A Hybrid Approximation to the Marginal Likelihood

Eric Chuu, Debdeep Pati, Anirban Bhattacharya

Texas A&M University Department of Statistics

15 Second Summary

- Computing the marginal likelihood is computationally challenging, particularly when the dimension of the parameter space is large.
- Existing methods [3] are known to be slow and potentially inaccurate when MCMC samples are few in number or non-exact.
- Our algorithm uses MCMC samples to learn a high probability partition of the parameter space and then forms a deterministic approximation over each of these partitions.

Problem Setup

Given data y, a likelihood function $p(y \mid u)$ indexed by u from a d-dimensional parameter space $\mathcal{U}$, and a prior distribution $p(u)$, we can write the marginal likelihood or evidence as,

$$p(y) = \int_{\mathcal{U}} p(y \mid u)\, p(u)\, du.$$

Definitions and Notation

Let γ be a probability density defined on $\mathbb{R}^d$ given by

$$\gamma(u) = \frac{e^{-\Phi(u)}\, \pi(u)}{\mathcal{Z}}, \quad u \in \mathcal{U} \subseteq \mathbb{R}^d.$$

Typically, $\Phi(\cdot)$ corresponds to a negative log-likelihood function and $\pi(\cdot)$ a prior distribution, so $\gamma(\cdot)$ is the corresponding posterior distribution. The marginal likelihood has the following form,

$$\mathcal{Z} = \int_{\mathcal{U}} e^{-\Psi(u)}\, du, \tag{1}$$

where $\Psi(u) = \Phi(u) - \log \pi(u))$ is the negative log-posterior.

Our Approach

Generally, we can evaluate Ψ, but are unable to compute the integral in Eq. (1). Provided that we can sample from γ, we propose a two-step approach for solving this problem:

Step 1: Obtain a partition of the parameter space that identifies regions of the posterior that have posterior mass.

Step 2: Approximate Ψ over each of these partition sets

Using these steps together provides a way to approximate $\mathcal{Z}$ by computing a simplified version of the integral over partition sets of the parameter space that have ideally taken into account the assumed non-uniform nature of the posterior distribution.

Method

- **Step 1: High Probability Partitioning of the Parameter Space**
  - Using samples u_j from γ, form $(u_j, \Psi(u_j))$, $1 \leq j \leq J$
  - Using $(u_j, \Psi(u_j))$ as covariate-response pairs as input to a regression tree model, we can a dyadic partition of $\mathcal{U}$
  - Define the compactification, A, of the parameter space $\mathcal{U}$ to be a bounding box using the range of posterior samples,

$$A = \otimes_l \left[\min \left\{ u_j^{(l)} \right\}, \max \left\{ u_j^{(l)} \right\} \right], \quad 1 \leq j \leq J, 1 \leq l \leq d,$$

where $u_j^{(l)}$ is the lth component of u_j.

- **Step 2: Piecewise Constant Approximation to Ψ**

$$\widehat{\Psi}(u) = \sum_{k=1}^{K} c_k^\star \cdot \mathbb{1}_{A_k}(u),$$

where $\mathcal{A} = \{A_1, \ldots, A_K\}$ is a partition of A, i.e., $A = \cup_{k=1}^K A_k$ and $A_k \cap A_{k'} = \varnothing$ for all $k \neq k'$, and $c_k^\star$ is a representative value of Ψ within the partition set A_k.

From step 1, we have rectangular partition sets of the form: $A_k = \prod_{l=1}^d [a_k^{(l)}, b_k^{(l)}]$. This leads to the *Hybrid Approximation*,

$$\int_A e^{-\Psi(u)}\, du \approx \int_A e^{-\widehat{\Psi}(u)}\, du = \sum_{k=1}^{K} e^{-c_k^\star} \cdot \mu(A_k).$$

Here, $\mu(B) = \int_B 1\, du$ is the d-dimensional volume of a set B.

