Workshop at the 7th Symposium on Advances in Approximate Bayesian Inference (non-archival), 2025 1–24

Stacking Variational Bayesian Monte Carlo

Francesco Silvestrin FRANCESCO.SILVESTRIN@HELSINKI.FI
Chengkun Li CHENGKUN.LI@HELSINKI.FI
Luigi Acerbi LUIGI.ACERBI@HELSINKI.FI
Department of Computer Science, University of Helsinki, Helsinki, Finland

Abstract

Variational Bayesian Monte Carlo (VBMC) is a sample-efficient method for approximate Bayesian inference with computationally expensive likelihoods. While VBMC's local surrogate approach provides stable approximations, its conservative exploration strategy and limited evaluation budget can cause it to miss regions of complex posteriors. In this work, we introduce Stacking Variational Bayesian Monte Carlo (S-VBMC), a method that constructs global posterior approximations by merging independent VBMC runs through a principled and inexpensive post-processing step. Our approach leverages VBMC's mixture posterior representation and per-component evidence estimates, requiring no additional likelihood evaluations while being naturally parallelisable. We demonstrate S-VBMC's effectiveness on two synthetic problems designed to challenge VBMC's exploration capabilities and two real-world applications from computational neuroscience, showing substantial improvements in posterior approximation quality across all cases.[1]

1. Introduction

Bayesian inference provides a powerful framework for parameter estimation and uncertainty quantification, but it is usually intractable, requiring approximate inference techniques (Brooks et al., 2011; Blei et al., 2017). Many scientific and engineering problems involve black-box models (Sacks et al., 1989; Kennedy and O'Hagan, 2001), where likelihood evaluation is time-consuming and gradients cannot be easily obtained, making traditional approximate inference approaches computationally prohibitive.

A promising approach to tackle expensive likelihoods is to construct a statistical surrogate model that approximates the target distribution, similar in spirit to surrogate approaches to global optimisation using Gaussian processes (Williams and Rasmussen, 2006; Garnett, 2023). However, attempting to build a single global surrogate model may lead to numerical instabilities and poor approximations when the target distribution is complex or multi-modal, without ad hoc solutions (Wang and Li, 2018; Järvenpää et al., 2020; Li et al., 2024). Local or constrained surrogate models, while more limited in scope, tend to be more stable and reliable in practice (El Gammal et al., 2023; Järvenpää and Corander, 2024).

Variational Bayesian Monte Carlo (VBMC; Acerbi, 2018) exemplifies this local approach, using active sampling to train a Gaussian process surrogate for the unnormalised log-posterior on which it performs variational inference. VBMC adopts a conservative exploration strategy that yields stable, local approximations (Acerbi, 2019). Compared to other surrogate-based approaches, the method offers a versatile set of features: it returns the approximate posterior as a tractable distribution (a mixture of Gaussians); it provides

1. This is a workshop paper. An extended version is forthcoming: https://arxiv.org/abs/2504.05004.

© F. Silvestrin, C. Li & L. Acerbi.

a lower bound for the model evidence (ELBO) via Bayesian quadrature (Ghahramani and Rasmussen, 2002), useful for model selection; and it can handle noisy log-likelihood evaluations (Acerbi, 2020), which arise in simulator-based models through estimation techniques such as *inverse binomial sampling* (van Opheusden et al., 2020) and *synthetic likelihood* (Wood, 2010; Price et al., 2018). However, VBMC's limited sampling budget combined with its local exploration strategy can leave it vulnerable to potentially missing regions of the target posterior – particularly for distributions with distinct modes or long tails.

In this work, we propose a practical, yet effective approach to constructing global surrogate models while overcoming the limitations of standard VBMC by combining multiple local approximations. We introduce *Stacking Variational Bayesian Monte Carlo* (S-VBMC), a method for merging independent VBMC inference runs into a coherent global posterior approximation. Our approach leverages VBMC's unique properties – its mixture posterior representation and per-component Bayesian quadrature estimates of the ELBO – to combine and reweigh each component through a simple post-processing step.

Crucially, our method requires no additional evaluations of either the original model or the surrogate. This approach is easily parallelisable and naturally fits existing VBMC pipelines that already employ multiple independent runs (Huggins et al., 2023). While our method could theoretically extend to other variational approaches based on mixture posteriors, VBMC is uniquely suitable for it as re-estimation of the ELBO would otherwise become impractical with expensive likelihoods (see Section 3).

We first introduce variational inference and VBMC (Section 2), then present our algorithm for stacking VBMC posteriors (Section 3). We demonstrate the effectiveness of our approach through experiments on two synthetic problems and two real-world applications that are challenging for VBMC (Section 4). We conclude with closing remarks (Section 5). Appendix A contains supplementary materials, including a discussion of related work (A.1).

2. Background

Variational Inference. Consider a model with prior $p(\boldsymbol{\theta})$ and likelihood $p(\mathcal{D}|\boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \mathbb{R}^D$ is a vector of model parameters and $\mathcal{D}$ a specific dataset. Variational inference (Blei et al., 2017) approximates the true posterior $p(\boldsymbol{\theta}|\mathcal{D})$ with a parametric distribution $q_{\boldsymbol{\phi}}(\boldsymbol{\theta})$ by maximising the evidence lower bound (ELBO):

$$\mathrm{ELBO}(\boldsymbol{\phi}) = \mathbb{E}_{q_{\boldsymbol{\phi}}}\left[\log p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})\right] + \mathcal{H}\left[q_{\boldsymbol{\phi}}(\boldsymbol{\theta})\right], \tag{1}$$

where the first term is the expected log joint distribution (the joint being likelihood times prior) and the second term the entropy of the variational posterior. Maximising Eq. 1 is equivalent to minimising the Kullback-Leibler divergence between $q_{\boldsymbol{\phi}}(\boldsymbol{\theta})$ and the true posterior. The ELBO provides a lower bound on the log model evidence $\log p(\mathcal{D})$, with equality when the approximation matches the true posterior.

Variational Bayesian Monte Carlo (VBMC). VBMC is a sample-efficient technique to obtain a variational approximation with only a small number of likelihood evaluations, often of the order of a few hundreds. VBMC uses a Gaussian process (GP) as a surrogate of the log-joint, Bayesian quadrature to calculate the expected log-joint, and active sampling to decide which parameters to evaluate next (see Acerbi, 2018, 2020 for details). Crucially, VBMC performs variational inference on the surrogate, instead of the true, expensive model.

In VBMC, the variational posterior is defined as

$$q_{\boldsymbol{\phi}}(\boldsymbol{\theta}) = \sum_{k=1}^{K} w_k q_{k,\boldsymbol{\phi}}(\boldsymbol{\theta}), \tag{2}$$

where q_k is the k-th component (a multivariate normal) and w_k its mixture weight, with $\sum_{k=1}^{K} w_k = 1$ and $w_k \geq 0$. Plugging in the mixture posterior, the ELBO (Eq. 1) becomes:

$$\text{ELBO}(\boldsymbol{\phi}) = \sum_{k=1}^{K} w_k \mathbb{E}_{q_{k,\boldsymbol{\phi}}} \left[\log p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})\right] + \mathcal{H}\left[q_{\boldsymbol{\phi}}(\boldsymbol{\theta})\right] = \sum_{k=1}^{K} w_k I_k + \mathcal{H}\left[q_{\boldsymbol{\phi}}(\boldsymbol{\theta})\right] \tag{3}$$

where we defined the k-th component of the expected log-joint as:

$$I_k = \mathbb{E}_{q_{k,\boldsymbol{\phi}}} \left[\log p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})\right] \approx \mathbb{E}_{q_{k,\boldsymbol{\phi}}} \left[f(\boldsymbol{\theta})\right], \tag{4}$$

with $f(\boldsymbol{\theta}) \approx \log p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$ the GP surrogate of the log-joint. Eq. 4 has a closed-form Gaussian expression via Bayesian quadrature, which yields posterior mean I_k and covariance matrix $J_{kk'}$ (Acerbi, 2018). The entropy of a mixture of Gaussians does not have an analytical solution, but gradients can be estimated via Monte Carlo. Thus, using the posterior mean of Eq. 4 as a plug-in estimator for the expected log-joint of each component, Eq. 3 can be efficiently optimised via stochastic gradient ascent (Kingma and Ba, 2014).

3. Stacking VBMC

In this work, we introduce *Stacking VBMC* (S-VBMC), a novel approach to merge different variational posteriors obtained from different runs on the same model and dataset.

Given M independent VBMC runs, one obtains M variational posteriors $q_{\boldsymbol{\phi}_m}(\boldsymbol{\theta})$, each with K_m Gaussian components, as defined in Eq. 2, as well as M different $\mathbf{I}_m$ vectors, as per Eq. 4. Our approach consists of "stacking" the Gaussian components of all posteriors $q_{\boldsymbol{\phi}_m}(\boldsymbol{\theta})$ leaving all individual components parameters (means and covariances) unchanged, and reoptimising *all* the weights. Thus, given the stacked posterior

$$q_{\tilde{\boldsymbol{\phi}}}(\boldsymbol{\theta}) = \sum_{m=1}^{M} \sum_{k=1}^{K_m} \tilde{w}_{m,k} q_{k,\boldsymbol{\phi}_m}(\boldsymbol{\theta}), \tag{5}$$

we optimise the global evidence lower bound with respect to the weights $\tilde{\mathbf{w}}$,

$$\text{ELBO}_{\text{stacked}}(\tilde{\mathbf{w}}) = \sum_{m=1}^{M} \sum_{k=1}^{K_m} \tilde{w}_{m,k} I_{m,k} + \mathcal{H}\left[q_{\tilde{\boldsymbol{\phi}}}(\boldsymbol{\theta})\right]. \tag{6}$$

Notably, this optimisation can be performed as a pure post-processing step, requiring neither evaluations of the original likelihood $p(\mathcal{D}|\boldsymbol{\theta})$ nor of the surrogate models f_m, only that the estimates $I_{m,k}$ are stored, as in current implementations (Huggins et al., 2023).

