APPROXIMATE METHODS FOR MARGINAL LIKELIHOOD ESTIMATION

A Dissertation

by

ERIC JASON CHUU

Submitted to the Graduate and Professional School of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Anirban Bhattacharya
Co-Chair of Committee,	Debdeep Pati
Committee Members,	Yang Ni
	Yu Ding
Head of Department,	Brani Vidakovic

August 2022

Major Subject: Statistics

ABSTRACT

We consider the estimation of the marginal likelihood in Bayesian statistics, a essential and important task known to be computationally expensive when the dimension of the parameter space is large. We propose a general algorithm with numerous extensions that can be widely applied to a variety of problem settings and excels particularly when dealing with near log-concave posteriors. Our method hinges on a novel idea that uses MCMC samples to partition the parameter space and forms local approximations over these partition sets as a means of estimating the marginal likelihood. In this dissertation, we provide both the motivation and the groundwork for developing what we call the Hybrid estimator. Our numerical experiments show the versatility and accuracy of the proposed estimator, even as the parameter space becomes increasingly high-dimensional and complicated.

DEDICATION

To my family.

ACKNOWLEDGMENTS

First, I thank my advisor Dr. Anirban Bhattacharya for his gracious guidance and patience throughout my academic journey. His passionate instruction in the Bayesian statistics course during my first year of graduate studies had a profound impact on my academic journey, and I credit this class and of course Dr. Bhattacharya for sparking my interest in Bayesian inference. Throughout my research, no problem ever seemed impossible because of the cornucopia of knowledge that Dr. Bhattacharya possessed. Most importantly, I appreciate the trust he bestowed upon me. I never felt rushed or pressured to prove my worth, and I felt that my opinions and contributions were valued. I am incredibly indebted to Dr. Bhattacharya for my success and development as a statistician throughout the course of this program.

Second, I thank my co-advisor Dr. Debdeep Pati for so willingly taking on the role of aiding and advising me during my academic career. His boundless knowledge in seemingly all areas of statistics was incredibly humbling and kept me motivated to stay curious. I enjoyed many lighthearted and friendly conversation with him, and I am very fortunate to have experienced the mentorship of such an accomplished statistician.

I thank the rest of my committee members, Professor Yang Ni and Professor Yu Ding, for their time and consideration. Their insightful feedback during both my preliminary exam and final exam gave me valuable perspective that allowed me to further hone my research.

I am fortunate to have met many brilliant and kind people throughout the course of my five years in College Station. Patrick Ding has been my go-to person for all things statistics and life-related ever since we met during our department visit back in April 2017. I am glad that we were able to embark on this long, and at times seemingly never-ending, journey together. I will treasure our late night talks that would start out being about research but would often devolve into a discussion about tennis and PC builds.

I am incredibly happy that I had Naveed Merchant as my unexpected confidante and dearest friend during my time in College Station. I am going to miss our lunch breaks, extended walks

around campus, and unapologetically unfiltered rant sessions. Naveed was always so willing to listen and offer his thoughts on problems that I encountered in my research. Without Naveed, there would have been far fewer laughs in Blocker.

I would be remiss if I did not mention my friends Anthony Xue and Crystal Chung, who have undoubtedly had a fair share to endure during these past five years. Their unconditional friendship and support has brought me through many lows, and I owe a great deal of my happiness to them.

I thank Lynse Chock for being so patient with me and my seemingly never-ending list of things to do, both school and triathlon-related. She has put up with a lot from me, and her support has never ceased. I am extremely indebted to how supremely understanding she has been with me.

I thank my parents, Yan-Han Chuu and Wen-Guang Chuu, from the bottom of my heart for their endless support during my academic journey. They witnessed both the highest of highs and the lowest of lows, and their love and encouragement during this time was unwavering. I believe only they knew how emotionally draining parts of my journey were. Undoubtedly, I would not have made it even half this far without them.

Finally, I thank my sister and best friend, Jennifer Chuu. She paved the way for me, both in my growth as a person and as an academic. From her studies in mathematics and statistics, I not only developed my own passion for these subjects, but I also learned the importance of grit and hard work. She taught me how to find the strength to keep battling even when the trials and tribulations seemed insurmountable. I appreciate the compassion she had for me and the advice she gave for me during my journey, even when my cynicism was likely insufferable.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a dissertation committee consisting of Professor Anirban Bhattacharya and Professor Debdeep Pati as co-chairs, Professor Yang Ni of the Department of Statistics, and Professor Yu Ding of the Department of Industrial and Systems Engineering.

The contents of Chapter 3 include excerpts from "A hybrid approximation to the marginal likelihood" (Chuu et al., 2021). This paper was accepted and presented in the 24th International Conference on Artificial Intelligence and Statistics (AISTATS).

Funding Sources

My research was partially supported by research assistantships from NSF DMS (1854731, 1916371) and NSF CAREER (1653404).

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION AND LITERATURE REVIEW

1.1 Introduction

With both the rise in model complexity and the ease with which one can fit a multitude of different models, the challenge of choosing the optimal model among a group of candidate models that could also have generated the data becomes a crucial task. Due to how common the model selection problem is in statistical inference, numerous criteria have been developed for quantifying model uncertainty. Adopting the Bayesian paradigm provides a natural way of evaluating competing models that essentially reduces to computing the marginal likelihood. The vital role this quantity plays is also accompanied by a high degree of difficulty in its computation. Since the marginal likelihood is essentially an integral over a parameter space, this computational burden is compounded when the underlying parameter space is high-dimensional or complicated.

Another important inferential task in Bayesian statistics that inherently relies on the marginal likelihood is Bayesian model averaging. Often, we may have competing models whose posterior probabilities deem them to be equally good models. Choosing a single model over another could result in a loss of valuable information. One compromise we can make is to take into account each model's uncertainty and leverage a model averaging scheme that forms predictions using a weighted average of predictions under their corresponding models, where the weights are the posterior probabilities of the respective models.

Considering the prevalence and usefulness of model selection and model averaging, it is especially necessary to develop and employ methodologies that facilitate efficient computation and scalable inference. Therefore, one of the primary goals of this dissertation is to provide alternative solutions to the marginal likelihood estimation problem that address some of the computational shortcomings of existing methods. In doing so, we hope that our general framework can additionally serve as the foundation for further advancements in Bayesian computation problems. In the following section, we first review and provide some background on marginal likelihood estimation before stating the research questions and contributions.

1.1.1 Review of the marginal likelihood

The marginal likelihood, also called the model evidence, provides a way to quantify the probability of observing the data given a particular model. As such, accurate and efficient computation of the marginal likelihood is of paramount importance for reliable inference. Suppose we observe data y, for which we are considering competing models, $\mathcal{M}_1, \ldots, \mathcal{M}_s$, with corresponding parameters, $u_1, \ldots, u_s$ coming from a parameter space $\mathcal{U}$. Provided that the likelihood function, $p\left(y \mid u_r, \mathcal{M}_r\right)$, indexed by u_r, and the prior distribution for θ_r are both specified, then the posterior distribution of u_r conditional on model $\mathcal{M}_r$ is

$$p\left(u_r \mid y, \mathcal{M}_r\right) \propto \; p\left(y \mid u_r, \mathcal{M}_r\right) p\left(u_r \mid \mathcal{M}_r\right). \tag{1.1}$$

This proportionality, given from Bayes' theorem, is the foundation for many Bayesian computation algorithms and is often sufficient for many MCMC methods, such as the Metropolis-Hastings algorithm. In addition, while there are many well-established ways to obtain samples from posterior distributions that do not require the normalizing constant, there are instances where the accurate calculation of the normalizing constant is crucial. In our work, we are interested in the marginal likelihood of $\mathcal{M}_r$, which is nothing but the normalizing constant of the posterior distribution given in Eq. (1.1). This is defined as the following integral over the parameter space $\mathcal{U}$,

$$p\left(y \mid \mathcal{M}_r\right) = \int_{\mathcal{U}} p\left(y \mid u_r, \mathcal{M}_r\right) p\left(u_r \mid \mathcal{M}_r\right) \, du_r. \tag{1.2}$$

To extend the discussion on model comparison from the previous section, we highlight that the posterior odds ratio and Bayes factor (Jeffreys, 1939) are pivotal quantities that provide an intuitive way to evaluate and choose between competing models. Namely, for two competing models,

$\mathcal{M}_1, \mathcal{M}_2$, the posterior odds ratio is defined as

$$\frac{p\left(\mathcal{M}_1 \mid y\right)}{p\left(\mathcal{M}_2 \mid y\right)}=\frac{p\left(y \mid \mathcal{M}_1\right)}{p\left(y \mid \mathcal{M}_2\right)} \frac{p\left(\mathcal{M}_1\right)}{p\left(\mathcal{M}_2\right)}. \tag{1.3}$$

This can be interpreted as the product of the Bayes factor and the prior odds. In particular, the Bayes factor for models $\mathcal{M}_1$ and $\mathcal{M}_2$ is given as the following ratio of the respective marginal likelihoods,

$$BF_{1,2}=\frac{p\left(y \mid \mathcal{M}_1\right)}{p\left(y \mid \mathcal{M}_2\right)}, \tag{1.4}$$

which offers the interpretation of favoring $\mathcal{M}_1$ when $BF_{1,2}$ is large, and favoring $\mathcal{M}_2$ when $BF_{1,2}$ is small. Therefore, in order to perform this comparison, we require a method for computing the marginal likelihood.

1.1.2 Research question and contribution

Barring specific conjugate settings, the marginal likelihood is analytically intractable in practice, so for most statistical models, accurately calculating the marginal likelihood poses a computationally challenging problem. Since numerical integration ceases to be an efficient solution beyond moderate dimensions, Markov Chain Monte Carlo (MCMC) algorithms offer a viable solution to deal with large-scale, complicated inference problems. In much of the existing literature devoted to estimating this quantity, the recurring idea is to use MCMC samples from the posterior distribution given in Eq. (1.1) to form an asymptotically unbiased estimator of the model evidence in Eq. (1.2). With an abundance of literature devoted to modifying and extending MCMC algorithms in order to meet the needs of the ever-expanding array of statistical models, there is certainly no shortage of ways to handle complex probability distributions and models that are encumbered by a large number of parameters.

Common algorithms for approximating the marginal likelihood include Laplace's method (Tierney and Kadane, 1986), the Adjusted Harmonic Mean estimator (Newton and Raftery, 1994; Lenk, 2009), Corrected Arithmetic Mean estimator (Pajor, 2017), Annealed Importance Sampling (Neal, 2001), Chib's method (Chib, 1995; Chib and Jeliazkov, 2001), (Warp) Bridge Sampling (Meng and

Wong, 1996; Meng and Schilling, 2002), and Nested Sampling (Skilling, 2006). More recently, there have been developments in approximate inference techniques that make way for new methods for estimating the marginal likelihood (Rezende and Mohamed, 2015; Salimans et al., 2015; Salimans and Knowles, 2013). We review some of these methods in Chapter 2. Another resource for a thorough comparison of existing marginal likelihood estimation methods can be found in the review by Friel and Wyse (2012).

While these aforementioned MCMC algorithms are typically easy to set up and provide useful theoretical guarantees, scenarios where the target distribution is highly nontrivial raise issues about both convergence time and accuracy. In addition to the time considerations, the accuracy of the methods as we move to higher dimensions is also of concern. Forming a Monte Carlo approximation to the marginal likelihood in addition to running an MCMC algorithm to obtain posterior samples can quickly accrue error as the dimension of the parameter space increases. Consequently, these algorithms, which may theoretically yield unbiased estimates, may require an exceedingly large number of high-quality samples from the target distribution in practice in order to actually form accurate estimates.

As a partial remedy for some of these concerns, there is extensive literature devoted to addressing the additional issues that arise when using these algorithms, such as different annealing schemes (Beskos et al., 2014), optimal temperature ladder in thermodynamic integration (Oates et al., 2016), and the effects of grid size in path sampling (Dutta and Ghosh, 2013). However, these model settings tend to be problem-specific, further limiting the practicality of some of these approximation schemes. Paired with the increasing complexity and dimensions of modern statistical problems, all of these problems together illustrate the need for more general and scalable methodologies that are less reliant on the MCMC samples themselves.

In contrast to these methods, we propose a novel approach which can be thought of as a hybrid between probabilistic and deterministic procedures. A high level view of our method can be broken down into two major steps:

1. The MCMC samples are used to learn a high-probability partition of the parameter space $\mathcal{U}$.

2. With this partition, we then make a deterministic approximation to the log posterior defined on each of the partition sets.

In essence, we seek to exploit the assumption that the posterior distribution will be far from a uniform looking distribution and instead exhibit concentration around some parameter. Then, learning a partition of the parameter space from the MCMC samples that can identify areas of high posterior mass by carving up these regions more finely yields a benefit that is two-fold. First, we are better equipped to make more precise and refined approximations to the log posterior over each of these regions. More importantly, this partitioning routine redirects our attention away from regions that have little to no contribution to the posterior distribution such that these less finely partitioned regions of the parameter space require fewer associated log-posterior estimates, which saves a tremendous amount of time and computation. Given the use of a probabilistic procedure in step 1, coupled with a deterministic calculation in step 2, we refer to the resulting approximation to the marginal likelihood as the *Hybrid estimator*.

Moreover, our contribution provides a way to bypass the need for the large number of posterior samples that is typically required for accurate estimation of the marginal likelihood. Recall that the typical guarantees for MCMC-based estimates of the marginal likelihood are asymptotic in the number of posterior samples. In many applications, however, evaluating the likelihood can be extremely time consuming, so collecting lots of posterior samples in such cases is prohibitively expensive in both time and computation. Our approach instead seeks to use the MCMC samples holistically to learn a skeleton of the posterior distribution in the form of the aforementioned high-probability partition, a process that we show empirically to be more resistant to issues involving the number and quality of the available MCMC samples. Ultimately, these steps result in a framework for computing the model evidence in high dimensional problems that is both scalable and robust.

After establishing the general framework of novel marginal likelihood estimation approach elicited above, we also discuss extensions and modifications to the Hybrid estimator to demonstrate the ease with which it can be adapted to multiple problem settings. By incorporating recent advancements in high-dimensional approximate integration techniques and making mild assumptions about the shape of the target distribution, we further showcase the Hybrid estimator's widespread applicability and its ability to compete with state of the art algorithms in problems such as Gaussian graphical models and factor models.

Finally, we note that while there exist numerous algorithms that address the marginal likelihood estimation problem, very few of them have publicly available implementations that are both efficient and easy to use. Indeed, there is a general scarcity of packages that can be easily utilized for Bayesian computation. Those that are available tend to require considerable prior knowledge about the algorithm details and are not user-friendly. With this in mind, we provide implementations of our methodology in the form of two `R` packages, both of which are publicly available on Github. These packages prioritize practical convenience so that our proposed methods can be more seamlessly integrated into to a variety of problem settings. Details regarding installation, general usage, and working examples can be found in Sections B.1 and B.2.

1.1.3 Roadmap

In Chapter 2, we review some of the marginal likelihood estimation literature in detail and briefly discuss some of the implementation details for competing algorithms to highlight the differences between our proposed method and other existing methods. In Chapter 3, we present a novel algorithm for marginal likelihood estimation that addresses some of the shortcomings of MCMC-based approximation methods. In Chapter 4, we focus on the marginal likelihood calculation for probability densities that observe a specific shape and extend the algorithm developed in the previous chapter. Here, we also demonstrate the modified algorithm on a diverse array of problems, with a heavy focus on high-dimensional Gaussian graphical models. In Chapter 5, we conclude our work and elaborate on future directions.

2. A REVIEW OF MARGINAL LIKELIHOOD ESTIMATION METHODS

2.1 Introduction

In this chapter, we review some of the popular methods for marginal likelihood estimation and discuss relevant implementation details. The two main goals of this concise literature review is to provide context for some of methodological decisions in our proposed solution, as well as to highlight any shortcomings in these existing methods that we attempt to address and improve upon in the methodology presented in this dissertation.

2.2 Laplace's method

Laplace's method, used in Tierney and Kadane (1986) to compute posterior quantities, assumes that the posterior distribution is unimodal and highly peaked around its mode so that a normal distribution to approximate the posterior distribution. With a large enough sample size and a suitably simple posterior distribution, this assumption is not unreasonable and has been shown to produce accurate estimates in practice. However, in many problem settings that we investigate in this dissertation, we will see that this assumption is too restrictive and that most target distributions are highly non-Gaussian. To construct this estimator, first define the log posterior to be $\ell\left(u\right) = \log\left(p\left(y \mid u\right)p\left(u\right)\right)$. Then, we can use a Taylor expansion about the posterior mode $u^\star$ to obtain a quadratic approximation of $\ell\left(u\right)$,

$$\ell\left(u\right) \approx \ell\left(u^\star\right) + \nabla\ell\left(u^\star\right)'\left(u - u^\star\right) + \frac{1}{2}\left(u - u^\star\right)'\nabla^2\ell\left(u^\star\right)\left(u - u^\star\right).$$

Since the marginal likelihood is simply $p\left(y\right) = \int e^{\ell(u)}du$, we can exponentiate the quadratic approximation above and integrate the resulting normal density to obtain the following final approximation,

$$p\left(y\right) \approx \left(2\pi\right)^d |\hat{\Sigma}|^{1/2} p\left(y \mid u^\star\right)p\left(u^\star\right),$$

where $\hat{\Sigma} = -\nabla^2 \ell(u^\star)$. The upshot of Laplace's method is that it requires fairly little in order to obtain an approximation, with the main requirement being the ability to compute the gradient vector and Hessian matrix of ℓ evaluated at the posterior mode. The posterior mode itself can easily be found through an iterative maximization routine such a Newton's method, which contributes minimally to the computational overhead since we already assume the ability to evaluate the gradient and Hessian. While the assumptions of the Laplace estimator are often too restrictive to directly apply to most problem settings, we note that these ideas can be localized to regions of the posterior and potentially adapted to higher-dimensional problems, as we will see in Chapter 4.

2.3 Harmonic mean estimator

Another frequently referenced estimator is the harmonic mean estimator from Newton and Raftery (1994). Using the following identity,

$$1 = \int p(u)\, du = p(y) \int \frac{1}{p(y \mid u)} p(u \mid y)\, du,$$

we can take the following expectation with respect to the posterior distribution $u \mid y$,

$$\frac{1}{p(y)} = \int \frac{1}{p(y \mid u)} p(u \mid y)\, du = \mathbb{E}_{p(u|y)} \left[\frac{1}{p(y \mid u)} \right].$$

Then, with $u_1, \ldots, u_J$ drawn from the target (posterior) distribution $\gamma(u) = p(u \mid y)$, we can approximate the quantity above using the following importance sampling estimator

$$p(y) \approx \left[\frac{1}{J} \sum_{i=1}^{J} \frac{1}{p(y \mid u_j)} \right]^{-1},$$

which is denoted as the harmonic mean estimator of the marginal likelihood. While it is straightforward to compute this estimator because it only requires likelihood evaluations, it is known to have many problems, such as infinite variance and a lack of sensitivity to the prior choice. Moreover, Raftery et al. (2007) note that the harmonic mean estimator overestimates the marginal likelihood, which is consistent with what we repeatedly observe in our numerical experiments and also characterized as simulation pseudo-bias (Lenk, 2009).

2.4 Arithmetic mean estimator

A similar method that also makes use of importance sampling is the arithmetic mean estimator. The idea is to take draws from the prior $u_j \sim p(u)$ and form the estimator

$$p(y) \approx \frac{1}{J} \sum_{i=1}^{J} p(y \mid u) \to \int p(y \mid u) \, p(u) \, du. \tag{2.1}$$

While this is unbiased, the estimator is extremely ineffective when the posterior distribution is more concentrated relative to the prior. This would render most of the points drawn useless because their likelihood contributions would be close to zero, thus requiring a prohibitively large number of prior draws in order to form an accurate estimate. A modification to this estimator, called the corrected arithmetic mean estimator (Pajor, 2017), seeks to target regions of the parameter space that have higher likelihood. In particular, for $A \subseteq \mathcal{U}$ and $P(A), P(A \mid y) < \infty$, we have $P(A \mid y) = \int_A p(u \mid y) \, du$. Then, we have the following expression for the marginal likelihood

$$p(y) = \frac{1}{P(A \mid y)} \int_{\mathcal{U}} p(y \mid u) \, \mathbb{1}_A(u) \, p(u) \, du = \frac{1}{P(A \mid y)} \mathbb{E}_{p(u)} \left[p(y \mid u) \, \mathbb{1}_A \right]$$

where the expectation is taken with respect to the prior distribution. We can then select an importance function and derive the corrected arithmetic mean estimator as follows,

$$p(y) \approx \frac{1}{\hat{P}(A \mid y)} \frac{1}{J} \sum_{j=1}^{J} \frac{p(y \mid u_j) \, p(u_j) \, \mathbb{1}_A(u_j)}{s(u_j)}$$

where u_j is drawn from the importance sampling distribution. Clearly, in order for this to be effective, A must cover a region for which $P(A \mid y)$ is sufficiently large in order to achieve large values of $p(y \mid u)$. In practice, $\hat{P}(A \mid y)$ is estimated through posterior samples, and the samples from the importance function should be easily obtainable. In many of the simulation studies that follow, we include the corrected arithmetic mean estimator in the results because of its accuracy when the problem setting is not too complicated.

2.5 Chib's method

Another popular method for estimating the marginal likelihood is Chibs's method (Chib, 1995), which makes use of the following identity

$$p\left(y\right) = \frac{p\left(y \mid u^{\star}\right) p\left(u^{\star}\right)}{p\left(u^{\star} \mid y\right)},$$

where $u^{\star} \in \mathcal{U}$. Taking the logarithm, we have

$$\log p\left(y\right) = \log p\left(y \mid u^{\star}\right) + \log p\left(u^{\star}\right) - \log p\left(u^{\star} \mid y\right).$$

By assumption (and in most setups), the prior and likelihood, $\log p\left(u^{\star}\right)$ and $\log p\left(y \mid u^{\star}\right)$, can be directly evaluated. The last term $\log p\left(u^{\star} \mid y\right)$ presents a more challenging quantity which can be approximated using the output of a Gibbs sampler. In particular, for $u = \left(u_1, \ldots, u_d\right)' \in \mathbb{R}^d$, we have the following factorization $p\left(u \mid y\right) = p\left(u_1 \mid u_{2:d}, y\right) \cdots p\left(u_2 \mid u_{3:d}, y\right) \cdots p\left(u_d \mid y\right)$, where each factor in the product can be estimated using its corresponding Gibbs sampler output. The downside of this method is that it requires the full conditional distributions for the parameters, in addition to the MCMC samples, which can be cumbersome as the dimension of the parameter space grows. There exists an extension of this method from Chib and Jeliazkov (2001) that increases the flexibility of the algorithm to allow for Metropolis-Hastings output, rather than a Gibbs sampler, but in the following analyses we provide only limited investigation into this method.

2.6 Annealed importance sampling

Annealed importance sampling (AIS) (Neal, 2001) leverages ideas from tempering schemes and importance sampling to obtain an estimate for the marginal likelihood. We first consider a general setting where importance sampling can be used to estimate the ratio of normalizing constants, $\mathcal{Z}_f, \mathcal{Z}_g$ of densities f and g, respectively,

$$\frac{\mathcal{Z}_f}{\mathcal{Z}_g} = \frac{1}{\mathcal{Z}_g}\int f\left(u\right)du = \int \frac{f\left(u\right)}{g\left(u\right)}\frac{g\left(u\right)}{Z_g}du = \mathbb{E}_g\left[w\left(u\right)\right] \approx \frac{1}{J}\sum_{i=1}^{J} w\left(u_j\right)$$

Here, the importance weights are the ratio of the unnormalized densities, $w\left(u_j\right) = f\left(u_j\right)/g\left(u_j\right)$. The AIS method proposes the use of a tempering scheme $\{t_1, \ldots, t_m\}$, where $0 = t_1 < t_2 < \cdots < t_m = 1$, and $p_j\left(u \mid y\right) \propto p\left(u\right)^{1-t_j} p\left(u \mid y\right)^{t_j}$. In addition, there must also be transition kernels, T_j, from p_j to p_{j+1}, with invariant p_j, for $j = 1, \ldots, m-1$. Similarly, define the reverse transition kernel to be $\tilde{T}_j$. From this setup, we see that p_1 is the prior distribution and p_m is the posterior distribution. Defining f and g to be

$$\begin{aligned} f\left(u_1, \ldots, u_m\right) &= p_m\left(u_m\right)\tilde{T}_{m-1}\left(u_m, u_{m-1}\right) \times \cdots \times \tilde{T}_1\left(u_2, u_1\right), \\ g\left(u_1, \ldots, u_m\right) &= p_1\left(u_1\right)T_1\left(u_1, u_2\right) \times \cdots \times T_{m-1}\left(u_{m-1}, u_m\right), \end{aligned}$$

we have the interpretation that f transitions from the (target) posterior distribution p_m to the prior distribution p_1, whereas g transitions from the prior distribution to posterior distribution. Eventually, we can show that by using these definitions for f and g, the marginal likelihood can be estimated by taking the following average of the importance weights,

$$p\left(y\right) \approx \frac{1}{J}\sum_{i=1}^{J} w\left(u_j\right).$$

While this method has seen success in cases where the posterior distribution is complicated and in modern applications like general additive models and variational autoencoders (Wu et al., 2017), the AIS method has more model and hyperparameter settings that require careful tuning, which ultimately contribute to the time complexity of the overall algorithm. The choice of the transition kernel and the choice of the temperature ladder from $\{t_1, \ldots, t_m\}$ are both also important. With this multitude of considerations in mind, the task of adapting and modifying this algorithm to different problems becomes slightly cumbersome.

2.7 Nested Sampling

As seen in the case of the arithmetic mean estimator, we can view the marginal likelihood written in Eq. (2.1) as an expectation of the likelihood taken with respect to the prior distribution. Skilling (2006) proposed the nested sampling method which can be derived by first writing the marginal likelihood as

$$p\left(y\right) = \int p\left(y \mid u\right) p\left(u\right) du = \int p\left(y \mid u\right) dX,$$

where $dX = p\left(u\right) du$. Define the function

$$X\left(\lambda\right) = \int_{p(y|u)>\lambda} p\left(u\right) du,$$

which is monotonic decreasing from 1 to 0, so the inverse function exists. It is then easier to integrate the inverse of $X\left(\lambda\right)$, rather than to integrate over the parameter space, which may be high-dimensional. If we denote the inverse function as $q\left(X\right)$ so that $q\left(X\left(\lambda\right)\right) = \lambda$, then we have

$$p\left(y\right) = \int_0^1 q\left(X\right) dX \approx \sum_{i=1}^{I} \left(X_i - X_{i+1}\right) p\left(y \mid u_j\right).$$

The final approximation is the result of a numerical approximation to the one-dimensional integral, where $0 < X_I < \cdots < X_2 < X_1 < 1$. Note that q is typically intractable and must be approximated in most cases. Despite the simplified form of the integral, Nested sampling has substantial computational costs associated with obtaining prior samples that satisfy $p\left(y \mid u\right) > \lambda$, for a given value of λ. This constrained sampling procedure often relies on MCMC sampling, which can further exacerbate the time complexity. Extensions such as nested importance sampling (Chopin and Robert, 2010) seek to remedy this by introducing instrumental prior and likelihood functions, $\{\tilde{p_u}, \tilde{L}\}$ that make generating samples under the constraint $\tilde{L}\left(u\right) > \lambda$ easier than under the original prior and likelihood.