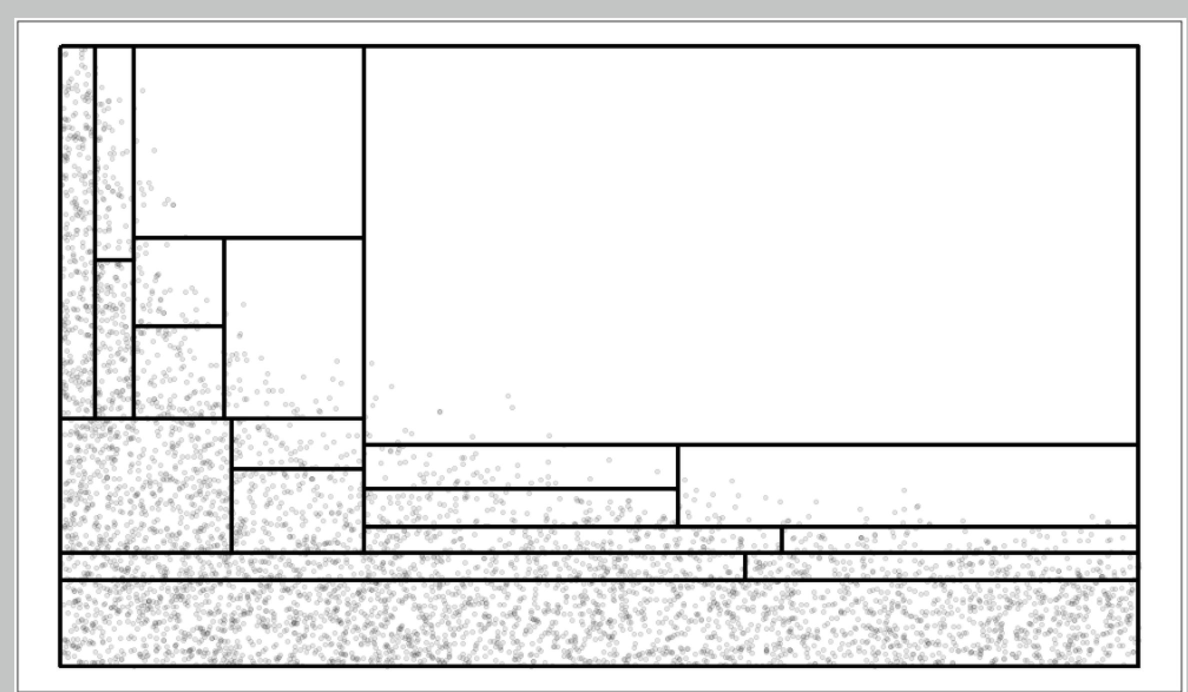

Figure: We draw samples from a density of the form $\gamma(u) \propto \exp(-n u_1^2 u_2^4)\pi(u)$, where $u \in [0,1]^2$ and $\pi(\cdot)$ is the uniform measure on $[0,1]^2$. Using these samples as input for CART [1], we form the following partition over the parameter space.

Experiments

- **Truncated Multivariate Normal Model**
  - Linear regression with a truncated multivariate normal (tMVN) prior on β
- **Unrestricted Covariance Matrices**
  - For data $x_1, \ldots, x_n \overset{\text{iid}}{\sim} N_d(0, \Sigma)$, where $\Sigma \in \mathbb{R}^{d \times d}$, we have the following likelihood

$$L(\Sigma) = (2\pi)^{-nd/2} \det(\Sigma)^{-n/2} e^{-\operatorname{tr}(\Sigma^{-1} S)/2},$$

where $S = \sum_{i=1}^n x_i x_i'$. Consider a conjugate inverse-Wishart (IW) prior on Σ, $\mathcal{W}^{-1}(\Lambda, \nu)$.

- **Gaussian Graphical Models**
  - Consider data $x_1, \ldots, x_n \overset{\text{iid}}{\sim} N_d(0, \Omega)$, where Ω is a sparse precision matrix. A probabilistic framework for learning the dependence structure and the graph G requires a prior distribution for (Ω, G). Conditional on G, we consider the hyper-inverse Wishart (HIW) prior [2] for Ω
- **Approximate MCMC Samples**
  - Linear regression with a multivariate normal-inverse-gamma (MVN-IG) prior on (β, σ^2)
  - We sample from a *mean field approximation* of the posterior distribution