Our stacking method hinges on the key feature of VBMC of providing accurate estimates $I_{m,k}$. While in principle Eq. 6 could apply to any collection of variational posterior mixtures, without an efficient way of calculating each I_k (Eq. 4), optimisation of the stacked ELBO would require many likelihood evaluations, which would be prohibitive for problems with expensive, black-box likelihoods. Figure 1 shows an example of two separate posteriors and the stacked result. In the following, we demonstrate the efficacy of this approach.

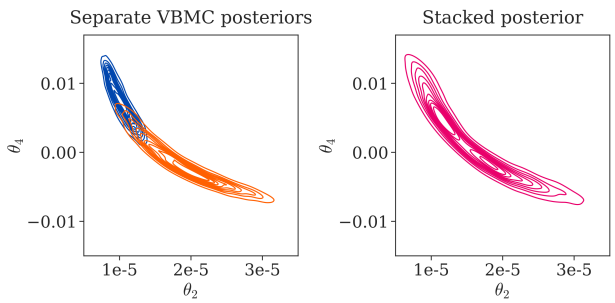


Figure 1: Two separate VBMC posteriors (left) and stacked posterior after running S-VBMC (right) for a neuronal model with real data (see Section 4); showing the marginal distribution of two out of the 5 model parameters.

4. Experiments

Procedure. We tested our method on two synthetic problems, designed to be particularly challenging for VBMC, as well as on two real-world datasets and models (see Appendix A.2 for full descriptions). We considered both noiseless problems (exact estimation) and noisy problems where Gaussian noise with $\sigma = 3$ is applied to each log-likelihood measurement, emulating what practitioners might find when estimating the likelihood via simulation (van Opheusden et al., 2020). For each benchmark, we performed 100 VBMC converging runs with default settings and random uniform initialisation within plausible parameter bounds (Acerbi, 2018). To investigate the effect of combining a different number of posteriors, we then randomly sampled and stacked with S-VBMC a varying number of runs (between 2 and 40) ten times each, and computed the median and interquartile range for all metrics.

Following Acerbi (2020); Li et al. (2024), we use three main metrics for evaluating the posterior approximation of our algorithm: the absolute difference between the true log marginal likelihood (LML) and its variational approximation (the ELBO); the mean marginal total variation distance (MMTV) between the approximate posterior and ground truth; and the "Gaussianised" symmetrised KL divergence (GsKL) between variational posterior and ground truth (see Appendix A.3 for a detailed description).

We used black-box variational inference (BBVI; Ranganath et al., 2014) as a baseline for all our benchmark problems. The target density evaluation budget for BBVI is $2000(D + 2)$ for noiseless problems and $3000(D + 2)$ for noisy problems, which correspond to the maximum number of evaluations used in total by 40 VBMC runs (see Appendix A.4).

Our results are described below and reported in full in Appendix A.5, with example visualisations of posterior approximations in Appendix A.6. Computational costs are briefly discussed in Appendix A.7.

Synthetic problems. The first synthetic target consists of a $2D$ Gaussian mixture model (GMM) with 20 components clustered around four distant centroids. We expected VBMC to discover only one of the clusters in each run. The second synthetic target (ring) consists of a very narrow ring-shaped distribution in two dimensions. We expected VBMC to only cover part of it in each run due to the limited budget (50) of Gaussian components.

Results in Figure 2 and Table A.1 show that merging more posteriors leads to a steady improvement in the GsKL and MMTV metrics, which measure the quality of the posterior

approximation. Remarkably, S-VBMC proves to be robust to noisy targets, with minimal differences between noiseless and noisy settings. S-VBMC outperforms the BBVI baseline and regular VBMC on the ring-shaped synthetic target. The BBVI baseline performs well and only marginally worse compared to S-VBMC only on the GMM problem, where it effectively managed to capture the four clusters (see Figure A.1 for a visualisation). As expected by design, individual VBMC runs tended to explore the two synthetic target distributions only partially. Still, the random initialisations allowed different runs to discover different portions of the posterior, allowing the merging process to cover the whole target (see Figure A.1).

Finally, we observe that, in noisy settings, while the ELBO keeps increasing, the ΔLML error (difference between ELBO and true log marginal likelihood) initially decreases but then increases again as further components are added, a point which we will discuss later.

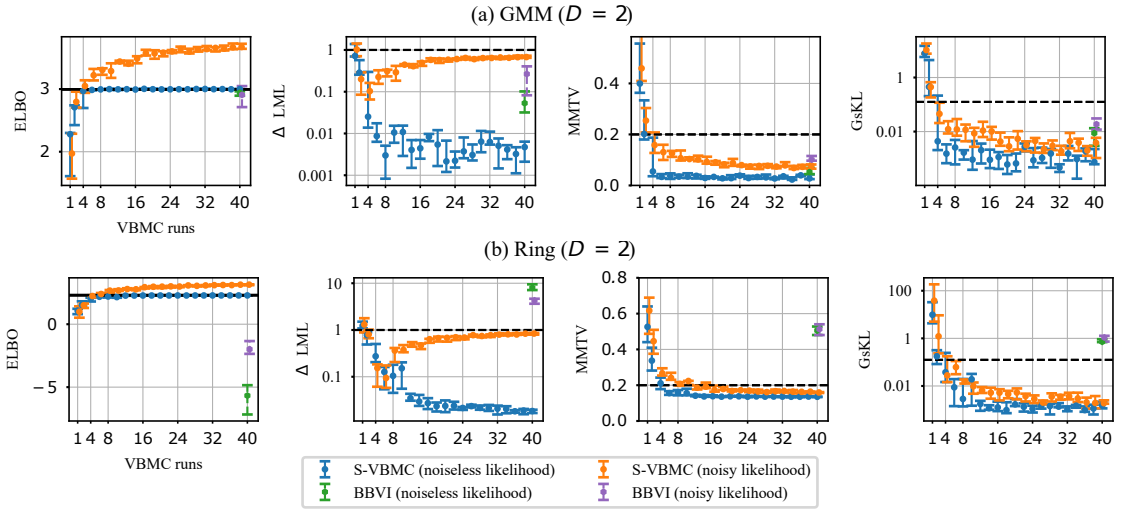


Figure 2: Synthetic problems. Metrics plotted as a function of the number of VBMC runs stacked (median and interquartile range). Likelihood evaluations are noiseless (blue) or noisy with $\sigma = 3$ log-likelihood noise (orange). The best BBVI results for noiseless and noisy likelihood are shown in green and purple. The black horizontal line in the ELBO panels represents the ground-truth LML while the dashed lines on ΔLML, MMTV and GsKL denote desirable thresholds for each metric (good performance is below the threshold; see Appendix A.3).

Real-world problems. Finally, we tested VBMC on two real-world models and datasets. First, we fitted the 5 biophysical parameters of a morphologically detailed neuronal model of hippocampal pyramidal cells (similar to Szoboszlay et al., 2016 for cerebellar Golgi cells) to experimental data consisting of a detailed three-dimensional reconstruction and electrophysiological recordings (Golding et al., 2005) of one of such cells. Then, we fitted a 6-parameter model of multisensory causal inference (Körding et al., 2007) to human behavioural data from a visuo-vestibular task (subject S1 from Acerbi et al., 2018), assuming log-likelihood measurement noise ($\sigma = 3$). This model describes how participants judge whether visual and vestibular motion cues share a common cause, incorporating sensory noise parameters and decision rules to account for participant responses in different experimental conditions.

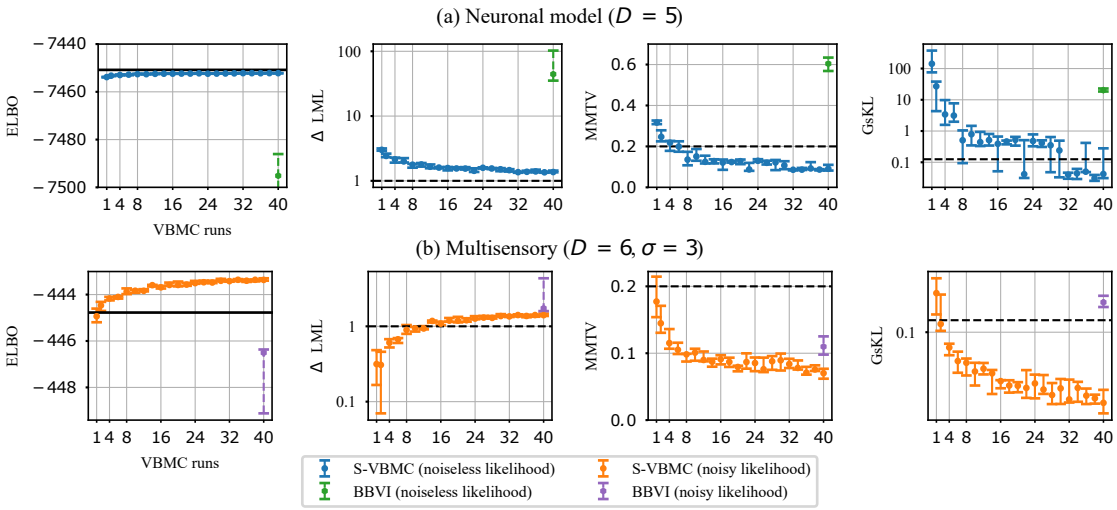


Figure 3: Real-world problems. Metrics plotted as a function of the number of VBMC runs stacked (median and interquartile range). Metrics plotted as a function of the number of VBMC runs stacked (median and interquartile range). Likelihood evaluations are noiseless (blue) or noisy with $\sigma = 3$ log-likelihood noise (orange). The best BBVI results for noiseless and noisy likelihood are shown in green and purple. See Figure 2 caption for additional details.

