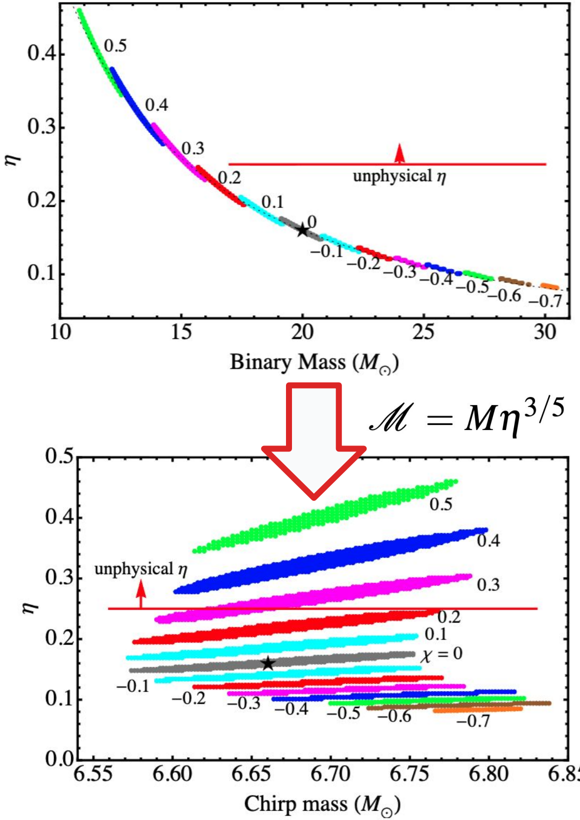
under some metric on parameter space.

- **Problem**: Param inference is very difficult along the degenerate direction.

Canonical examples

- CMB degeneracy
- weak lensing
- gravitational waves

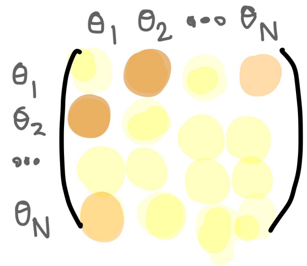




Learn Degeneracies in 3 steps

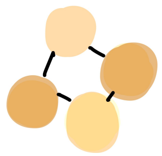
1) **Mutual Information (MI) Screening**

narrow down the parameter combinations



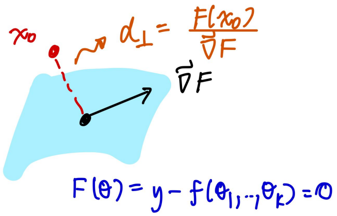
2) **Symbolic Fitting**

assume separable and fit one component at a time



3) **Diagnostics**

shows top symbolic equations and diagnostic visualizations

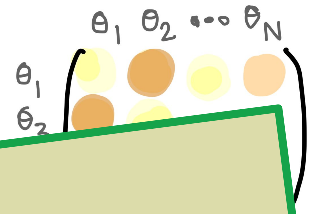


Rev. 04.26.2026

Learn Degeneracies in 3 steps

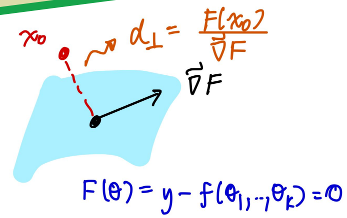
1) **Mutual Information (MI) Screening**

narrow down the parameter combinations



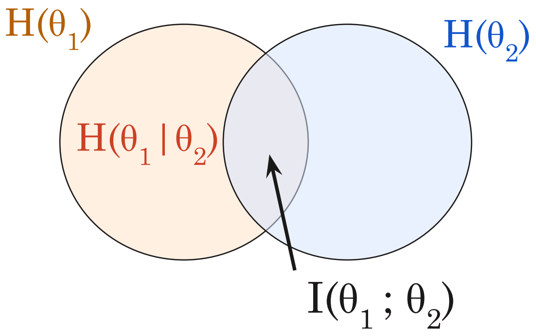

```
detector = DegenDetector(samples, param_names)
result = detector.search_couplings(coupling_depth=2,niterations=200)

runner = DiagnosticsRunner(result_dir)
runner.run()
```

op symbolic equations and diagnostic
visualizations



Rev. 04.26.2026

1) MI Screening



- MI measures the reduction in uncertainty about θ_1 given knowledge of θ_2.

