Comparing Methods for Estimating Marginal Likelihood in Symbolic Regression

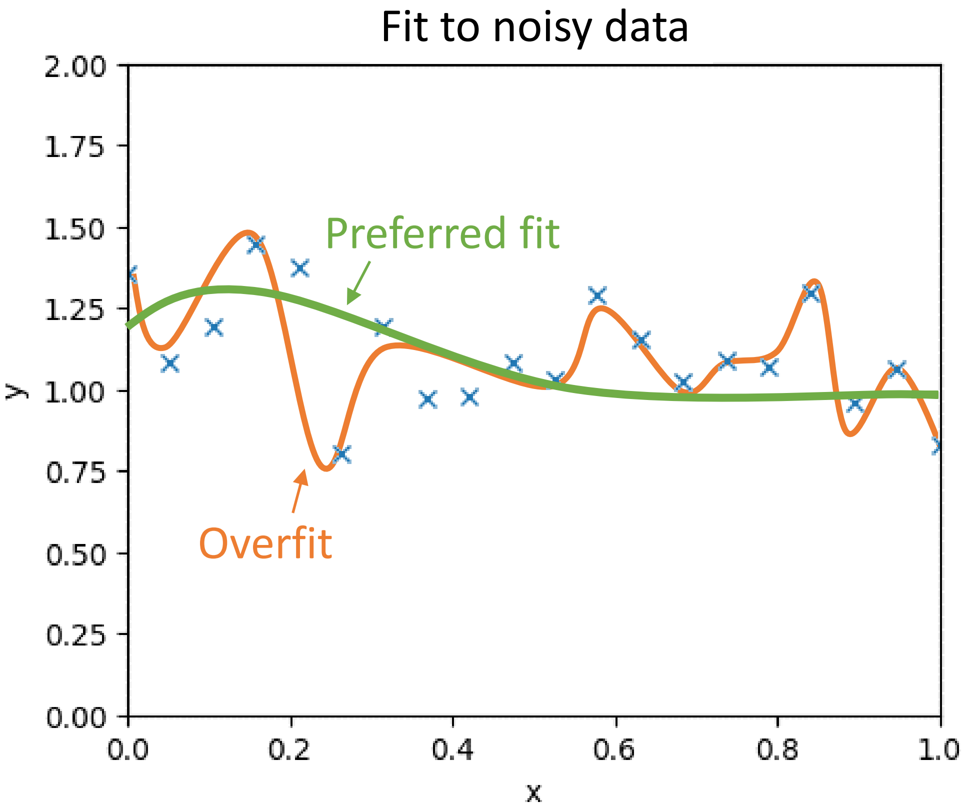
Patrick Leser, Geoffrey Bomarito
NASA Langley Research Center

Gabriel Kronberger
University of Applied Sciences Upper Austria

Fabrício Olivetti De França
Universidade Federal do ABC

Motivation

- Symbolic Regression (SR) has a proclivity for overfitting when data is scarce and noisy
- Bayesian model selection has been shown to help reduce bloat and improve generalizability in Genetic Programming based SR (GPSR)[1]
  - Quantifies uncertainty due to scarce, noisy data
  - Is based on model evidence, which implicitly penalizes parametric complexity
- How can model evidence be estimated in practice?
  - Laplace approximation
  - Sequential Monte Carlo (SMC) sampling



1. Bomarito, Leser, Strauss, Garbrecht, and Hochhalter. 2022. Bayesian model selection for reducing bloat and overfitting in genetic programming for symbolic regression. GECCO '22

What is model evidence?

- Anatomy of Bayes' theorem

Likelihood Prior

Posterior

$$\pi(\boldsymbol{\theta}|\boldsymbol{d}, f) = \frac{\pi(\boldsymbol{d}|\boldsymbol{\theta}, f)\pi(\boldsymbol{\theta}|f)}{\pi(\boldsymbol{d}|f)}$$

Model evidence

(a.k.a. marginal likelihood)

$\boldsymbol{d}$: Data
f: Model
$\boldsymbol{\theta}$: Model parameters

- Evidence: probability of data given a model

$$\pi(\boldsymbol{d}|f) = \int_{\mathbb{R}^p} \pi(\boldsymbol{d}|\boldsymbol{\theta}, f)\pi(\boldsymbol{\theta}|f)\, d\boldsymbol{\theta}$$

<u>Problem</u>:
Improper prior for SR!