2.8 Bridge sampling estimator

The final estimator that we review is the bridge sampling estimator, which we discuss in greater detail because of the practicality of the algorithm, both in how general it is and how easily it can be integrated into an existing problem. These are qualities that our proposed marginal likelihood estimation scheme strives toward. As is commonly the case in Bayesian inference, we consider densities that are known up to a normalizing constant, $p_i\left(u\right) = q_i\left(u\right)/c_i, u \in \mathcal{U}_i \subset \mathbb{R}^d, i = 1, 2$. In the original bridge sampling algorithm from Meng and Wong (1996), the quantity of interest is the ratio of these normalizing constants, c_1/c_2. They make use of the following identity to form an importance sampling ratio,

$$\frac{c_1}{c_2} = \frac{E_{q_2}\left[q_1\left(u\right)h\left(u\right)\right]}{E_{q_1}\left[q_2\left(u\right)h\left(u\right)\right]}.$$

Here h is the *bridge function* defined on the common support of p_1 and $p_2, \mathcal{U}_1 \cap \mathcal{U}_2$, that satisfies

$$0 < \left|\int_{\mathcal{U}_1 \cap \mathcal{U}_2} h\left(u\right)p_1\left(u\right)p_2\left(u\right)du\right| < \infty.$$

While the target quantity in the seminal paper is different, a small modification allows us to use the same algorithm to approximate the normalizing constant of a single density instead. We first note the following identity:

$$\begin{aligned}
p\left(y\right) &= \frac{\int p\left(y \mid u\right)p\left(u\right)h\left(u\right)g\left(u\right)du}{\int p\left(y \mid u\right)p\left(u\right)h\left(u\right)g\left(u\right)du} \cdot p\left(y\right) \\
&= \frac{\int p\left(y \mid u\right)p\left(u\right)h\left(u\right)g\left(u\right)du}{\int \frac{p(y|u)p(u)}{p(y)}h\left(u\right)g\left(u\right)du} \\
&= \frac{\int p\left(y \mid u\right)p\left(u\right)h\left(u\right)g\left(u\right)du}{\int p\left(u \mid y\right)h\left(u\right)g\left(u\right)du} \\
&= \frac{\mathbb{E}_{g(u)}\left[h\left(u\right)p\left(y \mid u\right)p\left(u\right)\right]}{\mathbb{E}_{p(u|y)}\left[h\left(u\right)g\left(u\right)\right]}.
\end{aligned}$$

Note that we are working with two unnormalized distributions, so we can draw equivalence to the setup in Meng and Wong (1996) by taking $q_1(u) \equiv p\left(y \mid u\right)$ and $q_2\left(u\right) \equiv g\left(u\right)$, which are the posterior and proposal distributions, respectively. Then, we can form the following bridge sampling estimator for the marginal likelihood:

$$p\left(y\right) = \frac{\mathbb{E}_{g(u)}\left[h\left(u\right)p\left(y \mid u\right)p\left(u\right)\right]}{\mathbb{E}_{p(u|y)}\left[h\left(u\right)g\left(u\right)\right]} \approx \frac{\dfrac{1}{n_2}\displaystyle\sum_{j=1}^{n_2} h\left(\tilde{u}_j\right)p\left(y \mid \tilde{u}_j\right)p\left(\tilde{u}_j\right)}{\dfrac{1}{n_1}\displaystyle\sum_{i=1}^{n_1} h\left(u_i^\star\right)g\left(u_i^\star\right)}.$$

Here, h denotes the bridge function and g denotes the proposal function. In addition, the collections $\left\{u_1^\star, \ldots, u_{n_1}^\star\right\}$ and $\left\{\tilde{u}_1, \ldots, \tilde{u}_{n_2}^\star\right\}$ are the n_1 and n_2 samples from the posterior distribution $p\left(u \mid y\right)$ and the proposal distribution $g\left(u\right)$, respectively. Since the posterior distribution is known up to a normalizing constant, the product $p\left(y \mid u\right)p\left(u\right)$ can be evaluated. Overstall and Forster (2010) recommend a normal distribution with its first two moments matching those of the posterior distribution for the proposal distribution g so that it can be easily sampled from and evaluated. The optimal bridge function given in Meng and Wong (1996) is

$$h\left(u\right) = C\left(s_1 p\left(y \mid u\right)p\left(u\right) + s_2 p\left(y\right)g\left(u\right)\right)^{-1}$$

where $s_1 = n_1/n, s_2 = n_2/n$, where $n = n_1 + n_2$. Since the bridge function is itself a function of $p\left(y\right)$, the bridge sampling algorithm applies an iterative updating scheme that runs until convergence.

$$\hat{p}\left(y\right)^{(t+1)} = \frac{\dfrac{1}{n_2}\displaystyle\sum_{i=1}^{n_2}\frac{p\left(y \mid \tilde{u}_j\right)p\left(\tilde{u}_j\right)}{s_1 p\left(y \mid \tilde{u}_j\right)p\left(\tilde{u}_j\right) + s_2\hat{p}\left(y\right)^{(t)}g\left(\tilde{u}_j\right)}}{\dfrac{1}{n_1}\displaystyle\sum_{i=1}^{n_1}\frac{g\left(u_i^\star\right)}{s_1 p\left(y \mid u\right)p\left(u_i^\star\right) + s_2\hat{p}\left(y\right)^{(t)}g\left(u_i^\star\right)}}$$

Accuracy of the estimate is dependent on the number of posterior samples and on the overlap between the posterior and proposal distribution. The Bridge sampling algorithm essentially maximizes the overlap between these two distributions during the iterative updates of the marginal likelihood. More details for this algorithm are available in Gronau et al. (2017).

As elicited above, we see that by using the pre-defined proposal distribution from Overstall and Forster (2010) and using the optimal bridge function, the bridge sampling algorithm significantly reduces the burden of the practitioner. In fact, other than the posterior samples, only the definition of the likelihood function and prior are necessary. As stated by the authors of the `bridgesampling` package, the implementation is intended to be a black box algorithm for computing the marginal likelihood. Given its success in many different statistical models and its ease of use, we use the bridge sampling estimator as the primary competitor in the following examples.

3. A HYBRID APPROXIMATION TO THE MARGINAL LIKELIHOOD[*]

3.1 Introduction

Before presenting the Hybrid estimator, we first introduce some preliminary notation that allows our problem setup to be more easily generalized. Suppose γ is a probability density with respect to the Lebesgue measure on $\mathbb{R}^d$ given by

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

In Bayesian inference, $\Phi\left(\cdot\right)$ is typically taken to be a negative log-likelihood function and $\pi\left(\cdot\right)$ is a prior distribution on u, thus making $\gamma\left(\cdot\right)$ the corresponding posterior distribution. However, this interpretation is not necessary for our approach. The marginal likelihood is defined as the normalizing constant of γ, which we can write as the following integral,

$$\mathcal{Z} = \int_{\mathcal{U}} e^{-\Psi(u)}\,du. \tag{3.1}$$

Here, $\Psi\left(u\right) = \Phi\left(u\right) + \left(-\log \pi\left(u\right)\right)$ is the negative log-posterior. Since the objective function Ψ is typically complicated, and the space over which we are integrating tends to be high-dimensional, the ensuing calculations end up being computationally expensive. As stated before, while we can evaluate Ψ, we are typically unable to compute the integral in Eq. (3.1). We can address this problem using two sub-routines that both work to reduce the computational burden of the overall problem. First, we find a partition of the parameter space that gives more attention to (i.e, more finely partitions) regions of the posterior distribution that have high posterior mass. Next, we propose a suitable approximation for Ψ that allows for easier evaluation of the integral over each of the partition sets learned from the previous step. These steps used in conjunction with each other

[*] Part of this chapter is reprinted with permission from “A hybrid approximation to the marginal likelihood” by Eric Chuu, Debdeep Pati, and Anirban Bhattacharya, 2021. Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, 130:3214-3222, Copyright 2021 by the authors.

give us 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. Contrast this methodology with traditional quadrature methods which may needlessly target regions of the parameter space that have little to no posterior concentration, resulting in unnecessary function evaluations that grow more expensive as the dimension of the parameter space increases.

3.2 Deterministic approximation

We first elaborate on our strategy to replace Ψ with an approximation $\widehat{\Psi}$. Our starting point is the following observation: fix $q \in (0,1)$ small and let $A \subseteq \mathcal{U}$ be a compact subset with $\gamma\left(A\right) \geq (1-q)$. Rearranging this equation, one obtains

$$(1-q) \leq \gamma\left(A\right) = \mathcal{Z}^{-1} \int_A e^{-\Psi(u)} du \leq 1,$$

leading to the two-sided bound

$$\int_A e^{-\Psi(u)}\, du \leq \mathcal{Z} \leq \frac{1}{1-q} \int_A e^{-\Psi(u)}\, du. \tag{3.2}$$

We then make the following approximation

$$\log \mathcal{Z} \approx F_A := \log \left[\int_A e^{-\Psi(u)}\, du \right]. \tag{3.3}$$

From Eq. (3.2), it is immediate that

$$|\log \mathcal{Z} - F_A| \leq \log\left(\frac{1}{1-q}\right) \approx q,$$

for q small. Henceforth, we aim to estimate the quantity F_A. This initial approximation step can be thought of as compactifying the parameter space to reduce its entropy. Even if $\mathcal{U}$ itself is compact, γ can be highly concentrated in a region A with vol$\left(A\right) \ll$ vol$\left(\mathcal{U}\right)$, particularly when the posterior exhibits concentration (Ghosal and Van Der Vaart, 2007), so it is judicious to eliminate such low posterior probability regions.

Having compactified the integral domain, our general plan is to replace Ψ with a suitable approximation $\widehat{\Psi}$ on the compact set A. In this chapter, we specifically focus on a piecewise constant approximation of the form

$$\widehat{\Psi}\left(u\right)=\sum_{k=1}^{K}c_{k}^{\star}\cdot\mathbb{1}_{A_{k}}\left(u\right), \tag{3.4}$$

where $\mathcal{A}=\{A_1,\ldots,A_K\}$ is a partition of A, i.e., $A=\bigcup_{k=1}^{K}A_k$ and $A_k\cap A_{k'}=\emptyset$ for all $k\neq k'$, and $c_k^{\star}$ is a representative or candidate value of Ψ within the partition set A_k. To simplify the ensuing calculations, we further restrict ourselves to dyadic partitions in the following discourse so that each of the partition sets is rectangular, $A_k=\prod_{l=1}^{d}[a_k^{(l)},b_k^{(l)}]$. Since the representative value $c_k^{\star}$ is constant in u and A_k can be broken down into d one-dimensional rectangles, we can write the following approximation

$$\int_{A}e^{-\Psi(u)}\,du\approx\int_{A}e^{-\widehat{\Psi}(u)}\,du=\sum_{k=1}^{K}e^{-c_{k}^{\star}}\cdot\mu\left(A_{k}\right), \tag{3.5}$$

which conveniently simplifies to a summation over each of the partition sets. Here, $\mu\left(B\right)=\int_{B}1\,du$ denotes the d-dimensional volume of a set B. We eventually define

$$\widehat{F}_{A}:=\log\left[\int_{A}e^{-\widehat{\Psi}(u)}\,du\right]=\log\left[\sum_{k=1}^{K}e^{-c_{k}^{\star}}\cdot\mu\left(A_{k}\right)\right] \tag{3.6}$$

to be our estimator of F_A, and hence of $\log\mathcal{Z}$. The choice of the piecewise constant approximation is motivated both by its approximation capabilities (Binev et al., 2005) as well as the analytic tractability of the approximating integral in Eq. (3.6). We remark here that the integral remains tractable if a piecewise linear approximation is employed, suggesting a natural generalization of our estimator.

Since F_A is a non-linear functional of Ψ, it is reasonable to question the validity of the approximation in Eq. (3.5), or equivalently, the approximation of F_A with $\widehat{F}_A$ — even if $\widehat{\Psi}$ is a good approximation to Ψ, it is not immediately clear if the same should be true of $\widehat{F}_A$. Using an interpolation trick, we show below that the approximation error $|\widehat{F}_A - F_A|$ can be bounded in terms of a specific distance between $\widehat{\Psi}$ and Ψ. Define

$$F\left(t\right) = \log\left[\int_A e^{-\left(t\Psi(u)+(1-t)\widehat{\Psi}(u)\right)}\,du\right], \quad t \in [0,1].$$

Clearly, $F\left(0\right) = \widehat{F}_A$ and $F\left(1\right) = F_A$, so that

$$F_A - \widehat{F}_A = F\left(1\right) - F\left(0\right) = \int_0^1 F'\left(t\right)dt.$$

Computing F', we get

$$\begin{aligned} F'\left(t\right) &= \frac{-\int_A \left(\Psi\left(u\right) - \widehat{\Psi}\left(u\right)\right) e^{-\left(t\Psi(u)+(1-t)\widehat{\Psi}(u)\right)}\,du}{\int_A e^{-\left(t\Psi(u)+(1-t)\widehat{\Psi}(u)\right)}\,du} \\ &= -\mathbb{E}_{U\sim\pi_t}\left(\Psi\left(U\right) - \widehat{\Psi}\left(U\right)\right), \end{aligned}$$

where π_t is the probability density on A given by

$$\pi_t\left(u\right) \,\propto\, e^{-\left(t\Psi(u)+(1-t)\widehat{\Psi}(u)\right)}, \quad u \in A.$$

Using the integral representation, we can now bound the approximation error,

$$|F_A - \widehat{F}_A| \leq \sup_{t\in[0,1]} \left|\mathbb{E}_{U\sim\pi_t}\left(\Psi\left(U\right) - \widehat{\Psi}\left(U\right)\right)\right|.$$

Interestingly, note that $\pi_1 \,\propto\, \gamma\mathbb{1}_A$ is our target density restricted to A, and $\pi_0\left(u\right) \,\propto\, e^{-\widehat{\Psi}(u)}\,\mathbb{1}_A\left(u\right)$ has normalizing constant $\widehat{F}_A$. The collection of densities $\{\pi_t\}$ can therefore be obtained by continuously interpolating between π_0 and π_1. Piecing together the various approximations, we arrive at the following result.

Proposition 1. *For any compact subset $A \subseteq \mathcal{U}$, we have*

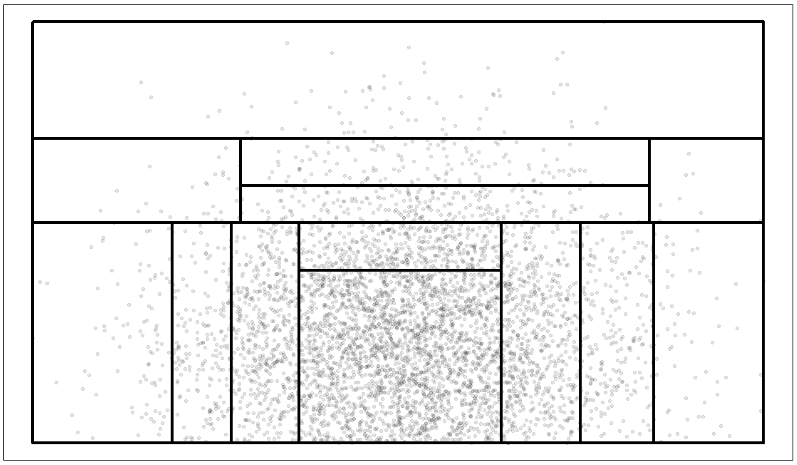
$$|\widehat{F}_A - \log \mathcal{Z}| \leq \sup_{t \in [0,1]} \left| \mathbb{E}_{U \sim \pi_t} \big(\Psi(U) - \widehat{\Psi}(U) \big) \right| + \log \left(\frac{1}{\nu(A)} \right).$$

Here, ν denotes the Lebesgue measure on $\mathbb{R}^d$. The first term in the right hand side above can be further bounded by $\|\Psi - \widehat{\Psi}\|_\infty := \sup_{u \in A} |\Psi(u) - \widehat{\Psi}(u)|$. This conclusion is not restricted to the piecewise constant approximation and can be used for other approximations, such as the piecewise linear one.

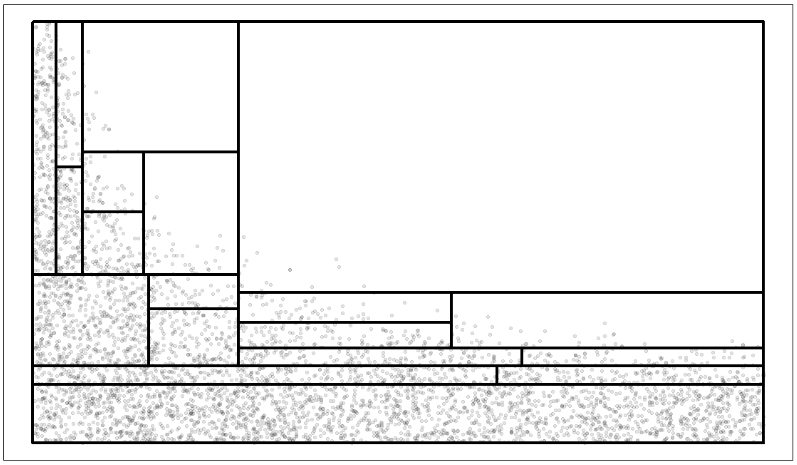
3.3 High probability partitioning of the parameter space

Next, we address the task of obtaining a suitable partition of the parameter space. Clearly, traditional quadrature methods would render this method ineffective, requiring the number of function evaluations to grow exponentially with d. Furthermore, with a posterior distribution that exhibits any degree of concentration, there will indubitably be regions of $\mathcal{U}$ where the posterior probability is close to 0. From a computationally mindful standpoint, it makes sense to then focus on more finely partitioned regions of $\mathcal{U}$ that have high posterior probability. With this in mind, we turn to using samples from γ to obtain such a partition. Specifically, let $u_1, \ldots, u_J$ be samples from γ, e.g., the output of an MCMC procedure. We treat $\{(u_j, \Psi(u_j))\}_{j=1}^J$ as covariate-response pairs and feed them to a tree-based model such as CART (Breiman, 1984), implemented in the `R` package `rpart` (Therneau and Atkinson, 2019), to obtain a dyadic partition. While the MCMC samples are typically used to construct Monte Carlo averages, we instead use them to construct a high probability partition of the parameter space. We assume the capability to evaluate Ψ, which is a very mild assumption since obtaining samples from γ using even a basic sampler like Metropolis–Hastings requires evaluating Ψ. Finally, the above procedure implicitly suggests the compactification A to be a bounding box using the range of posterior samples,

$$A = \bigotimes_{1 \leq l \leq d} \left[\min_{1 \leq j \leq J} \left\{ u_j^{(l)} \right\}, \max_{1 \leq j \leq J} \left\{ u_j^{(l)} \right\} \right].$$



(a) Truncated Normal density: $\gamma\left(u\right)\;\propto\;\mathcal{N}_2\left(0,\sigma^2\lambda^{-1}I_2\right)\cdot\mathbb{1}_{[0,\infty)^d}$



(b) A density of the form: $\gamma\left(u\right)\;\propto\;\exp\left(-nu_1^2u_2^4\right)\pi\left(u\right)$, where $u\in[0,1]^2$

Figure 3.1: Top: bivariate normal distribution truncated to the first orthant. Bottom: a density of the form $\gamma\left(u\right)\;\propto\;\exp\left(-nu_1^2u_2^4\right)\pi\left(u\right)$, where $u\in[0,1]^2$ and $\pi\left(\cdot\right)$ is the uniform measure on $[0,1]^2$. For this simulation, $n=1000$. Both plots show 5000 MCMC samples drawn from γ, overlayed with the resulting partition extracted from a CART model fitted to the covariate-response pairs $\left(u,\Psi\left(u\right)\right)$. The decision tree algorithm finely partitions high-probability regions of the parameter space. Adapted with permission from "A hybrid approximation to the marginal likelihood" by Eric Chuu, Debdeep Pati, and Anirban Bhattacharya, 2021. Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, 130:3214-3222, Copyright 2021 by the authors.

This compactification procedure allows us to simplify ensuing calculations by restricting our focus to regions of the parameter space that exhibit some degree of posterior concentration.

While our initial development of this methodology relied heavily on the `rpart` package for both the tree building algorithm and the partition set corresponding to the fitted decision tree, we eventually created an independent tree building package that more directly facilitates the overall needs of the main marginal likelihood estimation algorithm. Specifically, the `rpart` package does not conveniently return the partition sets as a data structure that we can use, leading to substantial post-processing of the fitted tree object before we could proceed with the rest of the Hybrid algorithm. We highlight a few of the features of our tree building algorithm and package. Most importantly, our implementation directly produces the partition of the bounding box A using the range of the input. In addition, we also improve the tree building algorithm to incorporate a backtracking algorithm, which reduces the computational overhead of the already-expensive recursive routines associated with building decision trees. These modifications ultimately result in a runtime improvement that is approximately ten times that of the `rpart` package.

3.4 Partitioning in two dimensions

Before moving into higher dimensions, we provide an illustration of the process described in the previous section in 2 dimensions, where the partitioning can be easily visualized. Suppose γ is a density on $\mathbb{R}^2$ supported on $\mathcal{U} \subseteq \mathbb{R}^2$, and we are able to draw samples $u_j \sim \gamma$ for $j = 1, \ldots, J$. Forming the pairs, $\{(u_j, \Psi(u_j))\}_{j=1}^{J}$, we then fit a CART model to these points and extract the decision rules, which form a dyadic partition of the aforementioned bounding box $A \subseteq \mathcal{U}$. Denote the partition as $\mathcal{A} = \{A_1, \ldots, A_K\}$, where each $A_k = \prod_{l=1}^{d} [a_k^{(l)}, b_k^{(l)}]$ is a d-dimensional hyperrectangle. Plotting the sampled points and overlaying the partition sets learned from the regression tree, we observe in Figure 3.1 that areas of $\mathcal{U}$ with a high concentration of points coincide with regions that are more finely partitioned by the regression tree model. Taking γ to be a posterior distribution, we are provided with the interpretation that the decision tree is able to target areas of the posterior distribution that have greater posterior mass, which was a desirable trait that we mentioned in the motivation for this approach and proves helpful in producing a better approximation. Equipped with the partition $\mathcal{A}$, we need only to determine the representative point of each partition set in order to form the approximation to Ψ.

Recall that the CART model fits a constant for each point within a given partition set. At any given stage, the CART model will search for the optimal predictor value, $u = (u_1, u_2)$, on which to partition the remaining points such that the sum of squares error (SSE) between the response, $\Psi(u)$, and the predicted constant is minimized. In particular, to partition data into two regions A_1 and A_2, the objective function is given as

$$SSE = \sum_{u_i \in A_1} \left(\Psi(u_i) - c_1\right)^2 + \sum_{u_i \in A_2} \left(\Psi(u_i) - c_2\right)^2 . \tag{3.7}$$

Upon minimization of the SSE, the resulting partition sets A_1 and A_2 have fitted values c_1 and c_2, respectively. For each partition set $A_k \in \mathcal{A}$, an intuitive first choice for the representative point $c_k^\star$ is the fitted value for A_k returned by the tree-fitting algorithm, and indeed if we were to follow this two-step process of using CART to obtain both the partition and the fitted values for each of the partition sets and then plug these into Eq. (3.5), we obtain a valid approximation to the marginal likelihood.

3.4.1 Conjugate 2-d example

As a brief illustration of our this initial version of the approximation scheme, we consider the simple conjugate normal model, where the data $y_1, \ldots, y_n$, conditional on the parameters $(\mu, \sigma^2) \in \mathbb{R}^2$, are drawn from a normal distribution, $y_{1:n} \mid \mu, \sigma^2 \sim \mathcal{N}(\mu, \sigma^2)$. We consider the hierarchical prior, $\mu \mid \sigma^2 \sim \mathcal{N}(m_0, \sigma^2/w_0)$, $\sigma^2 \sim \mathcal{IG}(r_0/2, s_0/2)$, where $\mathcal{IG}(\cdot, \cdot)$ denotes the inverse-gamma distribution. In order to compute the Hybrid estimator, we require samples from the posterior distribution and a way to evaluate Ψ. In this example, the posterior distribution of $u = (\mu, \sigma^2)$ is known: $\mu \mid \sigma^2, y_{1:n} \sim \mathcal{N}(m_n, \sigma^2/w_n)$ and $\sigma^2 \mid y_{1:n} \sim \mathcal{IG}(r_n/2, s_n/2)$, so we can draw exact posterior samples. Since the likelihood and prior are specified, the evaluation of Ψ is straightforward. Namely,

Table 3.1: Normal inverse-gamma example. We report the mean, standard deviation, average error (AE, truth - estimated), and the root mean squared error (RMSE), taken over 100 replications. Each replication has 50 observations and 1000 posterior samples. The true log marginal likelihood is -113.143. Estimators include the Harmonic Mean estimator (HME), Corrected Arithmetic Mean estimator (CAME), Bridge Sampling estimator (BSE), and the Hybrid estimator (HybE). Adapted with permission from "A hybrid approximation to the marginal likelihood" by Eric Chuu, Debdeep Pati, and Anirban Bhattacharya, 2021. Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, 130:3214-3222, Copyright 2021 by the authors.

	TRUTH	**HME**	**CAME**	**BSE**	**HYB**
MEAN	-113.143	-104.762	-112.704	-113.143	-113.029
SD	0	0.733	0.048	0.006	0.025
AE	0	-8.381	-0.439	0	-0.114
RMSE	0	8.431	0.441	0.006	0.117

$$\Psi\left(\mu, \sigma^2\right) = -\log\left\{\left[\prod_{i=1}^{n} \mathcal{N}\left(y_i \mid \mu, \sigma^2\right)\right] \times \mathcal{N}\left(\mu \mid m_0, \sigma^2/w_0\right) \times \mathcal{IG}\left(\sigma^2 \mid r_0/2, s_0/2\right)\right\}$$

With this architecture in place, we feed the pairs, $\{(u_j, \Psi(u_j))\}_{j=1}^{J}$, through CART to obtain a partition over the parameter space and each partition set's representative point. Then, we use Eq. (3.5) to compute the final approximation. Table 3.1 shows results for the Hybrid estimator and a number of other competing methods. Here, the true log marginal likelihood can be computed in closed form, so we have direct comparisons to the ground truth. All estimators except for the Harmonic Mean estimator give accurate approximations to the log marginal likelihood. While the Hybrid estimator delivers fairly accurate results in this simple example, in the next section we discuss how more careful consideration of the scale of the target quantity can lead to a more appropriate choice for each partition set's representative point. Model details including hyperparameter values for this experiment are given in Section B.3.1.

3.5 Algorithm description

Until this point, the representative point within each partition has simply been the fitted value for each partition as given from the CART model. When $\{(u_j, \Psi(u_j))\}_{j=1}^{J}$ is fed into the tree, it attempts to optimize the sum of squared errors as in Eq. (3.7). Note, however, that our eventual objective is to best approximate the functional $\log \int_A e^{-\Psi}$, and it is not unreasonable to suspect that the optimal value for Ψ within A_k chosen by the regression tree model, which has no knowledge of the functional, may not be a suitable choice for our end goal, especially for higher dimensions. Indeed, even for slightly larger p for $\mu \in \mathbb{R}^p$ in the conjugate example in the previous section, we observed a sharp decrease in accuracy in the Hybrid estimator that quickly took it out of contention. Simulations in other higher dimensional examples confirm this trend, and these inaccuracies become more evident for nontrivial examples and more complicated choices of γ. Before suggesting a remedy, we offer some additional understanding into the approximation mechanism that guides us toward an improved choice. To that end, write $\widehat{F}_A$ from Eq. (3.6) as

$$\begin{aligned}\widehat{F}_A &= \log \left[\sum_{k=1}^{K} e^{-c_k^\star} p_k \right] + \log \mu\left(A\right) \\ &:= \widehat{G} + \log \mu\left(A\right),\end{aligned}$$

where $\mu\left(B\right) = \mathrm{vol}\left(B\right)$ is the Lebesgue measure of a Borel set B, and $p_k := \mu\left(A_k\right)/\mu\left(A\right)$. We can also write $F_A = G + \log \mu\left(A\right)$, with

$$\begin{aligned}G &:= \log \left[\frac{1}{\mu\left(A\right)} \int_A e^{-\Psi(u)} du \right] \\ &= \log \left[\sum_{k=1}^{K} p_k \frac{1}{\mu\left(A_k\right)} \int_{A_k} e^{-\Psi(u)} du \right] \\ &= \log \left[\sum_{k=1}^{K} e^{-c_k} p_k \right],\end{aligned}$$

where

$$e^{-c_k} = \frac{1}{\mu\left(A_k\right)} \int_{A_k} e^{-\Psi(u)} du = \mathbb{E}_{U_k \sim \text{Unif}\left(A_k\right)} \left[e^{-\Psi(U_k)} \right].$$

Thus, for $\widehat{G}$ to approximate G, we would ideally like to have each $c_k^\star$ chosen so that $e^{-c_k^\star}$ targets e^{-c_k}. Importantly, the above exercise suggests the appropriate scale to perform the approximation – rather than working in the linear scale as in Eq. (3.7), it is potentially advantageous to work in the exponential scale.