Intuition:

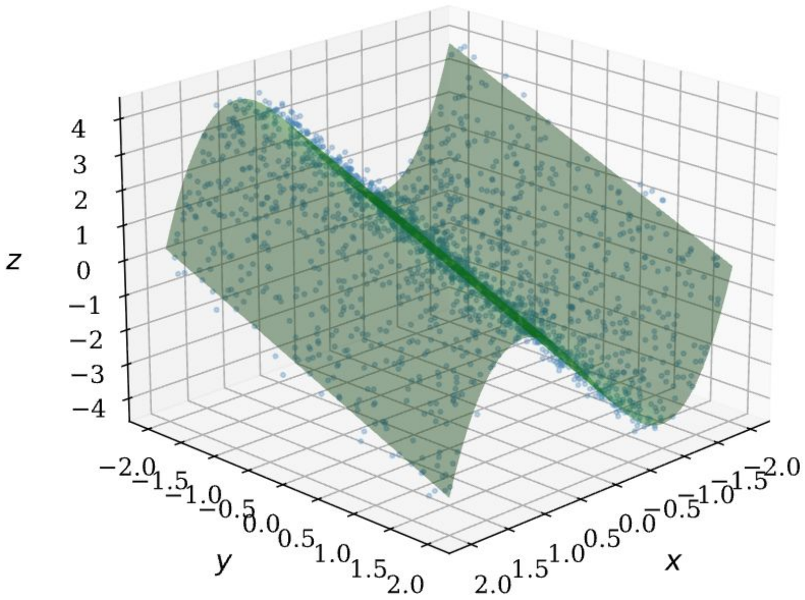
- Entropy as a proxy for "surprise": $$H(\theta) = -\sum_{\theta \in \Theta} \mathcal{P}(\theta) \log(\mathcal{P}(\theta))$$

- Conditional entropy: $$H(\theta_1 \mid \theta_2) = -\sum_{\theta_2} \mathcal{P}(\theta_2) \sum_{\theta_1} \mathcal{P}(\theta_1 \mid \theta_2) \log(\mathcal{P}(\theta_1 \mid \theta_2))$$

- Mutual information: $$I(\theta_1; \theta_2) = H(\theta_1) - H(\theta_1 \mid \theta_2)$$

MI matrix effectively isolates the degeneracy.