- Bayes Factor: relative probability of two models given the data

$$B = \frac{\pi(f_0|\boldsymbol{d})}{\pi(f_1|\boldsymbol{d})} = \frac{\pi(\boldsymbol{d}|f_0)\cancel{\pi(f_0)}}{\pi(\boldsymbol{d}|f_1)\cancel{\pi(f_1)}}$$

Fractional Bayes Factor

A normalized version of the Bayes Factor that works with improper priors

Bayes Factor: $B = \dfrac{c_1 \int_{\mathbb{R}^{p_1}} \pi(\boldsymbol{d}|\boldsymbol{\theta}_1, f_1) h(\boldsymbol{\theta}_1|f_1) d\boldsymbol{\theta}}{c_2 \int_{\mathbb{R}^{p_2}} \pi(\boldsymbol{d}|\boldsymbol{\theta}_2, f_2) h(\boldsymbol{\theta}_2|f_2) d\boldsymbol{\theta}}$

Fractional Bayes Factor[1]: $B_\gamma = \dfrac{q_0(\gamma)}{q_1(\gamma)}$

Normalized Marginal Likelihood (NML):

$$q_j(\gamma) = \frac{\int_{\mathbb{R}^p} \pi(\boldsymbol{d}|\boldsymbol{\theta}, f_j)\pi(\boldsymbol{\theta}|f_j)\, d\boldsymbol{\theta}}{\int_{\mathbb{R}^p} \pi(\boldsymbol{d}|\boldsymbol{\theta}, f_j)^{\gamma}\pi(\boldsymbol{\theta}|f_j)\, d\boldsymbol{\theta}} = \frac{\text{Evidence}}{\text{Evidence w/ } \gamma \in [0,1]}$$

(Simulates using a portion of data for normalization)

- For uniform improper priors $\pi(\theta|f) \propto 1$, unspecified normalizing constants appear in the standard Bayes Factor.
- The fractional Bayes Factor results in these constants canceling, enabling model comparison.

1. O'Hagan, Anthony. "Fractional Bayes factors for model comparison." *Journal of the Royal Statistical Society: Series B (Methodological)* 57.1 (1995): 99-118.

Estimating NML – Laplace Approximation

- Approximates posterior with a multivariate Gaussian distribution

$$\pi(\boldsymbol{\theta}|\boldsymbol{d}, f) \approx \hat{\pi}(\boldsymbol{\theta}|\boldsymbol{d}, f) = \mathcal{N}(\boldsymbol{\theta}^*, \Sigma)$$

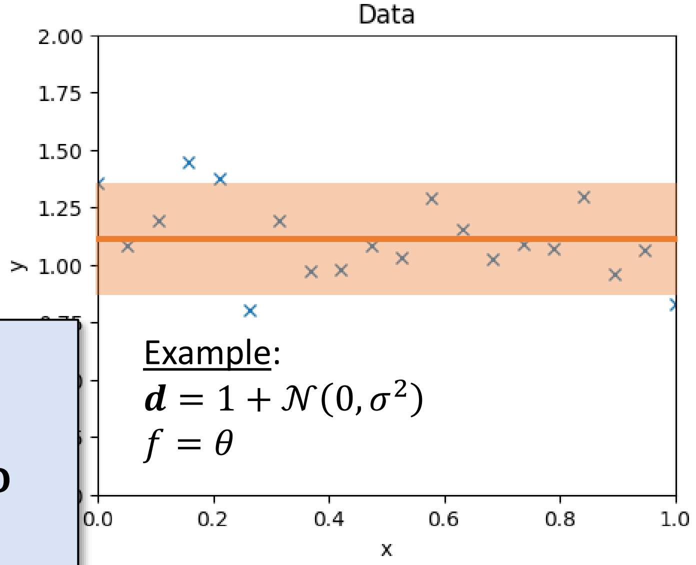
- The mean vector is the *maximum a posteriori* (MAP) estimate, which is equivalent to maximum likelihood in our case:

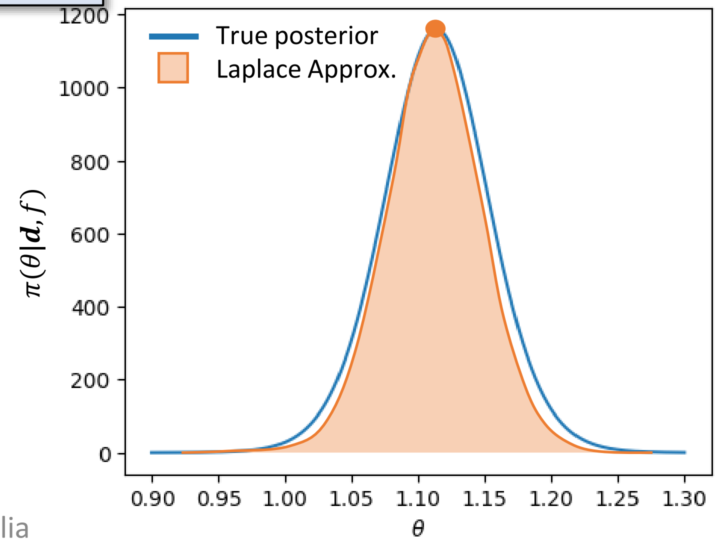
$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, \pi(\boldsymbol{d}|\boldsymbol{\theta}, f$$

- The normalized marginal log likeli

$$\boxed{\log \hat{q}_j(\gamma) = (1-\gamma) \log \pi \quad 2}$$

Note:
- **This is fast**
- **Sensitive to methodology used to solve the optimization problem**





Driving Questions

1. What if posterior is not represented well with a Gaussian? (e.g., multimodal, nonlinear posteriors)
2. Do these cases occur in GPSR?