The results in Figure 3 and Table A.2 confirm our earlier findings of improvements across the posterior metrics. We also find that S-VBMC is robust to noisy targets for real data, with performance that improves with increasing number of stacked runs in the multisensory model problem, and consistently better than standard VBMC and the BBVI baseline.

ELBO estimation bias. Our results show that, with noisy log-likelihood problems, merging more VBMC runs leads to a positive bias build-up in the estimated ELBO. This likely occurs because all $I_{m,k}$ are noisy estimates of the true expected log-joint contributions, causing S-VBMC to overweigh the *most overestimated* mixture components – an effect that increases with the number of components M (see Appendix A.8). This explanation is supported by the fact that we do not observe ELBO bias in noiseless log-likelihood problems, where the amount of noise in the aforementioned estimates is negligible. While this bias surprisingly does not affect other posterior quality metrics, which keep improving (or plateau) with increasing M, it should be considered when using $\text{ELBO}_{\text{stacked}}$ for model comparison. Future work should investigate bias sources and potential debiasing techniques.

5. Conclusions

In this work, we introduced S-VBMC, an approach for merging independent VBMC runs in a principled way to yield a global posterior approximation. We showed its effectiveness on challenging synthetic and real-world problems, as well as its robustness to noise. We briefly discussed the positive bias in the ELBO estimation introduced (or amplified) by the stacking process, leaving further investigation for future work.

Acknowledgments

This work was supported by Research Council of Finland (grants 358980 and 356498). The authors wish to thank the Finnish Computing Competence Infrastructure (FCCI) for supporting this project with computational and data storage resources.

The model referred to as the "neuronal model" in the main text and appendices was developed by Dániel Terbe, Balázs Szabó and Szabolcs Káli (HUN-REN Institute of Experimental Medicine, Budapest, Hungary), using data collected by Miklós Szoboszlay in Zoltán Nusser's laboratory (HUN-REN Institute of Experimental Medicine, Budapest, Hungary). The authors thank these researchers for making their data and model available for the work described in this paper.

References

Luigi Acerbi. Variational Bayesian Monte Carlo. *Advances in Neural Information Processing Systems*, 31:8222–8232, 2018.

Luigi Acerbi. An exploration of acquisition and mean functions in Variational Bayesian Monte Carlo. In *Symposium on Advances in Approximate Bayesian Inference*, pages 1–10. PMLR, 2019.

Luigi Acerbi. Variational Bayesian Monte Marlo with noisy likelihoods. *Advances in Neural Information Processing Systems*, 33:8211–8222, 2020.

Luigi Acerbi, Kalpana Dokka, Dora E Angelaki, and Wei Ji Ma. Bayesian comparison of explicit and implicit causal inference strategies in multisensory heading perception. *PLoS Computational Biology*, 14(7):e1006110, 2018.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng. *Handbook of Markov Chain Monte Carlo*. CRC press, 2011.

Kenneth P Burnham and David R Anderson. *Model selection and multimodel inference: a practical information-theoretic approach*. Springer Science & Business Media, 2003.

Trevor Campbell and Xinglong Li. Universal boosting variational inference. *Advances in Neural Information Processing Systems*, 32, 2019.

Daniel A De Souza, Diego Mesquita, Samuel Kaski, and Luigi Acerbi. Parallel MCMC without embarrassing failures. *International Conference on Artificial Intelligence and Statistics*, pages 1786–1804, 2022.

Jonas El Gammal, Nils Schöneberg, Jesús Torrado, and Christian Fidler. Fast and robust Bayesian inference using Gaussian processes with GPry. *Journal of Cosmology and Astroparticle Physics*, 2023(10):021, October 2023. ISSN 1475-7516. doi: 10.1088/1475-7516/2023/10/021.

Daniel Foreman-Mackey. Corner.py: Scatterplot matrices in Python. *Journal of Open Source Software*, 1(2):24, June 2016. ISSN 2475-9066. doi: 10.21105/joss.00024.

Roman Garnett. *Bayesian optimization*. Cambridge University Press, 2023.

Zoubin Ghahramani and Carl Rasmussen. Bayesian Monte Carlo. *Advances in Neural Information Processing Systems*, 15, 2002.

Nace L Golding, Timothy J Mickus, Yael Katz, William L Kath, and Nelson Spruston. Factors mediating powerful voltage attenuation along CA1 pyramidal neuron dendrites. *The Journal of Physiology*, 568(1):69–82, 2005.

Fangjian Guo, Xiangyu Wang, Kai Fan, Tamara Broderick, and David B Dunson. Boosting variational inference. *arXiv preprint arXiv:1611.05559*, 2016.

Michael Hines, Andrew P Davison, and Eilif Muller. NEURON and Python. *Frontiers in Neuroinformatics*, 3:391, 2009.

Michael L Hines and Nicholas T Carnevale. The NEURON simulation environment. *Neural Computation*, 9(6):1179–1209, 1997.

Bobby Huggins, Chengkun Li, Marlon Tobaben, Mikko J. Aarnos, and Luigi Acerbi. PyVBMC: Efficient Bayesian inference in Python. *Journal of Open Source Software*, 8(86):5428, 2023. doi: 10.21105/joss.05428. URL https://doi.org/10.21105/joss.05428.

Marko Järvenpää and Jukka Corander. Approximate Bayesian inference from noisy likelihoods with Gaussian process emulated MCMC. *Journal of Machine Learning Research*, 25(366):1–55, 2024.

Marko Järvenpää, Michael U Gutmann, Aki Vehtari, Pekka Marttinen, et al. Parallel Gaussian process surrogate Bayesian inference with noisy likelihood evaluations. *Bayesian Analysis*, 2020.

Marc C Kennedy and Anthony O'Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3):425–464, 2001.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *Proceedings of the 3rd International Conference on Learning Representations*, 2014.

Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. *Proceedings of the 2nd International Conference on Learning Representations*, 2013.

Konrad P Körding, Ulrik Beierholm, Wei Ji Ma, Steven Quartz, Joshua B Tenenbaum, and Ladan Shams. Causal inference in multisensory perception. *PLoS One*, 2(9):e943, 2007.

Chengkun Li, Grégoire Clarté, Martin Jørgensen, and Luigi Acerbi. Fast post-process Bayesian inference with variational sparse Bayesian quadrature, 2024. URL https://arxiv.org/abs/2303.05263.

Andrew C Miller, Nicholas J Foti, and Ryan P Adams. Variational boosting: Iteratively refining posterior approximations. In *International Conference on Machine Learning*, pages 2420–2429. PMLR, 2017.

Willie Neiswanger, Chong Wang, and Eric Xing. Asymptotically exact, embarrassingly parallel MCMC. *Uncertainty in Artificial Intelligence - Proceedings of the 30th Conference, UAI 2014*, 11 2013.

Leah F Price, Christopher C Drovandi, Anthony Lee, and David J Nott. Bayesian synthetic likelihood. *Journal of Computational and Graphical Statistics*, 27(1):1–11, 2018.

Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. In *Artificial Intelligence and Statistics*, pages 814–822. PMLR, 2014.

Jerome Sacks, Susannah B Schiller, and William J Welch. Designs for computer experiments. *Technometrics*, 31(1):41–47, 1989.

Steven L Scott, Alexander W Blocker, Fernando V Bonassi, Hugh A Chipman, Edward I George, and Robert E McCulloch. Bayes and big data: The consensus Monte Carlo algorithm. In *Big Data and Information Theory*, pages 8–18. Routledge, 2022.

Miklos Szoboszlay, Andrea Lőrincz, Frederic Lanore, Koen Vervaeke, R Angus Silver, and Zoltan Nusser. Functional properties of dendritic gap junctions in cerebellar Golgi cells. *Neuron*, 90(5):1043–1056, 2016.

Bas van Opheusden, Luigi Acerbi, and Wei Ji Ma. Unbiased and efficient log-likelihood estimation with inverse binomial sampling. *PLoS Computational Biology*, 16(12):e1008483, 2020.

Hongqiao Wang and Jinglai Li. Adaptive Gaussian process approximation for Bayesian inference with expensive likelihood functions. *Neural Computation*, 30(11):3072–3094, 2018.

Xiangyu Wang, Fangjian Guo, Katherine A Heller, and David B Dunson. Parallelizing MCMC with random partition trees. *Advances in Neural Information Processing Systems*, 28, 2015.

Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

Simon N Wood. Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, 466(7310):1102–1104, 2010.

Yuling Yao, Aki Vehtari, and Andrew Gelman. Stacking for non-mixing Bayesian computations: The curse and blessing of multimodal posteriors. *Journal of Machine Learning Research*, 23(79):1–45, 2022.

Appendix A.