$$x^3 - 3x + y + z = 0$$

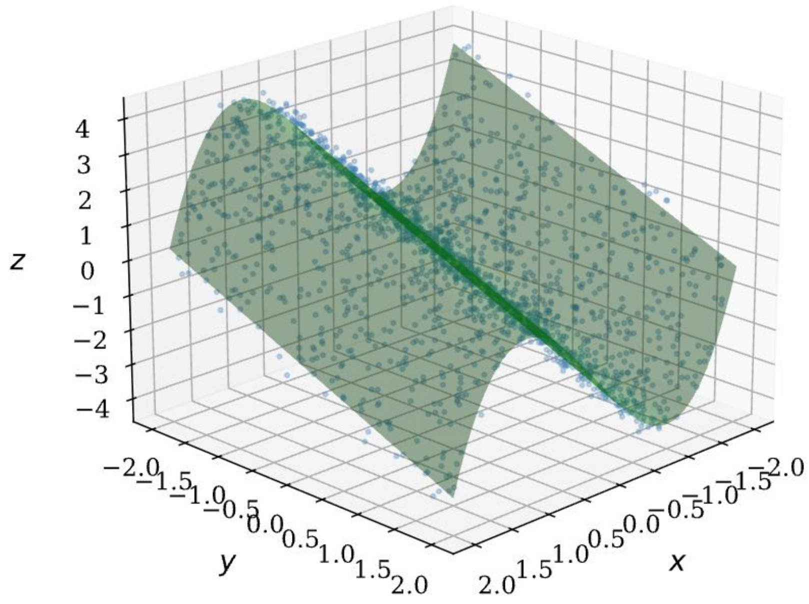


	x	y	z	a	b	c	d
x	0.00	0.00	**0.48**	0.02	0.00	0.00	0.00
y	0.00	0.00	**0.47**	0.02	0.01	0.00	0.00
z	**0.48**	**0.47**	0.00	0.00	0.00	0.00	0.00
a	0.02	0.02	0.00	0.00	0.04	0.00	0.02
b	0.00	0.01	0.00	0.04	0.00	0.01	0.01
c	0.00	0.00	0.00	0.00	0.01	0.00	0.00
d	0.00	0.00	0.00	0.02	0.01	0.00	0.00

Pairwise MI matrix (nats)

Rev. 04.26.2026

$$x^3 - 3x + y + z = 0$$



	x	y	z	a	b	c	d
x	0.00	0.00	**0.48**	0.02	0.00	0.00	0.00
y	0.00	0.00	**0.47**	0.02	0.01	0.00	0.00
z	**0.48**	**0.47**	0.00	0.00	0.00	0.00	0.00
a	0.02	0.02	0.00	0.00	0.04	0.00	0.02
b	0.00	0.01	0.00	0.04	0.00	0.01	0.01
c	0.00	0.00	0.00	0.00	0.01	0.00	0.00
d	0.00	0.00	0.00	0.02	0.01	0.00	0.00

Pairwise MI matrix (nats)

2) Symbolic Fit

- **But…it's still a k-dimensional fitting problem.**
  - The search space of potential formulas grows combinatorially as $\mathcal{O}(k^k)$ with input features.

<u>Solution:</u> We *split* into k independent one-dimensional symbolic regression problems by modeling the degeneracy as the level set of a *separable* function.

$$\underbrace{g_1(\theta_{j_1}) + g_2(\theta_{j_2}) + \cdots + g_k(\theta_{j_k})}_{\text{univariate functions to be fitted}} = c$$

Q: How do we fit these component functions?

(1) normalize using the z-score $\tilde{\theta}_\ell = \dfrac{\theta_\ell - \mu_\ell}{\sigma_\ell}$ and initialize $\left[\begin{array}{l} g_\ell(\tilde{\theta}_\ell) = \tilde{\theta}_\ell \\ c = \text{mean}(\sum_\ell \tilde{\theta}_\ell) \end{array}\right.$

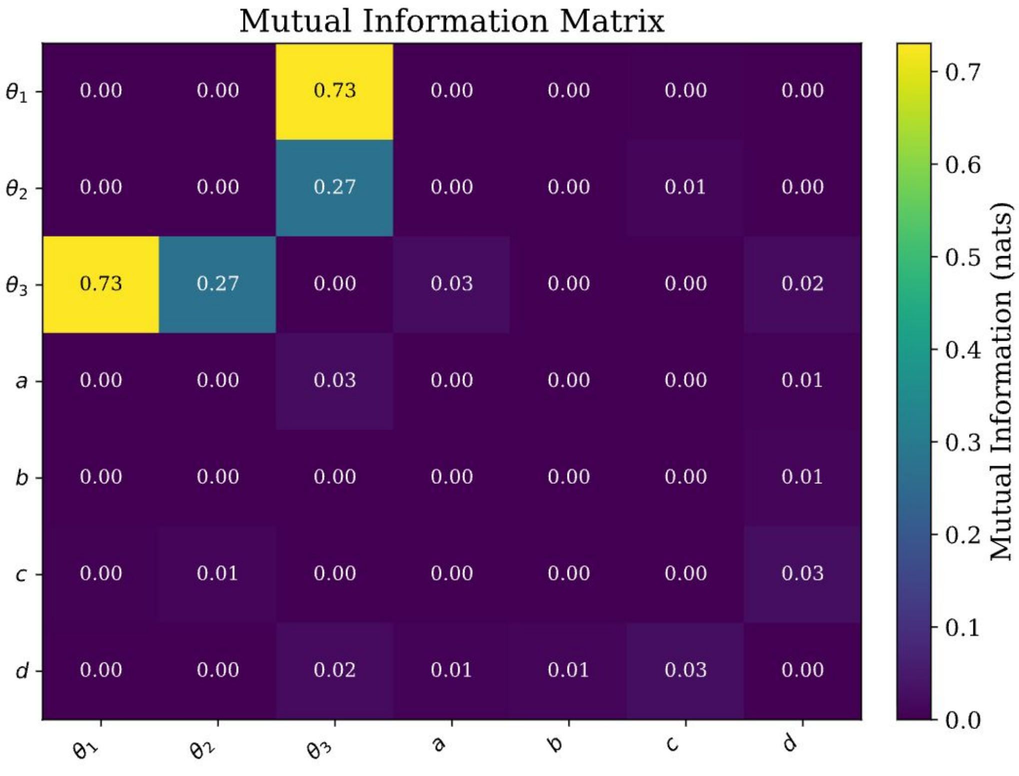
(2) Repeat for l = 1, …, k until convergence