Estimating NML – Sequential Monte Carlo (SMC)

- A method for drawing samples from posterior (like Markov chain Monte Carlo)
- SMC targets a series of distributions transitioning from the **prior** (easy to sample from) to the posterior (unknown):

Transition governed by likelihood annealing

$\pi_t \propto$ … $\ldots, T$

Note:
- **This is NOT fast**
- **Less sensitive to initialization**
- **No assumptions regarding posterior distribution required**

- The normalizing constant fo…

$\mathcal{Z}$ … $\boldsymbol{\theta}$

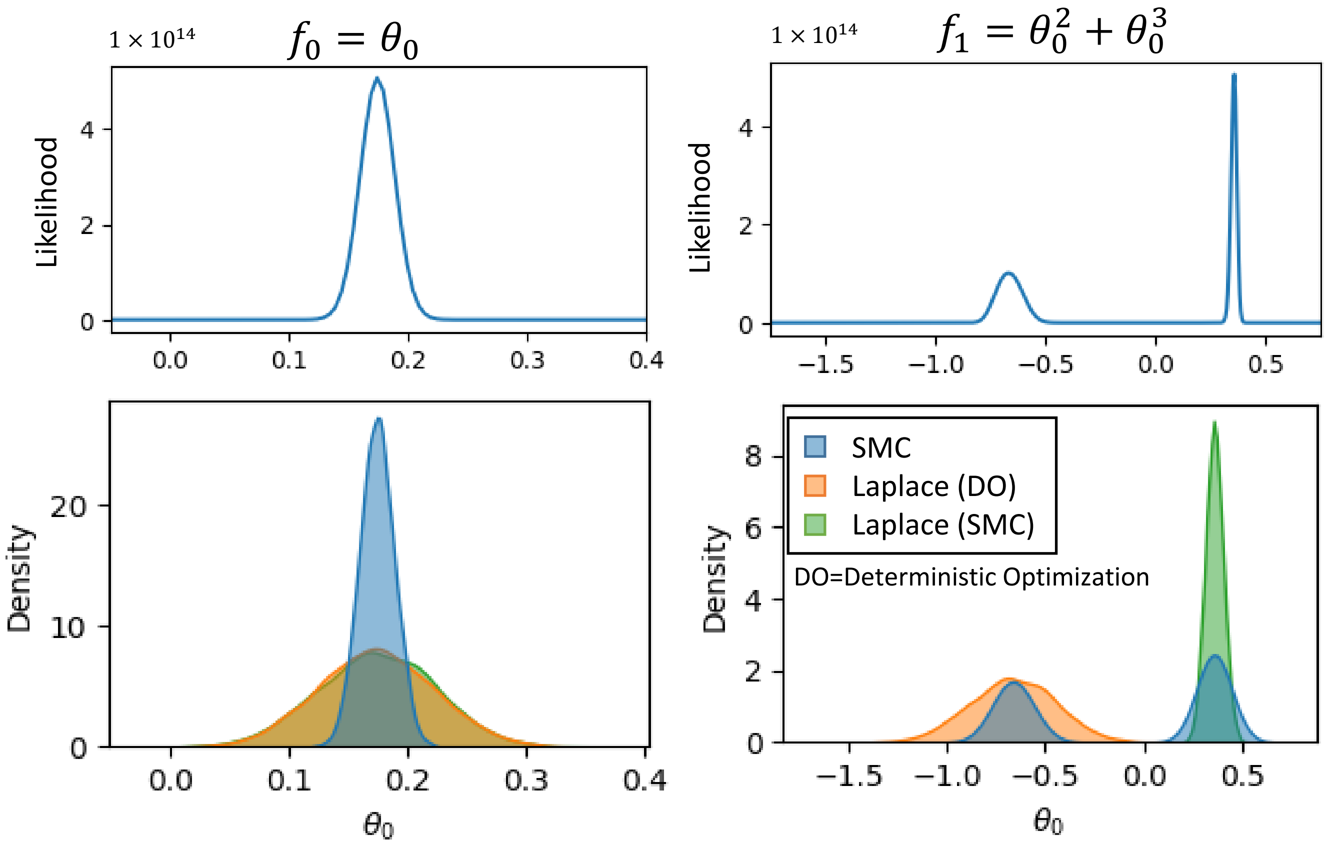
- Noting the similarity to NML formula, set:

$$\boldsymbol{\phi} = \{\phi_t\}_{t=1}^{T} = \{0, \ldots, \gamma, \ldots, 1\}$$

- Therefore, the NMLL is a natural byproduct of a single SMC run:

$$\boxed{\log \bar{q}_j(\gamma) = \mathcal{Z}_j^{\phi_T} - \mathcal{Z}_j^{\gamma}}$$

