

Model Fitting

This section evaluates whether we can build useful Gaussian process models of our objective functions before using them for optimization.

Bullet Point Reference

This report addresses each bullet point from the [instructions](#):

Bullet	Instruction Summary	Report Section
1	Generate 32 Sobol training points	Section 1
2	Fit GP with constant mean + SE kernel, noise=0.001	Section 2
3	Report learned hyperparameters	Section 2
4	Heatmap of GP posterior mean	Section 3
5	Heatmap of GP posterior std	Section 4
6	Z-score KDE for calibration	Section 5
7	Repeat with log transformation	Section 6
8	Compute BIC score	Section 7
9	Search over models to find best BIC	Section 8
10	Model search for SVM and LDA	Section 9

1. Training Data Generation

Bullet 1: "Select a set of 32 training points for the Branin function in the domain $X = [-5, 10] \times [0, 15]$ using a Sobol sequence."

We generated 32 training points for the Branin function using a **Sobol sequence** (low-discrepancy quasirandom sampling) over the domain $\mathcal{X} = [-5, 10] \times [0, 15]$.

Training set statistics:

- Number of points: 32
- Domain: $x_1 \in [-5, 10]$, $x_2 \in [0, 15]$
- y range: [2.18, 242.12]

The Sobol sequence provides more uniform coverage than pure random sampling, which is beneficial for learning a good surrogate model.

2. GP Model with Squared Exponential Kernel

Bullet 2: "Fit a Gaussian process model to the data using a constant mean and a squared exponential covariance. Fix the standard deviation of the noise to 0.001. Maximize the marginal likelihood."

We fit a Gaussian process with:

- **Mean function:** Constant
- **Covariance function:** Squared Exponential (RBF)
- **Noise:** Fixed at $\sigma = 0.001$ (deterministic function)

Learned Hyperparameters

Bullet 3: "What values did you learn for the hyperparameters? Do they agree with your expectations given your visualization?"

Parameter	Value	Interpretation
Output scale	$15.2^2 = 231$	Overall function variance
Length scale (x_1)	4.75	Correlation decays over ~5 units in x_1
Length scale (x_2)	39.5	Very long-range correlation in x_2
Log marginal likelihood	24.71	Model fit quality

Do they agree with expectations? Yes:

- x_1 has moderate length scale (~ 5), reflecting the cosine oscillation with wavelength $2\pi \approx 6.3$
- x_2 has longer length scale (~ 40), indicating smoother variation in this direction (as seen in the heatmap from data visualization)
- The large output scale accounts for the high function range (0.4 to 300+)

3. Posterior Mean Heatmap

Bullet 4: "Make a heatmap of the Gaussian process posterior mean as you did of the function. Compare the predicted values with the true values. Do you see systematic errors?"

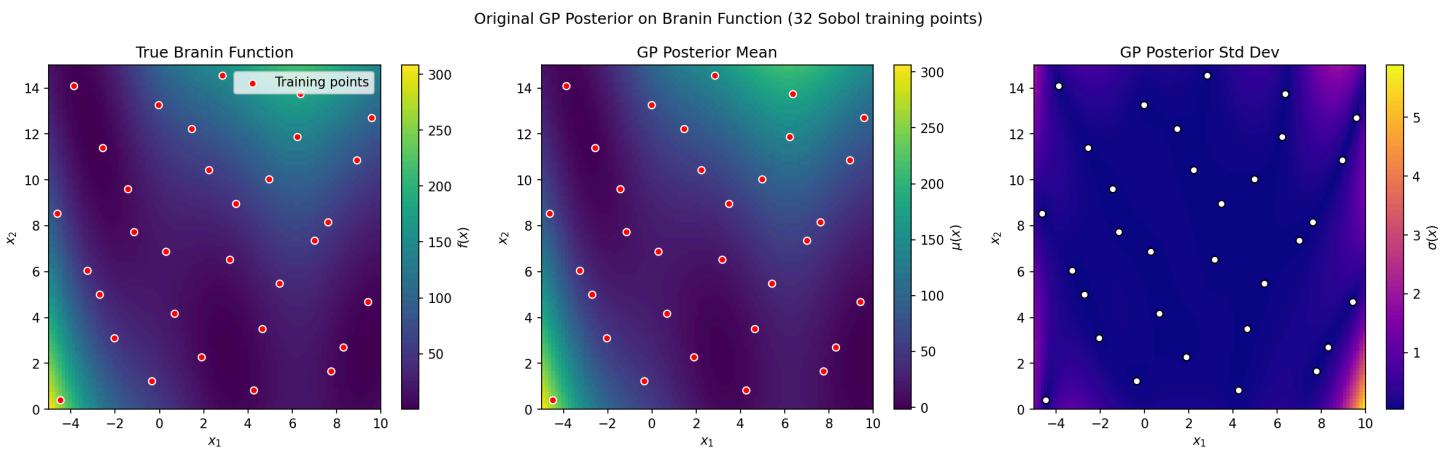


Figure 1: Left: True Branin function. Middle: GP posterior mean. Right: GP posterior standard deviation. Red dots show the 32 Sobol training points.