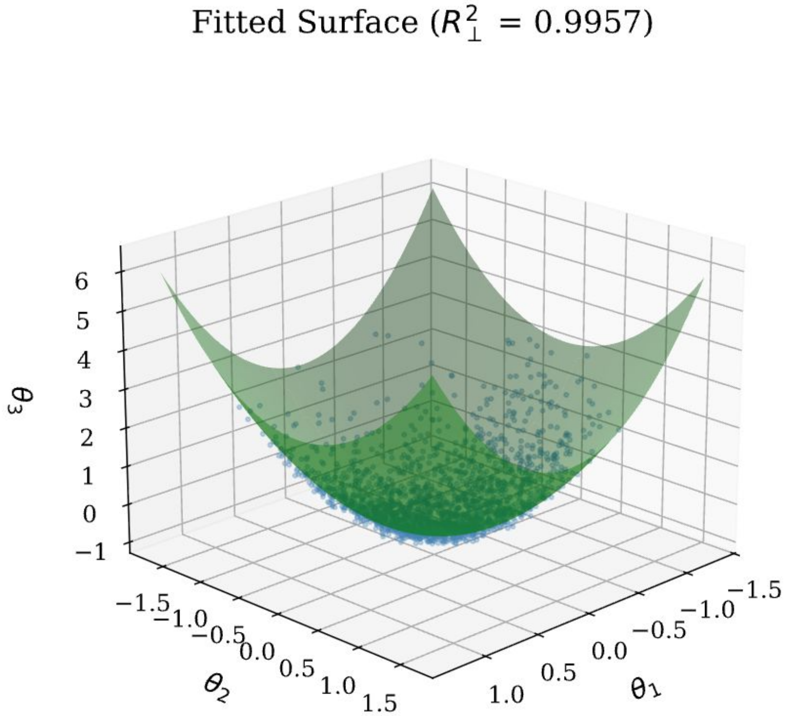
- ➢ Do symbolic regression for the pairs $\left(g_\ell(\tilde{\theta}_\ell),\ \boxed{c - \sum_{m \neq \ell} g_m(\tilde{\theta}_m)}\right)$

regression target!