Numerical Experiments: Multimodal Toy Problem

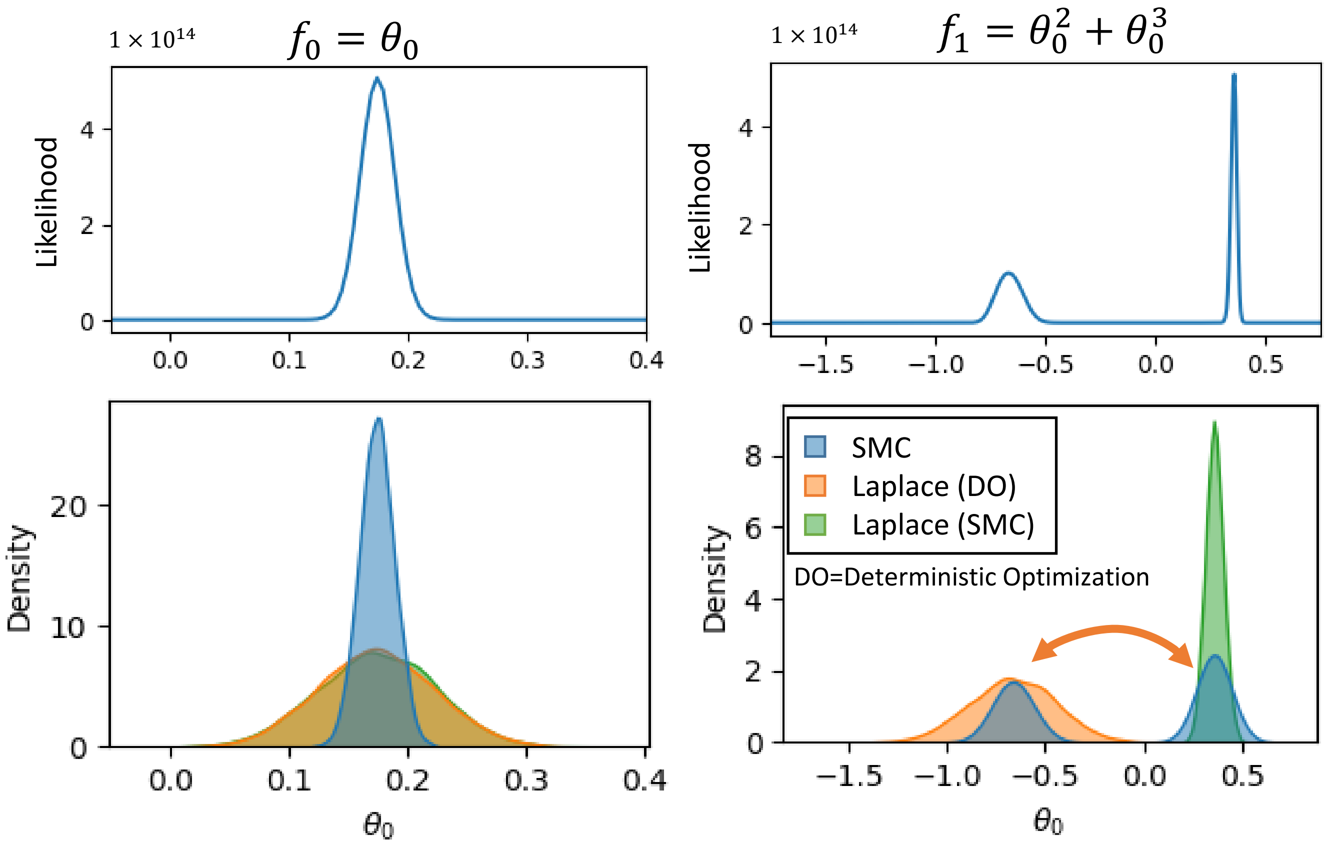


NMLL estimation using various methods (+/- standard deviation)

Target	MC	SMC	Laplace (DO)	Laplace (SMC)
$\log q_0(\gamma)$	28.66 ± 0.02	28.71 ± 0.07	28.82 ± 0.00	28.82 ± 0.0001
$\log q_1(\gamma)$	27.88 ± 0.04	28.11 ± 0.04	28.04 ± 0.76	28.82 ± 0.0002