This appendix provides additional details and analyses to complement the main text, included in the following sections:

- A brief overview of relevant existing work, A.1
- Model descriptions, A.2
- Metrics description, A.3
- Black-box variational inference implementation, A.4
- Additional experiment results, A.5
- Example posterior visualisations, A.6
- A brief discussion on the computational overhead of S-VBMC, A.7
- Further discussion of the ELBO bias mentioned in Section 4, A.8

A.1. Related work

Our work addresses the challenge of building global posterior approximations by combining local solutions from the VBMC framework (Acerbi, 2018, 2019, 2020). While the idea of combining posterior distributions has been explored before, previous approaches differ substantially in their goals and methodology. Yao et al. (2022) propose a similar "stacking" approach, but focus on optimising predictive performance through a leave-one-out strategy, whereas S-VBMC optimises the ELBO on the full dataset, allowing treatment of the log-joint as a black box. Other relevant approaches include *variational boosting* (Guo et al., 2016; Miller et al., 2017; Campbell and Li, 2019), which sequentially builds a mixture posterior by running variational inference multiple times on the whole dataset, and *embarrassingly parallel* Markov Chain Monte Carlo (MCMC) (Neiswanger et al., 2013; Wang et al., 2015; Scott et al., 2022; De Souza et al., 2022), which combines parallel "sub-posteriors" obtained from data subsets. Our method differs from variational boosting through its inherent parallel and surrogate-based approach, offering significant computational advantages, and from embarrassingly parallel inference methods by using the complete dataset in each run, thus remaining robust to individual run failures.

A.2. Model descriptions

GMM target. Our synthetic GMM target consists of a mixture of 20 bivariate Gaussian components arranged in four distinct clusters. The cluster centroids were positioned at $(-8, -8)$, $(-7, 7)$, $(6, -6)$ and $(5, 5)$. Around each centroid, we placed five Gaussian components with means drawn from $\mathcal{N}(\boldsymbol{\mu}_c, \mathbf{I})$, where $\boldsymbol{\mu}_c$ is the respective cluster centroid and $\mathbf{I}$ is the 2×2 identity matrix. Each component was assigned unit marginal variances and a correlation coefficient of ±0.5 (randomly selected with equal probability). This configuration produces an irregular mixture structure that requires a substantial number of components to approximate accurately. All components were assigned equal mixing weights of $1/20$. The resulting distribution is illustrated in Figure A.1 (top panels).

Ring target. Our second synthetic target is a ring-shaped distribution defined by the probability density function

$$p_{\text{ring}}(\theta_1, \theta_2) \propto \exp\left(-\frac{(r-R)^2}{2\sigma^2}\right) \tag{A.1}$$

where $r = \sqrt{(\theta_1 - c_1)^2 + (\theta_2 - c_2)^2}$ represents the radial distance from centre (c_1, c_2), R is the ring radius, and σ controls the width of the annulus. We set $R = 8$, $\sigma = 0.1$, and centred the ring at $(c_1, c_2) = (1, -2)$. The small value of σ produces a narrow annular distribution that challenges VBMC's exploration capabilities. The resulting distribution is shown in Figure A.1 (bottom panels).

Neuronal model. Our first real-world problem involved fitting five biophysical parameters of a detailed compartmental model of a hippocampal CA1 pyramidal neuron. The model was constructed based on experimental data comprising a three-dimensional morphological reconstruction and electrophysiological recordings of neuronal responses to current injections. The deterministic neuronal responses were simulated using the NEURON simulation environment (Hines and Carnevale, 1997; Hines et al., 2009), applying current step inputs that matched the experimental protocol. The model's parameters characterise key biophysical properties: intracellular axial resistivity (θ_1), leak current reversal potential (θ_2), somatic leak conductance (θ_3), dendritic conductance gradient (θ_4, per μm), and a dendritic surface scaling factor (θ_5). Based on independent measurements of membrane potential fluctuations, observation noise was modelled as a stationary Gaussian process with zero mean and a covariance function estimated from the data. The covariance structure was captured by the product of a cosine and an exponentially decaying function. For a similar approach applied to cerebellar Golgi cells, see Szoboszlay et al. (2016).

Multisensory causal inference model. Perceptual causal inference involves determining whether multiple sensory stimuli originate from a common source, a problem of particular interest in computational cognitive neuroscience (Körding et al., 2007). Our second real-world problem involved fitting a visuo-vestibular causal inference model to empirical data from a representative participant (S1 from Acerbi et al., 2018). In each trial, participants seated in a moving chair reported whether they perceived their movement direction (s_{vest}) as congruent with an experimentally-manipulated looming visual field (s_{vis}). The model assumes participants receive noisy sensory measurements, with vestibular information $z_{\text{vest}} \sim \mathcal{N}(s_{\text{vest}}, \sigma^2_{\text{vest}})$ and visual information $z_{\text{vis}} \sim \mathcal{N}(s_{\text{vis}}, \sigma^2_{\text{vis}}(c))$, where σ^2_{vest} and σ^2_{vis} represent sensory noise variances. The visual coherence level c was experimentally manipulated across three levels (c_{low}, c_{med}, c_{high}). The model assumes participants judge the stimuli as having a common cause when the absolute difference between sensory measurements falls below a threshold κ, with a lapse rate λ accounting for random responses. The model parameters $\boldsymbol{\theta}$ comprise the visual noise parameters $\sigma_{\text{vis}}(c_{\text{low}})$, $\sigma_{\text{vis}}(c_{\text{med}})$, $\sigma_{\text{vis}}(c_{\text{high}})$, vestibular noise σ_{vest}, lapse rate λ, and decision threshold κ (Acerbi et al., 2018).

A.3. Metrics description

Following Acerbi (2020); Li et al. (2024), we evaluate our method using three metrics:

1. The absolute difference between true and estimated log marginal likelihood (ΔLML), where values < 1 are considered negligible for model selection (Burnham and Anderson, 2003).

2. The mean marginal total variation distance (MMTV), which measures the average (lack of) overlap between true and approximate posterior marginals across dimensions:

$$\mathrm{MMTV}(p, q) = \frac{1}{2D} \sum_{d=1}^{D} \int_{-\infty}^{\infty} |p_d(x_d) - q_d(x_d)| \, dx_d, \tag{A.2}$$

   where p_d and q_d denote the marginal distributions along the d-th dimension.

3. The "Gaussianised" symmetrised KL divergence (GsKL), which evaluates differences in means and covariances between the approximate and true posterior:

$$\mathrm{GsKL}(p, q) = \frac{1}{2D} \left[D_{\mathrm{KL}} \left(\mathcal{N}[p] || \mathcal{N}[q] \right) + D_{\mathrm{KL}} (\mathcal{N}[q] || \mathcal{N}[p]) \right], \tag{A.3}$$

   where $\mathcal{N}[p]$ denotes a Gaussian with the same mean and covariance as p.

We consider MMTV < 0.2 and GsKL $< \frac{1}{8}$ as target thresholds for reasonable posterior approximation (Li et al., 2024). Ground-truth values are obtained through numerical integration, extensive MCMC sampling, or analytical methods as appropriate for each problem.

A.4. Black-box variational inference implementation

Our implementation of black-box variational inference (BBVI) follows Li et al. (2024). For gradient-free black-box models, we cannot use the reparameterisation trick (Kingma and Welling, 2013) to estimate ELBO gradients. Instead, we employ the score function estimator (REINFORCE; Ranganath et al., 2014) with control variates to reduce gradient variance.

The variational posterior is parameterised as a mixture of Gaussians (MoG) with either $K = 50$ or $K = 500$ components, matching the form used in VBMC. We initialise component means near the origin by adding Gaussian noise ($\sigma = 0.1$) and set all component variances to 0.01. We optimise the ELBO using Adam (Kingma and Ba, 2014) with stochastic gradients, performing a grid search over Monte Carlo sample sizes $\{1, 10, 100\}$ and learning rates $\{0.01, 0.001\}$. We select the best hyperparameters based on the estimated ELBO.

For fair comparison with VBMC, we set the target evaluation budget to $2000(D + 2)$ and $3000(D + 2)$ evaluations for noiseless and noisy problems respectively, matching the maximum evaluations used by 40 VBMC runs in total.

A.5. Additional experiment results

We present a comprehensive comparison of S-VBMC against VBMC and BBVI in Tables A.1 and A.2, complementing the visualisations in Figures 2 and 3. For both synthetic problems (Table A.1) and real-world problems (Table A.2), S-VBMC generally demonstrates consistently improved posterior approximation metrics compared to both baselines. However, we observe an increase in ΔLML error with larger numbers of stacked runs in problems with noisy targets. This increase likely stems from the accumulation of ELBO estimation bias, a phenomenon we analyse in detail in Appendix A.8.