3.5.1 Choosing the representative point

Based on the above discussion, we define a family of objective functions

$$Q_k\left(c\right) = \sum_{u \in A_k} \frac{\left| e^{-\Psi(u)} - e^{-c} \right|}{e^{-\Psi(u)}}, \quad c \in A_k, \tag{3.8}$$

one for each partition set A_k returned by the tree, and set $c_k^\star = \text{argmin}_c Q_k\left(c\right)$. We experimented with a number of different criteria and objective functions before zeroing in on the above relative error criterion in the exponential scale. Conveniently, minimizing (3.8) is a weighted ℓ_1 problem and admits a closed-form solution.

Thus, our overall algorithm can be summarized as follows. We obtain (posterior) samples $u_1, \ldots, u_J$ from γ, and feed the collection of pairs $\left\{\left(u_j, \Psi\left(u_j\right)\right)\right\}_{j=1}^{J}$ through a regression tree to partition the bounding box A defined by the range of the range of the samples. Then, rather than using the default fitted values for each partition set returned by the tree, we take the representative value $c_k^\star$ within each A_k as the minimizer of Q_k. Each $c_k^\star$ is then used to compute $\widehat{F}_A$ as in (3.6) – note that $\widehat{F}_A$ can be stably computed using the log-sum-exp trick. Finally, we declare $\widehat{F}_A$ as $\log \widehat{\mathcal{Z}}$, our estimator of $\log \mathcal{Z}$.

It is worth noting that incorporating this additional optimization to find $c_k^\star$ may seem roundabout and can be bypassed by directly modifying the objective function in the CART algorithm to one that operates on an exponential scale, but preliminary results from doing this did not lead to promising results, so we decided to adhere to the default objective function for the tree building procedure. However, our implementation of the tree building algorithm allows for the objective function to be easily exchanged with other user-specified loss functions, so this is an area that can be more thoroughly explored.

3.6 Hybrid algorithm R package

We have also developed an R package, `hybrid`, which implements the Hybrid algorithm as described in this chapter. Provided with functions to sample from the target distribution and evaluate the negative log posterior distribution, the `hybrid` package goes through the steps in Algorithm 1 and performs the relevant calculations to produce the Hybrid estimate to the log normalizing constant. See Section B.1 for more details regarding general use of the `hybrid` package.

3.7 Experiments

In the following experiments, we present a variety of problem settings with the goal of showcasing the versatility of the Hybrid estimator. First, we consider the Bayesian linear regression model under different prior specifications where the true marginal likelihood is known, so we can easily verify the accuracy of any subsequent approximations. We then extend the application of the Hybrid estimator to examples for which the parameter of interest is a $p \times p$ covariance matrix, thus demonstrating the applicability of our methodology even when the parameter space is non-Euclidean. Recall that one of the motivations for developing the Hybrid estimator was the potential scarcity and low-quality nature of samples when the parameter space is nontrivial or difficult to learn. In order to recreate a similar situation, albeit on a lower and more reproducible scale, we examine the performance of the Hybrid estimator alongside competing methods using posterior samples that are few in number, and in some cases, non-exact. Finally, we investigate the marginal likelihood estimation problem in the context of factor models, a setup that is common in many fields of study but has the problem that it does not admit an analytic form of the marginal likelihood.

Algorithm 1: Hybrid Algorithm

```
Input  :  Sampler for the target distribution $\gamma$, method for evaluating $\Psi$, the negative
          log posterior
Output:   Estimate of the logarithm of the normalizing constant of $\gamma$
Sample $u_1, \ldots, u_J \sim \gamma$
Fit a CART model, $\mathcal{T}$, to $(u_1, \Psi(u_1)), \ldots, (u_J, \Psi(u_J))$
Extract the partition $\mathcal{A} = \{A_1, \ldots, A_K\}$ from $\mathcal{T}$ of the bounding box $A$ of $\mathcal{U}$
for $k \in \{1, \ldots, K\}$ do
    $c_k^\star \leftarrow \text{argmin}_{c \in A_k} \log Q_k(c)$
    $\widehat{\mathcal{Z}}_k \leftarrow e^{-c_k^\star} \prod_{l=1}^{d} \left(b_k^{(l)} - a_k^{(l)}\right)$
end
return $\log \widehat{\mathcal{Z}} = \texttt{log-sum-exp}\left(\log \hat{\mathcal{Z}}_1, \ldots, \log \hat{\mathcal{Z}}_K\right)$
```

In addition to the Hybrid estimator (HybE), we examine the following additional estimators: Bridge Sampling estimator (BSE), Warp Bridge Sampling estimator (WBSE), Harmonic Mean estimator (HME), and Corrected Arithmetic Mean estimator (CAME). The BSE and WBSE results are obtained using the `bridgesampling` package (Gronau et al., 2020). Corresponding calculations and formulae for posterior parameters and analytical marginal likelihoods are given in Section B.

3.7.1 Bayesian linear regression

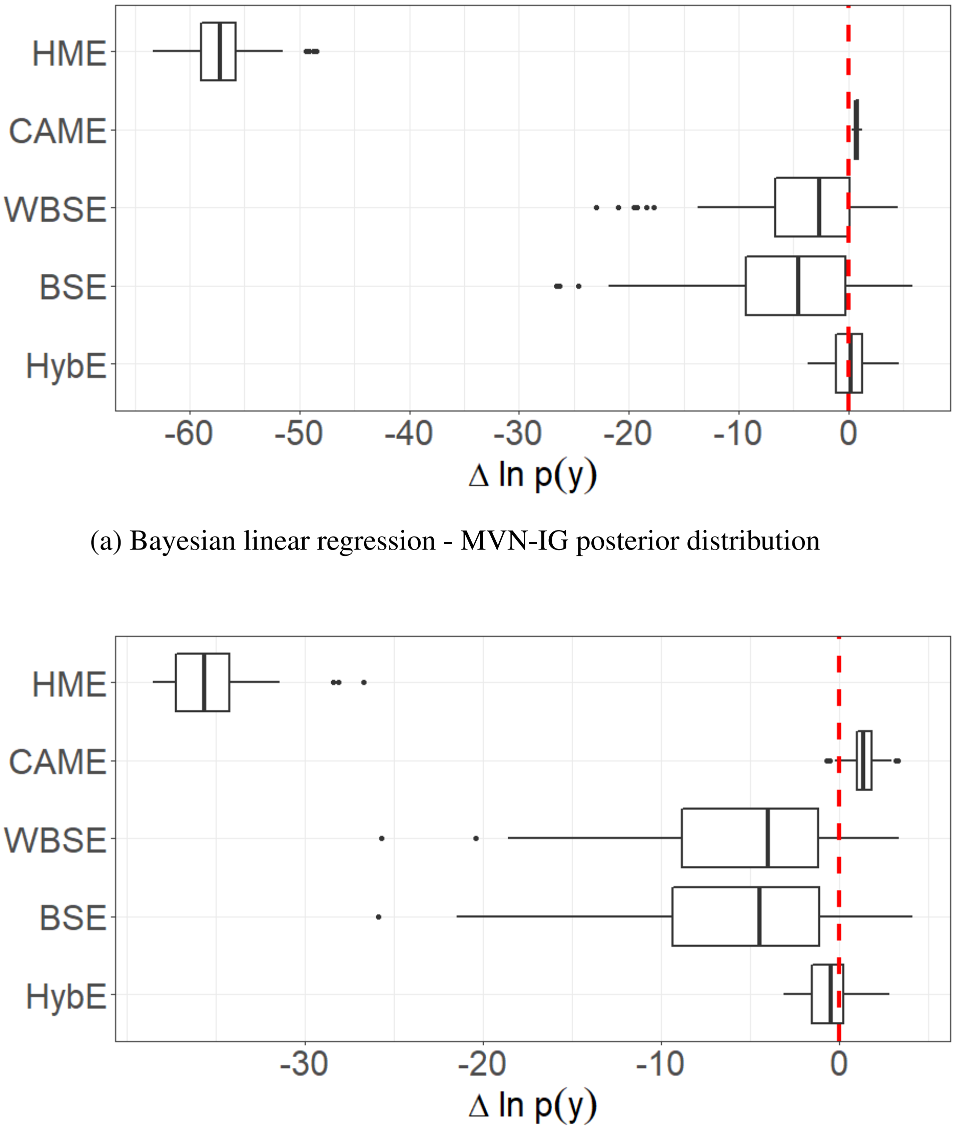
Consider the usual setup of the linear regression model,

$$y = X\beta + \varepsilon, \quad y \in \mathbb{R}^n, \quad X \in \mathbb{R}^{n \times d}, \quad \beta \in \mathbb{R}^d, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_n).$$

In the next two examples, we consider different prior distributions on the parameters β and σ^2.

3.7.1.1 Multivariate normal inverse-gamma model

We assume a multivariate normal inverse-gamma (MVN-IG) prior on (β, σ^2), where $\beta \mid \sigma^2 \sim \mathcal{N}_d(\mu_\beta, \sigma^2 V_\beta)$, $\sigma^2 \sim \mathcal{IG}(a_0, b_0)$. Given this choice of the prior, the posterior distribution also



(a) Bayesian linear regression - MVN-IG posterior distribution

(b) Bayesian linear regression - TN posterior distribution

Figure 3.2: Boxplots of the error (truth - estimate) for the log marginal likelihood in the MVN-IG (left, true $\log p(y)$: -303.8482) and truncated MVN (right, true $\log p(y)$: -250.2755) examples. Both examples correspond to 20-dimensional parameter spaces. Results are reported over 100 simulations, with 100 observations. Estimates are based on 50 MCMC samples. Adapted with permission from "A hybrid approximation to the marginal likelihood" by Eric Chuu, Debdeep Pati, and Anirban Bhattacharya, 2021. Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, 130:3214-3222, Copyright 2021 by the authors.

follows a MVN-IG distribution: $\beta \mid \sigma^2, y \sim \mathcal{N}\left(\mu_n, \sigma^2 V_n\right), \sigma^2 \mid y \sim \mathcal{IG}\left(a_n, b_n\right)$, where the posterior parameters μ_n, V_n, a_n, b_n are known and given in Section B.3.2. In our simulations, we take $d = 19$, so that $u = (\beta, \sigma^2) \in \mathbb{R}^{20}$. Since the log marginal likelihood in this example is well known, we can evaluate each of the estimates against the true value. In Figure 3.2, we plot the errors for each of the estimators when only 50 posterior samples are used for each approximation. The accuracy and standard error of the Hybrid estimator are clearly superior compared to the well-established estimators. We remind readers that the relatively poor results of the BSE and WBSE are unsurprising because the number of available posterior samples is so few, and if we were to increase the number of samples, we fully expect the MCMC-based methods to produce more reliable estimates.

3.7.1.2 Truncated multivariate normal model

Next, we remain in the linear regression setting, but we fix σ^2 and place a multivariate normal prior on β truncated to the first orthant. In particular, $\beta \sim \mathcal{N}_d\left(0, \sigma^2\lambda^{-1}I_d\right) \cdot \mathbb{1}_{[0,\infty)^d}$, where σ^2, λ are known. This produces a posterior distribution of the form,

$$\beta \mid y \sim \mathcal{N}_d\left(\beta \mid Q^{-1}b, Q^{-1}\right) \cdot \mathbb{1}_{[0,\infty)^d},$$

where $Q = \frac{1}{\sigma^2}\left(X'X + \lambda I_d\right)$ and $b = \frac{1}{\sigma^2}X'y$. Then, the marginal likelihood can be written as

$$\begin{aligned} p\left(y\right) &= \int_R \mathcal{N}\left(y \mid X\beta, \sigma^2 I_n\right) 2^{-d} \mathcal{N}\left(\beta \mid 0, \sigma^2\lambda^{-1}I_d\right)\, d\beta \\ &= C \cdot \int_R \det\left(Q\right)^{\frac{1}{2}} e^{-\frac{1}{2}\left(\beta - Q^{-1}b\right)' Q \left(\beta - Q^{-1}b\right)} d\beta. \end{aligned}$$

Here, $R = [0, \infty)^\infty$, and by keeping track of the constants from the likelihood and the truncated normal prior, we have

$$C = 2^d \left(2\pi\right)^{-\frac{n}{2}} \left(\sigma^2\right)^{-\frac{1}{2}(n+d)} \tau^{\frac{d}{2}} e^{-\frac{1}{2\sigma^2}y'y} e^{\frac{1}{2}\eta' Q^{-1}\eta} |Q|^{-\frac{1}{2}}.$$

Note that in this case, however, the integral is not analytically available and prevents the marginal likelihood from being easily computed. To address this, Botev (2016) uses a minimax tilting method to calculate the normalizing constant of truncated normal distributions and shows that the proposed estimator has the vanishing relative error property (Kroese et al., 2011). In light of this, we accept Botev's estimator as the true marginal likelihood in the following experiments. The `TruncatedNormal` package (Botev and Belzile, 2019) not only implements Botev's estimator, but it also provides samples from truncated normal distributions, so posterior samples from $\beta \mid y$ are readily available.

In Figure 3.2, we present the simulation results for the case when $d = 20$. Each approximation uses 50 MCMC samples, and we compare the results against the true log marginal likelihood. Once again, the Hybrid estimator outperforms the other estimators and reinforces its ability to deal with a scarcity of posterior samples. Provided with a sufficiently large number of samples, however, the BSE and CAME are both eventually able produce more accurate results than the Hybrid estimator.

3.7.2 Approximate posterior samples

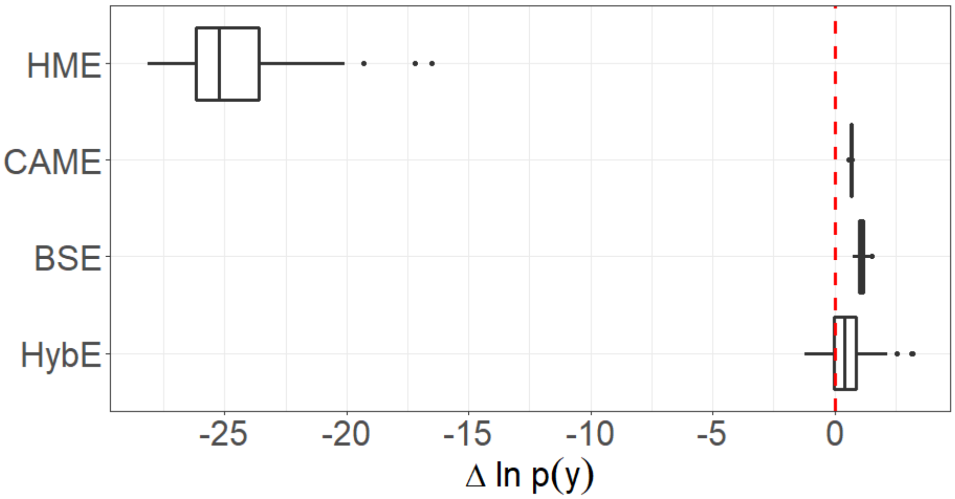
Up until now, we have assumed that asymptotically exact samples from the posterior distribution are available to be used as input for all of the marginal likelihood estimation schemes. In fact, for all previous numerical experiments, we have used samples drawn from the exact posterior distribution, ridding us of the need for burn-in or thinning. While MCMC algorithms can eventually provide us with exact posterior samples, it is not uncommon to have a target distribution that is too complex and difficult for traditional MCMC methods to learn within a reasonable time constraint. As a result, obtaining exact posterior samples may pose as big of an issue as the approximation procedure itself. Seeing as approximate inference techniques, such as variational Bayes, have been gaining traction as attractive alternatives for dealing with intractable distributions, it is worth considering the quality of marginal likelihood estimation algorithms that use samples drawn from these approximate distributions. As noted by Bishop (2016); Blei et al. (2017), factorized approximations such as those frequently used in variational methods tend to underestimate the variance of the true posterior. Consequently, the samples drawn from these approximate distributions may be inexact and could likely result in unpredictable behavior in algorithms that use them for further approximations. On the other hand, the Hybrid algorithm does not directly use these samples in the approximation to the negative log posterior, and instead they are only involved in the partitioning scheme. The hope is then that the partitioning scheme is not as sensitive to small inaccuracies in the posterior samples, but rather uses them to loosely learn the shape of the posterior distribution.

As a demonstration, we revisit the MVN-IG example in Section 3.7.1.1 and consider the case where $(\beta, \sigma^2) \in \mathbb{R}^{10}$. Clearly, the posterior distribution is tractable in this case and can easily be sampled from, but for the sake of demonstration, we work instead with a mean field approximation $q\left(\beta, \sigma^2\right)$ to the posterior distribution $p\left(\beta, \sigma^2 \mid y\right)$. In particular, suppose q factorizes over the parameters, so that $q\left(\beta, \sigma^2\right) = q\left(\beta\right) q\left(\sigma^2\right)$, where

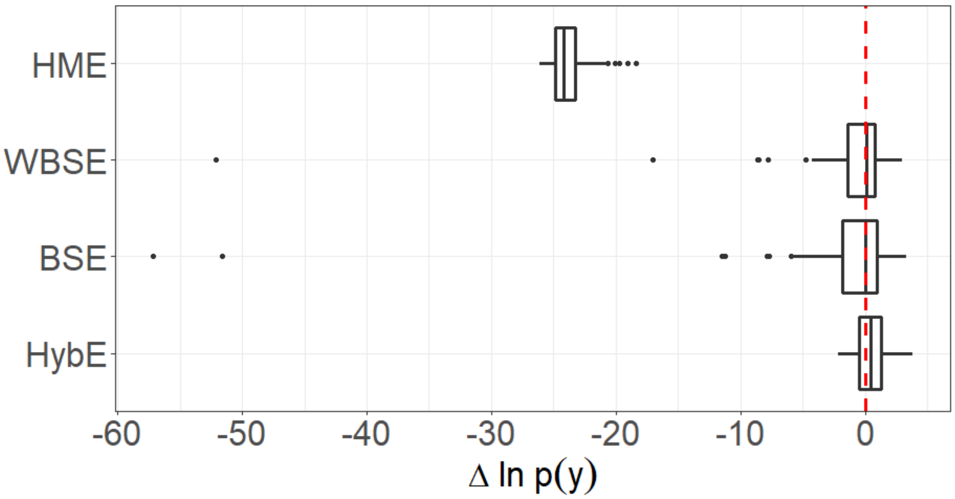
$$q\left(\beta\right) \equiv \prod_{i=1}^{3} \mathcal{N}_3\left(\mu_n^{(i)}, \sigma_0^2 V_n^{(i)}\right), \; q\left(\sigma^2\right) \equiv \mathcal{IG}\left(a_n, b_n\right).$$

Here, we have split the original 9-dimensional normal distribution for β into a product of 3-dimensional normal distributions, with the mean and covariance components extracted from the true posterior parameters. In particular, $\mu_n^{(1)} = (\mu_{n1}, \mu_{n2}, \mu_{n3})'$, $\mu_n^{(2)} = (\mu_{n4}, \mu_{n5}, \mu_{n6})'$, $\mu_n^{(3)} = (\mu_{n7}, \mu_{n8}, \mu_{n9})'$. Each $V_n^{(i)}$ is defined as the corresponding 3×3 block matrix in V_n, and σ_0^2 is the posterior mean of σ^2. We take $q\left(\sigma^2\right)$ to be the exact posterior distribution of σ^2, so a_n and b_n are identical to those defined in Section 3.7.1.1.

In the top boxplot of Figure 3.3, we observe that even with non-exact posterior samples, the Hybrid approximation produces accurate estimates, with an average error of 0.449 over 100 replications, compared to average errors of 0.698 and 1.035 for the CAME and BSE, respectively. While the latter two estimators have lower variance than the Hybrid estimator, neither of the former covers the true marginal likelihood. The results indicate that the BSE and CAME are very sensitive to the exactness of the samples, which should not be surprising because these algorithms use the MCMC samples directly in the calculation of the estimator, so any inaccuracies present in the samples themselves will likely manifest in the final estimator as well. From the results, we



(a) Bayesian linear regression - approximate MVN-IG posterior distribution



(b) Unrestricted covariance matrices - IW posterior distribution

Figure 3.3: Boxplots of the error (truth - estimate) for the MVN-IG example with approximate posterior samples $(\beta, \sigma^2) \in \mathbb{R}^{10}$ (top, $\log p\,(y)$: -147.3245) and the unrestricted covariance matrix example (bottom, true $\log p\,(y)$: -673.7057). For the MVN-IG example, we used 100 observations and 100 approximate posterior samples drawn from the mean field approximate posterior distribution. For the inverse-Wishart example, we consider 4×4 covariance matrices with 10 free parameters. Results are reported over 100 simulations, with 100 observations and 25 MCMC samples. Adapted with permission from "A hybrid approximation to the marginal likelihood" by Eric Chuu, Debdeep Pati, and Anirban Bhattacharya, 2021. Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, 130:3214-3222, Copyright 2021 by the authors.

confirm the Hybrid estimator holistic use of these samples to learn the posterior distribution and identify regions of concentration allows for small perturbations and inaccuracies in the samples and still yields a robust approximation.

3.7.3 Unrestricted covariance matrices

The examples have thus far dealt with parameters in Euclidean space. For the following example, we move beyond the usual Euclidean space and consider parameters in $\mathbb{R}^{p \times p}$. In particular, let $x_1, \ldots, x_n \overset{\text{iid}}{\sim} \mathcal{N}_p\left(0, \Sigma\right)$, where $\Sigma \in \mathbb{R}^{p \times p}$. Then the likelihood can be written as follows,

$$L\left(\Sigma\right) = \left(2\pi\right)^{-np/2} \det\left(\Sigma\right)^{-n/2} e^{-\operatorname{tr}\left(\Sigma^{-1} S\right)/2}, \tag{3.9}$$

where $S = \sum_{i=1}^{n} x_i x_i'$. For simplicity, we consider a conjugate inverse-Wishart (IW) prior, $\mathcal{W}^{-1}(\Lambda, \nu)$, for Σ, which has the following density,

$$\pi\left(\Sigma\right) = C_{\Lambda,\nu} \det\left(\Sigma\right)^{-(\nu+d+1)/2} e^{-\operatorname{tr}(\Sigma^{-1}\Psi)/2}, \quad C_{\Lambda,\nu} = \frac{\det\left(\Lambda\right)^{\nu/2}}{2^{\nu p/2}\,\Gamma_p\left(\nu/2\right)}. \tag{3.10}$$

Here, Λ is a positive definite $p \times p$ matrix, $\nu > p - 1$ is the degrees of freedom, and $\Gamma_p\left(\cdot\right)$ is the multivariate gamma function. Consequently, the posterior distribution of Σ is $\mathcal{W}^{-1}\left(\Lambda + S, \nu + n\right)$, and we can compute the marginal likelihood in closed form. Observe that in this case, where the underlying parameter space is a sub-manifold of $\mathbb{R}^{p \times p}$ and extremely non-Gaussian, the Laplace approximation does not hold (even asymptotically), so the need for a method that can handle a variety of problem settings is especially apparent.

Note that despite being able to sample from the posterior distribution, we cannot yet carry out the Hybrid approximation algorithm. As mentioned previously, since posterior samples are drawn from a sub-manifold of $\mathbb{R}^{p \times p}$, if we were to proceed as usual to obtain a partition over $\mathbb{R}^{p \times p}$, there would be no guarantee that a given point within the partition could be reconstructed to form a valid covariance matrix. As such, we circumvent this issue by working with an alternative representation of the covariance matrix. In particular, we take the Cholesky factorization of the covariance matrix $\Sigma = TT'$, where T is a lower triangular matrix with positive diagonal entries, t_{jj} for $j = 1, \ldots, p$. Because we are now working with T instead of Σ, the log-likelihood and prior must be appropriately modified so that we can form the covariate-response pairs $(L, \Psi(L))$ to be used by the Hybrid algorithm.

Under this transformation, we can define $\Psi(T) = -\log L(T) - \log \pi(T)$, where we can use Eq. (3.9) to rewrite the likelihood in terms of the Choleksy factor T,

$$L(T) = (2\pi)^{-np/2} \det(T)^{-n} \, e^{-\operatorname{tr}\left((TT')^{-1}S\right)/2}.$$

Conveniently, the determinant of the Jacobian matrix J of this transformation is well-known and given as

$$|J| = 2^p \prod_{j=1}^{p} t_{jj}^{p+1-j}.$$

By the change of variable formula, the induced prior on T is

$$\pi(T) = C_{\Lambda,\nu} \det(T)^{-(\nu+p+1)} \exp\left\{-\operatorname{tr}\left((TT')^{-1}\Lambda\right)/2\right\} \cdot 2^p \prod_{j=1}^{p} t_{jj}^{p+1-j}.$$

Obtaining posterior samples of T is trivial, as we can simply draw Σ from the inverse-Wishart distribution $\mathcal{W}^{-1}(\Lambda + S, \nu + n)$, and then take the corresponding lower Cholesky factor. With this general setup in place, it is worth noting that even with another prior on Σ, we can carry out the entire algorithm, provided that we have a way to sample from the posterior of Σ and a way to compute the Jacobian of the transformation.

In the bottom boxplot of Figure 3.3, we present the results for which each approximation uses 25 MCMC samples. The boxplot of the approximations' errors reinforce the robustness of the Hybrid estimator, which produces accurate and low-variance estimates. Although the BSE and WBSE both cover the true log marginal likelihood value, it is apparent that these estimators suffer from stability and convergence issues that are not present in the Hybrid estimator.

3.7.4 Hyper-inverse Wishart induced Cholesky factor density

In the following examples, we extend the previous analysis of unrestricted covariance matrices to a graphical modeling context. Relevant definitions and basic graph theory concepts can be found in Appendix C. Broadly speaking, Gaussian graphical models (GGM) are popular tools to learn the dependence structure among variables of interest. In particular, let $G = (V, E)$ be an undirected decomposable graph with vertex set $V = \{1, \dots, p\}$ and edge set E. Define $\mathbb{S}^p$ as the set of symmetric $p \times p$ matrices and $\mathbb{S}^p_{\succ 0}$ as the cone of positive definite $p \times p$ matrices in $\mathbb{S}^p$. Let $X \sim \mathcal{N}(\mu, \Sigma)$, $\Sigma^{-1} \in \mathbb{S}^p_{\succ 0}(G)$, where

$$\mathbb{S}^p_{\succ 0}(G) = \{M = (M_{ij}) \in \mathbb{S}^p_{\succ 0} \mid M_{ij} = 0, \forall (i,j) \notin E\}.$$

Then, X satisfies the GGM with graph G, where G dictates the conditional dependence structure and restricts the sparse concentration matrix $\Omega = (\omega_{ij})_{p \times p} = \Sigma^{-1}$ so that $(i,j) \in E$ if and only if $\omega_{ij} = 0$, and $x^{(i)}$ and $x^{(j)}$ are conditionally independent if and only if $\omega_{ij} = 0$. In other words, if the variables i and j do not share an edge in a graph G, then $\omega_{ij} = 0$. Hence, an undirected graphical model corresponding to G restricts the inverse covariance matrix Ω to a linear subspace of the cone of positive definite matrices. A probabilistic framework for learning the dependence structure and the graph G requires specification of a prior distribution for (Ω, G). Conditional on G, the hyper-inverse Wishart (HIW) distribution (Diaconis et al., 1979) for $\Sigma = \Omega^{-1}$ and the corresponding induced class of distributions (Roverato, 2000) for Ω are attractive choices for priors.