- ➢ **<u>Goal:</u>** want to minimize MSE for all the N pairs of the dataset

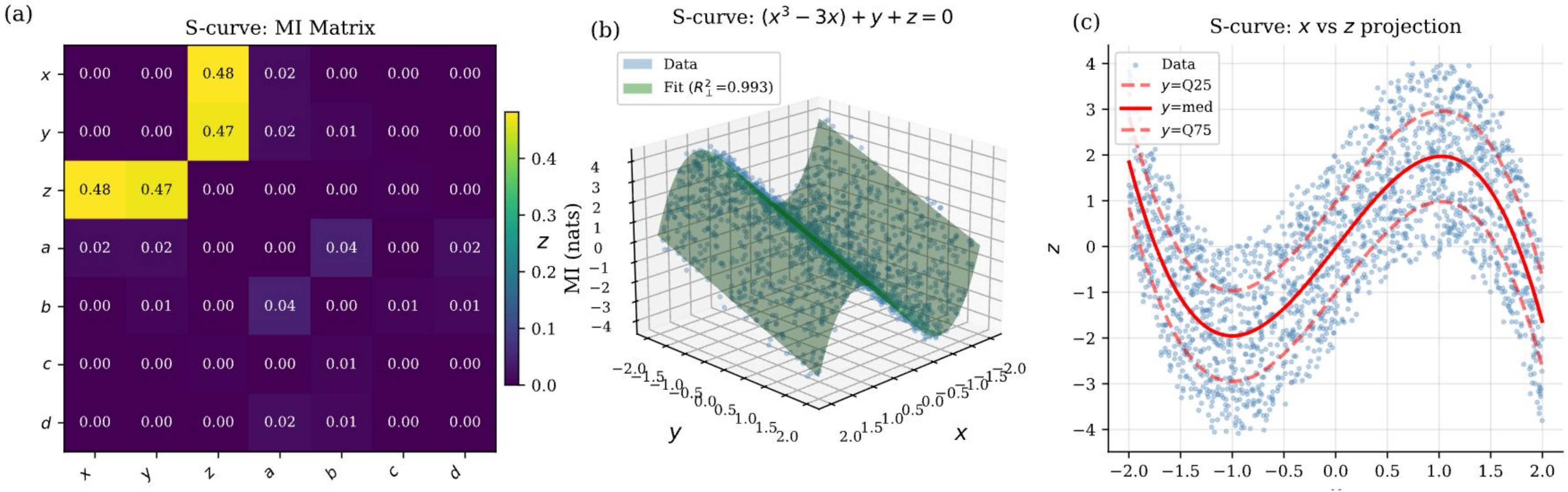
Benchmark Results





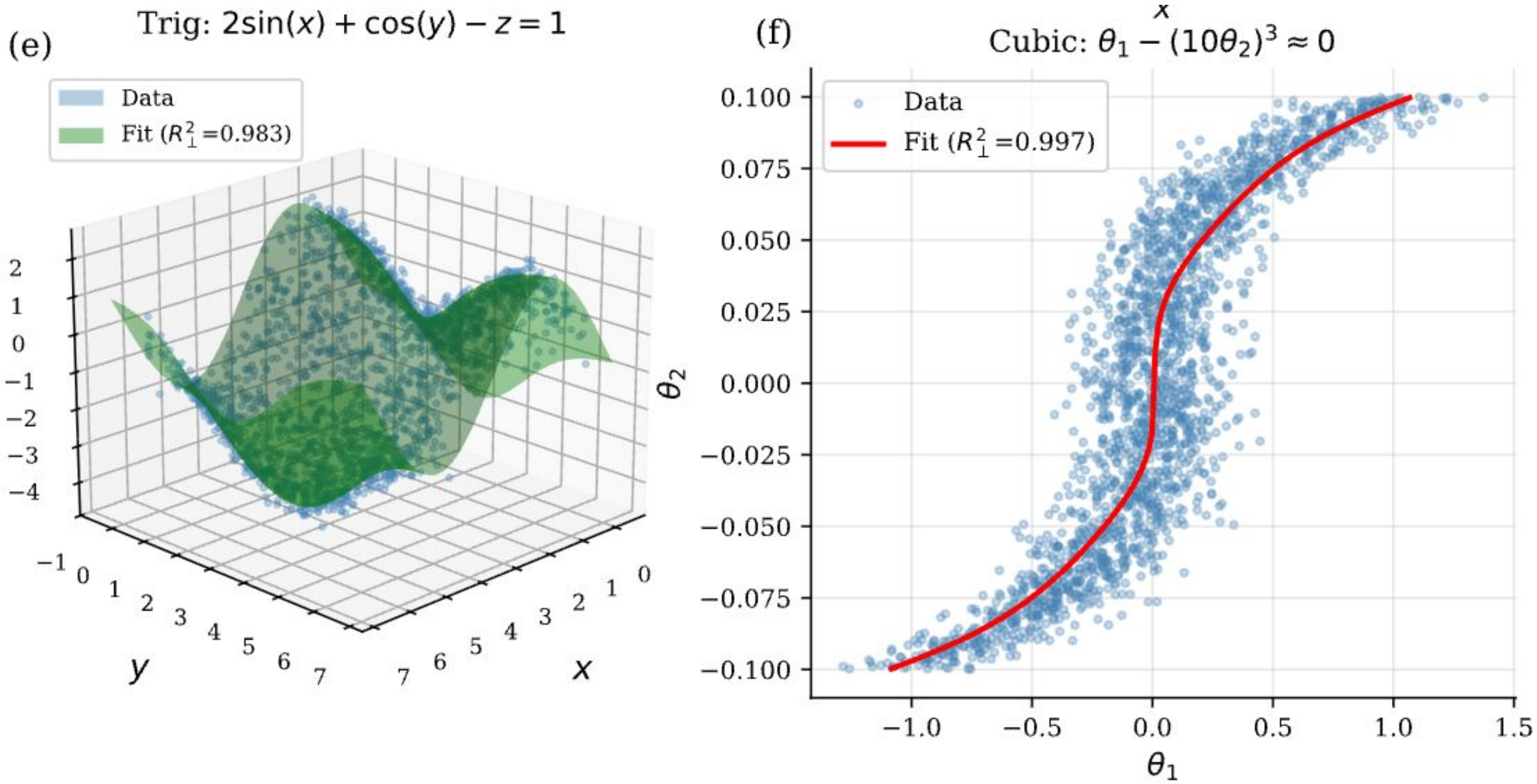
Rev. 04.26.2026

Benchmark Results



Rev. 04.26.2026

Benchmark Results

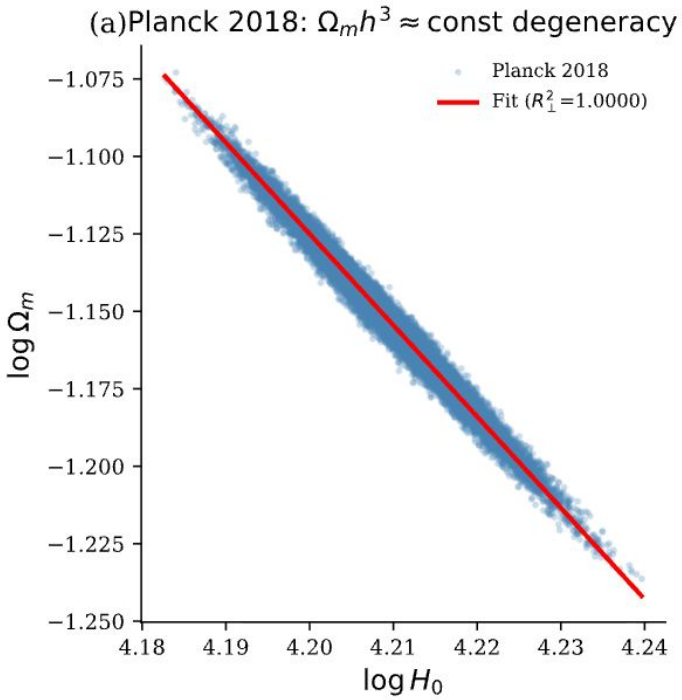


Rev. 04.26.2026

Science Experiment

- **DegenLogMode:** Although the power law degeneracy is non-separable in the original space, it can be recovered as the level set in log space.

$\Omega_m\, h^3 \approx \text{const} \;\longrightarrow\; \mathbf{\log \Omega_m + 3 \log h \approx const}$



$$\boxed{123.97 \log H_0 + 42.07 \log \Omega_m = \text{const}}$$

Rev. 04.26.2026

Conclusion

- We developed **DegenDetector**, a new degeneracy detection framework that combines mutual information ranking and alternating symbolic regression to find coupled parameters and express that degeneracy as an interpretable equation.
- On synthetic benchmarks and Planck 2018 posteriors, we recover the true functional form of degeneracies with orthogonal $R^2 > 0.98$ across all cases without domain-specific input.
- The main limitation of our method is the assumption of separability, which could be extended in future works to non-separable functional forms using multivariate symbolic regression or reparameterizations.

Questions?

Hit me up at t.chaipat@columbia.edu