	Benchmarks					
	GMM			**Ring**		
Algorithm	**ΔLML**	**MMTV**	**GsKL**	**ΔLML**	**MMTV**	**GsKL**
			Noiseless			
BBVI, MoG ($K = 50$)	0.059 [0.030,0.072]	0.059 [0.038,0.077]	0.0083 [0.0015,0.014]	8.0 [7.2,9.5]	0.51 [0.48,0.53]	0.72 [0.70,0.92]
BBVI, MoG ($K = 500$)	0.053 [0.032,0.10]	0.052 [0.044,0.069]	0.0087 [0.0030,0.013]	8.3 [7.2,10.]	0.47 [0.46,0.49]	0.67 [0.58,0.79]
VBMC	0.72 [0.7,1.4]	0.4 [0.36,0.56]	7.9 [5.8,15]	1.2 [1.1,1.5]	0.53 [0.44,0.64]	9.8 [4.3,33]
S-VBMC (10 runs)	**0.011** [0.006,0.015]	**0.035** [0.03,0.042]	**0.0017** [0.00092,0.002]	0.15 [0.055,0.2]	**0.16** [0.14,0.18]	**0.019** [0.0015,0.032]
S-VBMC (20 runs)	**0.0054** [0.0017,0.012]	**0.028** [0.024,0.031]	**0.00061** [0.00027,0.0014]	**0.024** [0.017,0.031]	**0.14** [0.14,0.14]	**0.0017** [0.0014,0.0028]
			Noisy ($\sigma = 3$)			
BBVI, MoG ($K = 50$)	**0.23** [0.13,0.34]	0.13 [0.097,0.17]	0.030 [0.010,0.095]	4.3 [3.6,4.7]	0.51 [0.48,0.54]	1.1 [0.73,1.3]
BBVI, MoG ($K = 500$)	**0.27** [0.082,0.40]	0.10 [0.097,0.12]	**0.019** [0.012,0.031]	4.7 [4.1,5.4]	0.93 [0.92,0.94]	48. [32.,49.]
VBMC	1 [0.71,1.4]	0.46 [0.41,0.59]	10 [7.6,18]	1.3 [0.89,1.8]	0.62 [0.49,0.69]	38 [5,1.9e+02]
S-VBMC (10 runs)	**0.29** [0.18,0.42]	0.11 [0.098,0.14]	**0.012** [0.0044,0.018]	**0.38** [0.33,0.48]	0.22 [0.2,0.22]	**0.0097** [0.0047,0.014]
S-VBMC (20 runs)	0.57 [0.49,0.62]	**0.083** [0.071,0.089]	**0.0028** [0.0019,0.0061]	0.68 [0.56,0.71]	**0.17** [0.17,0.18]	**0.0053** [0.0027,0.0079]

Table A.1: Comparison of S-VBMC, VBMC, and BBVI performance on synthetic benchmark problems. Values show median with interquartile ranges in brackets. Bold entries indicate best median performance; multiple entries are bolded when interquartile ranges overlap with the best median.

	Benchmarks					
	Multisensory model ($\sigma = 3$)			**Neuronal model**		
Algorithm	**ΔLML**	**MMTV**	**GsKL**	**ΔLML**	**MMTV**	**GsKL**
BBVI, MoG ($K = 50$)	1.7 [1.6,4.3]	0.11 [0.098,0.13]	0.17 [0.16,0.20]	44. [35.,1.0e+02]	0.60 [0.57,0.63]	20. [18.,23.]
BBVI, MoG ($K = 500$)	1.8 [1.6,2.3]	0.31 [0.28,0.32]	0.53 [0.50,0.54]	1.7e+02 [1.4e+02,2.4e+02]	0.67 [0.65,0.69]	21. [18.,25.]
VBMC	**0.32** [0.17,0.48]	0.18 [0.15,0.21]	0.21 [0.14,0.27]	3 [2.9,3.2]	0.32 [0.31,0.33]	1.4e+02 [75,3.8e+02]
S-VBMC (10 runs)	0.94 [0.85,0.98]	0.1 [0.089,0.11]	**0.048** [0.037,0.057]	1.8 [1.7,1.9]	**0.15** [0.13,0.19]	**0.79** [0.35,1.5]
S-VBMC (20 runs)	1.2 [1.1,1.3]	**0.079** [0.073,0.081]	**0.037** [0.033,0.039]	**1.5** [1.5,1.6]	**0.13** [0.11,0.14]	**0.52** [0.34,0.65]

Table A.2: Comparison of S-VBMC, VBMC, and BBVI performance on neuronal and multisensory causal inference models.

A.6. Example posterior visualisations

Figure A.1 illustrates how S-VBMC significantly improves the result of a single VBMC run, capturing a larger portion of the target posterior mass as more runs are stacked together.

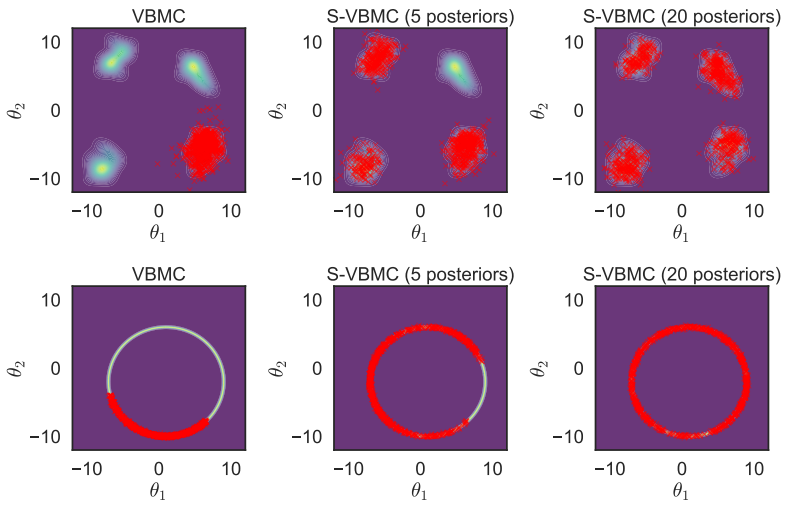
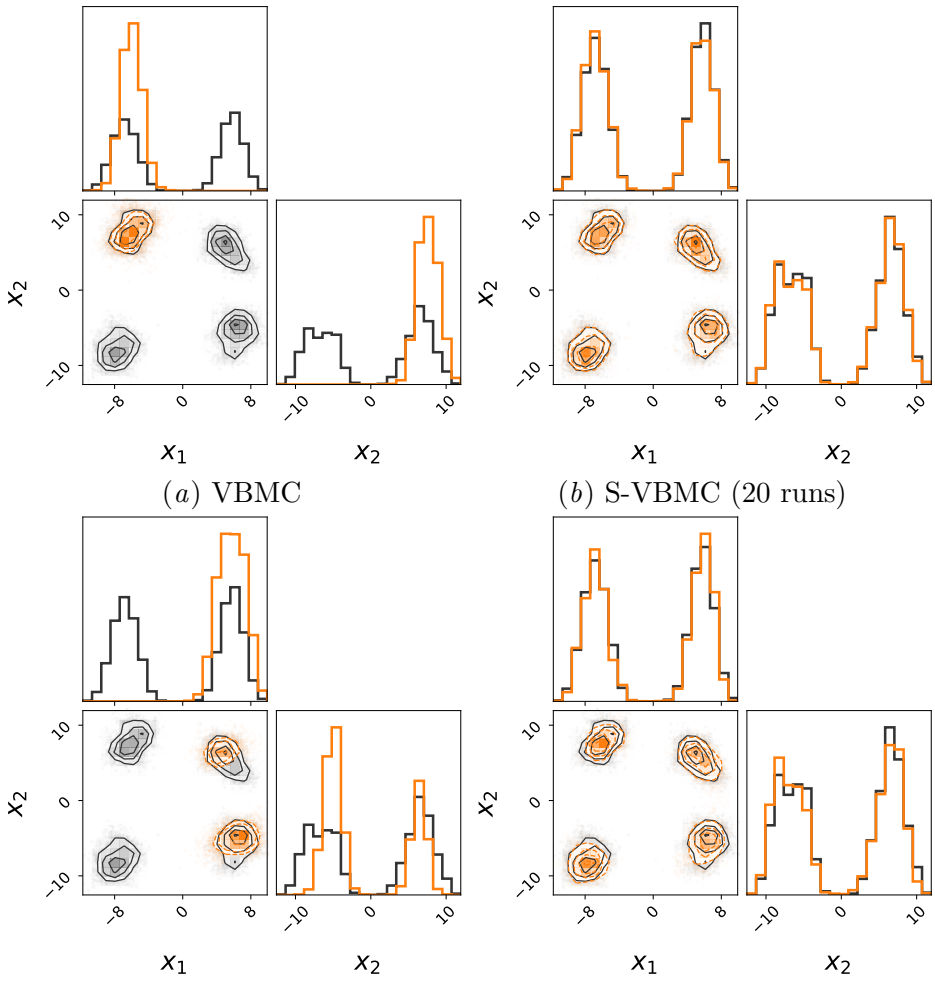


Figure A.1: Examples of overlap between the ground truth and the posterior when combining different numbers of VBMC runs. The red points indicate samples from the posterior approximation, with the target density depicted with colour gradients in the background.

We further use 'corner plots' (Foreman-Mackey, 2016) to visualise exemplar posterior approximations from different algorithms, including S-VBMC, VBMC and BBVI. These plots depict both one-dimensional marginal distributions and all pairwise two-dimensional marginals of the posterior samples. Example results are shown in Figures A.2, A.3, A.4, and A.5, where orange contours and points represent posterior samples obtained from different algorithms while the black contours and points represent ground truth samples. S-VBMC consistently improves the posterior approximations over standard VBMC and generally outperforms BBVI, showing a closer alignment with the target posterior.