Given G, we place a hyper-inverse Wishart prior $\mathrm{HIW}_G(\delta, \Lambda)$ on $\Omega = \Sigma^{-1}$, where $\delta > 2$ is the degrees of freedom and $\Lambda \in \mathbb{S}^p_{\succ 0}$ is fixed. The HIW distribution is defined over the cone of $d \times d$ positive definite matrices, with the corresponding density:

$$f(\Omega \mid G) \propto |\Omega|^{(\delta-2)/2} \exp\left(-\operatorname{tr}(\Omega\Lambda)/2\right). \tag{3.11}$$

When G is decomposable, an alternative parameterization is given by the Cholesky decomposition of $\Sigma^{-1} = \Omega = \phi'\phi$. Provided that the vertices of $G = (V, E)$ are enumerated according to a *perfect vertex elimination scheme*, Section 2 of (Roverato, 2000) tells us that the upper triangular matrix ϕ observes the same sparsity as Ω. While the likelihood function is identical to the one given in Eq. (3.9), we are instead working with the Cholesky factor of the inverse covariance matrix, so the likelihood of ϕ is given as

$$L(\phi) = (2\pi)^{-np/2} \det(\phi)^n \, e^{-\operatorname{tr}(\phi'\phi S)/2}, \quad S = \sum_{i=1}^{n} x_i x_i'. \tag{3.12}$$

Note that $\det(\Sigma^{-1}) = \det^2(\phi)$ and $\det(\phi) = \prod_{i=1}^{p} \phi_{ii}$. In order to complete the definition of $\Psi(\phi)$, we need only compute the induced prior on the nonzero elements of ϕ, which, together with the likelihood in Eq. (3.12), gives us an explicit expression for the negative log posterior. From Roverato (2000), the determinant of the Jacobian matrix J of the transformation $\Omega \to \phi$ is given by

$$|J| = 2^p \prod_{i=1}^{p} \phi_{ii}^{\nu_i + 1},$$

where the i-th row of ϕ has exactly $\nu_i + 1$ many nonzero elements. From Dawid and Lauritzen (1993), we know that the distribution of Σ has the strong hyper-Markov property, so we can ascertain the mutual independence of the rows of ϕ, provided that the vertices are in perfect vertex elimination scheme. In addition, the induced distributions of the diagonal elements ϕ_{ii} and the off-diagonal elements ϕ_{rs}, with $r < s, (r, s) \in E$ (Roverato, 2000, Theorem 4) allow us to specify the joint density of the free elements of ϕ as follows,

$$\pi(\phi) = \left[\prod_{i=1}^{p} \frac{2^{-(\delta+\nu_i)/2}}{\Gamma((\delta+\nu_i)/2)} \phi_{ii}^{\delta+\nu_i-2} e^{-\frac{1}{2}\phi_{ii}^2} (2\phi_{ii})\right] \times \left[\prod_{(r,s):s>r,(v_s,v_r)\in E} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\phi_{rs}^2}\right]. \tag{3.13}$$

Since the HIW distribution is conjugate for Ω, the posterior distribution of Ω is also conveniently known, $\mathrm{HIW}_G(\delta + n, \Lambda + S)$. Putting all of these ideas together, we can easily supply the necessary ingredients for the Hybrid estimation framework by drawing samples from the posterior

Table 3.2: Mean and standard deviation of the estimates for the hyper-inverse Wishart model, taken over 100 replications. Here, we consider 5×5 covariance matrices with 10 free parameters. Each replication has 100 observations and 25 posterior samples. The true log marginal likelihood is -506.306. Adapted with permission from "A hybrid approximation to the marginal likelihood" by Eric Chuu, Debdeep Pati, and Anirban Bhattacharya, 2021. Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, 130:3214-3222, Copyright 2021 by the authors.

		TRUTH	**HME**	**WBSE**	**HYB**
	MEAN	-506.306	-486.675	-501.225	-507.760
$p = 5$	SD	0	0.895	9.815	1.362
$d = 10$	AE	0	-19.631	-5.081	1.454
	RMSE	0	19.699	11.009	1.988

distribution, taking the upper Cholesky factor of each sample, and computing the negative log posterior $\Psi(\phi)$ using the likelihood in Eq. (3.12) and the prior in Eq. (3.13). While this procedure appears to be quite simple, the implications are significant in that if we have a different prior on Σ for which we can do the posterior computation, all other aspects of the algorithm would remain the same. All that is required is a way to sample from the posterior of Ω and a expression for the Jacobian of the corresponding transformation.

For the numerical experiments, we take $\delta = 3, B = I_5$ and present the results in Table 3.2. In this example, we consider 5×5 precision matrices that have 10 free elements, where each marginal likelihood estimate uses only 25 posterior samples. Even in the case of a relatively low-dimensional parameter space, we observe that without a large number of samples, traditional methods fail to deliver accurate results and often have high variance. In contrast, the Hybrid estimator retains its ability to produce reliable results that do not exhibit high variance that we see in the WBSE. However, as the number of MCMC samples increases, the WBSE expectedly stabilizes and eventually beats the Hybrid estimator. In order to have a baseline for evaluating these marginal likelihood estimates, we must also compute the true normalizing constant of HIW density. Since G is decomposable, all of its prime components are complete and are thus cliques. As pointed out by Roverato (2000), Σ can be written as a sequence of clique marginal matrices,

$\{\Sigma_C : C \in \mathscr{C}\}$, where $\mathscr{C}$ is the collection of cliques. As a result, the HIW density factorizes over these cliques, with each Σ_C following an inverse-Wishart distribution, $\mathcal{W}^{-1}(\delta, \Lambda_C)$. Here, Σ_C and Λ_C refer to submatrices which can be constructed by taking the entries of Σ and Λ that correspond to the nodes in C. As seen in the previous section, the inverse-Wishart density has a closed-form normalizing constant, so the normalizing constant for the original HIW density can also be derived analytically. More details about the calculation of the HIW normalizing constant can be found in Section C.2.

3.7.5 Factor models

In this section, we consider the issue of computing the marginal likelihood for factor models, a problem setting for which the quantity of interest is intractable. Factor analysis provides a tool for investigating multivariate dependence among variables in terms of latent factors that are typically fewer in number than the number of observed variables. This type of analysis has found its way into areas such as psychological research when constructing scales for measuring attitudes, perceptions, motivations, etc. (Ford et al., 1986), financial analysis (Aguilar and West, 2000; Pitt and Shephard, 1999), and gene expression problems (Sabatti and James, 2006; Pournara and Wernisch, 2006). As a result, factor models are now being used in high-dimensional modeling and increasingly complex situations, thus making scalable inference in this domain important. In Bayesian model selection in factor analysis, the main inferential goal deals with the uncertainty quantification associated with the number of latent factors in a multivariate factor model. This ultimately comes down to comparing models that differ only in the number of factors, a comparison that hinges on accurate calculation of the marginal likelihood. Since the marginal likelihood for factor models is not easily computed and typically requires approximations, MCMC methods are frequently employed. While there is no shortage of MCMC methods that target these posterior probabilities (Lopes and West, 2004; Polasek, 1997), the computational costs associated with both running these MCMC algorithms for high-dimensional problems and obtaining (a sufficient number of) reliable MCMC samples cannot be ignored. In an effort to more aptly model these problems that have parameters that frequently concentrate in lower-dimensional regions of an otherwise high-dimensional ambient space, sparse factor models (West, 2002), which provide a more scalable modeling framework that incorporate dimension reduction, have also been an area of active research and development.

As shown in numerous previous problem settings, our proposed Hybrid approximation scheme lends itself as a potential solution. We remark that while the following exposition largely deals with a general factor model setup, they Hybrid methodology can easily be adapted to incorporate the aforementioned advancements in factor modeling.

3.7.5.1 Setup and notation

Let y_i denote the m-dimensional observed variable, and suppose the factor $f_i \sim \mathcal{N}_k(0, I_k)$, with $k \leq m$. The k-factor model states:

$$y_i = \beta f_i + \epsilon_i, \tag{3.14}$$

where $\epsilon_i \sim \mathcal{N}(0, \Omega)$, $\Omega = \operatorname{diag}(\omega_1^2, \ldots, \omega_m^2)$, ϵ_i and f_j are independent, and the f_i's are independent for $1 \leq i, j \leq n$. Here, $\beta \in \mathbb{R}^{m \times k}$ is an unknown loading matrix. Let $Y = (y_1, \ldots, y_n)'$ be the $n \times m$ matrix of observations. Similarly, let $F = (f_1, \ldots, f_n)'$, and $E = (\epsilon_1, \ldots, \epsilon_n)'$. We then have the following conditional density of y given F, β, Ω,

$$p(y \mid F, \beta, \Omega) \propto |\Omega|^{-n/2} \exp\left(-0.5 \operatorname{tr}\left(\Omega^{-1}\epsilon\epsilon'\right)\right), \tag{3.15}$$

where we have defined $\Sigma = \Omega + \beta\beta'$. Integrating out the latent factors F, we arrive at the following density:

$$p(y \mid \beta, \Omega) = |\Sigma|^{-n/2} \exp\left(-0.5 \operatorname{tr}\left(\Sigma^{-1}y'y\right)\right). \tag{3.16}$$

With this model setup, we see that conditional on the factors, each observation y_i is independent, with $y_i \sim \mathcal{N}(0, \Sigma)$. Therefore, the dependence structure among the m components, given by Σ, is explained by the common factors. As such, this factor model setup intuitively facilitates dimension reduction in terms of the number of parameters in consideration. In particular, by learning the parameters (β, Ω) instead of learning Σ directly, we have at most $mk + m$ free elements, which is typically fewer than the $O\left(m^2\right)$ free elements in Σ. This reduction is especially significant in the case of the sparse factor model where $k \ll m$.

Using the likelihood in Eq. (3.16), and suitably defined prior distributions, $p\left(\beta\right), p\left(\Omega\right)$, we have the following expression for the marginal likelihood:

$$p\left(y\right) = \int p\left(y \mid \beta, \Omega\right) p\left(\beta\right) p\left(\Omega\right) d\beta \, d\Omega. \tag{3.17}$$

In the next section, we present a commonly used prior for (β, Ω) and also discuss a modification of this prior that improves the interpretability of the subsequent inference for the model selection and model comparison problem.

3.7.5.2 Prior specification

For identifiability of the loading matrix β, we adopt a convention similar to that of Geweke and Zhou (1995); Aguilar and West (2000); Lopes and West (2004), and consider β to be (block) lower triangular with positive diagonal entries. This ensures that β is full rank matrix and uniquely identified by $\beta\beta'$. Bhattacharya and Dunson (2011) and Ghosh and Dunson (2009) provide alternative priors for the loading matrix that incorporate sparsity and provide appealing theoretical properties, but in this discussion, we focus on the prior specification as detailed in Lopes and West (2004). However, similar to before in the graphical modeling examples, the following procedures can be easily adapted to other priors, so the following analysis is not limited to the prior we discuss.

Initially, we assume independent priors for the lower triangular terms, where

$$\beta_{ij} \sim \begin{cases} \mathcal{N}\left(0, C_0\right) & i > j \\ \mathcal{N}\left(0, C_0\right) \mathbb{1}\left(\beta_{ij} > 0\right) & i = j. \end{cases}$$

Here, $C_0 > 0$ is a hyperparameter. The variances ω_i^2 are assumed independent of β and mutually independent, with the following inverse-gamma distribution,

$$\omega_i^2 \sim \mathcal{IG}\left(\nu/2, \nu s^2/2\right), \tag{3.18}$$

with hyperparameters $\nu, s > 0$. However, as noted by Leung and Drton (2014), the induced prior on $\beta\beta'$, as given above, is not order-invariant, so the resulting inference may differ depending on how the variables are ordered. With slight changes to this prior, we can achieve inference that is invariant in the arrangement of the variables and therefore have more accurate quantities for model comparison.

Using the proposal from Leung and Drton (2014), we adopt the following *order-invariant prior distribution* on β, which allows for inference that remains unchanged upon reordering of the variables while maintaining the identifiability constraint. In particular, suppose we use a spherical normal prior, $\beta \sim \mathcal{N}_{n\times k}\left(0, C_0 I_m \otimes I_k\right)$, where $m \geq k$. This distribution is invariant under permutation of the rows of β, so the induced prior on Σ is similarly invariant. Then, we perform the following LQ decomposition, $\beta = LQ$, where L is lower triangular and Q is orthogonal. If we substitute this expression for β back into the definition of Σ, we see that we can essentially work with L, rather than β, since $\beta\beta' = LQQ'L' = LL'$. In order to use the Hybrid estimation framework, we require an expression for the negative log posterior in terms of L. Since the likelihood function remains unchanged, we need only to derive an expression for the induced prior on L. Starting with the proposed prior on β, we can write the density as follows,

$$p\left(\beta\right) \propto \exp\left(-\frac{1}{2}\operatorname{tr}\left(\beta\beta'\right)\right) \tag{3.19}$$

Using the results from Chapter 2 of Muirhead (2005), we know that Jacobian of the transformation $\beta \mapsto LQ$ is $d\beta = \prod_{i=1}^{k-1} L_{ii}^{k-i}\left(dL\right)\left(Q'dQ\right)$, where $\left(Q'dQ\right)$ is the Haar measure on the set of all $k \times k$ matrices. Substituting this into the joint density on (L, Q), we can deduce that L and Q are independent, which then produces the following induced prior on L,

$$p(L) \propto \exp\left(-\frac{1}{2}\operatorname{tr}(LL')\right)\prod_{i=1}^{k} L_{ii}^{k-i}. \tag{3.20}$$

More specifically, the density of the diagonal terms of L can be written as:

$$p(L_{ii}) = \chi_{k-i+1}\left(\frac{L_{ii}}{\sqrt{C_0}}\right) = \left(\frac{L_{ii}}{\sqrt{C_0}}\right)^{k-i}\exp\left(-\frac{L_{ii}^2}{2C_0}\right)2^{-\frac{k-i+1}{2}+1}\Gamma^{-1}\left(\frac{k-i+1}{2}\right),$$

where χ_{k-i+1} is the Chi distribution with degrees of freedom $k-i+1$. The density of the off-diagonal terms of L can be written as:

$$p(L_{ij}) = \mathcal{N}(0, C_0) = (2\pi C_0)^{-1/2}\exp\left(-\frac{1}{2C_0}L_{ij}^2\right). \tag{3.21}$$

Define $p = (m-k)k + k(k+1)/2$ to be the number lower triangular and diagonal elements of β. Then, in total, we have $d = p + m$ many parameters, where m denotes the number of diagonal elements in Ω. Therefore, the marginal likelihood is an integral over d-dimensional space.

3.7.5.3 Incorporation of the Hybrid estimator

One major convenience of Leung's modification is that its accompanying Gibbs sampler only slightly differs from that of Lopes and West (2004) in the full conditional distribution of $\beta_i, i \leq k$. More importantly, we are able to work in terms of (L, Ω), which together form Σ. Similar to how we adapted other examples where the parameter is not intrinsically a vector, we perform a modified vectorization of (β, Ω) by extracting the lower triangular elements of β and the diagonal elements of Ω and concatenating these into a d-dimensional vector u. Then, the negative log posterior of u can be written as

$$\begin{aligned}\Psi(u) &= -\left(\log p(f_L(u)) + \log p(f_\Omega(u)) + \ell(f_L(u), f_\Omega(u))\right)\\ &= -\left(\log p(L) + \log p(\Omega) + \ell(L, \Omega)\right), \end{aligned} \tag{3.22}$$

which can easily be evaluated by reconstructing β and Ω from the input vector u i.e., $f_L(u) = L, f_\Omega(u) = \Omega$ for suitably defined functions f_L, f_Ω. Using Eq. (3.16), we also have the following log-likelihood

$$\ell(\Omega, L;\, y_{1:n}) = -\frac{nm}{2}\log(2\pi) - \frac{n}{2}\log|\Omega + LL'| - \frac{1}{2}\sum_{i=1}^{n} y_i'\left(\Omega + LL'\right)^{-1} y_i. \qquad (3.23)$$

Since we can draw posterior samples of (β, Ω), and we can evaluate the negative log posterior distribution using Eq (3.22), then we are adequately equipped with the input for the Hybrid estimation algorithm.

Before presenting the numerical experiments, we comment briefly on other computational considerations. While the Gibbs sampler above accurately provides posterior samples, the convergence time becomes a concern in high-dimensional applications. There has subsequently been a push toward developing approximate (variational) inference algorithms that make small sacrifices in the overall accuracy in order to achieve computational efficiency. In the context of sparse factor models, Foo and Shim (2021) propose using a mean field approximation to the posterior distribution as an alternative to MCMC sampling, and their empirical results demonstrate a substantial speedup in convergence. Should a practitioner decide to employ one of these approximate inference algorithms to bypass the need to wait for MCMC convergence, they would be able do so without modification to any step in the Hybrid approximation framework, other than replacing the MCMC samples with the approximate posterior samples from the variational distribution.

3.7.5.4 Model selection exercise

We use the same setup as in Section 6.1 of Lopes and West (2004), and we generate $n = 100$ observations from the one-factor models with

$$\beta' = (0.995, 0.975, 0.949, 0.922, 0.894, 0.866, 0.837)\,,$$
$$\operatorname{diag}(\Omega) = (0.01, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30)\,,$$

Table 3.3: Comparison of k-factor model selection based on the marginal likelihood computed for each of the candidate models. The datasets are generated with $n = 100$ and each estimate uses 1,000 MCMC samples.

METHOD	$k = 1$	$k = 2$	$k = 3$
RJMCMC	1000	0	0
HYB	935	357	292
$\hat{p}_H$	428	258	314
$\hat{p}_{LM}$	1000	0	0
$\hat{p}_{OPT}$	1000	0	0

so that $y_i \sim \mathcal{N}(0, \Omega + \beta\beta')$. For each of these datasets, we fit various k-factor models, for $k = 1, 2, 3$, and compute the corresponding marginal likelihoods. Based on the value of the marginal likelihoods, we then select the k-factor model that has the highest (relative) posterior probability. Iterating through this process, we keep track of how many times we select each model and are subsequently able to determine which marginal likelihood estimation method most frequently picks the correct factor model. For this analysis, we use the order-invariant prior setup outlined in the preceding sections, for which we also supply the necessary groundwork to facilitate its use with the Hybrid estimation framework. For the other estimators included in this experiment, we use the original prior setup as presented in Lopes and West (2004). These estimators are the bridge sampling estimator with the optimal bridge function p_{OPT}, the Laplace-Metropolis estimator p_{LM}, the harmonic mean estimator p_H, and the RJMCMC estimator developed for factor models from Lopes and West (2004).

While the Hybrid estimator reports slightly less accurate results than most of the other methods, we highlight that our prior choice in the Hybrid estimator allows for more reliable inference that does not depend on the ordering of the variables, an important quality that is overlooked in the other estimators. We also note that the software used to report many of the other estimators in Table 3.3 experiment are not publicly available as packages, so reproducing this analysis for other problems presents a significant roadblock. On the other hand, following the steps in Section

3.7.5.3 to reparametrize the problem and using the `hybrid` package makes the marginal likelihood estimation task straightforward.

3.8 Chapter summary

In this chapter, we developed a novel algorithm that combines a variety of ideas, both probabilistic and deterministic, to efficiently estimate the marginal likelihood. By first using a regression tree to identify high-probability regions of the parameter space and then leveraging numerical integration ideas to obviate the need to trust the quality of the MCMC samples, we are able to construct an approximation that scales well with both the dimension and the complexity of the parameter space. We again emphasize that in the experiments provided in this section, we seek to mimic scenarios for which posterior sampling is highly expensive and/or mixing is poor. From the simulation studies, we see that the Hybrid estimator is both accurate and reliable in these problematic situations. By considering a small number of samples as the input for these marginal likelihood estimation algorithms, we provide a realistic scenario for the regime in which we wish to operate. Therefore, our contribution is multifaceted and bears practical value in that even in higher dimensions and in instances where generating MCMC samples is itself a bottleneck, the Hybrid estimator successfully delivers promising results. Equally important as these results is the availability of the Hybrid algorithm in the `R` package `hybrid`, which can be used to reproduce the experiments in this chapter.

Furthermore, the Hybrid approximation scheme outlined in this chapter lays the groundwork for future work in a number of possible directions. One area of potential refinement is the construction of the partition of the parameter space. While we use CART for its convenience and interpretability, we find that the default objective function for CART is unsuitable for determining the representative point of each partition set, and we have to solve an additional optimization problem to obtain these points. Instead of this roundabout two-step approach, where we use CART to learn the partition and the objective function in Eq. (3.8) to identify representative points, we can investigate alternative objective functions that can better target the desired objective function.

Another aspect of the Hybrid algorithm that can be further developed is the current formulation of the local approximation to Ψ in each of the partition sets. The piecewise constant approximation in Eq. (3.4), though providing encouraging results, is a rather simplistic way to approximate Ψ, particularly when the target distribution is highly nontrivial. In the next chapter, we modify the constant approximation to incorporate higher order terms via a local Taylor expansion so that piecewise linear and quadratic terms contribute to the approximation to Ψ. While we mention in Section 3.2 that a piecewise linear estimator would similarly result in a tractable integral calculation, additional quadratic terms would require more delicate handling. Consequently, a sizeable portion of the extended methodology is devoted to dealing with these higher order terms.

4. HYBRID-EP METHOD FOR MARGINAL LIKELIHOOD ESTIMATION OF UNIMODAL DENSITIES

4.1 Introduction

The exposition on the Hybrid estimator in the previous chapter sets the foundation for a promising way to estimate the marginal likelihood, demonstrating that our proposed estimator can compete with other well-known estimators in a variety of problem settings. Moreover, the generality of the methodology makes it convenient to make refinements to the algorithm so that it can be applied to more specific problem setups. In this chapter, we propose a few modifications to the piecewise approximation to Ψ to form an estimator that is more suitable for tackling higher-dimensional problems. Further, we restrict our scope to target the normalizing constants of a specific class of densities—unimodal densities that are approximately log-concave around the mode. Indeed, the problems of sampling from and integrating log-concave functions have seen no shortage of work from Applegate and Kannan (1991) and Lovász and Vempala (2007). In the paper from Lovasz and Vempala (2006), they frame the integration aspect of their algorithm as a generalized version of simulated annealing. Brosse et al. (2018) also present an approach for computing the normalizing constant of log-concave densities that uses Gaussian annealing and the Unadjusted Langevin algorithm. Even with the plethora of available methods, the sampling-based nature of most of these algorithms makes them susceptible to prohibitively expensive computations that are only exacerbated in higher dimensions. In the following discussions, we call the estimator that arises from this chapter's methodology the *Hybrid-EP estimator* (HYB-EP) and the original estimator from the previous chapter's methodology the *vanilla Hybrid estimator* (HYB).

For the initial development, we do not restrict our analysis to posterior distributions and illustrate the modified algorithm in the general case of log-concave target densities. Namely, let γ be a probability density with respect to the Lebesgue measure on $\mathbb{R}^d$ of the form $\gamma\left(u\right) = \mathcal{Z}^{-1}e^{-\Psi(u)}$ for $u \in \mathcal{U} \subseteq \mathbb{R}^d$, where $\Psi : \mathcal{U} \to \mathbb{R}$ is a continuously differentiable strictly convex function.

Adhering to the notation in the previous chapter allows us to write the normalizing constant of γ as the following integral:

$$\mathcal{Z} = \int_{\mathcal{U}} e^{-\Psi(u)} du < +\infty, \quad \mathcal{U} \subseteq \mathbb{R}^d. \tag{4.1}$$

The difference here is our assumption of the shape of Ψ, but we emphasize that the log-concavity assumption on γ is only used to ensure that the Hessian of the objective function is positive definite everywhere and can in fact be relaxed to unimodal densities that are approximately log-concave in a suitable neighborhood around the mode. In particular, this relaxed assumption is widely satisfied by Bayesian posteriors in regular parametric models and due to the Bernstein-von Mises phenomenon, most regular posterior distributions are approximately log-concave with sufficiently large sample size; see Section 4.6 for more details. Next, we present a high-level overview of the algorithm before delving into more specific details. Readers may notice that many of the initial steps overlap with those stated previously in the vanilla Hybrid methodology, but we include a complete outline of the proposed algorithm in its entirety to avoid any ambiguity.

Like before, we work with a compactification of the parameter space and make the following approximation to the normalizing constant of γ, $\int_A e^{-\Psi(u)} du$, where A is a compact subset of the parameter space $\mathcal{U}$. Recall that the motivation for this compactification is to eliminate low probability regions of the domain whose contribution to the integral in Eq. (3.1) is negligible. This is particularly useful in a Bayesian context where the posterior concentrates with increasing sample size (Ghosal and Van Der Vaart, 2007; Kleijn and van der Vaart, 2006). Next, proceed by partitioning A into d-dimensional rectangles, i.e., $\mathcal{A} = \{A_1, \ldots, A_K\}$ such that $A = \bigcup_{k=1}^K A_k$ and $A_k \cap A_{k'} = \emptyset$ for all $k \neq k'$. We then propose a piecewise approximation to Ψ defined over each partition set:

$$\Psi(u) \approx \widehat{\Psi}(u) = \sum_{k=1}^{K} \widehat{\Psi}_k(u) \mathbb{1}_{A_k}(u). \tag{4.2}$$

The general form of this piecewise approximation is identical to the one used in the vanilla Hybrid estimator, but we shall see below that our choice for $\widehat{\Psi}_k$ differs in that we no longer employ a constant approximation to Ψ. Using the form of the approximation in Eq. (4.2) gives us the resulting general approximation to $\mathcal{Z}$,

$$\mathcal{Z} \approx \int_A e^{-\Psi(u)}\, du \approx \int_A e^{-\widehat{\Psi}(u)}\, du = \sum_{k=1}^{K} \int_{A_k} e^{-\widehat{\Psi}_k(u)} du. \tag{4.3}$$

We elaborate in Section 4.2 on the choice for $\widehat{\Psi}_k$. As mentioned in the development of the vanilla Hybrid estimator, the goal of the partitioning step is to divide up the parameter space in a way such that we more finely partition regions for which γ has higher mass, thus allowing for more precise approximations where γ exhibits concentration. By sampling $u_j \sim \gamma$, evaluating $\Psi\left(u_j\right)$, and fitting a CART model (Breiman, 1984) to the pairs $\left\{\left(u_j, \Psi\left(u_j\right)\right)\right\}_{j=1}^{J}$, we can learn a dyadic partition of $\mathcal{U}$. We obtain the compact subset of the parameter space $\mathcal{U}$ by bounding the partition from CART using the range of the samples from γ. Clearly, we make the assumption that we can sample from γ and evaluate Ψ, but both of these are basic requirements in many MCMC-based algorithms.

Below, we highlight three fundamental features of the Hybrid-EP algorithm that distinguish it from the vanilla Hybrid algorithm, followed by a detailed discussion of each of these steps. See Algorithm 2 for a formal statement of the Hybrid-EP estimation procedure.

- In Eq. (4.2), we take $\widehat{\Psi}_k$ to be a local approximation to Ψ constructed using a second order Taylor expansion of Ψ around a representative point $u_k \in A_k$.

- The second order Taylor approximation introduces a quadratic term that is accompanied by additional computational challenges, as it complicates the calculation of the integrals in Eq. (4.3). We tackle this issue with an efficient Expectation Propagation (EP) algorithm that targets high-dimensional Gaussian integrals.