Reference

Numerical Experiments: Multimodal Toy Problem



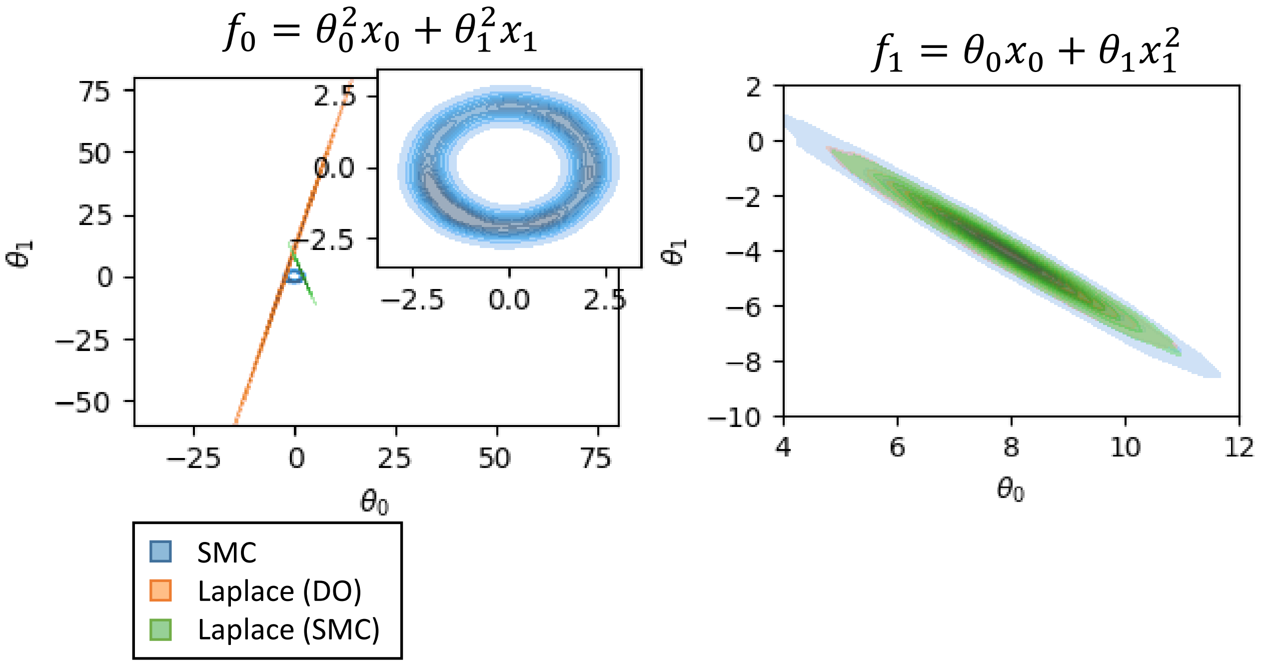
- Laplace produces very consistent results *if correct mode is found*
- The correct mode is found more often when using a global optimizer like SMC
- SMC is consistently accurate but has larger estimator variance than Laplace
- Laplace based on deterministic optimization (DO) can be biased if a local optima is used instead of MAP

NMLL estimation using various methods (+/- standard deviation)

Target	MC	SMC	Laplace (DO)	Laplace (SMC)
$\log q_0(\gamma)$	28.66 ± 0.02	28.71 ± 0.07	28.82 ± 0.00	28.82 ± 0.0001
$\log q_1(\gamma)$	27.88 ± 0.04	28.11 ± 0.04	28.04 ± 0.76	28.82 ± 0.0002