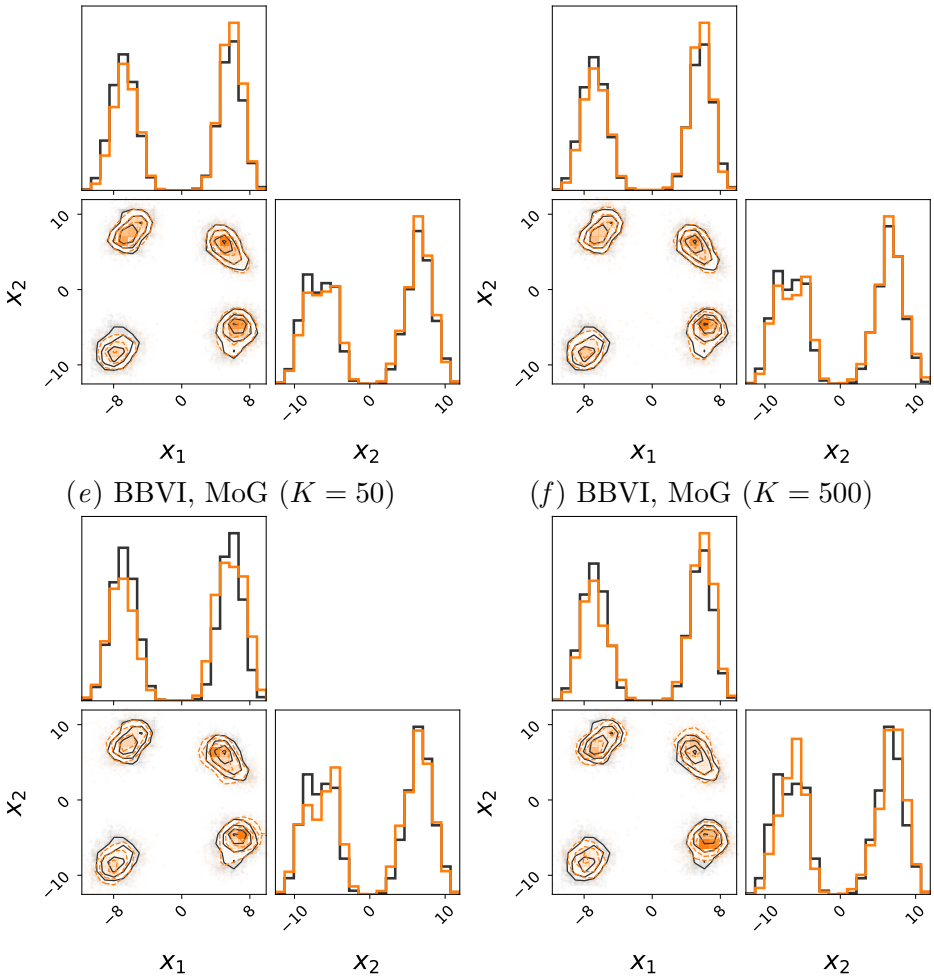


(a) VBMC

(b) S-VBMC (20 runs)

(c) VBMC (noisy)

(d) S-VBMC (20 runs, noisy)



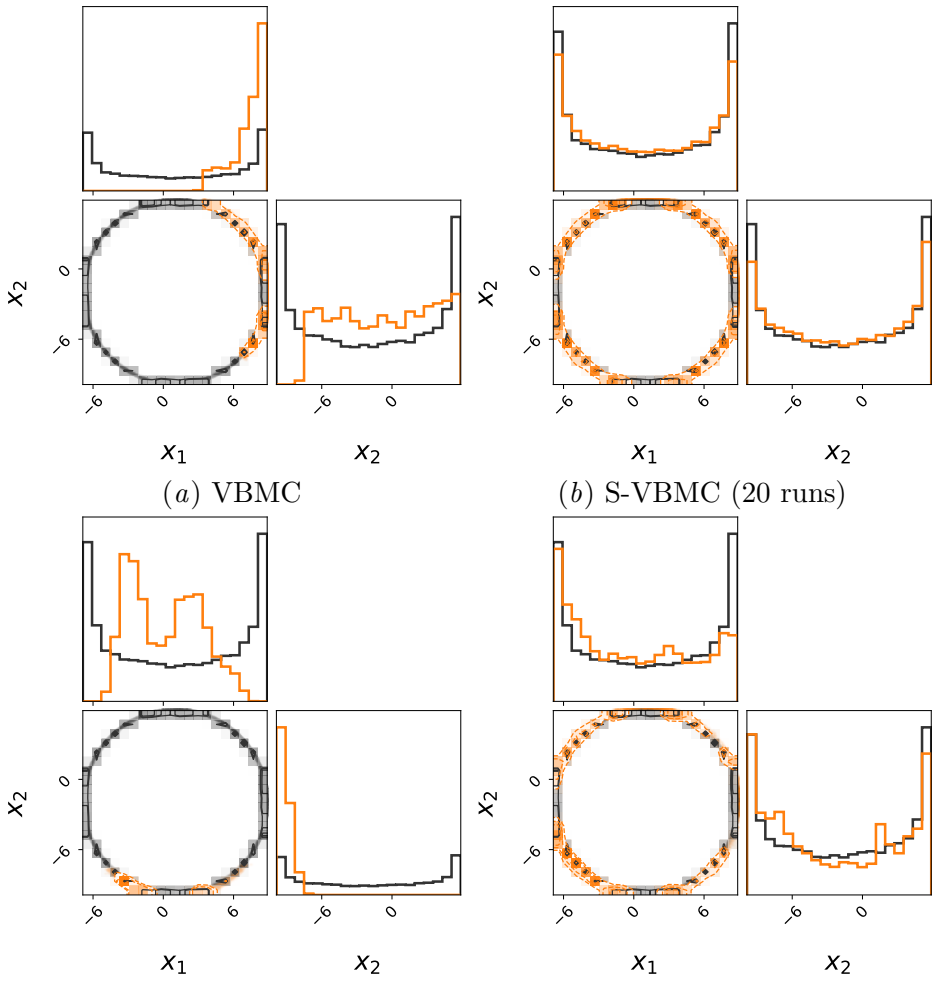
(e) BBVI, MoG ($K = 50$)

(f) BBVI, MoG ($K = 500$)

(g) BBVI, MoG ($K = 50$), noisy

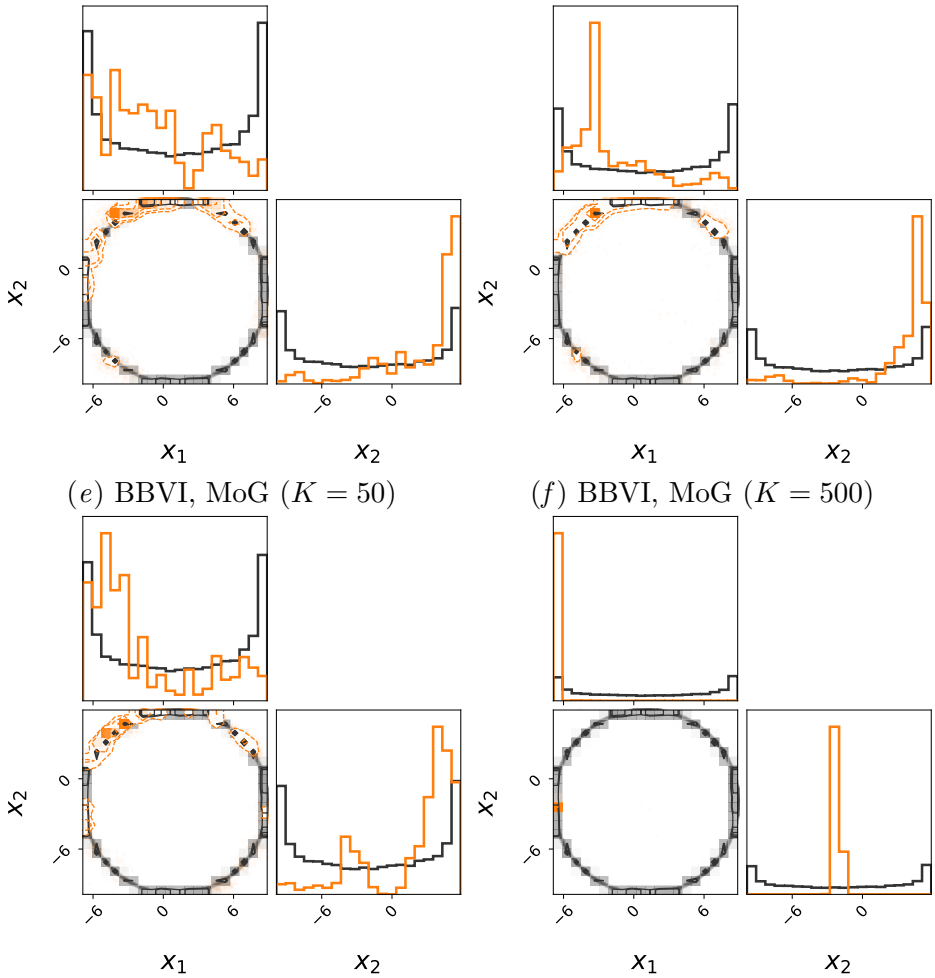
(h) BBVI, MoG ($K = 500$), noisy

Figure A.2: GMM ($D = 2$) example posterior visualisation.



(a) VBMC (b) S-VBMC (20 runs)

(c) VBMC (noisy) (d) S-VBMC (20 runs, noisy)



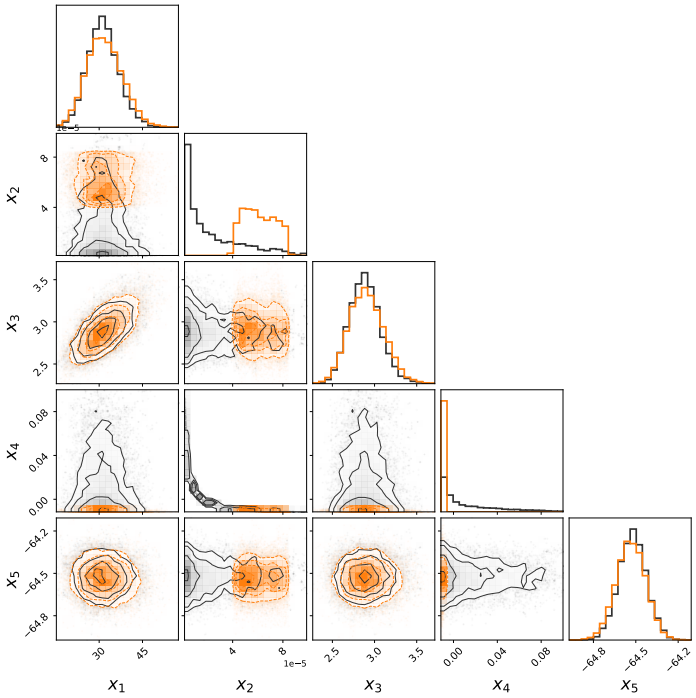
(e) BBVI, MoG ($K = 50$)