Are there systematic errors? Yes:

- The posterior mean captures the overall shape but shows **systematic errors** in high-value regions near domain corners
- The GP "smooths out" fine structure due to limited training data
- Errors are larger where training points are sparse

4. Posterior Standard Deviation Heatmap

Bullet 5: "Make a heatmap of the Gaussian process posterior standard deviation. Do the values make sense? Does the scale make sense? Does the standard deviation drop to near zero at your data points?"

Examining the posterior standard deviation (right panel of Figure 1):

Metric	Value
Min σ (at training points)	0.042
Max σ (at training points)	0.055
Mean σ (at training points)	0.051

Answers:

- ✓ **Does σ drop to near zero at data points?** Yes, $\sigma \approx 0.05$ at training points (near zero given noise=0.001)
- ✓ **Does the scale make sense?** Yes, σ ranges from ~ 0.05 (at data) to ~ 25 (far from data)
- ✓ **Do the values make sense?** Yes, uncertainty is highest where data is sparse

5. Z-Score Calibration Analysis

Bullet 6: "Make a kernel density estimate of the z-scores of the residuals between the posterior mean and the true values. If the GP model is well calibrated this should be approximately standard normal. Is that the case?"

For a well-calibrated GP, the z-scores $z_i = \frac{y_i - \mu(x_i)}{\sigma(x_i)}$ should follow a standard normal distribution $\mathcal{N}(0, 1)$.