Reference

Numerical Experiments: Nonlinear Toy Problem

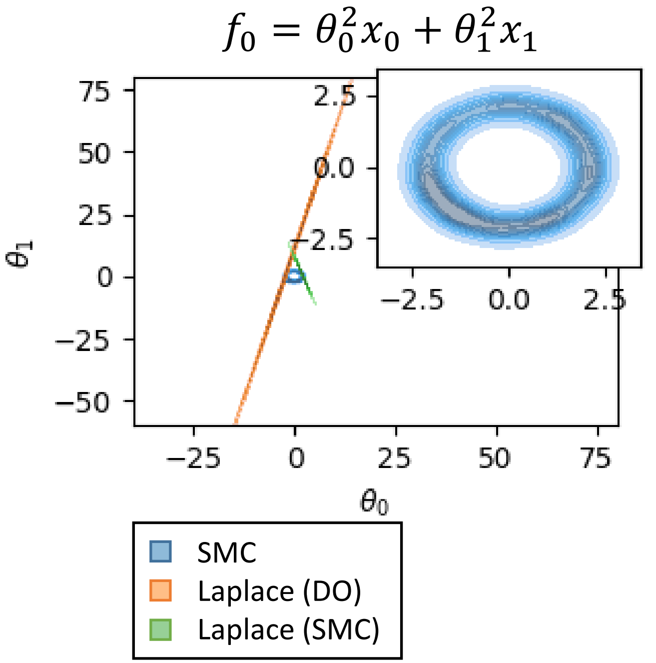


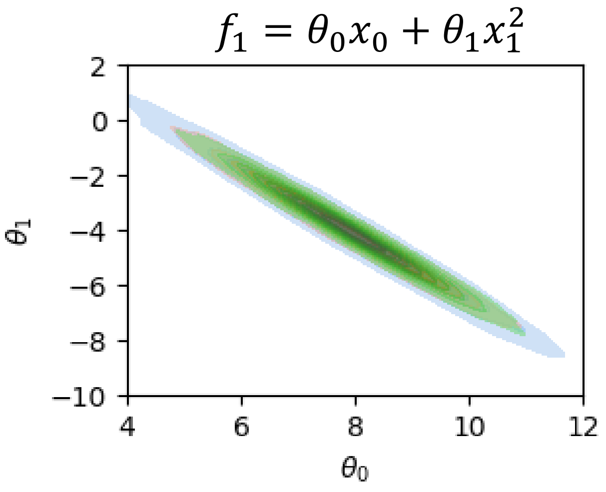
NMLL estimation using various methods (+/- standard deviation)

Target	MC	SMC	Laplace (DO)	Laplace (SMC)
$\log q_0(\gamma)$	-22.38 ± 0.02	-22.13 ± 0.12	-23.48 ± 6.11	-22.62 ± 0.0001
$\log q_1(\gamma)$	-21.69 ± 0.04	-21.57 ± 0.18	-20.57 ± 0.00	-20.57 ± 0.002

Reference

Numerical Experiments: Nonlinear Toy Problem





- Laplace approximates the ring distribution with a tangent gaussian
- Laplace is very consistent but biased
- SMC is again more accurate albeit with larger variance than Laplace
- Laplace is less accurate even in a case that is unimodal and approximately Gaussian

NMLL estimation using various methods (+/- standard deviation)

Target	MC	SMC	Laplace (DO)	Laplace (SMC)
$\log q_0(\gamma)$	-22.38 ± 0.02	-22.13 ± 0.12	-23.48 ± 6.11	-22.62 ± 0.0001
$\log q_1(\gamma)$	-21.69 ± 0.04	-21.57 ± 0.18	-20.57 ± 0.00	-20.57 ± 0.002