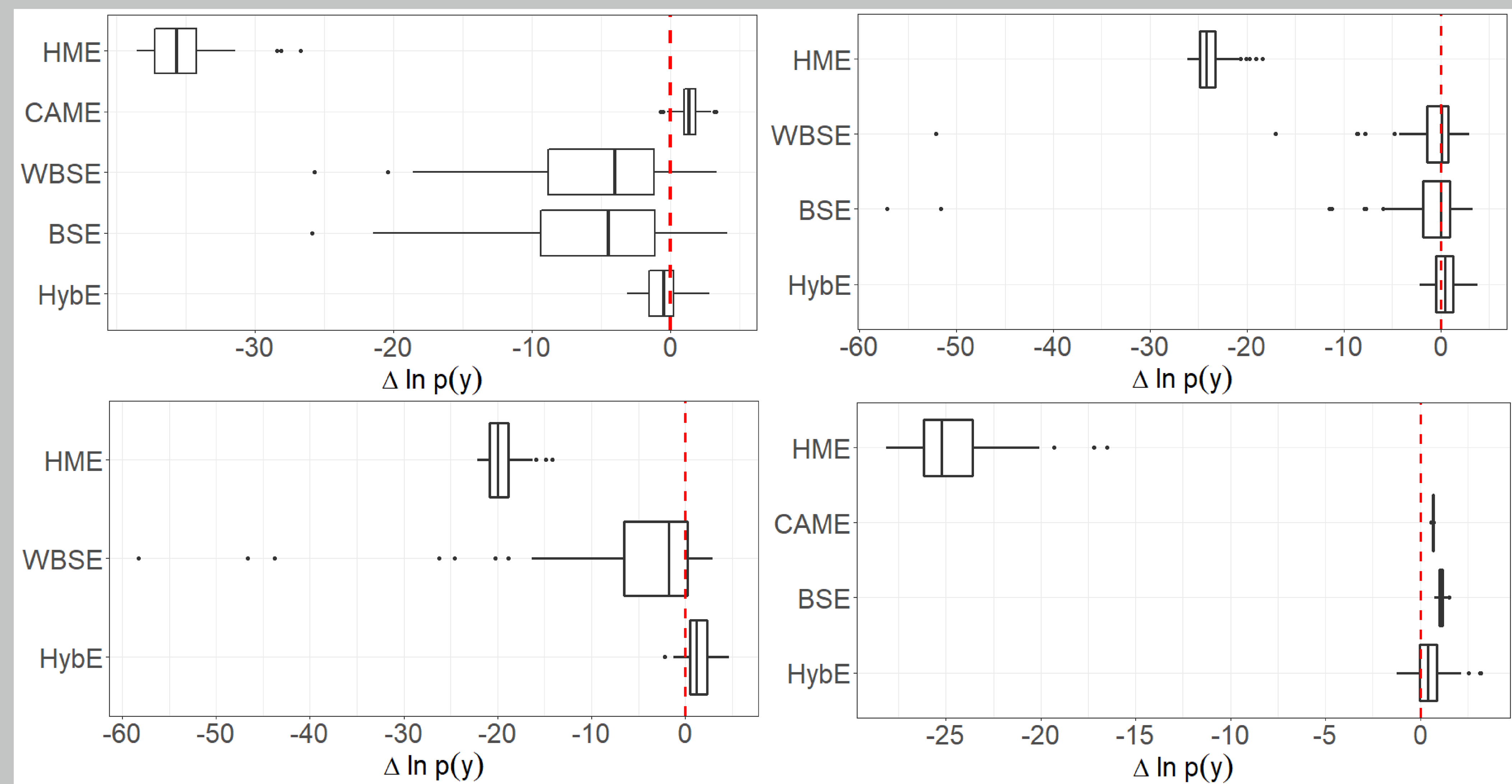

Figure: Boxplots of the error (truth - estimate). For the tMVN example (top left), $\beta \in \mathbb{R}^{20}$. For the IW example (top right), $\Sigma \in \mathbb{R}^{4 \times 4}$, with 10 free parameters. For the HIW example (bottom left), $\Omega \in \mathbb{R}^{5 \times 5}$, with 10 free parameters. For the approximate MVN-IG example (bottom right), $(\beta, \sigma^2) \in \mathbb{R}^{10}$. Other than the Hybrid Estimator (HybE), we considered various competing estimators: Harmonic Mean Estimator (HME), Corrected Arithmetic Mean Estimator (CAME), (Warped) Bridge Sampling Estimator (WBSE, BSE).

References

[1] L. Breiman, "Classification and regression trees," 1984.

[2] A. P. Dawid and S. L. Lauritzen, "Hyper markov laws in the statistical analysis of decomposable graphical models," *The Annals of Statistics*, pp. 1272–1317, 1993.

[3] N. Friel and J. Wyse, "Estimating the evidence - a review," *Statistica Neerlandica*, vol. 66, no. 3, p. 288–308, 2012.

[ericchuu, anirbanb, debdeep]@stat.tamu.edu