Z-Scores: Original Branin Data

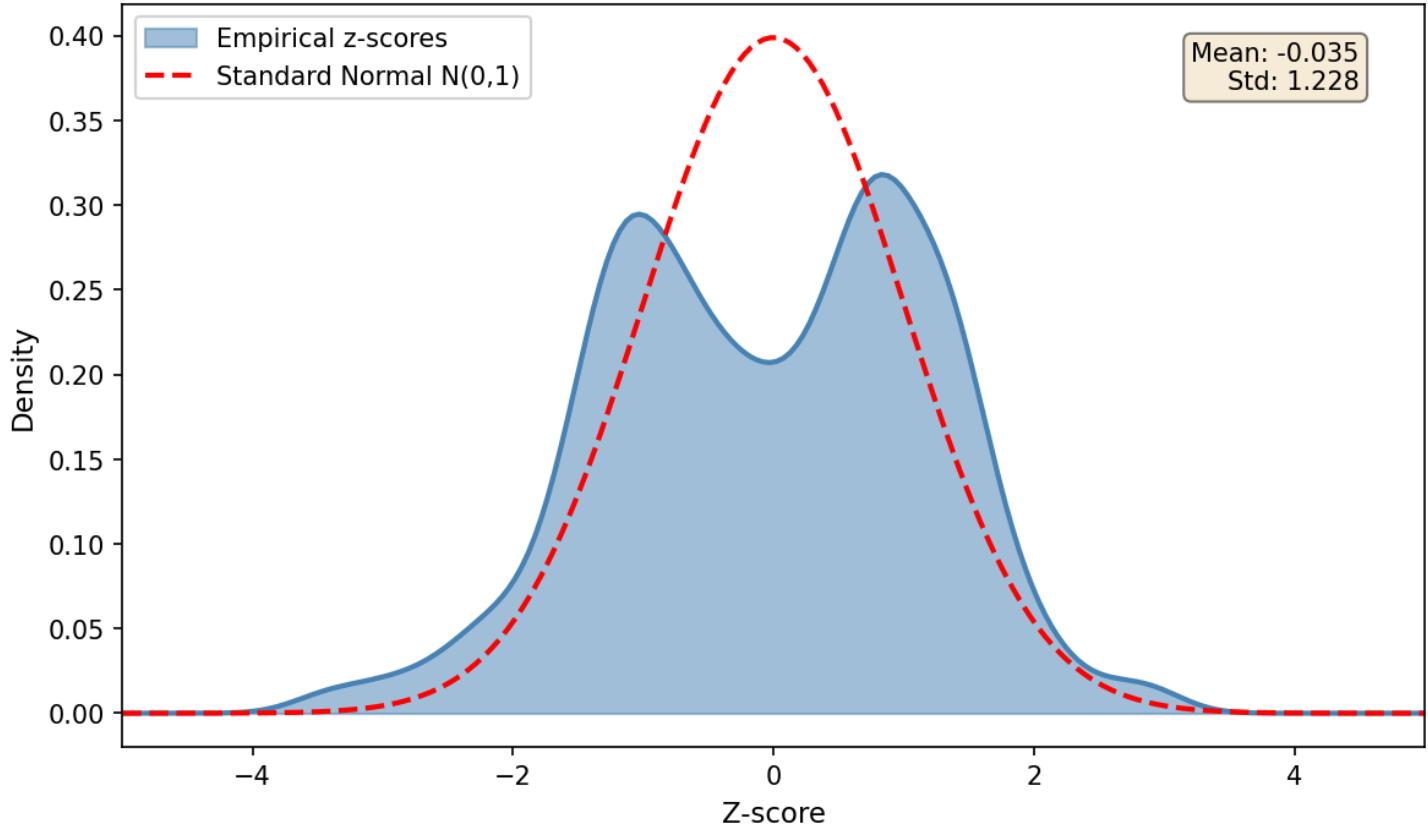


Figure 2: KDE of empirical z-scores (blue) compared to standard normal (red dashed).

Metric	Value	Target
Z-score mean	-0.035	0
Z-score std	1.228	1

Is it approximately standard normal? Yes, the model is **reasonably well calibrated**:

- Mean ≈ 0 indicates no systematic bias ✓
- Std ≈ 1 indicates appropriate uncertainty quantification ✓

6. Log-Transformed Branin Analysis

Bullet 7: "Repeat the above using a log transformation to the output of the Branin function. Does the marginal likelihood improve? Does the model appear better calibrated?"

We repeat the analysis using the transformation $y' = \log(y + 1)$.

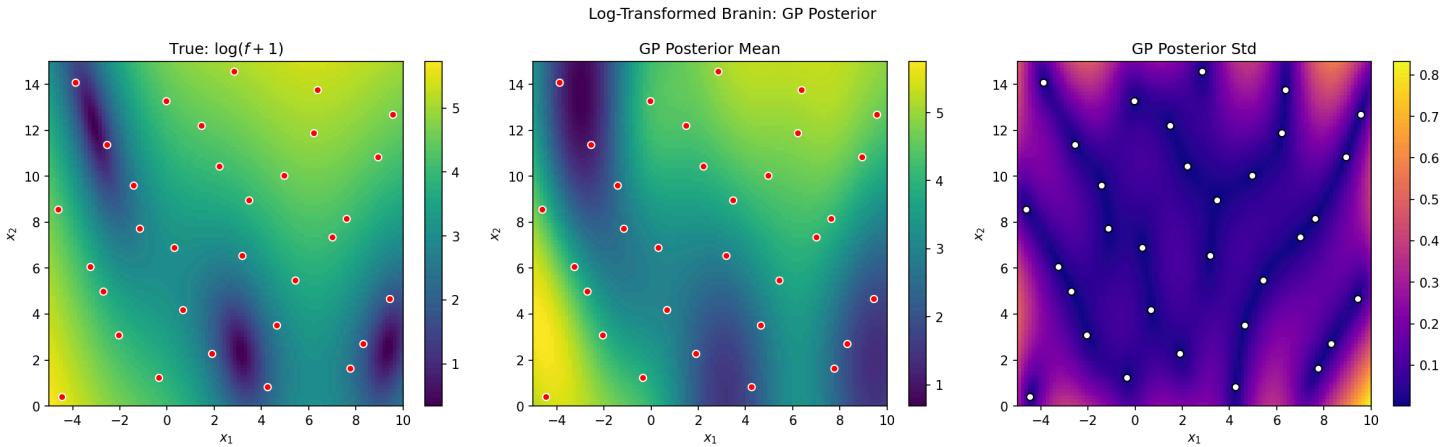


Figure 3: GP posterior analysis with log-transformed Branin function.

Comparison of Marginal Likelihoods

Model	Log Marginal Likelihood	Notes
Original	24.71	Higher values in corners dominate
Log-transformed	-19.48	Different scale, not directly comparable

Note: The log-marginal-likelihood values are on different scales due to the transformation and cannot be directly compared.