- We exploit the unimodality of γ to find suitable points within each partition set around which we perform the Taylor expansion required for $\widehat{\Psi}_k$.

Each of the three steps above is nuanced in that they appear to be quite simple to implement and easy to integrate into the existing algorithm, but they also carry with them substantial computational demands that require careful attention as well.

4.2 Local approximation using a Taylor expansion

We revisit Eq. (4.2) and consider the following piecewise quadratic approximation to Ψ,

$$\Psi\left(u\right) \approx \sum_{k} \left[\Psi\left(u_k\right) + \left(u - u_k\right)' \nabla \Psi\left(u_k\right) + \frac{1}{2}\left(u - u_k\right)' \nabla^2 \Psi\left(u_k\right)\left(u - u_k\right)\right] \mathbb{1}_{A_k}\left(u\right), \tag{4.4}$$

where u_k is a representative point of A_k. This iterates on the idea of the piecewise constant approximation performed in Chapter 3, but introduces higher order terms, with the hope that these lead to increased accuracy. Upon exponentiation of the approximation in Eq. (4.4), we observe that the second exponential in the summation below is proportional to a Gaussian density,

$$e^{-\widehat{\Psi}(u)} = \sum_{k} \exp\Big\{ -\Psi(u_k) + u_k'\lambda_k - \tfrac{1}{2}u_k' H_k u_k \Big\} \exp\Big\{ -\tfrac{1}{2}u' H_k u + (H_k u_k - \lambda_k)' u \Big\} \mathbb{1}_{A_k}\left(u\right).$$

To ease notation, we have taken $\lambda_k := \nabla \Psi\left(u_k\right)$ and $H_k := \nabla^2 \Psi\left(u_k\right)$, which represent the gradient vector and the Hessian matrix of Ψ evaluated at u_k, respectively. Since γ is assumed to be log-concave, the Hessian matrix is positive definite, and hence H_k is invertible. Integrating the approximation above and keeping track of the normalizing constants, we propose the following approximation to $\mathscr{Z}$:

$$\begin{aligned}
\int_A e^{-\Psi(u)} du &\approx \int_A e^{-\widehat{\Psi}(u)} du \\
&= \int_A \left[\sum_k C_k \mathcal{N}\left(u \mid H_k^{-1} b_k, H_k^{-1}\right)\right] du \\
&= \sum_k C_k \cdot \int_{A_k} \mathcal{N}\left(u \mid H_k^{-1} b_k, H_k^{-1}\right) du && (4.5) \\
&=: \widehat{\mathscr{Z}}_{\text{HYB-EP}}. && (4.6)
\end{aligned}$$

Here, C_k stores the normalizing constants for each of the K Gaussian densities, and b_k, m_k are known quantities:

$$C_k = \exp\left(-\Psi\left(u_k\right) + \lambda_k' u_k - \frac{1}{2} u_k' H_k u_k + \frac{1}{2} m_k' H_k m_k\right),$$
$$b_k = H_k u_k - \lambda_k, \quad m_k = H_k^{-1} b_k.$$

Provided that we can compute the Gaussian integrals in the summation in Eq. (4.5) and determine suitable points $u_k \in A_k$, then $\widehat{\mathcal{Z}}_{\text{HYB-EP}}$ is a tractable estimator for $\mathcal{Z}$. In the next two sections, we provide scalable methods to accomplish both of these tasks.

4.3 Estimating truncated Gaussian probabilities

Despite the prevalence of Gaussian densities in statistical modeling, Gaussian probabilities are difficult to compute, as they typically require integration over high-dimensional spaces. Some established methods rely on numerical integration (Genz, 1992), but these prove to be prohibitively expensive (in the number of points required) and inefficient beyond low-dimensional settings. This has led to the development of more scalable integration techniques, such as Expectation Propagation (Minka, 2013), which is widely used to compute approximate integrals. To better understand how the Expectation Propagation (EP) algorithm can be applied to the intractable Gaussian integral in the previous section, we lay out some preliminary groundwork for the EP algorithm. Starting with the Gaussian distribution $p_0\left(x\right) = \mathcal{N}\left(x \mid m, K\right)$, we define the following unnormalized truncated distribution

$$p\left(x\right) = \begin{cases} p_0\left(x\right), & x \in \mathcal{A} \\ 0, & \text{otherwise.} \end{cases}$$

The Gaussian probability of interest can be written as:

$$F\left(\mathcal{A}\right) = \int_{\mathcal{A}} p_0\left(x\right) dx = \int p\left(x\right) dx. \tag{4.7}$$

A natural candidate for a distribution $q(x)$ that can replace the intractable distribution $p(x)$ is one that minimizes the Kullback-Leibler (KL) divergence between p and q, denoted $D(p \,||\, q)$, as this provides a quantification of the quality of q as an approximation with respect to the target distribution p. However, since p intractable, we cannot directly evaluate $D(p \,||\, q)$, and thus the corresponding minimization is problematic. The first step is to then replace $p(x)$ with the product of a prior distribution $p_0(x)$ and factors $t_i(x)$, such that

$$p(x) = p_0(x) \prod_i t_i(x).$$

We make a simplifying assumption about the structure of $t_i(x_i)$ so that the integration boundaries are not functions of x. In particular, let $t_i(x) = \mathbb{1}\{a_i < x_i < b_i\}$, where a_i, b_i are simply the lower and upper bounds of integration. Then, the target integral in Eq. (4.7) can be written as

$$F(\mathcal{A}) = \int p(x)\, dx = \int p_0(x) \prod_{i=1}^{d} t_i(x_i)\, dx_i.$$

We proceed to approximate each of the intractable factors t_i with a tractable, unnormalized Gaussian $\tilde{t}_i(x)$, which produces the final approximation q of p. More specifically, we take q to mirror the product form of p,

$$q(x) = p_0(x) \prod_i \tilde{t}_i(x) = p_0(x) \prod_i \tilde{Z}_i \mathcal{N}\left(x \mid \tilde{\mu}_i, \tilde{\sigma}_i^2\right) = Z\mathcal{N}(x \mid \mu, \Sigma),$$

where the parameters of these unnormalized Gaussian distributions, $\{\tilde{\mu}_i, \tilde{\sigma}_i^2, \tilde{Z}_i\}$, admit closed form updates that are the result of an iterative moment matching scheme. From this, we observe that by estimating the normalizing constant of q, we also obtain an approximation for the normalizing constant of the target distribution p. See equations (21), (22), and (23) in Cunningham et al. (2013) for the closed form updates for each of these parameters and more details regarding the relevant notation. Essentially, the EP algorithm iteratively constructs the approximating distribution $q(x)$ to minimize $D\left(t_i q^{\setminus i} \,||\, \tilde{t}_i q^{\setminus i}\right)$, which in turn approximately minimizes $D(p \,||\, q)$. Here,

$q^{\setminus i}(x) = q(x) \,/\, \tilde{t}_i(x)$ is defined as the cavity distribution. By running the EP algorithm to convergence, we can calculate the following mean and covariance parameters of the normal distribution q,

$$\mu = \Sigma \left(K^{-1} m + \sum_{i=1}^{d} \frac{\tilde{\mu}_i}{\tilde{\sigma}_i^2} e_i \right), \quad \Sigma = \left(K^{-1} + \sum_{i=1}^{d} \frac{1}{\tilde{\sigma}_i^2} c_i c_i' \right)^{-1},$$

where e_i is the i-th standard basis vector. With this, we also obtain a closed form expression for the normalizing constant of q,

$$\begin{aligned} \log Z = &-\frac{1}{2}\left(m' K^{-1} m + \log|K|\right) + \sum_{i=1}^{d}\left(\log \tilde{Z}_i - \frac{1}{2}\left(\frac{\tilde{\mu}_i^2}{\tilde{\sigma}_i^2} + \log \tilde{\sigma}_i^2 + \log(2\pi)\right)\right) \\ &+ \frac{1}{2}\left(\mu' \Sigma^{-1} \mu + \log|\Sigma|\right), \end{aligned}$$

which in turn approximates the Gaussian probability $F(\mathcal{A})$ in Eq. (4.7). It is worth noting that the algorithm is still valid for arbitrary factors $t_i(x)$, albeit with different parameter updates. Our simplistic choice of $t_i(x)$ to be the indicator function defined over the constant lower and upper limits of integration reflects the needs of the Hybrid-EP estimator and the rectangular nature of the partition sets.

In summary, the EP framework boils down to two main approximations. The first idea is to choose an approximating $q(x)$ from a tractable (Gaussian) family that closely resembles p, such that $D(p \,||\, q)$ is minimized. The intractability of p paves the way for the second approximation, for which we instead work with a simplified representation of p so that the problem reduces to locally minimizing $D\left(t_i q^{\setminus i} \,||\, \tilde{t}_i q^{\setminus i}\right)$, for $i = 1, \ldots, d$. This can be done iteratively with the Expectation Propagation Multivariate Gaussian Probability (EPMGP) algorithm described in Section 2.1.1 in Cunningham et al. (2013).

With this, we revisit the problem stated in the Hybrid-EP routine. As noted in Eq. (4.5), a prerequisite for computing $\widehat{\mathscr{Z}}_{\text{HYB-EP}}$ is the Gaussian integral,

$$\int_{A_k} \mathcal{N}\left(u \mid H_k^{-1} b_k, H_k^{-1}\right) \, du \tag{4.8}$$

which is nothing but a truncated Gaussian probability. While this integral is typically intractable, our adaptation of the EPMGP algorithm conveniently allows us to approximate this quantity. In particular, recall that A_k is the d-dimensional hyperrectangle of the form, $A_k = \prod_{l=1}^{d} [a_k^{(l)}, b_k^{(l)}]$. Taking $p_0(u) \equiv \mathcal{N}\left(u \mid H_k^{-1} b_k, H_k^{-1}\right)$ and $t_i = \mathbb{1}\left\{a_k^{(i)} < u_i < b_k^{(i)}\right\}$, for $i = 1, \ldots, d$, we observe that the target quantity in Eq. (4.8) is exactly the integral given in Eq. (4.7). Thus, we can directly use the EPMGP algorithm described above from Cunningham et al. (2013) to obtain an estimate for the Gaussian probability in Eq. (4.8). While algorithms such as the minimax tilting method (Botev, 2016) and elliptical slice sampling (Murray et al., 2010) method also solve the problem of estimating truncated Gaussian probabilities, our empirical studies suggest that EPMGP tends to converge more quickly and provide more reliable results in higher dimensions. In addition, as noted by Cunningham et al. (2013) and further verified by our own experiments, the EPMGP algorithm performs exceptionally well when the constraint set is rectangular, which coincides exactly with our setup.

4.4 Selecting the candidate point in each partition set

The final piece of the Hybrid-EP estimator that requires addressing is the representative point u_k used in the piecewise Taylor approximation $\widehat{\Psi}$ in Eq. (4.4). In the vanilla Hybrid estimator, recall that we solve a minimization problem over each of the partition sets to obtain this point. We offer a slightly different solution for the Hybrid-EP methodology that leverages the assumption of the shape of the target distribution. In our unimodal setup, a natural choice for each partition set's representative point is one that is closest to the global mode of γ. More specifically, if u_0 is the global mode, then $u_k = \operatorname{argmin}_{u \in A_k} ||u, u_0||_1$. This minimization has $O\left(n_k\right)$ time, where n_k is the number of points within the partition set A_k. Note that we can easily obtain the global mode of γ using Newton's method for root finding with little additional computational effort because we already have expressions for the gradient, Hessian, and inverse-Hessian of the objective function as part of the approximation in Eq. (4.6).

4.5 Hybrid-EP algorithm R package

With this architecture in place, we are equipped with all of the necessary to compute the Hybrid-EP estimator. In Algorithm 2 below, we again rely on our independently developed CART algorithm for the tree building and partitioning routines. While there is a MATLAB implementation available for the EPMGP algorithm, we contributed to the development of a more efficient C++ implementation, ultimately resulting in the `R` package `rcpp-epmgp` (Ding, 2020). We have woven together these optimally implemented subroutines in the `hybrid` R package, which also contains the implementation of the vanilla Hybrid estimator. By supplying a sampler for the target distribution and functions to evaluate Ψ and its gradient and Hessian, practitioners can easily obtain an approximation to the log marginal likelihood without the burdens tuning hyperparameters and monitoring convergence. In order to maintain the numerical precision in these calculations, we operate in the log scale and utilize the log-sum-exp trick for further stability. Readers can refer to Section B.1 for more detailed instructions regarding the installation and use of the `hybrid` package in practice.

4.6 Extension to regular posterior distributions

We now focus on a Bayesian setup and delineate mild conditions on the likelihood, prior, and data-generating mechanism for the Hybrid-EP algorithm to be applicable. The conditions are formulated in a non-asymptotic manner under possible model misspecification, akin to Spokoiny (2012a). We do not require the posterior to be log-concave. We also refrain from assuming a Bernstein–von Mises (BvM) phenomenon; i.e., the posterior asymptotically assuming a Gaussian shape (Ghosh et al., 2007). Instead, we only require a local quadratic bracketing of the log-likelihood and posterior concentration (Kleijn and van der Vaart, 2006).

Algorithm 2: Hybrid-EP

```
Input  : Sampler for the target distribution $\gamma$, methods for evaluating $\Psi$, $\nabla\Psi$, $\nabla^2\Psi$,
         where $\Psi$ is the negative log posterior
Output: Estimate of the logarithm of the normalizing constant of $\gamma$
Sample $u_1, \ldots, u_J \sim \gamma$
Fit a CART model, $\mathcal{T}$, to $(u_1, \Psi(u_1)), \ldots, (u_J, \Psi(u_J))$
Extract the partition $\mathcal{A} = \{A_1, \ldots, A_K\}$ from $\mathcal{T}$ of the bounding box $A$ of $\mathcal{U}$
Calculate the global mode, $u_0$, of $\gamma$
for $k \in \{1, \ldots, K\}$ do
    $u_k \leftarrow \text{argmin}_{u \in A_k} ||u - u_0||_1$
    $\lambda_k \leftarrow \nabla\Psi(u_k)$
    $H_k \leftarrow \nabla^2\Psi(u_k)$
    $C_k \leftarrow (2\pi)^{d/2}|H_k|^{-1/2} \exp\left(\frac{1}{2}\left(u_k' H_k^{-1} u_k - 2\lambda_k' u_k + \lambda_k' H_k^{-1} \lambda_k\right)\right)$
    $b_k \leftarrow H_k u_k - \lambda_k$
    $G_k \leftarrow \int_{A_k} \mathcal{N}\left(u \mid H_k^{-1} b_k, H_k^{-1}\right) du$
    $\widehat{\mathcal{Z}}_k \leftarrow C_k \cdot G_k$
end
return $\log \widehat{\mathcal{Z}} = \texttt{log-sum-exp}\left(\log \hat{\mathcal{Z}}_1, \ldots, \log \hat{\mathcal{Z}}_K\right)$
```

Suppose data Y are modeled as $Y \mid \theta \sim \mathbb{P}_\theta$; Y may denote n iid/independent samples. For each $\theta \in \Theta \subseteq \mathbb{R}^d$, assume $\mathbb{P}_\theta$ admits a density $p_\theta = (d\mathbb{P}_\theta/d\mu)$ with respect to a common σ-finite measure μ on the sample space $\mathcal{Y}$. Let $\pi(\cdot)$ be a continuous proper prior on Θ and let $\gamma(\cdot)$ denote the corresponding posterior distribution so that $\gamma(\theta) = e^{\ell(\theta)}\pi(\theta)/\mathcal{Z}$, with $\ell(\theta) = \log p_\theta(Y)$ the log-likelihood function and $\mathcal{Z}$ the posterior normalizing constant. We operate in a misspecified framework allowing the true data distribution $\mathbb{P}$ to lie outside the model class $\{\mathbb{P}_\theta : \theta \in \Theta\}$. Without loss of generality, assume $\mathbb{P} \ll \mu$ and let $p = d\mathbb{P}/d\mu$. We reserve $\mathbb{E}$ to denote an expectation with respect to $\mathbb{P}$. Let

$$\theta^* := \arg\min_{\theta\in\Theta} D\left(p\,||\,p_\theta\right) = \arg\max_{\theta\in\Theta} \mathbb{E}\ell\left(\theta\right)$$

be the closest Kullback–Leibler (KL) point to the truth inside the parameter space, with $D\left(p\,||\,q\right) = E_p\left(\log p/q\right)$ the KL divergence between densities p and q. In a misspecified setting, the pseudo-true parameter θ^* plays the role of the true parameter in well-specified models. For any $\theta, \theta^\dagger \in \Theta$, let $\ell\left(\theta, \theta^\dagger\right) := \ell\left(\theta\right) - \ell\left(\theta^\dagger\right)$ denote the log-likelihood ratio. Let $\ell_r\left(\theta\right) := \ell\left(\theta\right) - \mathbb{E}\ell\left(\theta\right)$ and

$$B^* := \{\theta \in \Theta\,:\,|\theta_j - \theta_j^*| \leq r_j\,\forall\,j \in [d]\}$$

be a rectangular neighborhood of θ^*. We now lay down the main assumptions. Throughout the following discourse, $C, C_1, C_2, \ldots$ denote global positive constants.

Assumption 1 (Posterior concentration). *There exist constants* $\eta, \delta \in (0, 1/4)$ *such that*

$$\mathbb{P}\big\{\gamma\left(B^*\right) \geq 1 - \eta\big\} \geq 1 - \delta.$$

Assumption 2 (Local quadratic bracketing). *There exists a partition* $\{B_k\}_{k=1}^K$ *of* B^* *into rectangular sets, points* $\theta_k^* \in B_k$*, positive definite matrices* $\{H_k\}_{k=1}^K$*, and constants* $c_k \in (1/2, 1)$*, such that*

$$|\mathbb{E}\ell\left(\theta_k^*, \theta^*\right)| \leq C_1 d, \forall\,k, \tag{4.9}$$

and

$$\left(\theta - \theta_k^*\right)' H_k\left(\theta - \theta_k^*\right)/(2c_k) \geq -\mathbb{E}\,\ell\left(\theta, \theta_k^*\right) \geq \left(\theta - \theta_k^*\right)' H_k\left(\theta - \theta_k^*\right)/2, \;\forall\,\theta \in B_k. \tag{4.10}$$

Assumption 3 (Stochastic component of the likelihood ratio). *There exists a positive constant*

C_2 *and* $\tilde{\delta} \in (0, 1/4)$ *such that*

$$\mathbb{P}\left\{\sup_{\theta\in B^*} |\ell_r(\theta) - \ell_r(\theta^*)| \leq C_2\, d\right\} \geq 1 - \tilde{\delta}.$$

Assumption 1 requires the posterior to place sufficient mass around the pseudo-true parameter θ^*; note that no assumptions regarding its shape is made. Conditions for posterior concentration in misspecified models can be found in Kleijn and van der Vaart (2006); see also De Blasi and Walker (2013); Sriram et al. (2013); Ramamoorthi et al. (2015); Atchadé (2017); Bhattacharya et al. (2019). Assumptions 2 and 3 together pose mild additional conditions on $\ell(\theta, \theta^*)$ in the neighborhood B^*. We separate the conditions into stochastic and deterministic components by writing

$$\ell(\theta, \theta^*) = \mathbb{E}\ell(\theta, \theta_k^*) + \mathbb{E}\ell(\theta_k^*, \theta^*) + \ell_r(\theta) - \ell_r(\theta^*)$$

within B_k. Assumption 1 posits that $-\mathbb{E}\,\ell(\theta, \theta_k^*)$ can be locally bracketed by a quadratic form in $(\theta - \theta_k^*)$ inside the partition set B_k. Since $-\mathbb{E}\ell(\theta, \theta^*) \geq 0$, it remains positive in a neighborhood of θ^* by continuity, and hence (4.10) is not vacuous. If $\theta \mapsto \mathbb{E}\ell(\theta)$ is twice differentiable, a natural choice is to perform a local Taylor expansion around θ_k^* and set H_k to the Hessian matrix. Contrast this to the global quadratic expansion in Cavanaugh and Neath (1999), which is among the most general derivations of the Laplace approximation. Assumption 2 controls the supremum of the centered empirical process $\ell_r(\cdot)$ over the set B^*. Such probabilistic bounds can be derived using standard chaining arguments; see Talagrand (2006); Boucheron et al. (2013); Dirksen (2015); Vershynin (2018) and van de Geer (2006); Spokoiny (2012b) in statistical contexts. More details can be found in the Appendix A.

Theorem 1. *Under assumptions 1–3, with* $\mathbb{P}$*-probability at least* $(1 - \delta - \tilde{\delta})$*,*

$$e^{\ell(\theta^*)}\, e^{C_3 d} \sum_{k=1}^{K} \big[\inf_{\theta\in B_k} \pi(\theta)\big] h_k \gamma_{1k} \leq \mathcal{Z} \leq \frac{e^{\ell(\theta^*)}\, e^{C_4 d}}{1-\eta} \sum_{k=1}^{K} \big[\sup_{\theta\in B_k} \pi(\theta)\big] h_k \gamma_{2k},$$

where

$$
\begin{aligned}
h_k &= [\det(H_k)]^{-1/2}, \\
\gamma_{1k} &= \int_{B_k} \mathcal{N}_d\left(\theta; \theta_k^*, c_k H_k^{-1}\right) d\theta, \\
\gamma_{2k} &= \int_{B_k} \mathcal{N}_d\left(\theta; \theta_k^*, H_k^{-1}\right) d\theta.
\end{aligned}
$$

The tight two-sided bound in Theorem 1 offers a non-asymptotic generalization to the classical Laplace approximation. Indeed, with $K = 1$, it correctly recovers the $-d \log n/2$ BIC penalty. More importantly, its derivation offers insights into the population quantities that the Hybrid-EP algorithm targets. The concentration of the posterior narrows down the activity to B^*, and the local quadratic bracketing condition in Assumption 2 helps bound $\mathbb{E}\ell\left(\theta, \theta_k^*\right)$ by quadratic terms from both sides within B_k. Taking care of the stochastic component using the deviation bound in Assumption 3, one obtains the two-sided bound in terms of the rectangular Gaussian probabilities γ_k. The Hybrid-EP method uses data-driven estimates for each of these components to provide a computable bound in the spirit of Theorem 1. Specifically, the posterior samples are first used to determine B^* and then to obtain the partition sets B_k^*, and the EP algorithm approximates the γ_ks after a careful choice of the θ_k^*s.

4.7 Numerical experiments

In this chapter's experiments, we target unimodal densities. One key difference from the previous chapter's examples is that in most of the following simulations, the true normalizing constant is unknown and must be estimated. As a result, there is often no baseline that can be used for evaluation, so we rely on the results of other well-established estimators. We first evaluate the performance of the Hybrid-EP estimator in a commonly presented logistic regression model example. Next, we expand the repertoire of examples beyond the usual linear/logistic regression model setup by incorporating Gaussian graphical models (GGM), where the parameter space is non-Euclidean. Estimating normalizing constants of GGMs (based on Wishart distributions) is an active area of research (Lauritzen, 1996), motivated by the need to approximate marginal likelihoods for graph selection. Difficulties arise when the underlying graph is non-decomposable, and barring special cases (Uhler et al., 2016), no closed form analytic solutions are available. Dedicated sampling based approaches (Atay-Kayis and Massam, 2005) successfully estimate the normalizing constants in lower dimensions, which give us a strong baseline to use for comparison. However, these methods often fail in higher dimensions, making the problem of estimating normalizing constants in non-decomposable graphical models extremely challenging. As we shall see in Section 4.7.2.2, the Hybrid-EP estimator is successful in providing accurate approximations even in these challenging settings, making it a valuable independent contribution in the graphical modeling literature.

For the graphical model examples, we use the `BDgraph` package (Mohammadi and Wit, 2019) which implements the algorithm from Atay-Kayis and Massam (2005), which we refer to as the GNORM algorithm/estimator. As mentioned previously, the GNORM estimator is a specialized technique for graphical models that simplifies the structure of the integral and is widely accepted as a state of the art method for computing normalizing constants for the G-Wishart density. We provide an overview of this algorithm in Section C.4. We also use the `bridgesampling` package (Gronau et al., 2020) for its implementation of both the bridge sampling estimator (BSE) and the warp bridge sampling estimator (WBSE). The specificity (as well as the non-Gaussian nature) of the problem prevents some of the more traditional normalizing constant estimation methods from being directly adapted, so there are fewer competing estimators in the graphical models simulations.

Another difference in the setup of the experiments in Section 3.7 of the previous chapter compared to that of this chapter is that the former focused on the case where a limited number of MCMC samples is available. In the following simulations, however, we instead seek to highlight both the Hybrid-EP estimator's versatility when applied to problems that have parameter spaces whose shapes are irregular (highly non-Gaussian) and ability to scale well with the dimension of the parameter space.

4.7.1 Competing logistic regression models

First, we consider competing logistic regression models for the Pima Indians dataset, where the binary response is an indicator of diabetes for $n = 532$ Pima Indian women. Then, we have the response-covariate pairs $(y_i, x_i) \in \{0, 1\} \times \mathbb{R}^{d+1}$ for $i = 1, \ldots, n$, with

$$p_i = \mathrm{pr}\left(y_i = 1 \mid x_i\right) = \frac{e^{x_i'\theta}}{1 + e^{x_i'\theta}}.$$

This gives us the following likelihood function,

$$p\left(y \mid \theta\right) = \prod_{i=1}^{n} p_i^{y_i} \left(1 - p_i\right)^{1-y_i}.$$

For the parameter θ, we assume a Gaussian prior, $\theta \sim \mathcal{N}\left(0, (\tau I_d)^{-1}\right)$. Then we have the following intractable marginal likelihood,

$$\mathcal{Z} = \int_{\mathbb{R}^d} e^{\sum_i y_i x_i'\theta - \sum_i \log\left(1+e^{x_i'\theta}\right)} \left(2\pi\right)^{-d/2} \tau^{d/2} e^{-\frac{\tau}{2}\theta'\theta} d\theta.$$

With this setup, we consider the model selection problem between the following two competing models,

$$\begin{aligned}
\mathcal{M}_1 &= \text{logit}\left(p\right) = 1 + \texttt{NP} + \texttt{PGC} + \texttt{BMI} + \texttt{DP}, \\
\mathcal{M}_2 &= \text{logit}\left(p\right) = 1 + \texttt{NP} + \texttt{PGC} + \texttt{BMI} + \texttt{DP} + \texttt{AGE},
\end{aligned}$$

with four and five predictors, respectively. The definitions and additional context of each of the predictors used in the logistic regression models above can be found in Section B.3.6. For evaluating these two models, we need to form the Bayes Factor, $\mathrm{BF}_{1,2} = p\left(y \mid \mathcal{M}_1\right)/p\left(y \mid \mathcal{M}_2\right)$. We demonstrate the use of the Hybrid-EP estimator to compute the marginal likelihood of $\mathcal{M}_1, \mathcal{M}_2$. Defining $X = (x_1, \ldots, x_n)'$ to the $n \times (d+1)$ design matrix, we can write the negative log posterior as:

$$\Psi(\theta) = -y'X\theta + \sum_{i=1}^{n} \log\left(1 + e^{x_i'\theta}\right) + \frac{d}{2}\log(2\pi) - \frac{d}{2}\tau + \frac{\tau}{2}\theta'\theta. \tag{4.11}$$

Using Eq. (4.11), we can compute the gradient and Hessian of $\Psi(\theta)$, so we have all the necessary functions to proceed with the Hybrid-EP algorithm. In order to obtain the MCMC samples, we rely on the sampler implemented in Friel and Wyse (2012). We then compare the performance of the Hybrid-EP estimator with other popular methods, such as the Laplace method (L), Laplace at the Maximum a Posteriori (L-MAP), Chib's method (C), Annealed Importance Sampling (AIS), Power Posterior (PP), Brosse's estimator (AV), and the Hybrid-EP estimator (HYB-EP). We include the AV estimator from Brosse et al. (2018) as another method for estimating the normalizing constant of log-concave densities. The AV estimator, whose implementation is available at `https://github.com/nbrosse/normalizingconstant`, makes refinements to existing algorithms and provides appealing theoretical bounds and guarantees. In our experiments, we use each of these estimators to compute the marginal likelihood for both $\mathcal{M}_1$ and $\mathcal{M}_2$ and summarize the results in the boxplot shown in Figure 4.1. For fair evaluation, each of the MCMC-based methods uses the same 200,000 posterior samples. See Section 4.2 of Friel and Wyse (2012) for a more complete picture of the exact experimental setup and the hyperparameter settings. Given the large quantity of MCMC samples, we report experimental results from 10 replications. Aside from the AIS estimator, all of the other estimators have low variance and produce similar results. In particular, L, L-MAP, PP, and HYB-EP all roughly agree on the marginal likelihood values and have extremely low variance.