Reference

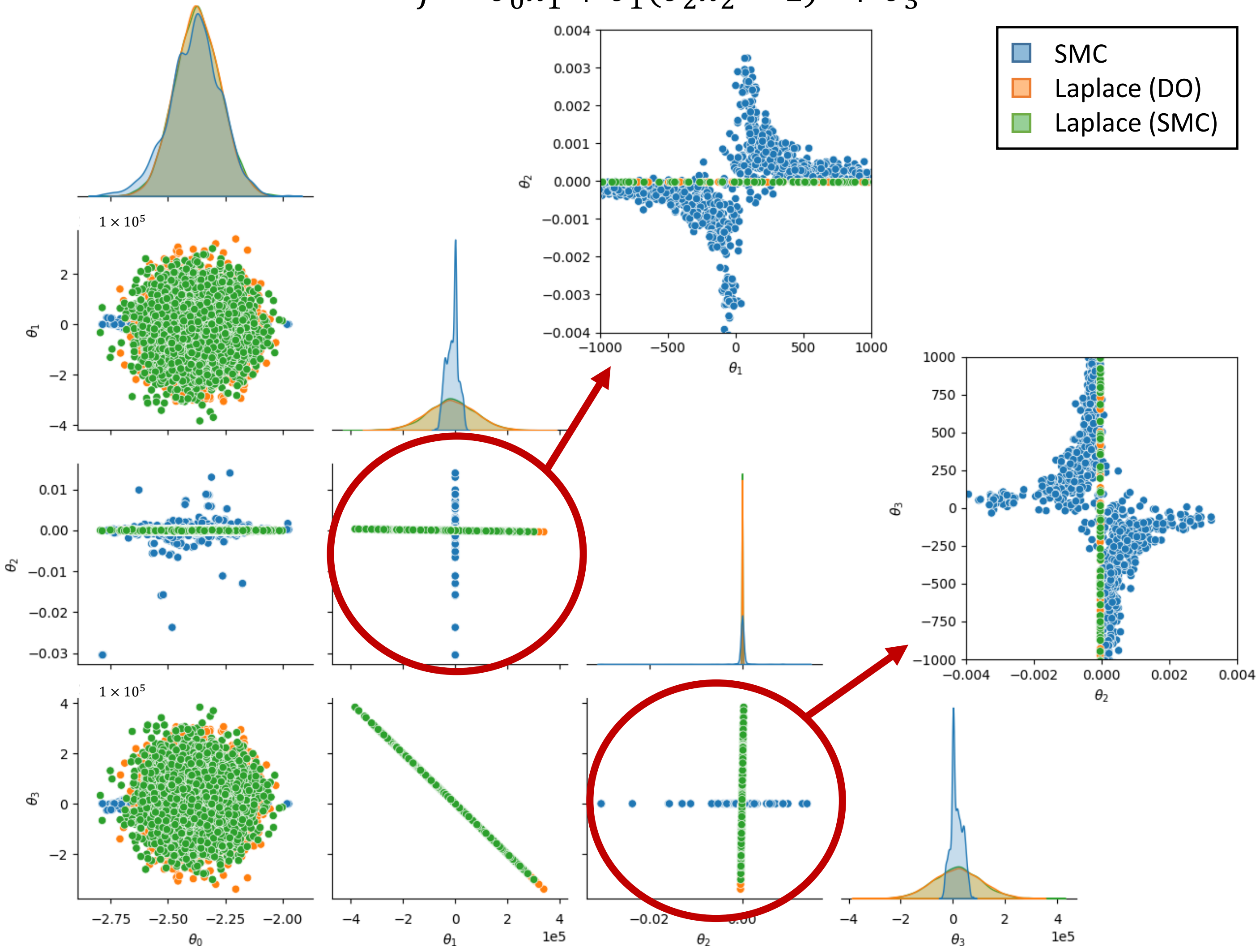
Real World Examples

- Expressions produced by GPSR (Operon) applied to the Feynman benchmarks in SRBENCH[1,2,3]
- Results presented:
    1. Single example highlighting existence of non-Gaussian, nonlinear posteriors
    2. Summary of NMLL predictions across entire set of 43 expressions

1. La Cava, et al. "Contemporary symbolic regression methods and their relative performance." arXiv:2107.14351 (2021)
2. Orzechowski, et al. "Where are we now? A large benchmark study of recent symbolic regression methods." Proc. of GECCO (2018)
3. Udrescu and Tegmark. "AI Feynman: A physics-inspired method for symbolic regression." Science Advances, (2020)

Real World Example: Single Expression

$$f = \theta_0 x_1 + \theta_1(\theta_2 x_2 - 1)^2 + \theta_3$$



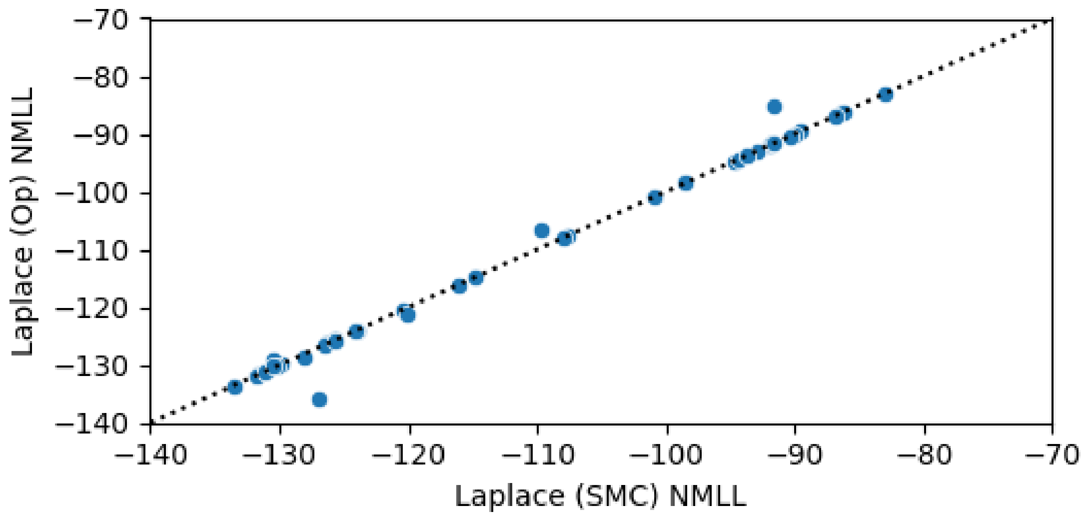
- True posterior is both multimodal and nonlinear
- No Monte Carlo reference available due to computational expense
- New initialization approach introduced
- SMC produces most consistent results for this case
- Laplace approximation exhibits:
  - Higher variance than SMC
  - Sensitivity to initial optimization