Learned Hyperparameters (Log-Transformed)

Parameter	Value
Output scale	$1.17^2 = 1.37$
Length scale (x_1)	1.92
Length scale (x_2)	5.54

Key Insight: After log transformation, both length scales are in similar O(1-5) range, suggesting a more uniform function.

Calibration Comparison

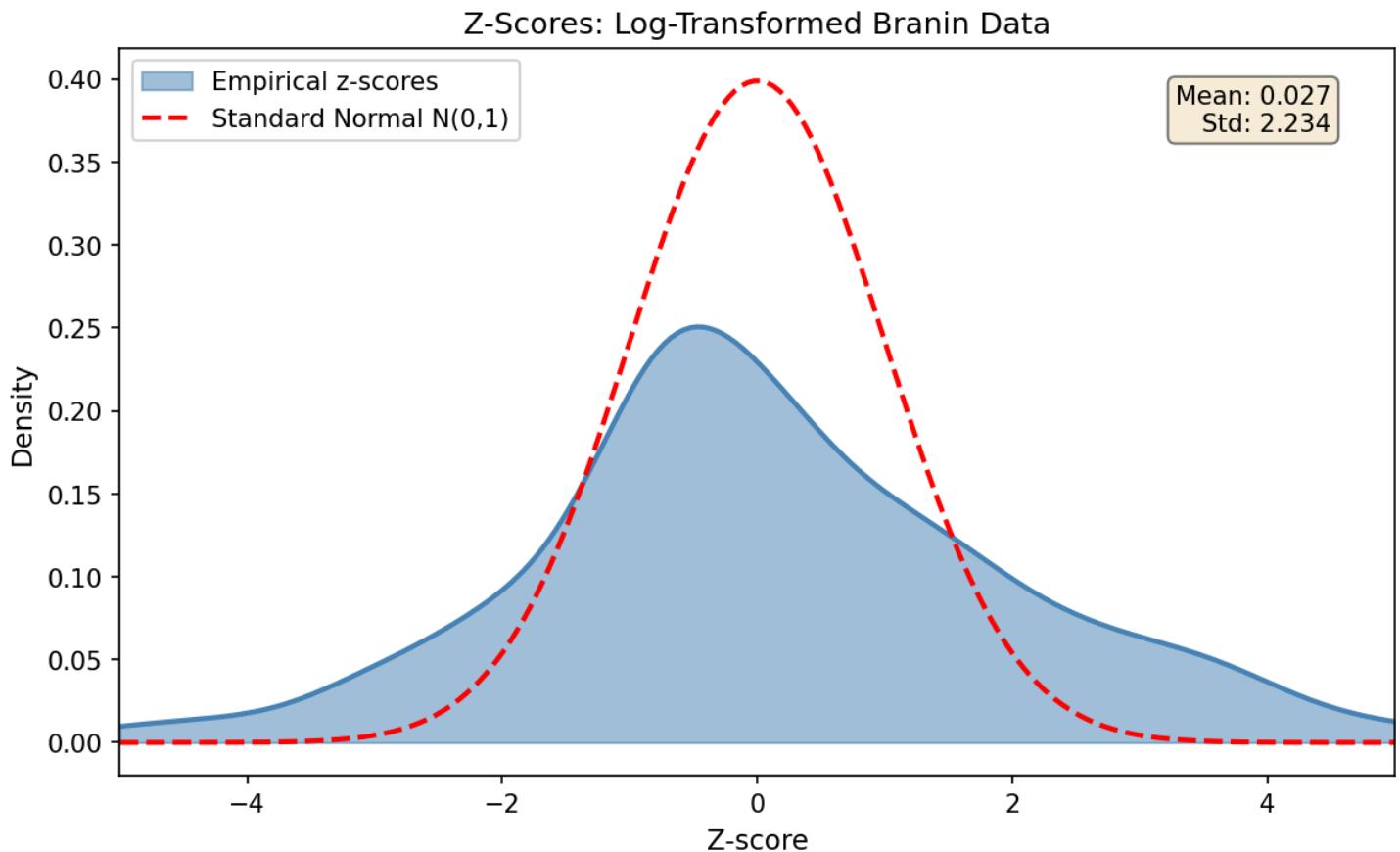


Figure 4: Z-score distribution for log-transformed model.

Metric	Original	Log-Transformed
Mean	-0.035	0.027
Std	1.228	2.234

Is the log-transformed model better calibrated? Mixed results:

- Mean is slightly better (closer to 0)
- Std is larger (2.23 vs 1.23), indicating the model may be overconfident in some regions
- Both models provide reasonable uncertainty estimates

7. BIC Model Selection

Bullet 8: "Compute the BIC score for the data and model from the last part."