4.7.2 Graphical models

In the following examples, we revisit the problem of estimating the normalizing constant for distributions in the context of Gaussian graphical models. Recall from Section 3.7.4 that $G = (V, E)$ denotes an undirected graph with vertex set $V = \{1, \ldots, p\}$ and edge set E. Further, we have X coming from a normal distribution $\mathcal{N}(\mu, \Sigma)$ satisfying the GGM with graph G, with

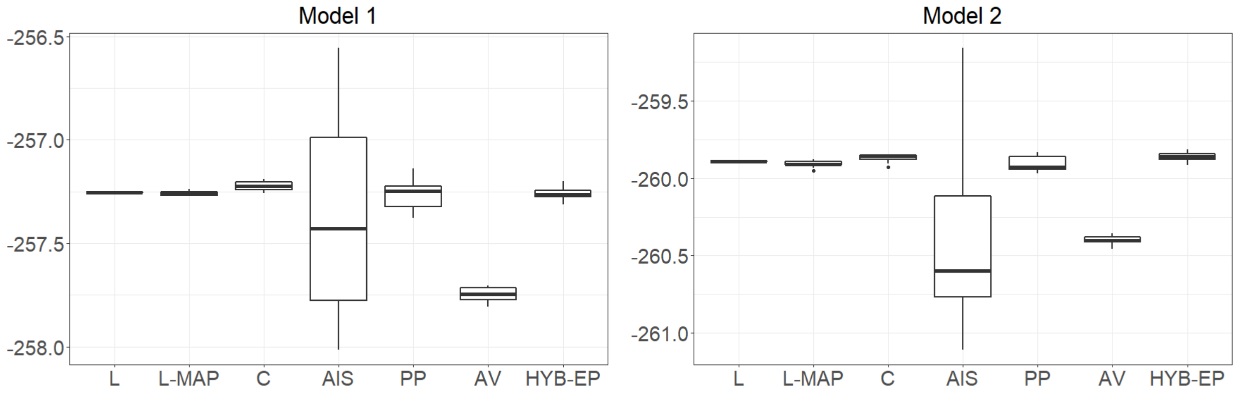


Figure 4.1: Boxplots of the log marginal likelihood for two competing logistic regression models for the Pima Indians dataset. The included methods are: the Laplace method (L), Laplace at the Maximum a Posteriori (L-MAP), Chib's method (C), Annealed Importance Sampling (AIS), Power Posterior (PP), Brosse et al. (2018) (AV), and Hybrid-EP (HYB-EP)

$\Sigma \in \mathbb{S}^p_{\succ 0}(G)$. In order to perform inference on the parameter $\Sigma^{-1} = \Omega$, we require a prior distribution for (Ω, G). Conditional on G, we consider two options for the prior. In Section 4.7.2.1, we assume G is a decomposable graph, so one option is to place a Hyper-Inverse Wishart (HIW) (Dawid and Lauritzen, 1993) prior on Ω. In Section 4.7.2.2, we seek to broaden the scope of examples and consider G to be a general (non-decomposable) graph, leading to a G-Wishart (GW) (Roverato, 2000) prior on Ω. In the following analyses, we make references to p, the cardinality of the vertex set of G, and d, the dimension of the parameter space, the latter of which coincides with the number of free (nonzero) elements on and above the diagonal of the adjacency matrix of G.

We emphasize that even though the setup of the graphical model examples in this section is quite similar to those shown in the previous chapter, the dimension of the parameter space here is notably higher, which causes more computational difficulties associated with marginal likelihood estimation. While there has been substantial work done in the realm of Gaussian graphical models that deals with inference for both decomposable (Giudici and Green, 1999) and non-decomposable (Dellaportas et al., 2003; Khare et al., 2015; Atay-Kayis and Massam, 2005) graphs, the need for algorithms that scale well in high dimensions is ever present. Many of the existing methods in literature are quite restrictive in their assumptions and only perform well in simple, low-dimensional problem settings. Also, the ground truth actually may not be analytically available when the underlying graph is non-decomposable. In such cases, dedicated methods, such as the importance sampling algorithm from Atay-Kayis and Massam (Atay-Kayis and Massam, 2005) can provide estimates, but these tend to only be reliable in moderate dimensions and have clear computational limitations. Given this, our development of a novel method that has proven to adequately handle high-dimensional problem settings is a valuable contribution to the graphical models literature and also adds a versatile solution that can be easily be adapted to a wide variety of graphical modeling contexts.

4.7.2.1 Hyper inverse-Wishart induced Cholesky factor density

Recall that for $x_1, \dots, x_n \overset{\text{iid}}{\sim} \mathcal{N}_p\left(0, \Sigma\right)$, the likelihood function can be written as follows,

$$L(\Sigma) = (2\pi)^{-np/2} \det\left(\Sigma\right)^{-n/2} e^{-\operatorname{tr}\left(\Sigma^{-1}S\right)/2}, \quad B = \sum_{i=1}^{n} x_i x_i'. \tag{4.12}$$

Given G, we place a HIW (δ, Λ) prior on $\Omega = \Sigma^{-1}$, where $\delta > 2$ is the degrees of freedom and $\Lambda \in \mathbb{S}^p_{\succ 0}$ is fixed. Like before, we work with the Cholesky factor ϕ, where $\phi'\phi = \Omega$. Then, the results from Section 3.7.4 carry over and we can once again write the likelihood function in terms of the upper Cholesky factor, ϕ, and derive the induced prior. Both of these are given in Eq. (3.12) and (3.13), respectively.

For this example, we set $\Lambda = I_p$. Like before, we can generate samples of Ω from the posterior distribution HIW$_G\left(\delta + n, I_p + B\right)$, extract the Cholesky factor ϕ and evaluate $\Psi\left(\phi\right)$ using Eq. (3.12) and (3.13). In addition to this, we also require expressions for the gradient and Hessian of $\Psi\left(\phi\right)$ in order to use the Hybrid-EP algorithm. The closed form expressions for these quantities, as well as their derivations are shown in Section C.3.

In our simulations, we try to emulate a high-dimensional setting by stacking the adjacency matrix of G_9 (shown in Figure 4.2) 8 and 10 times along the diagonal to construct larger graphs, G_{72} and G_{90}, with $d = 200, 250$ free elements, respectively. Using these large graphs, we draw data that satisfy their corresponding GGMs. We take $\delta = 3, n = 100$, and $\Lambda = I_p$, for $p = 72, 90$, and

Table 4.1: Mean, average error (AE), root mean squared error (RMSE) for the approximations to the log normalizing constant of the $\text{HIW}_G\,(\delta + n, I_p + B)$ distribution. G_{72} has 72 vertices with 200 parameters and G_{90} has 90 vertices and 250 parameters. We compare Hybrid-EP results with the Bridge sampling estimator (BSE) and Warp Bridge sampling estimator (WBSE). Estimates are reported over 100 replications, each using 1000 samples from the true posterior.

		TRUTH	**BSE**	**WBSE**	**HYB-EP**
$p = 72$ $d = 200$	Mean	-6231.297	-6230.561	-6230.594	-6231.7054
	AE	0	-0.7356	-0.7026	0.4086
	RMSE	0	1.9328	2.2161	0.4506
$p = 90$ $d = 250$	Mean	-7880.95	-7875.6947	-7875.7620	-7881.4622
	AE	0	-5.2554	-5.1880	0.5121
	RMSE	0	6.3054	6.0863	0.5574

compare the HYB-EP results with the ground truth (available for HIW distributions), GNORM, BSE, and WBSE. While the GNORM estimator gives accurate estimates for lower dimensions, it fails to produce sensible outputs for high-dimensional settings, so we exclude it from Table 4.1. From the simulation results, we see that when $d = 200$, BSE and WBSE are reasonably competitive with HYB-EP, but when $d = 250$, the quality of both bridge sampling estimators deteriorates. The relatively small error of the Hybrid-EP estimator in this high-dimensional graphical models setting is particularly encouraging, considering the prominence of the GNORM estimator in the graphical modeling literature. We attempted to incorporate the Nested sampling estimator as another competitor, but the computational overhead prevented us from obtaining meaningful results within reasonable time constraints.

4.7.2.2 G-Wishart prior for general graphs

Since the constraint that a graph be decomposable is quite restrictive, the extension to arbitrary graphs paves the way for a more general distribution that allows for the same analysis to be done in more diverse contexts. For a general graph G that may not be decomposable, a popular choice

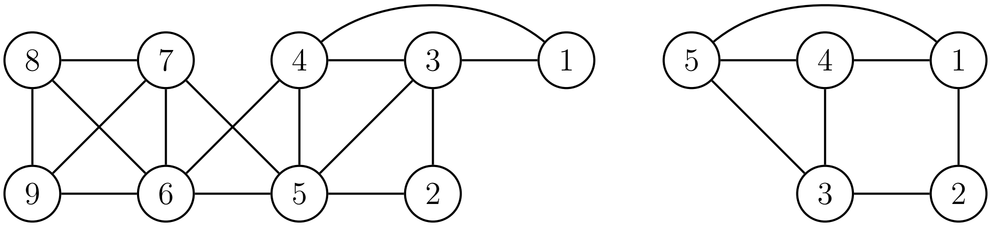


Figure 4.2: G_9 (left), a decomposable graph with vertices enumerated according to a perfect elimination ordering, and G_5 (right), an undirected, non-decomposable graph.

for the prior is the G-Wishart prior on Ω, GW (δ, Λ), which has the following density,

$$f\left(\Omega \mid G\right) \propto \left|\Omega\right|^{(\delta-2)/2} \exp\left(-\text{tr}\left(\Omega\Lambda\right)/2\right). \tag{4.13}$$

Here, Ω, Λ are $p \times p$ non-negative definite matrices, and $\delta > 2$ is the degrees of freedom. As a result, developing computational tools to more efficiently use this distribution for inference has been a popular research area, leading to various methods for reliably sampling from the G-Wishart distribution and doing model comparison and model search (Rajaratnam et al., 2008; Piccioni, 2000; Lenkoski, 2013). While the density expression is similar to that of the HIW density, the tractability of the normalizing constant that we previously enjoyed in the HIW example is no longer universally present because G is no longer assumed to be decomposable. Estimating this normalizing constant then becomes a computationally challenging problem. In the following simulations, we investigate two different problem settings. First, we make some simplifying assumptions that make the G-Wishart normalizing constant available in closed form in order to establish the accuracy of the Hybrid-EP estimator. After verifying the viability of our proposed solution, we then generalize the problem to encapsulate more typical settings where the normalizing constant is intractable so that we can better compare the Hybrid-EP estimator with GGM-specific algorithms.

4.7.2.3 Exact formula for normalizing constant of G-Wishart density

From Uhler et al. (2016), we know that for special cases, exact formulae for G-Wishart normalizing constants are available. These will serve as our ground truth to evaluate the Hybrid-EP estimator. In order for us to take advantage of these closed form normalizing constants, we initially restrict our focus to GW densities for which the scale matrix is diagonal. Instead of duplicating the setup in the previous section and incorporating the likelihood function so that we are dealing with a posterior distribution with a potentially complicated scale matrix, we modify the degrees of freedom and the scale matrix so that the resulting distribution mimics a posterior distribution, i.e., one that exhibits concentration and appears more Gaussian. This can be done by taking a larger value for δ and taking the scale matrix to be $\Lambda = nI_p$, where n is large. The normalizing constant that we wish to compute is of the form

$$C_G\left(\delta, \Lambda)\right) = 2^{\frac{1}{2}p\delta + |E|} \cdot I_G\left(\tfrac{1}{2}(\delta - 2), \Lambda\right), \tag{4.14}$$

$$I_G\left(\delta, \Lambda\right) = \int_{\mathbb{S}^p_{\succ 0}} |\Omega|^{\delta} \exp\left(-\text{tr}\left(\Omega\Lambda\right)\right)\, d\Omega, \tag{4.15}$$

which can be computed using Theorem 3.3 in Uhler et al. (2016). Here,

$$d\Omega = \prod_{i=1}^{p} d\omega_{ii} \cdot \prod_{i<j,(i,j)\in E} d\omega_{ij}.$$

With the parameters (δ, nI_p), we perform a change of variable $I_p \to nI_p$, to obtain the following normalizing constant:

$$C_G(\delta, nI_p) = 2^{\frac{1}{2}p\delta + |E|} \cdot I_G\left(\tfrac{1}{2}(\delta - 2), I_p\right) \cdot n^{-\frac{1}{2}p\delta - |E|}. \tag{4.16}$$

Then, for $G = G_5$ as shown in Figure 4.2, we can use the following formula given in Eq. (2.4) in Uhler et al. (2016) in conjunction with Eq. (4.16) to derive an *exact* expression for the normalizing constant of GW (δ, nI_5), conditional on the graph G_5,

$$I_{G_5}(\delta, I_5) = \pi^{7/2} \frac{\Gamma\left(\delta + \frac{5}{2}\right)}{\Gamma\left(\delta + 3\right)} \Gamma\left(\delta + 1\right) \Gamma(\delta + \tfrac{3}{2}) \left[\Gamma\left(\delta + 2\right)\right]^2 \Gamma(\delta + \tfrac{5}{2}).$$

4.7.2.4 G-Wishart induced Cholesky factor density

Like before, we require an expression for Ψ in order to use HYB-EP. We first establish some of the notation relevant to the GW density. For $G = (V, E)$, we follow the conventions in (Roverato, 2000) and define

$$\begin{aligned}\mathscr{V} &= \{(i, j) : i \leq j \text{ where } i = j, i \in V \text{ or } (i, j) \in E\},\\ \mathscr{W} &= \{(i, j) : i, j \in V, i \leq j\}.\end{aligned}$$

Then, let $\bar{\mathscr{V}} = \mathscr{W} \setminus \mathscr{V}$, and $A_{p\times p} = (a_{ij})$, where $a_{ij} = 0$ if $(i, j) \in \bar{\mathscr{V}}$ or if $i = j$, and $a_{ij} = 1$ otherwise. Then, let k_i be the number of 1's in the i-th column of A. Proceeding in a similar fashion as in the HIW example, we take the Cholesky decomposition of $\Omega = \phi'\phi$ and $\Lambda = (T'T)^{-1}$, where $T = (t_{ij})_{1\leq i\leq j\leq p}$. As an added step, we make a change of variable $\zeta = \phi T^{-1}$. The Jacobian of the first transformation $\Omega \to \phi$ is identical to the one given in Section 4.7.2.1, and the Jacobian of the second transformation $\phi \to \zeta$ is given by $\prod_{i=1}^{p} t_{ii}^{k_i+1}$. Putting this all together, we can rewrite the normalizing constant as an integral over the free variables of $\zeta = (\zeta_{ij})_{1\leq i\leq j\leq p}$,

$$\begin{aligned}C_G(\delta, \Lambda) = 2^p \prod_{i=1}^{p} (t_{ii}^2)^{(\delta+b_i-1)/2} \int \exp\left(-\frac{1}{2} \sum_{(i,j)\in\bar{\mathscr{V}}} \zeta_{ij}^2\right) \prod_{i=1}^{p} (\zeta_{ii}^2)^{(\delta+\nu_i-1)/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{p} \zeta_{ii}^2\right)\\ \times \exp\left(-\frac{1}{2} \sum_{(i,j)\in\mathscr{V}, i\neq j} \zeta_{ij}^2\right) \prod_{i=1}^{p} d\zeta_{ii} \prod_{(i,j)\in\mathscr{V}, i\neq j} d\zeta_{ij},\end{aligned} \tag{4.17}$$

where $b_i = \nu_i + k_i + 1$, $\nu_i = |\text{ne}(i) \cap \{i + 1, \ldots, p\}|$, and $\text{ne}\,(i) = \{j \in V : (i, j) \in E\}$. Taking log of the integrand above, we can write the following expression for $\Psi\,(\zeta)$,

$$\Psi\,(\zeta) = C + \sum_{i=1}^{p} (\delta + \nu_i - 1) \log \zeta_{ii} - \frac{1}{2}\sum_{i=1}^{p} \zeta_{ii}^2 - \frac{1}{2} \sum_{(i,j)\in\bar{\mathscr{V}}} \zeta_{ij}^2 - \frac{1}{2} \sum_{(i,j)\in\mathscr{V}, i\neq j} \zeta_{ij}^2, \tag{4.18}$$

where $C = p \log(2) + \sum_{i=1}^{p} (\delta + b_i - 1) \log t_{ii}$. Another key difference between the GW and HIW setups is how the non-free elements change the objective function Ψ. In the case of the HIW density, the Cholesky factor observes the same sparsity pattern as the adjacency matrix of the graph G, so the free elements are simply taken to be the nonzero elements in the upper Cholesky factor. This makes the evaluation of Ψ very simple because it is defined over the nonzero elements of the upper Cholesky factor. However, in the case of the GW distribution, the sparsity that we see in G is not necessarily reflected in the Cholesky factors. Therefore, Ψ is no longer a function of exclusively the nonzero elements of ζ. Indeed, upon inspection of the integral in Eq. (4.17), we observe that non-free elements, denoted as $\zeta_{ij}, (i, j) \in \bar{\mathscr{V}}$, have nonzero contribution to the objective function Ψ. From Lemma 2 in Atay-Kayis and Massam (2005), we know that the non-free elements of ζ are actually functions of the free elements. Furthermore, we can explicitly represent each of these non-free entries using the following recursive formula,

$$\zeta_{rs} = \sum_{j=r}^{s-1} \left(-\zeta_{rj} \frac{\lambda_{js}}{\lambda_{ss}} \right) - \sum_{i=1}^{r-1} \left(\frac{\zeta_{ir} + \sum_{j=i}^{r-1} \zeta_{ij} \frac{\lambda_{jr}}{\lambda_{rr}}}{\zeta_{rr}} \right) \left(\zeta_{is} + \sum_{j=i}^{s-1} \zeta_{ij} \frac{\lambda_{js}}{\lambda_{ss}} \right), \tag{4.19}$$

for $(r, s) \in \bar{\mathscr{V}}$ and $r < s$. As a result, these terms must also be accounted for when computing the gradient and Hessian of $\Psi(\zeta)$. Expressions for both of these quantities and the details for their calculations are in Appendix C.

In the following experiments, we form larger graphs by stacking G_5 along the diagonal 15, 17, and 19 times, which gives us 180, 204, and 228 free parameters, respectively. Taking $\delta = 100$ and $n = 100$ to help mimic a posterior distribution, we can then compute the normalizing constant directly using the formula derived in Eq. (4.16) and compare any subsequent approximations to this. From our results in Table 4.2, we observe that the Hybrid-EP estimator delivers accurate results even when $d > 200$. Similar to the results from the HIW example, BSE and WBSE take considerably longer to converge and falter as we move to larger graphs, while the Hybrid-EP estimator consistently delivers reliable estimates.

Since we are not dealing with a true posterior distribution in this example and the scale matrices are diagonal and thus much simpler in structure, GNORM performs exceedingly well. However, as seen in the analysis of the previous HIW example where we consider a true posterior distribution

Table 4.2: Mean, average error (AE), root mean squared error (RMSE) for the approximations to the log normalizing constant of the GW density for graphs with p vertices and d free elements. We compare HYB-EP with the Bridge sampling estimator (BSE) and Warp Bridge sampling estimator (WBSE). Estimates are reported over 100 replications, each using 1000 samples from the true posterior.

		TRUTH	**BSE**	**WBSE**	**HYB-EP**
$p = 75$ $d = 180$	Mean	-3973.049	-3972.7178	-3972.7340	-3973.5982
	AE	0	-0.3316	-0.3154	0.5488
	RMSE	0	1.4213	1.4177	0.5994
$p = 85$ $d = 204$	Mean	-4502.789	-4501.371	-4501.315	-4503.455
	AE	0	-1.4180	-1.4743	0.6653
	RMSE	0	2.5325	2.3042	0.7174
$p = 95$ $d = 228$	Mean	-5032.529	-5029.687	-5029.526	-5033.179
	AE	0	-2.8425	-3.0034	0.6499
	RMSE	0	3.6417	3.6321	0.7143

with a non-diagonal scale matrix, GNORM fails to converge. We investigate this problem setting more thoroughly in the next section, where we also provide a scalable method for computing the normalizing constant of the G-Wishart density with non-diagonal scale matrices and large non-decomposable graphs.

4.7.2.5 G-Wishart density for non-diagonal scale matrices

After demonstrating that the Hybrid-EP estimator indeed produces sensible estimates for cases where the normalizing constant can be analytically verified, we proceed with the general setting for which the scale matrix is non-diagonal. Note that the diagonal assumption from the previous section led to significant simplifications in the recursive formula for each of the ζ_{rs}. While Eq. (4.19) suggests a complicated relationship between free and non-free parameters, when Λ is diagonal, this term becomes

$$\zeta_{rs} = -\frac{1}{\zeta_{rr}} \sum_{k=1}^{r-1} \zeta_{kr}\zeta_{ks},$$

making the structure of ζ less cumbersome. Because the Hybrid-EP algorithm requires gradient and Hessian calculations of $\Psi(\zeta)$, the corresponding derivatives taken with respect to the free elements propagate through these recursive definitions. The diagonal assumption therefore eliminates the need for many of these calculations.

In the non-diagonal case, we do not benefit from any of the simplifying assumptions and must keep track of all of the terms in Eq. (4.19). Because the normalizing constant in this case is intractable, we rely on the GNORM estimator from Atay-Kayis and Massam (2005), which is widely accepted as a state of the art method for computing normalizing constants for G-Wishart densities. However, we will see that the GNORM estimates quickly fail to be a viable solution as the dimension of the graph increases. In order to combat the computational problems associated with these larger graphs, we make a small modification to the Hybrid-EP algorithm that breaks down the graph into subgraphs according to its topology using a junction tree representation. These subgraphs, while still potentially yielding intractable target quantities, pose much smaller scale problems to which we can apply our approximation scheme. By sidestepping the original problem and leveraging the attractive properties of a junction representation of a connected graph, we can achieve drastic computational speedup in the marginal likelihood calculation that would otherwise be prohibitively expensive.

4.7.2.6 Junction tree representation

We briefly discuss the properties of decomposable versus non-decomposable graphs and how their respective junction tree representations produce a viable method for simplifying the marginal likelihood calculation. Connected graphs can be decomposed, often in different ways, into sequences of interconnecting subgraphs separated by complete subgraphs. Such a decomposition is known as a junction tree representation of a graph, which defines an ordered sequence of subgraphs with a tree structure. Based on a specified (usually arbitrary) ordering of the nodes, a junction tree decomposition has the form

$$G \mapsto \mathcal{J}_G = \{P_1, S_2, P_2, S_3, \dots, P_{m-1}, S_m, P_m\}. \tag{4.20}$$

We highlight three important properties of this decomposition:

- Each *prime component* P_i, $i = 1, \ldots, m$, is a proper subgraph of G which may or may not be complete.
- Each *separator* S_i, $i = 2, \ldots, m$, is a complete subgraph of G, regardless of whether or not G is decomposable.
- S_i is the intersection of P_i with all the previous components $\{P_1, P_2, \ldots, P_{i-1}\}$, so that S_i separates the next component from the previous set.

Essentially, if we decompose a graph into subgraphs until there exists no further decomposition, then the resulting collection of subgraphs, $\{P_i : i = 1, \ldots, m\}$, is the set of prime components. Consequently, the junction tree representation is nothing but the set of the m prime component subgraphs of G linked by the sequence of $m - 1$ separating subgraphs, $\{S_2, \ldots, S_m\}$. One important concept to note is that the existence of a decomposition does not imply that a graph is decomposable. If *any* of the prime components found by this iterative decomposition procedure is not complete and cannot be further decomposed, then that component is non-decomposable, thereby making the entire graph non-decomposable (Fitch et al., 2014). From this definition, we see that a decomposable graph is one that can be successively decomposed into its cliques, i.e., P_i is complete for $i = 1, \ldots, m$.

With these concepts in place, we recall the original goal of marginal likelihood estimation, but also look to include the intermediate step of obtaining the junction tree representation of the graph. By working with the prime components instead of the original graph, we can overcome the complexity associated with high-dimensional graphs by working on the individual subgraphs, ultimately reducing the dimension of the original calculation. In addition to the computational benefit of working with lower dimensional graphs, we are also able to take advantage of the distributional properties of the complete prime components.

Like before, suppose we have data $X = \{x_1, \ldots, x_n\}$ satisfying the GGM with a general graph G. Since G is assumed to be connected, we can represent G with a junction tree, $\mathcal{J}_G$, just as we have done in Eq. (4.20). We denote the prime component sequence as $\mathscr{P}$ and the separator sequence as $\mathscr{S}$. In the case where all of the prime components are complete (cliques), we denote the clique sequence as $\mathscr{C}$. In the general case, however, the prime components may or may not be complete. In the following discourse, we denote cliques as C and prime components as P. As mentioned previously, regardless of whether or not a graph is decomposable, we can use the junction tree representation of G to write the joint density of X given Σ,

$$p\left(X \mid \Sigma\right) = \frac{\prod_{P \in \mathscr{P}} p\left(X_P \mid \Sigma_P\right)}{\prod_{S \in \mathscr{S}} p\left(X_S \mid \Sigma_S\right)}, \tag{4.21}$$

where Σ_P and Σ_S denote the corresponding sub-matrices of the covariance matrix for the prime components and separators, respectively. Note that this likelihood function factorizes over the prime components and separators.

Since the G-Wishart distribution is a generalization of the hyper inverse-Wishart distribution, we first revisit the case where G is decomposable, with a hyper inverse-Wishart prior, $\mathrm{HIW}_G\left(\delta, \Lambda\right)$, on Σ^{-1}. Since G is decomposable, the junction tree representation can be written as $\mathscr{J}_G = \{C_1, S_2, C_2, S_3, \ldots, C_{m-1}, S_m, C_m\}$, where the prime components are cliques. Then, the prior density factorizes over the cliques and separators,

$$p\left(\Sigma \mid G\right) = \frac{\prod_{C \in \mathscr{C}} p\left(\Sigma_C \mid G\right)}{\prod_{S \in \mathscr{S}} p\left(\Sigma_S \mid G\right)}. \tag{4.22}$$

The completeness of the prime components admits distributional properties that make the normalizing constants tractable. In particular, the prior density on Σ_C is inverse-Wishart,

$$p\left(\Sigma_C \mid \delta, \Lambda_C\right) \propto \left|\Sigma_C\right|^{-\frac{\delta+2|C|}{2}} \operatorname{etr}\Big\{ -\frac{1}{2}\Sigma_C^{-1}\Lambda_C \Big\}, \tag{4.23}$$

which has a known normalizing constant. See Section C.2 for more details. Putting this together with the likelihood given in Eq. (4.21), we deduce that the marginal likelihood also factorizes over the cliques and separators as follows,

$$
\begin{aligned}
p\left(X \mid G\right) &= \int_{\Sigma|G} p\left(X \mid G\right) p\left(\Sigma \mid G\right) d\Sigma \\
&= \left(2\pi\right)^{-\frac{np}{2}} \frac{h\left(G, \delta, \Lambda\right)}{h\left(G, \delta + n, \Lambda + B\right)} && (4.24) \\
&= \left(2\pi\right)^{-\frac{np}{2}} \frac{\prod_{C \in \mathscr{C}} w\left(C\right)}{\prod_{S \in \mathscr{S}} w\left(S\right)}. && (4.25)
\end{aligned}
$$

Here $B = \sum_i x_i x_i'$. In this case, the factorization of the likelihood and prior over the cliques yields a product of tractable normalizing constants. Therefore, for a decomposable graph, the hyper inverse-Wishart normalizing constants in Eq. (4.24) for the prior and posterior distributions are functions of the normalizing constants for the inverse-Wishart clique and separator densities, which have closed forms. Exact formulae for $h\left(G, \delta, \Lambda\right)$, $w\left(C\right)$ and $w\left(S\right)$ can be found in Eq. (C.5) and (C.6).