(f) BBVI, MoG ($K = 500$)

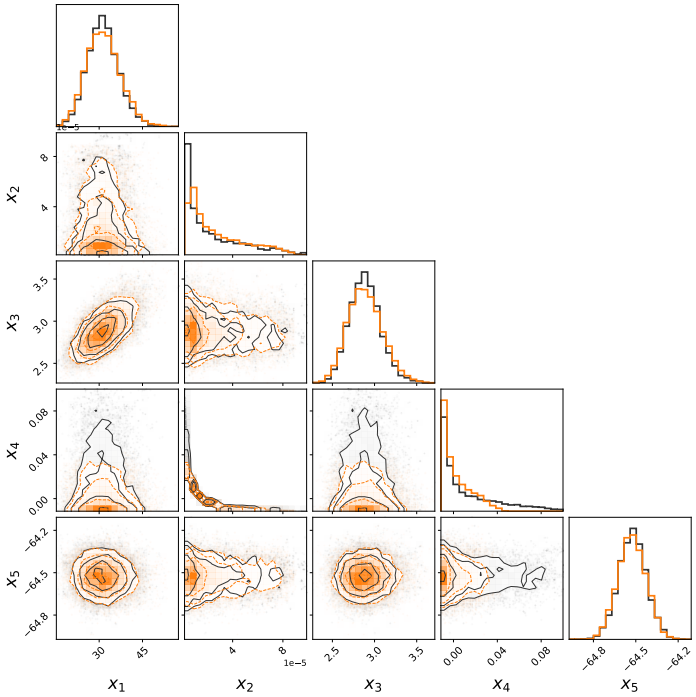
(g) BBVI, MoG ($K = 50$), noisy

(h) BBVI, MoG ($K = 500$), noisy

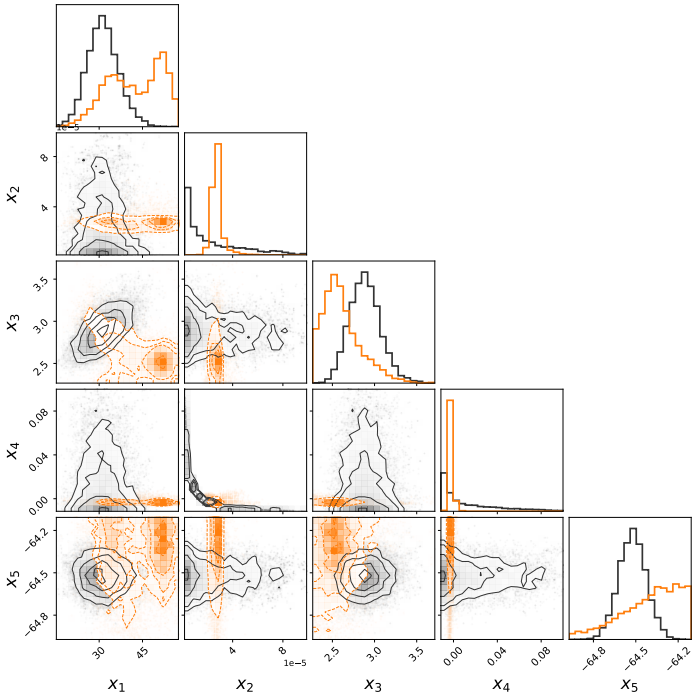
Figure A.3: Ring ($D = 2$) example posterior visualisation.



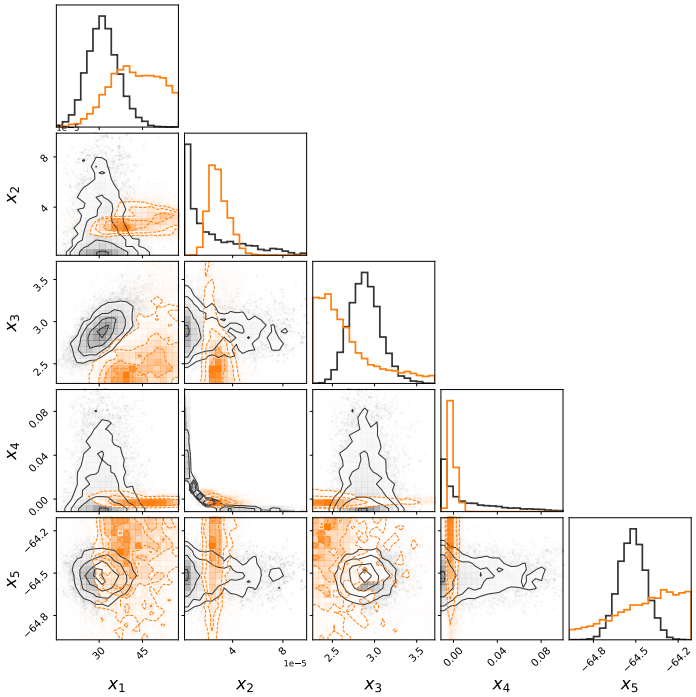
(a) VBMC



(b) S-VBMC (20 runs)

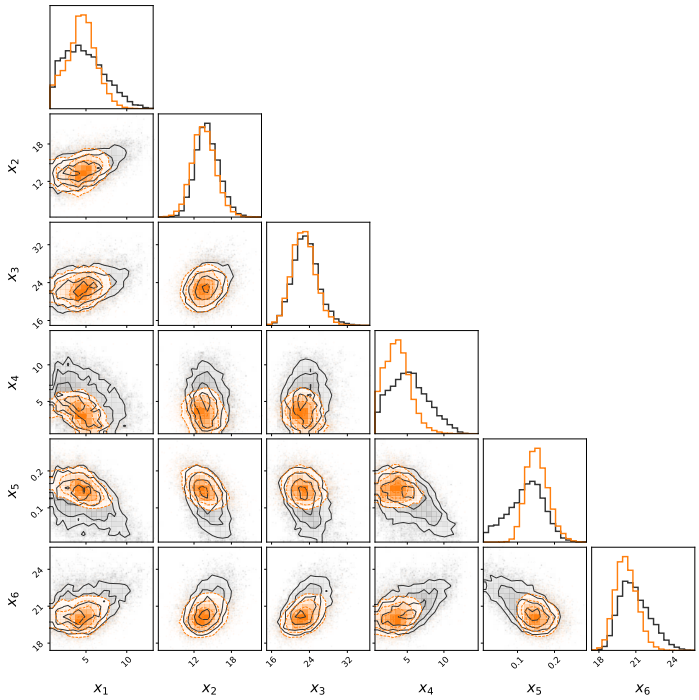


(c) BBVI, MoG ($K = 50$)

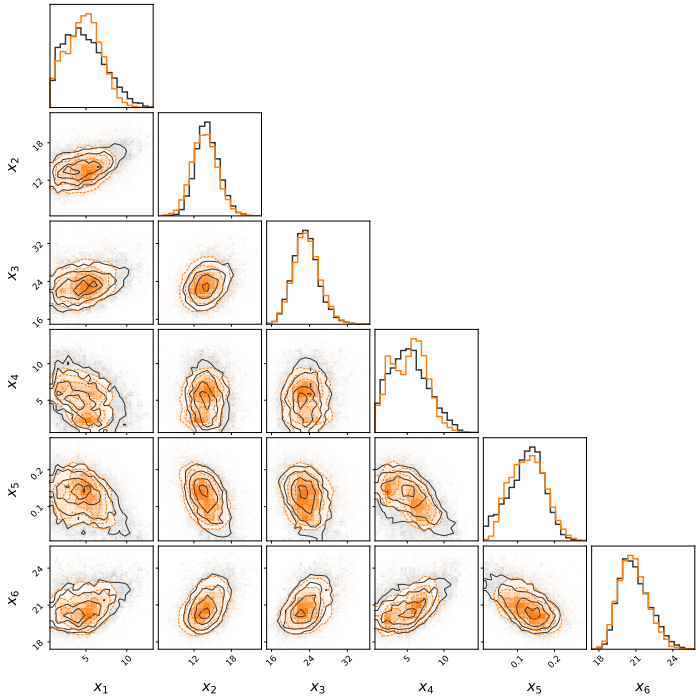


(d) BBVI, MoG ($K = 500$)

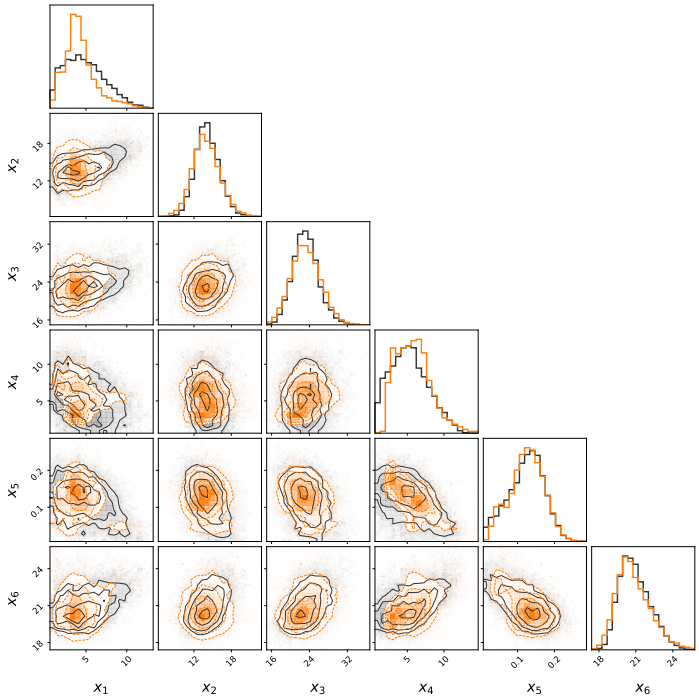
Figure A.4: Neuronal model ($D = 5$) example posterior visualisation.