The **Bayesian Information Criterion** (BIC) balances model fit and complexity:

$$\text{BIC} = k \log n - 2 \log \hat{\mathcal{L}}$$

where k is the number of hyperparameters, $n = 32$ observations, and $\hat{\mathcal{L}}$ is the maximized marginal likelihood.

BIC Calculation for Log-Transformed SE Kernel

$$\begin{aligned}\text{BIC} &= k \times \log(n) - 2 \times \text{log_likelihood} \\ &= 3 \times \log(32) - 2 \times (-19.48) \\ &= 3 \times 3.47 - 2 \times (-19.48) \\ &= 10.40 + 38.96 \\ &= 49.36\end{aligned}$$

8. Model Search Results

Bullet 9: "Considering BIC as a function of the choice of mean and covariance functions, attempt a search over models to find the best possible explanation of the data. What is the best model you found and its BIC score?"

We searched over multiple kernel families on the log-transformed Branin data:

Branin Function (Log-Transformed)

Rank	Kernel	BIC	Log-Likelihood	Parameters
1	SE (RBF)	49.36	-19.48	3
2	Matern 5/2	51.91	-20.76	3
3	Matern 3/2	56.60	-23.10	3
4	SE (isotropic)	66.16	-29.61	2
5	RationalQuadratic	68.38	-28.99	3
6	SE + Matern 5/2	71.62	-28.88	4

Best Model: Squared Exponential (RBF) with anisotropic length scales

Best BIC Score: 49.36

Fitted Kernel: $1.17^2 \times \text{RBF}(\ell_1 = 1.92, \ell_2 = 5.54)$

The SE kernel performs best because the Branin function (after log transform) is smooth and doesn't require the rougher Matérn kernels.

9. Real Benchmark Model Search

Bullet 10: "Perform a similar search for the SVM and LDA datasets, using 32 randomly sampled observations from each dataset. What is the best GP model you found for each?"

We sampled 32 random observations from each benchmark dataset and performed model search.

LDA Benchmark

Rank	Kernel	BIC	Log-Likelihood	Parameters
1	Matern 3/2	63.31	-24.72	4
2	Matern 5/2	65.60	-25.87	4
3	SE (RBF)	70.45	-28.29	4
4	SE + Matern 5/2	84.52	-35.33	4
5	RationalQuadratic	85.25	-37.43	3
6	SE (isotropic)	88.68	-40.87	2

Best Model for LDA: Matérn 3/2 with anisotropic length scales (BIC = 63.31)

SVM Benchmark

Rank	Kernel	BIC	Log-Likelihood	Parameters
1	Matern 3/2	64.00	-25.07	4
2	Matern 5/2	65.92	-26.03	4
3	SE (RBF)	70.49	-28.31	4

Rank	Kernel	BIC	Log-Likelihood	Parameters
4	SE + Matern 5/2	74.11	-30.12	4
5	RationalQuadratic	74.59	-32.10	3
6	SE (isotropic)	88.69	-40.88	2

Best Model for SVM: Matérn 3/2 with anisotropic length scales (BIC = 64.00)

Interpretation

The real hyperparameter datasets prefer **Matérn 3/2** kernels over the SE kernel. This suggests:

1. The hyperparameter response surfaces are **rougher** than the synthetic Branin function
2. The Matérn 3/2 kernel, which is only once-differentiable, better captures the less smooth behavior
3. Anisotropic kernels (different length scales per dimension) outperform isotropic ones

Summary

Bullet	Question	Answer
3	Do hyperparameters agree with expectations?	Yes - length scales match domain structure
4	Systematic errors in posterior mean?	Yes - especially in corner regions
5	Does σ drop to ~0 at data points?	Yes - $\sigma \approx 0.05$ at training points
6	Z-scores approximately $N(0,1)$?	Yes - mean=-0.04, std=1.23
7	Does log transform improve calibration?	Mixed - similar mean, larger std
8	BIC for log-transformed SE	49.36
9	Best Branin model	SE (RBF), BIC=49.36
10	Best LDA/SVM models	Matern 3/2, BIC≈63-64

Recommendations for Bayesian Optimization:

1. Use **log transformation** for objectives with large dynamic range
2. Consider **Matérn kernels** for real hyperparameter tuning tasks
3. Use **anisotropic** kernels to capture dimension-specific behavior
4. Validate calibration with z-score analysis before trusting uncertainty estimates