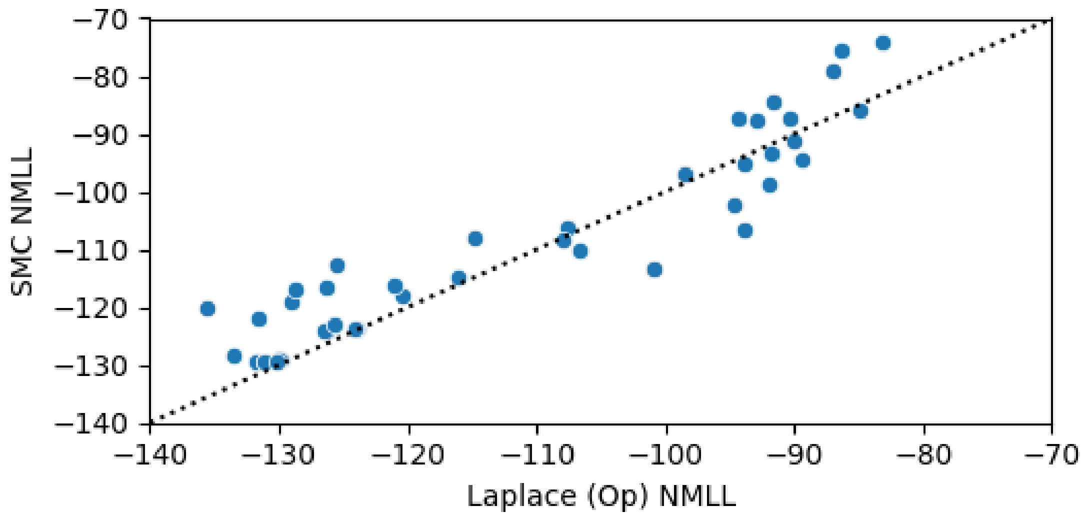
NMLL estimation using various methods (+/- standard deviation)

Target	SMC	Laplace (DO)	Laplace (Op)	Laplace (SMC)
$\log q(\gamma)$	-120.4 ± 4.1	-116.7 ± 9.2	-118.7 ± 8.9	-119.5 ± 8.6

Real World Example: Aggregated Results

- Operon’s maximum likelihood optimization is very close to the MAP produced by SMC
  - This is likely due to local optima lost to evolution (cost hidden by GPSR)
- NMLL produced by SMC and Laplace are correlated but different

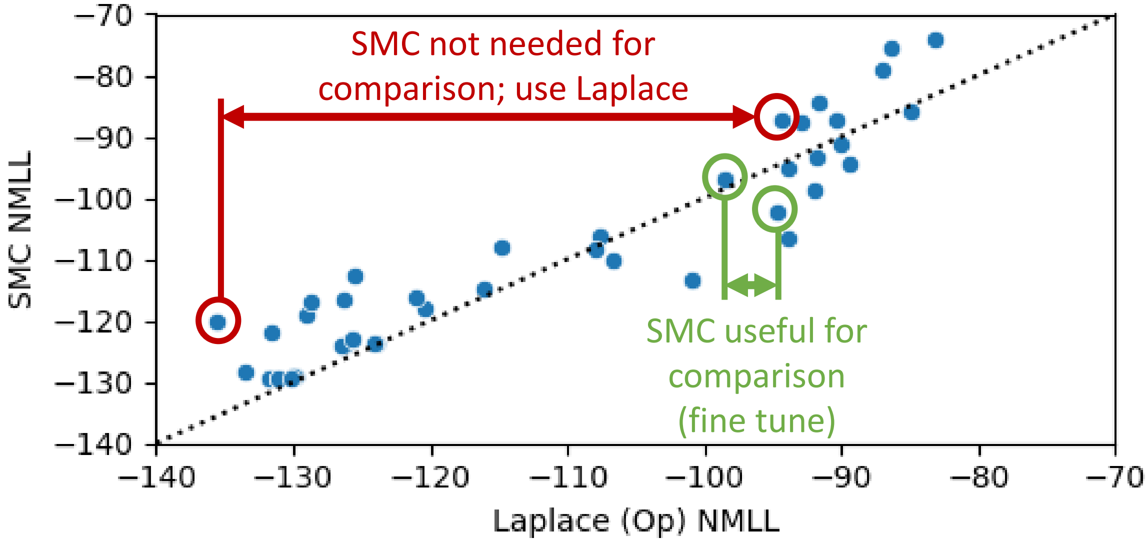




Conclusions

- Sequential Monte Carlo (SMC):
  - More accurate, robust NMLL estimates
  - More computationally expensive
  - Tunable precision/cost tradeoff
- Laplace Approximation:
  - Fast and consistent NMLL estimates
  - Potential for biased estimates in non-Gaussian cases (e.g., nonlinear, multimodal posteriors)
  - Dependent upon parameter optimization
- The types of expressions that exacerbate differences are present in GPSR

➢ A filtering-based approach could be useful in practice
  - Spend the extra time on SMC only when needed



Contact: patrick.e.leser@nasa.gov