We can easily generalize the calculations above to those of a non-decomposable graph G since the G-Wishart density in Eq. (4.13) has a similar form to the hyper inverse-Wishart density (up to a normalizing constant). As a result, the marginal likelihood calculation for the G-Wishart prior mirrors that of the HIW prior such that it also factorizes over the prime components and separators,

$$
p\left(\mathrm{X} \mid G\right) = \left(2\pi\right)^{-\frac{np}{2}} \frac{h\left(G, \delta, \Lambda\right)}{h\left(G, \delta + n, \Lambda + B\right)} = \left(2\pi\right)^{-\frac{np}{2}} \frac{\prod_{P \in \mathscr{P}} w\left(P\right)}{\prod_{S \in \mathscr{S}} w\left(S\right)}. \tag{4.26}
$$

The difference in this calculation is that the product in Eq. (4.25) is taken over the *cliques*, whereas in Eq. (4.26), the product is taken over general prime components, which may or may not be complete. Since the normalizing constants for the non-complete prime components are generally not available in closed form, they require estimation via MCMC methods, or in our case, the Hybrid-EP algorithm.

In summary, the representation of the marginal likelihood in Eq. (4.26) conveniently breaks up the original calculation involving the entire graph G into sub-problems that are intrinsically lower-dimensional and can be solved more efficiently. When $P \in \mathscr{P}$ is complete, we can rely on the closed form for the corresponding inverse-Wishart normalizing constant to compute $w(P)$. For non-complete $P \in \mathscr{P}$, we take the corresponding parameters, $\Sigma_P^{-1} = \Omega_P, B_P, \Lambda_P$, and write the normalizing constant for the non-complete prime component as the following integral

$$h(P, \delta, \Lambda_P) = \int |\Omega_P|^{\frac{\delta-2}{2}} \exp\left(-\operatorname{tr}(\Omega_P \Lambda_P)/2\right) d\Omega_P. \tag{4.27}$$

Although this quantity remains intractable, it is a lower-dimensional integral than before, and we can use the same procedure outlined in Section 4.7.2.4. The posterior normalizing constant $h(P, \delta + n, \Lambda_P + B_P)$ can be similarly computed.

The steps discussed above make up the Hybrid-EP + Junction Tree (JT) algorithm, which is concisely summarized in Algorithm 3 below. We briefly discuss some of the implementation details. Before proceeding with any normalizing constant calculations, we extract the junction tree representation of G using the algorithm from Jones et al. (2005). Then, we proceed to compute the normalizing constant associated to each of these prime components. Similarly, in the case that the prime component P_i is complete, we can easily compute the normalizing constant using the inverse-Wishart normalizing constant formula. In the non-complete case, we apply the Hybrid-EP algorithm to the subgraph P_i, which is ideally a much smaller graph than the original graph G, and thus has less computational overhead. Note that while it appears as if each prime component P_i has its own objective function Ψ_{P_i}, these are nothing but the original function Ψ defined over the subgraph P_i, so no additional functions need to be defined. In addition, the function g that extracts each of the free parameters from ζ simply vectorizes ζ and keeps only the elements that correspond to the indices of nonzero entries of P_i. We emphasize that for each prime component, after we apply the transformation, $\Omega \mapsto \zeta$, the process for estimating the marginal likelihood is identical to the rest of the general Hybrid-EP algorithm described in Algorithm 2. Essentially, we are applying the Hybrid-EP algorithm to each of the non-complete prime components to reduce the dimension of the problem and the associated algorithm runtime complexity. After iterating through each of

Table 4.3: Mean, SD, and relative runtime of the approximations to the log normalizing constant of the GW (δ, Λ) density for a graph with p vertices. Here, Λ is not diagonal. We compare the Hybrid-EP + Junction Tree algorithm with the Atay's GNORM approximation. Estimates are reported over 20 replications, each using 1000 samples from the corresponding G-Wishart distribution. The runtime of the GNORM algorithm is calculated relative to the runtime of the Hybrid-EP + JT algorithm.

# VERTICES	METHOD	MEAN	SD	RUNTIME
$p = 10$	HYBRID-JT	-2477.401	0.0032	1
	GNORM	-2468.192	0	0.0069
$p = 30$	HYBRID-JT	-7450.691	0.0125	1
	GNORM	-7427.999	0.7730	4.1393
$p = 40$	HYBRID-JT	-10030.95	0.0152	1
	GNORM	-10003.81	1.5961	9.85573
$p = 50$	HYBRID-JT	-12563.87	0.0135	1
	GNORM	-12528.42	2.3854	15.9585
$p = 60$	HYBRID-JT	-15170.05	0.0171	1
	GNORM	`-Inf`	—	15.1635

the prime components and separators and computing the corresponding normalizing constants, these individual approximations are summed together to produce the approximation to the log normalizing constant corresponding to the original graph G.

In the following experiments, we compare the accuracy and runtime of various algorithms for higher dimensional graphs. In terms of benchmarking accuracy, we do not have a ground truth available, but both of these methods have previously proven to be reliable and have reported similar estimates in the results shown in Table 4.2, so for this set of simulations, we evaluate the estimates against each other. Indeed, for the first four simulations, where $p = 10, 30, 40, 50$, very little separates the log normalizing constant estimates. Unsurprisingly, the GNORM estimator demonstrates its strength in relatively low dimensions with a runtime that is more than 100 times faster than that of the Hybrid-EP + JT algorithm for $p = 10$. However, the Hybrid-EP + JT algorithm quickly flips the script in all of the subsequent experiments and scales well as p grows. In the final experiment where $p = 60$, the GNORM estimator fails to give a finite estimate for the log normalizing constant, while the Hybrid-EP + JT algorithm continues to produce sensible estimates. In constructing these experiments, each of the graphs is randomly generated using a Bernoulli draw to determine the existence of an edge. In addition, we ensure that each graph's junction tree representation does not exclusively consist of cliques, as this would simply reduce the problem to a summation of closed form inverse-Wishart log normalizing constants, and we would not be able to fairly assess the approximating ability of the Hybrid-EP + JT algorithm.

Contrast these results with the examples from the previous section where the scale matrix was assumed to be diagonal and the dependence structure was simple. In that setting, the GNORM estimator outperformed all competing methods even in high dimensions. Evidently, the added complexity induced by a nontrivial dependence structure contributes to the computational burden that cannot be easily overcome using standard methods. While the GNORM estimator from Atay-Kayis and Massam (2005) remains a valuable tool that performs well for graphs that are simpler in structure and lower in dimension, it is clear that more scalable and robust solutions are needed in these nontrivial settings. By taking advantage of the junction tree representation of the graph and weaving in the general Hybrid-EP methodology, we can greatly simplify the normalizing constant calculation for GGMs. Further, the accuracy and efficiency of this proposed estimator make it a versatile and appealing alternative to other state of the art algorithms, even those developed specifically for graphical models.

4.7.2.7 Hybrid-EP + JT algorithm R package

Recognizing the intricacy of the Hybrid-EP + JT algorithm and the difficulty in having to manually combine the Hybrid-EP and the junction tree methodologies, we have developed a package that is meant specifically for estimating the normalizing constant of G-Wishart densities. The `R` package `graphml` serves as a black box method that performs all of the calculations in the Hybrid-EP + JT algorithm without any user input other than the adjacency matrix representation of the graph and the G-Wishart density parameters. See Section B.2 for more details regarding the installation and use of this package.

Algorithm 3: Hybrid-EP + Junction Tree

```
Input  : Graph $G$, prior parameters $(\delta, \Lambda)$, methods for evaluating $\Psi_G$, $\nabla\Psi_G$, $\nabla^2\Psi_G$,
         where $\Psi_G$ is the negative log posterior for a given graph $G$
Output: Estimate of the logarithm of the normalizing constant of $\mathrm{GW}_G\left(\delta, \Lambda\right)$
Obtain the junction tree representation of $G \mapsto \mathcal{J}_G = \{P_1, S_2, P_2, S_3, \ldots, P_{m-1}, S_m, P_m\}$
for $i \in \{1, \ldots, m\}$ do
    if $i > 1$ then
        $\log \widehat{\mathcal{Z}}_{S_i} \leftarrow \log h\left(S_i, \delta, \Lambda_{S_i}\right)$  /* $\mathcal{W}_{S_i}^{-1}(\delta, \Lambda_{S_i})$ normalizing constant */
    end
    if $P_i$ is complete then
        $\log \widehat{\mathcal{Z}}_{P_i} \leftarrow \log h\left(P_i, \delta, \Lambda_{P_i}\right)$  /* $\mathcal{W}_{P_i}^{-1}\left(\delta, \Lambda_{P_i}\right)$ normalizing constant */
    else
        Compute the Cholesky decomposition, $\Lambda_{P_i}^{-1} = T_i' T_i$
        for $j \in \{1, \ldots, J\}$ do
            Sample $\Omega_{(j)} \sim \mathrm{GW}_{P_i}\left(\delta, \Lambda_{P_i}\right)$
            Compute the Choleksy decomposition, $\Omega_{(j)} = \phi_{(j)}' \phi_{(j)}$
            $\zeta_{(j)} \leftarrow \phi_{(j)} T_i^{-1}$
            Extract the free parameters $u_j = g\left(\zeta_{(j)}\right)$
        end
        Fit a CART model, $\mathcal{T}_i$, to $(u_1, \Psi_{P_i}(u_1)), \ldots, (u_J, \Psi_{P_i}(u_J))$
        Extract the partition $\mathcal{A} = \{A_1, \ldots, A_K\}$ from $\mathcal{T}_i$ of the bounding box $A$ of $\mathcal{U}$
        Calculate the global mode, $u_0$, of $\Psi_{P_i}$
        for $k \in \{1, \ldots, K\}$ do
            $u_k \leftarrow \mathrm{argmin}_{u \in A_k} ||u - u_0||_1$
            $\lambda_k \leftarrow \nabla\Psi_{P_i}\left(u_k\right)$
            $H_k \leftarrow \nabla^2\Psi_{P_i}\left(u_k\right)$
            $C_k \leftarrow (2\pi)^{d/2} |H_k|^{-1/2} \exp\left(\frac{1}{2}\left(u_k' H_k^{-1} u_k - 2\lambda_k' u_k + \lambda_k' H_k^{-1} \lambda_k\right)\right)$
            $b_k \leftarrow H_k u_k - \lambda_k$
            $G_k \leftarrow \int_{A_k} \mathcal{N}\left(u \mid H_k^{-1} b_k, H_k^{-1}\right)\, du$
            $\widehat{\mathcal{Z}}_k \leftarrow C_k \cdot G_k$
        end
        $\log \widehat{\mathcal{Z}}_{P_i} \leftarrow$ log-sum-exp $\left(\log \hat{\mathcal{Z}}_1, \ldots, \log \hat{\mathcal{Z}}_K\right)$
    end
end
return $\log \widehat{\mathcal{Z}} = \sum_i \log \widehat{\mathcal{Z}}_{P_i} - \sum_i \log \widehat{\mathcal{Z}}_{S_i}$
```

4.8 Chapter Summary

In this chapter, we developed an extension of the vanilla Hybrid estimator for target distributions that are unimodal and approximately log-concave around the mode. By introducing higher-order terms into the approximation to the negative log posterior Ψ, we are able to make large strides in terms of both the accuracy of the estimator and the range of problems that we can tackle. With these added terms, the integration over the partition sets becomes more involved, and we subsequently need to bring in approximate integration techniques that are suitable for high dimensions. We emphasize that other than the unimodal assumption, the extra requirements for the Hybrid-EP estimator only consist of the additional gradient and Hessian functions. As such, the setup of the algorithm itself remains fairly general, so the Hybrid-EP algorithm can easily be used in a variety of problems. In our examples, we demonstrate the widespread applicability of our proposed estimator, ranging from simple problems such as the logistic regression example to more complicated and specialized problems like graphical model problems. In particular for the GGM examples, estimating the marginal likelihood is known to be a computationally challenging task and has various dedicated methods for this calculation.

Our empirical results indicate that the Hybrid-EP estimator excels even in high-dimensional settings and thus presents itself as a reliable and accurate estimator. This is particularly interesting in the case of non-decomposable graphical models, where there is a clear dearth of scalable methods. While a complete theoretical analysis of the Hybrid-EP estimator is an interesting avenue for future work, the minimal assumptions underlying Theorem 1 already offer some insights into the success of methodology in high-dimensional settings. In addition, it is especially noteworthy that the Hybrid-EP algorithm provides a high degree of automation in that beyond the hyperparameter settings in CART, there is nothing further to tune. Practitioners can easily employ this algorithm without being burdened by example-specific model settings. Furthermore, since the implementation of both the general Hybrid-EP method and the Hybrid-EP + JT method have an accompanying R package, the task of using these algorithms in practice becomes even more straightforward.

5. SUMMARY AND CONCLUSIONS

In this dissertation, we present a novel method for approximating the marginal likelihood that addresses many of the known practical issues associated with existing algorithms. While abundant research has been devoted to computing this typically intractable, high-dimensional integral, much of the work is centered around using MCMC samples from the target distribution to form an asymptotically unbiased estimator. In our experiments, we demonstrate how heavily some of these methods rely on having a large number of exact samples from the posterior distribution in order to form accurate estimates and reveal the consequences when these sample size requirements are not met. This often becomes a hindrance when the target distribution is highly nontrivial, as the accuracy and runtime issues become practical considerations. Even though there is no shortage of such estimators, some better suited for specific problem settings than others, there is a growing necessity for a more general method that is both robust to the number and quality of the MCMC samples and scalable with respect to the dimension of the parameter space. Our proposed Hybrid estimator and its extensions offers a general solution to the marginal likelihood estimation problem that addresses these issues.

Our approach breaks up the marginal likelihood estimation calculation into two separate steps, with the essential components of our method consisting of the decision tree partitioning scheme that identifies high probability regions of the parameter space and the piecewise estimator to the negative log posterior defined over each of the partition sets. This modularization allows for potential extensions and modifications to either of the steps with minimal effect on the rest of the estimation procedure. In Chapter 3, we present the vanilla Hybrid estimator, which uses a piecewise constant approximation to the negative log posterior. Chapter 4 extends this methodology to form the Hybrid-EP estimator to target unimodal densities by incorporating higher order terms and employing high-dimensional integration techniques. Through different examples and statistical models, we show that the Hybrid estimator performs well in scenarios where MCMC samples are scarce and non-exact, where the underlying parameter space is both high-dimensional and complex, and where there exist state of the art estimators that have been developed specifically for the problem.

One important aspect of our work that we highlight is the additional software contribution that efficiently implements the proposed methodologies. While there exist many marginal likelihood estimation algorithms, the ones that have accompanying software packages that can be conveniently applied to a multitude of problems are few. Those that are available are often complicated in nature (many hyperparameters to tune, extensive knowledge of the underlying algorithm, familiarity with the code base, etc.) or too specialized in their applications that they do not perform well under slight perturbations of the problem settings. In our experiments, we repeatedly demonstrate that the Hybrid estimator can be seamlessly integrated into a diverse assortment of statistical models. The architecture for this approximation framework is available through the `hybrid` `R` package (Chuu, 2022b). In most instances, the algorithms in this package can serve as black-box methods for computing normalizing constants. For the specific case of graphical models where we proposed the Hybrid-EP + JT algorithm in Section 4.7.2.6, we developed a separate package, `graphml` (Chuu, 2022a), which is not only easy to use, but also highly efficient.

5.1 Further Study

It is clear that in some instances, such as the factor model example, further work must be done in order to make the Hybrid estimator competitive with other state of the art methods. The obvious next step for this example is to adapt the factor model setup to the Hybrid-EP setup, which involves obtaining the gradient and Hessian of the negative log posterior of (β, Ω), but upon doing so, we run into issues with the Hessian matrix because the likelihood surface is not log-concave. We explored using a Cholesky decomposition of $\Sigma = \beta\beta' + \Omega$ so that we could reformulate the problem to work instead with the Cholesky factors, but the problem of obtaining an expression for the induced prior density prevented us from tractably evaluating the log posterior distribution. As a potential solution, we can turn to variational methods to approximate the induced prior here so that the Hybrid-EP method can properly evaluate the necessary functions.

Despite this issue, we emphasize that our proposed methods' versatility allows for model-specific adjustments to be done without substantial modification to the core algorithm itself. In particular, because our algorithm is parametrization-invariant and only requires a way of obtaining MCMC samples and an expression for the negative log posterior, future work that involves changes to any of the examples can be easily accommodated.

Another area that has potential for improvement is the partitioning scheme used to identify high-probability regions of the parameter space. Currently, the Hybrid approximation algorithm and its extensions make use of the CART decision tree algorithm that partitions the feature space. One obvious modification that is worth exploring is an alternative tree building routine such as BART (Chipman et al., 2010), whose sum-of-trees approach allows for each of the trees to target a part of the overall fit, rather than relying on a single tree to capture the relationship between the response and the predictors. Other extensions to BART that adapt to smoothness and sparsity (Linero and Yang, 2018) and fit piecewise linear functions at each of the leaves/terminal nodes instead of piecewise constants are also viable alternatives (Prado et al., 2021) that can better identify regions of interest in the predictor space. However, the potential accuracy and flexibility that we might gain from these extensions may be overshadowed by the additional computational costs associated with fitting more complex decision trees.

For applications in statistical physics or genomic analysis, data representation is frequently very rich and high-dimensional. In addition to potential sparsity, there could exist redundancies and dependence between the variables that further inflate the parameter space. For these cases where the data points in $\mathbb{R}^D$ have an intrinsic dimension that is significantly smaller than D, it may be beneficial to develop and incorporate a function estimation method that can exploit the fact that the covariates (or parameters) lie on a d-dimensional manifold of $\mathbb{R}^D$, where $d \ll D$. This paves the way for another angle worth pursuing—taking advantage of a potentially lower intrinsic dimension of the parameter space. While our empirical results demonstrate reliable results in high dimensions, we can further improve upon the speed and accuracy if we can appropriately reformulate the underlying problem so that we are operating in a lower dimensional setting. Random projection trees (Dasgupta and Freund, 2008; Kpotufe and Dasgupta, 2012), which account for the intrinsic low dimensional structure in data without needing to explicitly learn the structure, present an alternative direction that we can explore to deal with high-dimensional spaces more efficiently. The caveat of using the random projection tree splitting routine is that resulting partition sets are no longer rectangular, so our simplified representation of the integral would need to be revised.

Finally, we recognize that even though the Hybrid estimator empirically demonstrates its competitiveness in a variety of examples, many of the popular marginal likelihood estimation methods have strong theoretical guarantees that may make those methods more appealing. As such, the theoretical aspects of the Hybrid estimator remain an area of future work and development.

REFERENCES

Omar Alejandro Aguilar and Mike West. Bayesian dynamic factor models and portfolio allocation. *Journal of Business & Economic Statistics*, 18:338 – 357, 2000.

David Applegate and Ravi Kannan. Sampling and integration of near log-concave functions. In *Proceedings of the Twenty-Third Annual ACM Symposium on Theory of Computing*, STOC '91, page 156–163, New York, NY, USA, 1991. Association for Computing Machinery. ISBN 0897913973. doi: 10.1145/103418.103439. URL `https://doi.org/10.1145/103418.103439`.

Aliye Atay-Kayis and Hélène Massam. A monte carlo method for computing the marginal likelihood in nondecomposable gaussian graphical models. *Biometrika*, 92(2):317–335, 2005.

Y.A. Atchadé. On the contraction properties of some high-dimensional quasi-posterior distributions. *The Annals of Statistics*, 45(5):2248–2273, 2017.

Yannick Baraud. A Bernstein-type inequality for suprema of random processes with applications to model selection in non-Gaussian regression. *Bernoulli*, 16(4):1064–1085, 2010.

Alexandros Beskos, Dan O. Crisan, Ajay Jasra, and Nick Whiteley. Error bounds and normalising constants for sequential Monte Carlo samplers in high dimensions. *Advances in Applied Probability*, 46(1):279 – 306, 2014. doi: 10.1239/aap/1396360114. URL `https://doi.org/10.1239/aap/1396360114`.

A. Bhattacharya, D. Pati, and Y. Yang. Bayesian fractional posteriors. *The Annals Statistics*, 47 (1):39–66, 02 2019. doi: 10.1214/18-AOS1712.

Anirban Bhattacharya and David Dunson. Sparse bayesian infinite factor models. *Biometrika*, 98: 291–306, 01 2011. doi: 10.2307/23076151.

Peter Binev, Albert Cohen, Wolfgang Dahmen, Ronald DeVore, and Vladimir Temlyakov. Universal algorithms for learning theory part i: Piecewise constant functions. *Journal of Machine Learning Research*, 6:1297–1321, 2005.

Christopher M. Bishop. *Approximate Inference*, page 461–470. Springer-Verlag New York, 2016.

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, apr 2017. doi: 10.1080/01621459.2017.1285773. URL `https://doi.org/10.1080%2F01621459.2017.1285773`.

Z. I. Botev. The normal law under linear restrictions: simulation and estimation via minimax tilting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(1):125–148, 2016. doi: 10.1111/rssb.12162.

Zdravko Botev and Leo Belzile. *TruncatedNormal: Truncated Multivariate Normal and Student Distributions*, 2019. URL `https://CRAN.R-project.org/package=TruncatedNormal`. R package version 2.1.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

Leo Breiman. *Classification and regression trees*. Wadsworth International Group, 1984.

Nicolas Brosse, Alain Durmus, and Éric Moulines. Normalizing constants of log-concave densities. *Electronic Journal of Statistics*, 12(1):851 – 889, 2018. doi: 10.1214/18-EJS1411. URL `https://doi.org/10.1214/18-EJS1411`.

J.E. Cavanaugh and A.A. Neath. Generalizing the derivation of the Schwarz information criterion. *Communications in Statistics-Theory and Methods*, 28(1):49–66, 1999.

Siddhartha Chib. Marginal likelihood from the gibbs output. *Journal of the American Statistical Association*, 90(432):1313–1321, 1995. doi: 10.1080/01621459.1995.10476635.

Siddhartha Chib and Ivan Jeliazkov. Marginal likelihood from the metropolis–hastings output. *Journal of the American Statistical Association*, 96(453):270–281, 2001. doi: 10.1198/016214501750332848.

Hugh A. Chipman, Edward I. George, and Robert E. McCulloch. BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266 – 298, 2010. doi: 10.1214/09-AOAS285. URL `https://doi.org/10.1214/09-AOAS285`.

Nicolas Chopin and Christian P. Robert. Properties of nested sampling. *Biometrika*, 97(3):

741–755, 2010. ISSN 00063444, 14643510. URL http://www.jstor.org/stable/25734120.

Eric Chuu. *graphml: Hybrid approximation for computing normalizing constants of G-Wishart densities*, 2022a. R package version 1.0.

Eric Chuu. *hybrid: Hybrid approximation to the marginal likelihood*, 2022b. R package version 1.0.

Eric Chuu, Debdeep Pati, and Anirban Bhattacharya. A hybrid approximation to the marginal likelihood. In *AISTATS*, 2021.

John P. Cunningham, Philipp Hennig, and Simon Lacoste-Julien. Gaussian probabilities and expectation propagation, 2013.

Sanjoy Dasgupta and Yoav Freund. Random projection trees and low dimensional manifolds. *Proceedings of the fortieth annual ACM symposium on Theory of computing*, 2008.

A Philip Dawid and Steffen L Lauritzen. Hyper markov laws in the statistical analysis of decomposable graphical models. *The Annals of Statistics*, pages 1272–1317, 1993.

P. De Blasi and S. G. Walker. Bayesian asymptotics with misspecified models. *Statistica Sinica*, pages 169–187, 2013.

Petros Dellaportas, Paolo Giudici, and Gareth Roberts. Bayesian inference for nondecomposable graphical gaussian models. *Sankhyā: The Indian Journal of Statistics*, pages 43–55, 2003.

Persi Diaconis, Donald Ylvisaker, et al. Conjugate priors for exponential families. *The Annals of statistics*, 7(2):269–281, 1979.

Patrick Ding. *epmgp: EP for approximate truncated multivariate normal sampling and multivariate normal probability*, 2020. R package version 0.1.0.

S. Dirksen. Tail bounds via generic chaining. *Electronic Journal of Probability*, 20, 2015.

Ritabrata Dutta and Jayanta K. Ghosh. Bayes Model Selection with Path Sampling: Factor Models and Other Examples. *Statistical Science*, 28(1):95 – 115, 2013. doi: 10.1214/12-STS403. URL https://doi.org/10.1214/12-STS403.

A. Marie Fitch, M. Beatrix Jones, and Hélène Massam. The Performance of Covariance Selection

Methods That Consider Decomposable Models Only. *Bayesian Analysis*, 9(3):659 – 684, 2014. doi: 10.1214/14-BA874. URL `https://doi.org/10.1214/14-BA874`.

Yong See Foo and Heejung Shim. A comparison of bayesian inference techniques for sparse factor analysis, 2021. URL `https://arxiv.org/abs/2112.11719`.

J. Kevin Ford, Robert C. MacCallum, and Marianne Tait. The application of exploratory factor analysis in applied psychology: A critical review and analysis. *Personnel Psychology*, 39:291–314, 1986.

Nial Friel and Jason Wyse. Estimating the evidence - a review. *Statistica Neerlandica*, 66(3): 288–308, 2012. doi: 10.1111/j.1467-9574.2011.00515.x.

Alan Genz. Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics*, 1(2):141, 1992. doi: 10.2307/1390838.

John Geweke and Guofu Zhou. Measuring the pricing error of the arbitrage pricing theory. *Staff Report*, 1995.

Subhashis Ghosal and Aad Van Der Vaart. Convergence rates of posterior distributions for noniid observations. *The Annals of Statistics*, 35(1):192–223, 2007.

J.K. Ghosh, M. Delampady, and T. Samanta. *An introduction to Bayesian analysis: theory and methods*. Springer Science & Business Media, 2007.

Joyee Ghosh and David B. Dunson. Default prior distributions and efficient posterior computation in bayesian factor analysis. *Journal of Computational and Graphical Statistics*, 18(2):306–320, 2009. doi: 10.1198/jcgs.2009.07145. URL `https://doi.org/10.1198/jcgs.2009.07145`. PMID: 23997568.

P Giudici and PJ Green. Decomposable graphical Gaussian model determination. *Biometrika*, 86(4):785–801, 12 1999. ISSN 0006-3444. doi: 10.1093/biomet/86.4.785. URL `https://doi.org/10.1093/biomet/86.4.785`.

Peter J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 12 1995. ISSN 0006-3444. doi: 10.1093/biomet/82.4.711. URL `https://doi.org/10.1093/biomet/82.4.711`.

Quentin F. Gronau, Alexandra Sarafoglou, Dora Matzke, Alexander Ly, Udo Boehm, Maarten Marsman, David S. Leslie, Jonathan J. Forster, Eric-Jan Wagenmakers, and Helen Steingroever. A tutorial on bridge sampling, 2017. URL `https://arxiv.org/abs/1703.05984`.