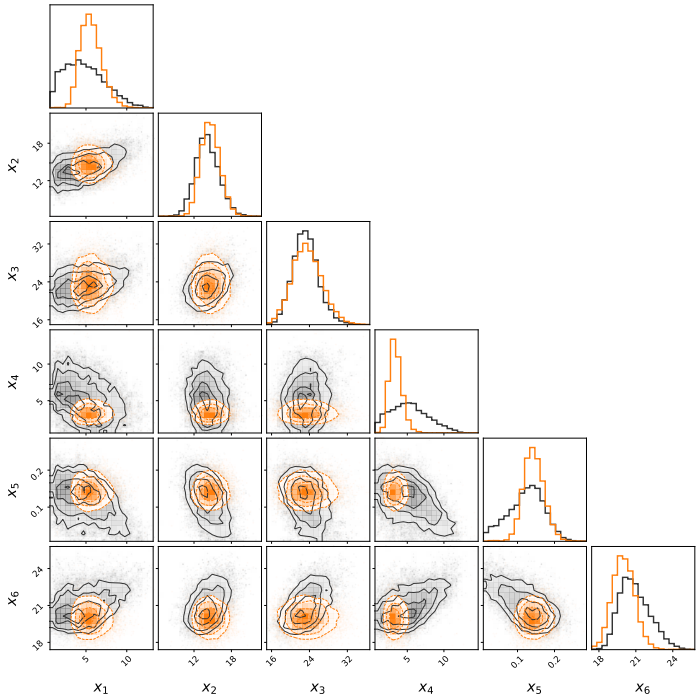
(a) VBMC



(b) S-VBMC (20 runs)



(c) BBVI, MoG ($K = 50$)



(d) BBVI, MoG ($K = 500$)

Figure A.5: Multisensory model ($D = 6, \sigma = 3$) example posterior visualisation.

A.7. Computational overhead

In this appendix we present details about the additional computational cost (quantified as compute time) introduced by S-VBMC on top of VBMC.

Figure A.6 illustrates how S-VBMC introduces a relatively small computational overhead, even when comparing the post-process cost of S-VBMC with the average cost of *one* VBMC run, under the idealized condition where the M VBMC runs happen all in parallel.[2] In particular, running our algorithm with $M \approx 10$ – which vastly improves the resulting posterior, as shown in Section 4 and Appendices A.5 and A.6 – still adds a small amount of post-processing time to VBMC for all our benchmark problems ($\approx$ 5-15% overhead).

Put together, our results confirm that S-VBMC yields high returns in terms of inference performance at a very marginal cost in terms of compute time.

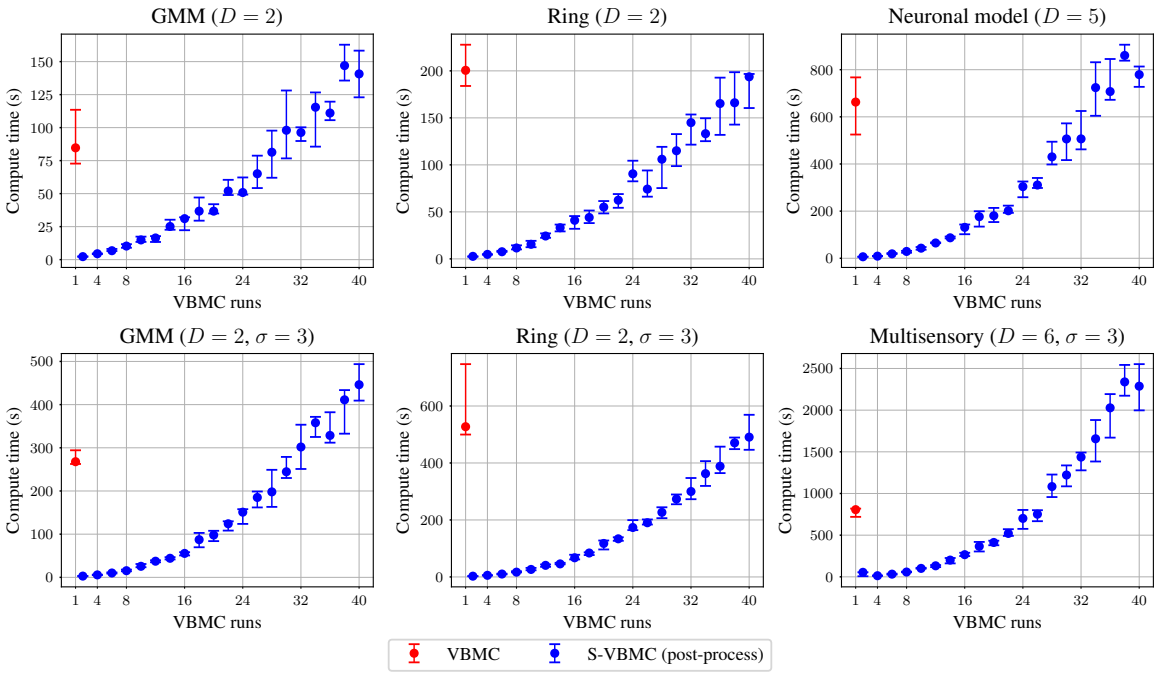


Figure A.6: Compute time of a single VBMC run (red) and post-processing time (i.e. computational overhead) of S-VBMC (blue) plotted as a function of the number of VBMC runs stacked (median and interquartile range). Each subplot represents a different benchmark problem. The compute times values plotted here were taken as we performed the experiments described in Section 4.

2. In practice, completing M VBMC runs will be more expensive due to additional parallelisation costs, making S-VBMC's relative overhead even smaller than what we report here.

A.8. ELBO bias

Here we analyse the ELBO overestimation observed in our results through a simplified example that illustrates one potential mechanism for this bias. While other factors may contribute, this analysis provides insight into why the bias tends to increase with the number of merged VBMC runs.

Consider M VBMC runs that return identical posteriors each with a single component. The stacked posterior takes the form:

$$q_{\tilde{\boldsymbol{\phi}}}(\boldsymbol{\theta}) = \sum_{m=1}^{M} \tilde{w}_m q_m(\boldsymbol{\theta}). \tag{A.4}$$

For each single-component posterior, the expected log-joint is approximated as

$$I_m = \mathbb{E}_{q\boldsymbol{\phi}_m}\left[\log p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})\right] \approx \mathbb{E}_{q_{\boldsymbol{\phi}_m}}\left[f_m(\boldsymbol{\theta})\right] \tag{A.5}$$

where $f_m(\boldsymbol{\theta})$ is the surrogate log-joint from the m-th VBMC run. Since all posteriors share identical parameters, their entropies are equal:

$$\mathcal{H}\left[q_{\boldsymbol{\phi}_1}(\boldsymbol{\theta})\right] = \mathcal{H}\left[q_{\boldsymbol{\phi}_2}(\boldsymbol{\theta})\right] = ... = \mathcal{H}\left[q_{\boldsymbol{\phi}_M}(\boldsymbol{\theta})\right]. \tag{A.6}$$

The stacked posterior is thus a mixture of identical components with different associated values I_m. The optimal mixture weights $\tilde{\mathbf{w}}$ depend solely on the noisy estimates of I_m:

$$\hat{I}_m = \mathbb{E}_{q_{\boldsymbol{\phi}_m}}\left[f_m(\boldsymbol{\theta})\right] = \mathbb{E}_{q_{\boldsymbol{\phi}_m}}\left[\log p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})\right] + \epsilon_m \tag{A.7}$$

where $\epsilon_m \sim \mathcal{N}(0, J_m)$ represents estimation noise with variance J_m. Since all posteriors are identical and derived from the same data and model, differences in expected log-joint estimates arise purely from noise deriving from the Gaussian process surrogates f_m.

Given that entropy remains constant under merging, optimising $\text{ELBO}_{\text{stacked}}$ reduces to selecting the posterior with the highest expected log-joint estimate. If we denote $\hat{I}_{\max} = \max_m \hat{I}_m$, the optimal ELBO becomes

$$\text{ELBO}^*_{\text{stacked}} = \hat{I}_{\max} + \mathcal{H}\left[q_{\tilde{\boldsymbol{\phi}}}(\boldsymbol{\theta})\right]. \tag{A.8}$$

Since the true expected log-joint is identical across posteriors, the optimisation selects the most overestimated value. The magnitude of this overestimation increases with both M and the observation noise for f_m, introducing a positive bias in $\text{ELBO}^*_{\text{stacked}}$ that grows with the number of stacked runs and is more substantial for surrogates of noisy log-likelihoods.

While this simplified scenario does not capture the complexity of practical applications – where posteriors have multiple, non-overlapping components – it illustrates a fundamental issue: if we model each $\hat{I}_{m,k}$ as the sum of the true $I_{m,k}$ and noise, the merging process will favour overestimated components, biasing the final $\text{ELBO}_{\text{stacked}}$ estimate upward.

This hypothesis is substantiated by our results, as we only observe a noticeable bias in problems with noisy targets, where levels of noise in the VBMC estimation of $I_{m,k}$ are non-negligible (note that VBMC outputs an estimate of such noise, see Section 2). Further work is needed to develop debiasing techniques to counteract this tendency.