Quentin F. Gronau, Henrik Singmann, and Eric-Jan Wagenmakers. bridgesampling: An R package for estimating normalizing constants. *Journal of Statistical Software*, 92(10):1–29, 2020. doi: 10.18637/jss.v092.i10.

Harold Jeffreys. *Theory of Probability*. Oxford, England: Clarendon Press, 1939.

Beatrix Jones, Carlos Carvalho, Adrian Dobra, Chris Hans, Chris Carter, and Mike West. Experiments in Stochastic Computation for High-Dimensional Graphical Models. *Statistical Science*, 20(4):388 – 400, 2005. doi: 10.1214/088342305000000304. URL `https://doi.org/10.1214/088342305000000304`.

Kshitij Khare, Bala Rajaratnam, and Abhishek Saha. Bayesian inference for gaussian graphical models beyond decomposable graphs, 2015. URL `https://arxiv.org/abs/1505.00703`.

B. J. K. Kleijn and A. W. van der Vaart. Misspecification in infinite-dimensional Bayesian statistics. *Ann. Statist.*, pages 837–877, 2006.

Samory Kpotufe and Sanjoy Dasgupta. A tree-based regressor that adapts to intrinsic dimension. *J. Comput. Syst. Sci.*, 78:1496–1515, 2012.

Dirk P. Kroese, Thomas Taimre, and Zdravko I. Botev. *Handbook of Monte Carlo methods*. Wiley-Blackwell, 2011.

Steffen L Lauritzen. *Graphical models*, volume 17. Clarendon Press, 1996.

Peter Lenk. Simulation pseudo-bias correction to the harmonic mean estimator of integrated likelihoods. *Journal of Computational and Graphical Statistics*, 18(4):941–960, 2009. doi: 10.1198/jcgs.2009.08022.

Alex Lenkoski. A direct sampler for g-wishart variates. *Stat*, 2, 12 2013. doi: 10.1002/sta4.23.

Dennis Leung and Mathias Drton. Order-invariant prior specification in bayesian factor analysis. *Statistics & Probability Letters*, 111, 09 2014. doi: 10.1016/j.spl.2016.01.006.

Christopher Liaw, Abbas Mehrabian, Yaniv Plan, and Roman Vershynin. A simple tool for bounding the deviation of random matrices on geometric sets. In *Geometric aspects of functional analysis*, pages 277–299. Springer, 2017.

Antonio R. Linero and Yun Yang. Bayesian regression tree ensembles that adapt to smoothness and sparsity. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80, 2018.

Hedibert Freitas Lopes and Mike West. Bayesian model assessment in factor analysis. *Statistica Sinica*, 14(1):41–67, 2004. ISSN 10170405, 19968507. URL `http://www.jstor.org/stable/24307179`.

Laszlo Lovasz and Santosh Vempala. Fast algorithms for logconcave functions: Sampling, rounding, integration and optimization. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '06, page 57–68, USA, 2006. IEEE Computer Society. ISBN 0769527205. doi: 10.1109/FOCS.2006.28. URL `https://doi.org/10.1109/FOCS.2006.28`.

László Miklós Lovász and Santosh S. Vempala. The geometry of logconcave functions and sampling algorithms. *Random Structures & Algorithms*, 30, 2007.

Xiao-Li Meng and Stephen Schilling. Warp bridge sampling. *Journal of Computational and Graphical Statistics*, 11(3):552–586, 2002. doi: 10.1198/106186002457.

Xiao-Li Meng and Wing Hung Wong. Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistia Sinica*, 6:831–860, 1996. URL `www.jstor.org/stable/24306045`.

Thomas P. Minka. Expectation propagation for approximate bayesian inference, 2013. URL `https://arxiv.org/abs/1301.2294`.

Reza Mohammadi and Ernst C. Wit. BDgraph: An R package for Bayesian structure learning in graphical models. *Journal of Statistical Software*, 89(3):1–30, 2019. doi: 10.18637/jss.v089.i03.

Robb J. Muirhead. *Aspects of Multivariate Statistical Theory*. Wiley-Interscience, 2005.

Iain Murray, Ryan Adams, and David MacKay. Elliptical slice sampling. In *Proceedings of* *the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 541–548. PMLR, 13–15 May 2010. URL `http://proceedings.mlr.press/v9/murray10a.html`.

Radford M. Neal. Annealed importance sampling. *Statistics and Computing*, 11(2):125–139, 2001. doi: 10.1023/a:1008923215028.

Michael A. Newton and Adrian E. Raftery. Approximate bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(1): 3–26, 1994. doi: 10.1111/j.2517-6161.1994.tb01956.x.

Chris J. Oates, Theodore Papamarkou, and Mark Girolami. The controlled thermodynamic integral for bayesian model evidence evaluation. *Journal of the American Statistical Association*, 111 (514):634–645, 2016. doi: 10.1080/01621459.2015.1021006.

Antony M. Overstall and Jonathan J. Forster. Default bayesian model determination methods for generalised linear mixed models. *Computational Statistics & Data Analysis*, 54(12):3269–3288, 2010. ISSN 0167-9473. doi: https://doi.org/10.1016/j.csda.2010.03.008. URL `https://www.sciencedirect.com/science/article/pii/S0167947310001106`.

Anna Pajor. Estimating the marginal likelihood using the arithmetic mean identity. *Bayesian Analysis*, 12(1):261–287, 2017. doi: 10.1214/16-ba1001.

Mauro Piccioni. Independence structure of natural conjugate densities to exponential families and the gibbs' sampler. *Scandinavian Journal of Statistics*, 27(1):111–127, 2000. ISSN 03036898, 14679469. URL `http://www.jstor.org/stable/4616594`.

Michael K. Pitt and Neil Shephard. *Time varying covariances: a factor stochastic volatility approach*, pages 547–570. Oxford University Press, Oxford, (edited by j.m. bernardo, j.o. berger, a.p. dawid and a.f.m smith) edition, 1999.

Wolfgang Polasek. *Factor analysis and outliers: a Bayesian approach*. Citeseer, 1997.

Iosifina Pournara and Lorenz Wernisch. Factor analysis for gene regulatory networks and transcription factor activity profiles. *BMC Bioinformatics*, 8:61 – 61, 2006.

Estevão B. Prado, Rafael A. Moral, and Andrew C. Parnell. Bayesian additive regression trees with model trees. *Statistics and Computing*, 31(3), may 2021. ISSN 0960-3174. doi: 10.1007/s11222-021-09997-3. URL `https://doi.org/10.1007/s11222-021-09997-3`.

Adrian E Raftery, Michael A Newton, Jaya M Satagopan, and Pavel N Krivitsky. Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. *Bayesian Statistics*, 8:1–45, 2007.

Bala Rajaratnam, Hélène Massam, and Carlos M. Carvalho. Flexible covariance estimation in graphical Gaussian models. *The Annals of Statistics*, 36(6):2818 – 2849, 2008. doi: 10.1214/08-AOS619. URL `https://doi.org/10.1214/08-AOS619`.

R. V. Ramamoorthi, K. Sriram, and R. Martin. On posterior concentration in misspecified models. *Bayesian Analysis*, 10(4):759–789, 2015.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1530–1538, Lille, France, 07–09 Jul 2015. PMLR.

Alberto Roverato. Cholesky decomposition of a hyper inverse wishart matrix. *Biometrika*, 87(1): 99–112, 2000.

Chiara Sabatti and Gareth M. James. Bayesian sparse hidden components analysis for transcription regulation networks. *Bioinformatics*, 22 6:739–46, 2006.

Tim Salimans and David A. Knowles. Fixed-Form Variational Posterior Approximation through Stochastic Linear Regression. *Bayesian Analysis*, 8(4):837 – 882, 2013. doi: 10.1214/13-BA858. URL `https://doi.org/10.1214/13-BA858`.

Tim Salimans, Diederik Kingma, and Max Welling. Markov chain monte carlo and variational inference: Bridging the gap. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1218–1226, Lille, France, 07–09 Jul 2015. PMLR.

John Skilling. Nested sampling for general bayesian computation. *Bayesian Analysis*, 1(4): 833–859, 2006. doi: 10.1214/06-ba127.

V. Spokoiny. Parametric estimation. finite sample theory. *The Annals of Statistics*, 40(6):2877–2909, 2012a.

V. Spokoiny. Supplement to parametric estimation. finite sample theory. *The Annals of Statistics*, 40(6):2877–2909, 2012b.

K. Sriram, R.V. Ramamoorthi, and P. Ghosh. Posterior consistency of Bayesian quantile regression based on the misspecified asymmetric Laplace density. *Bayesian Analysis*, 8(2):479–504, 2013.

M. Talagrand. *The generic chaining: upper and lower bounds of stochastic processes*. Springer Science & Business Media, 2006.

Terry Therneau and Beth Atkinson. *rpart: Recursive Partitioning and Regression Trees*, 2019. URL `https://CRAN.R-project.org/package=rpart`. R package version 4.1-15.

Luke Tierney and Joseph B. Kadane. Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81(393):82–86, 1986. doi: 10.1080/01621459.1986.10478240.

Caroline Uhler, Alex Lenkoski, and Donald Richards. Exact formulas for the normalizing constants of wishart distributions for graphical models, 2016.

S. van de Geer. *Empirical Processes in M-estimation*. Cambridge UP, 2006.

R. Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

Mike West. Bayesian factor regression models in the "large p, small n" paradigm. *Bayesian Stat.*, 7, 08 2002.

Yuhuai Wu, Yuri Burda, Ruslan Salakhutdinov, and Roger B. Grosse. On the quantitative analysis of decoder-based generative models. *ArXiv*, abs/1611.04273, 2017.

APPENDIX A

PROOF OF THEORETICAL REESULTS

A.1 Assumptions

We provide supplemental details regarding the assumptions stated in Section 4.6. First, we make an additional note about Assumptions 1 and 2. Given n samples and d free parameters, the posterior will concentrate in a $\sqrt{d/n}$ neighborhood of the pseudo-true parameter under mild assumptions, that is, we can take

$$B^* = \prod_{j=1}^{d} \left[\theta_j^* - C\sqrt{d/n},\, \theta_j^* + C\sqrt{d/n}\right].$$

Now, if

$$|\mathbb{E}\ell\left(\theta_k^*, \theta^*\right)| = -\mathbb{E}\ell\left(\theta_k^*, \theta^*\right) \lesssim n\|\theta_k^* - \theta^*\|^2,$$

then the second part of Assumption 2 is clearly satisfied. This is a mild assumption which is broadly satisfied. In particular, for well-specified models,

$$-\mathbb{E}\ell\left(\theta_k^*, \theta^*\right) = D_{KL}\left(p_{\theta^*} \,||\, p_{\theta_k^*}\right),$$

and the inequality requires the KL divergence to be bounded by a constant multiple of the squared ℓ_2 norm in a neighborhood of the truth. We now focus on Assumption 3. For the sake of concreteness, we focus on the generalized linear model (GLM) framework, where the response $y_i \in \mathbb{R}$ conditional on the covariates $x_i \in \mathbb{R}^d$ are independently distributed according to a GLM $P_{x_i'\beta}$ in canonical form, with the log-likelihood

$$\ell\left(\beta\right)=\log p_{\beta}\left(y\right)=\sum_{i=1}^{n}\big\{y_{i}x_{i}^{\prime}\beta-a\left(x_{i}^{\prime}\beta\right)\big\},$$

where $\beta\in\mathbb{R}^{d}$ is the unknown vector of regression parameters and $a:\mathbb{R}\to\mathbb{R}$ is a strictly convex partition function with first and second derivatives $a^{(1)}$ and $a^{(2)}$, respectively. The pseudo-true parameter β^{*} satisfies

$$\nabla\mathbb{E}\ell\left(\beta^{*}\right)=\sum_{i=1}^{n}\{\mathbb{E}y_{i}-a^{(1)}\left(x_{i}^{\prime}\beta^{*}\right)\}x_{i}=0_{d}. \tag{A.1}$$

The quantity $\ell_{r}\left(\beta\right)-\ell_{r}\left(\beta^{*}\right)$ appearing in Assumption 3 equals $\langle y-\mathbb{E}y,X\left(\beta-\beta^{*}\right)\rangle$ in the present context. Define an index set $\mathcal{T}=\{x\in\Re^{d}\,:\,\|x\|\leq1\}$, and a stochastic process $Z_{\alpha}=\langle y-\mathbb{E}y,X\alpha\rangle$ for $\alpha\in\mathcal{T}$. Observe that for any $\beta\neq\beta^{*}\in B^{*}$,

$$\begin{aligned}\big|\langle y-\mathbb{E}y,X\left(\beta-\beta^{*}\right)\rangle\big| &= \left|\langle y-\mathbb{E}y,\frac{X\left(\beta-\beta^{*}\right)}{\|\beta-\beta^{*}\|}\rangle\right|\|\beta-\beta^{*}\|\\ &\leq\left(\sup_{\alpha\in\mathcal{S}^{d-1}}\big|\langle y-\mathbb{E}y,X\alpha\rangle\big|\right)R\left(\frac{d}{n}\right)^{1/2},\end{aligned}$$

where $\mathcal{S}^{d-1}=\{x\in\Re^{d}\,:\,\|x\|=1\}$. Letting $\alpha_{0}=0_{d}$, we can thus bound

$$\sup_{\beta\in B^{*}}|\ell_{r}\left(\beta\right)-\ell_{r}\left(\beta^{*}\right)|\leq R\left(d/n\right)^{1/2}\,\big(\sup_{\alpha\in\mathcal{T}}|Z_{\alpha}-Z_{\alpha_{0}}|\big).$$

The verification of Assumption 3 thus requires control over the supremum of the stochastic process (Z_{α}), which in turn depends on the moment assumptions on the true data distribution.

As an illustrative example, assume that $(y-\mathbb{E}y)$ is a centered sub-Gaussian random variable (Vershynin, 2018), that is, there exists a constant $\tau>0$ such that for any $v\in\Re^{n}$,

$$\mathbb{E}\exp\langle y-\mathbb{E}y,v\rangle\leq\exp\left(\tau^{2}\|v\|^{2}/2\right).$$

If the coordinates y_i are independent, one may take $\tau = \max_i \|y_i - \mathbb{E}y_i\|_{\psi_2}$ to be the maximum of the sub-Gaussian norms of $(y_i - \mathbb{E}y_i)$; see Vershynin (2018) for the definition of the sub-Gaussian norm $\|\cdot\|_{\psi_2}$. However, independence is not necessary for the above condition to hold and it can be verified for various dependence structures. In particular, if y has a joint Gaussian distribution, then τ equals the largest eigenvalue of cov (y). Under the above sub-Gaussian assumption, the process (Z_α) has sub-Gaussian increments, since for any $\lambda \in \Re$,

$$\mathbb{E}e^{\lambda(Z_\alpha - Z_{\tilde{\alpha}})} \leq e^{\lambda^2\tau^2\|X\alpha - X\tilde{\alpha}\|^2/2} \leq e^{\lambda^2\tau^2\|X\|_2^2\|\alpha-\tilde{\alpha}\|^2},$$

where $\|X\|_2$ is the operator norm of X. For processes with sub-Gaussian increments, a convenient high-probability bound for the supremum was developed in Theorem 4.1 of Liaw et al. (2017) as a corollary to the more general tail bound of Dirksen (2015). In preparation for applying their bound, we have

$$\|Z_\alpha - Z_{\tilde{\alpha}}\|_{\psi_2} \leq \tau\|X\|_2\|\alpha - \tilde{\alpha}\|,$$

for any $\alpha, \tilde{\alpha} \in \mathcal{T}$. Also, diam $(\mathcal{T}) = \sup_{\alpha,\tilde{\alpha}\in\mathcal{T}} \|\alpha - \tilde{\alpha}\| \leq 2$ and the Gaussian width of $\mathcal{T}$, $\mathbb{E}\sup_{\alpha\in\mathcal{T}}\langle g, \alpha\rangle$ for $g \sim \mathcal{N}_d(0, I_d)$, is in the order of $d^{1/2}$. Thus, with probability at least $1 - e^{-d}$,

$$\sup_{\alpha\in\mathcal{T}} |Z_\alpha - Z_{\alpha_0}| \leq C\tau\|X\|_2\, d^{1/2}.$$

It then follows that with probability at least $1 - e^{-d}$,

$$\sup_{\beta\in B^*} |\ell_r(\beta) - \ell_r(\beta^*)| \leq Cd.$$

Alternatively, suppose $(y_i - \mathbb{E}y_i)$ are independent sub-exponential (Vershynin, 2018) random variables, so that there exist $g_i > 0$ and ν_i such that

$$\mathbb{E}e^{\lambda(y_i - \mathbb{E}y_i)} \leq e^{\lambda^2 \nu_i^2/2}, \quad |\lambda| < g_i, \quad i = 1, \ldots, n.$$

Fix λ such that $|\lambda| \leq \min_i g_i := \bar{g}^{-1}$. Under the above assumption, we have, for any $\alpha, \tilde{\alpha} \in \mathcal{T}$ that

$$\begin{aligned} \mathbb{E}e^{\lambda \frac{Z_\alpha - Z_{\tilde{\alpha}}}{\|X\alpha - X\tilde{\alpha}\|}} &= \prod_{i=1}^{n} \mathbb{E}e^{\lambda \frac{x_i'(\alpha - \tilde{\alpha})}{\|X\alpha - X\tilde{\alpha}\|}(y_i - \mathbb{E}y_i)} \\ &\leq e^{\lambda^2 \sum_{i=1}^{n} \frac{\nu_i^2 \{x_i'(\alpha - \tilde{\alpha})\}^2}{\|X\alpha - X\tilde{\alpha}\|^2}} \\ &\leq e^{\lambda^2 \nu^2/2} \\ &\leq e^{\lambda^2 \nu^2 / \{2(1 - |\lambda| \bar{g})\}}, \end{aligned}$$

where $\nu = \max_i \nu_i$. From the second to the third step, we used the fact that $|x_i'(\alpha - \tilde{\alpha})|/\|X\alpha - X\tilde{\alpha}\| \leq 1$. Hence, Z_α is a centered process on $\mathcal{T}$ with sub-exponential increments. Define the norm

$$d(\alpha_1, \alpha_2) = \|X\alpha_1 - X\alpha_2\|.$$

Clearly, $d(\alpha_1, \alpha_2) \leq \|X\|_2$ for $\alpha_1, \alpha_2 \in \mathcal{T}$. From Theorem 2.1 of Baraud (2010),

$$\mathbb{P}\Big[\sup_{\alpha \in \mathcal{T}} |Z_\alpha - Z_{\alpha_0}| > \|X\|_2 \sqrt{1 + x} + \bar{g}x\Big] \leq 2e^{-x},\ x > 0,$$

thereby verifying Assumption 3 by setting $x = d$.

A.1.1 Proof of Theorem 1

Let $\mathcal{Y}_g$ denote the subset of the sample space $\mathcal{Y}$ where the events in Assumptions 1 and 2 both hold. We shall work inside the set $\mathcal{Y}_g$, with $\mathbb{P}(\mathcal{Y}_g) \geq 1 - \delta - \tilde{\delta}$ by Bonferroni's inequality. We first prove the upper bound. By Assumption 1,

$$(1-\eta) \leq \gamma\left(B^*\right) = \frac{\int_{B^*} e^{\ell(\theta,\theta^*)}\,\pi\left(\theta\right)d\theta}{\int_\Theta e^{\ell(\theta,\theta^*)}\,\pi\left(\theta\right)d\theta}.$$

Rearranging terms, this gives

$$\log \mathcal{Z} \leq \ell\left(\theta^*\right) + \log\left(\frac{1}{1-\eta}\right) + \log\int_{B^*} e^{\ell(\theta,\theta^*)}\,\pi\left(\theta\right)d\theta.$$

We now bound the integral inside the logarithm in the right hand side of the above display. Write

$$\begin{aligned}
\int_{B^*} e^{\ell(\theta,\theta^*)}\,\pi\left(\theta\right)d\theta &= \int_{B^*} e^{\ell_r(\theta)-\ell_r(\theta^*)+\mathbb{E}\ell(\theta,\theta^*)}\pi\left(\theta\right)d\theta \\
&\leq e^{Cd}\int_{B^*} e^{\mathbb{E}\ell(\theta,\theta^*)}\pi\left(\theta\right)d\theta \\
&= e^{Cd}\sum_{k=1}^K\int_{B_k} e^{\mathbb{E}\ell(\theta,\theta_k^*)+\mathbb{E}\ell(\theta_k^*,\theta^*)}\pi\left(\theta\right)d\theta \\
&\leq e^{(C+C_1)d}\sum_{k=1}^K\big[\sup_{\theta\in B_k}\pi\left(\theta\right)\big]\int_{B_k} e^{-(\theta-\theta_k^*)'H_k(\theta-\theta_k^*)/2}d\theta \\
&= e^{(C+C_1)d}\sum_{k=1}^K\big[\sup_{\theta\in B_k}\pi\left(\theta\right)\big]\left(2\pi\right)^{d/2}\left|H_k\right|^{-1/2}\int_{B_k}\mathcal{N}_d\left(\theta;\theta_k^*,H_k^{-1}\right) \\
&= e^{(C+C_1+\log(2\pi)/2)d}\sum_{k=1}^K\big[\sup_{\theta\in B_k}\pi\left(\theta\right)\big]\left|H_k\right|^{-1/2}\gamma_{2k}.
\end{aligned}$$

Cascading through the previous inequalities delivers the upper bound. For the lower bound, we use

$$\begin{aligned}
\log \mathcal{Z} &= \ell\left(\theta^*\right) + \log\int_\Theta e^{\ell(\theta,\theta^*)}\,\pi\left(\theta\right)d\theta \\
&\geq \ell\left(\theta^*\right) + \log\int_{B^*} e^{\ell(\theta,\theta^*)}\,\pi\left(\theta\right)d\theta.
\end{aligned}$$

Thus, we need a lower bound to the same quantity we derived an upper bound for previously. Write

$$\begin{aligned}
\int_{B^*} e^{\ell(\theta,\theta^*)}\,\pi\left(\theta\right)d\theta &= \int_{B^*} e^{\ell_r(\theta)-\ell_r(\theta^*)+\mathbb{E}\ell(\theta,\theta^*)}\pi\left(\theta\right)d\theta \\
&\geq e^{-Cd}\int_{B^*} e^{\mathbb{E}\ell(\theta,\theta^*)}\pi\left(\theta\right)d\theta \\
&= e^{-Cd}\sum_{k=1}^{K}\int_{B_k} e^{\mathbb{E}\ell(\theta,\theta_k^*)+\mathbb{E}\ell(\theta_k^*,\theta^*)}\pi\left(\theta\right)d\theta \\
&\geq e^{-(C+C_1)d}\sum_{k=1}^{K}\big[\inf_{\theta\in B_k}\pi\left(\theta\right)\big]\int_{B_k} e^{-(\theta-\theta_k^*)'H_k(\theta-\theta_k^*)/(2c_k)}d\theta \\
&= e^{-(C+C_1)d}\sum_{k=1}^{K}\big[\inf_{\theta\in B_k}\pi\left(\theta\right)\big]\,(2\pi)^{d/2}\,c_k^{d/2}|H_k|^{-1/2}\int_{B_k}\mathcal{N}_d(\theta;\theta_k^*,c_kH_k^{-1}) \\
&\geq e^{\left\{-C-C_1+\log(2\pi)/2+\min_k(\log c_k)/2\right\}d}\sum_{k=1}^{K}\big[\inf_{\theta\in B_k}\pi\left(\theta\right)\big]\,|H_k|^{-1/2}\,\gamma_{1k}.
\end{aligned}$$

This proves the lower bound.

APPENDIX B

DETAILS FOR NUMERICAL EXPERIMENTS

B.1 General use of the **hybrid** package

For convenience, we have developed `hybrid` (Chuu, 2022b), an `R` package that allows practitioners to easily compute estimates of the marginal likelihood. In Figure B.1, we provide a snippet of code to demonstrate how both the vanilla Hybrid approximation from Chapter 3 and the Hybrid-EP approximation from Chapter 4 can be used in practice. For the vanilla Hybrid method, users only need to provide a way to evaluate the negative log posterior Ψ and a sampler for the target distribution γ. After drawing the samples using the user-defined `sample_post()` and evaluating them using the `hybrid::preprocess()` function, we can calculate the log marginal likelihood estimate with the `hybrid::hybml_const()` function.

For the Hybrid-EP method, users will need to supplement the input of the vanilla Hybrid method with function definitions for the gradient vector and Hessian matrix of Ψ. These function definitions, along with the posterior samples, are then passed into the `hybrid::hybml()` function. Optionally, a representative point (typically the global mode) can be supplied to the function call to play the role of u_0 as defined in Section 4.4, but in the case where no point is given, the implementation will take u_0 to be the point with highest posterior mass. Both functions perform the partitioning and integration calculations under the hood and return an approximation to the log marginal likelihood. We emphasize that beyond specifying the model and supplying a sampler, which is typically required in all other competing methods, there are no hyperparameters to tune and no problem-specific settings that require modification or attention. This makes our solution one of the few black box marginal likelihood estimation methods that has been empirically shown to scale well with dimension and accommodate complex parameter spaces.

One detail to be mindful of when writing the user-defined functions is that the posterior samples generated by the `sample_post()` function must be returned as a matrix, where each of the samples is stored row-wise as a d-dimensional vector. Here, d corresponds to the dimension of the parameter space. In addition, the input to the negative log posterior functions must be vectors, so for parameters that are not already vectors (covariance matrices, a group of vectors and scalars, etc.), a suitable concatenation and/or vectorization scheme must be defined to transform the model parameters. The repository that contains the source code for these algorithm implementations can be found at `https://github.com/echuu/hybrid`. The repository also contains installation instructions and a working example that demonstrates the use of both the vanilla Hybrid method and the Hybrid-EP method to compute the marginal likelihood for the Bayesian linear regression model with a multivariate normal inverse-gamma prior on the parameters (β, σ^2), as described in Section 3.7.1.1. This example, along with all of the other examples and simulation results in this dissertation can be reproduced using the `hybrid` package. For the graphical model examples, see Section B.2 for a dedicated package.

B.2 General use of the `graphml` package

While the graphical modeling examples can be adapted to be used with the `hybrid` package by following the code outline in Figure B.1, we have also developed a package specific to graphical models that further simplifies the process of weaving together the Hybrid-EP methodology with the junction tree representation of general graphs—as discussed in Section 4.7.2.6 and presented in Algorithm 3—because of the importance of the normalizing constant calculation in graphical modeling literature. With the `graphml` (Chuu, 2022a) package, we can thus easily employ the Hybrid-EP + JT methodology to compute the normalizing constant of the G-Wishart density given the adjacency matrix for a general graph, the scale matrix, the degrees of freedom, and the number of MCMC samples to be drawn from the corresponding G-Wishart density.

Along with the following snippet of code in Figure B.2 that demonstrates how the Hybrid-EP + JT algorithm can be used in practice, we also include the code necessary to obtain the GNORM approximation from Atay-Kayis and Massam (2005), as provided by the `BDgraph` package. Note that the arguments passed into `graphml::hybridJT()` are nearly identical to those passed into `BDgraph::gnorm()`, with the exception of an additional edge matrix argument which can be

```
#### ----------------- VANILLA HYBRID, HYBRID-EP DEMO  ----------------- ####

## In the functions below, ‘params’ is an object that stores
## any miscellaneous values (hyperparameters, sample size, dimensions)
## that may be necessary to compute the corresponding functions.

library(hybrid)

#### ----- The following 4 functions must be supplied by the user ----- ####

# sample_post(): returns a (J x d) matrix of samples from the
# target distribution
sample_post = function(J) { ... }

# psi(): returns the (scalar) negative log posterior evaluated at u
psi  = function(u, params) { ... }

# grad(): returns the (d x 1) gradient vector of psi evaluated at u
grad = function(u, params) { ... }

# hess(): returns the (d x d) Hessian matrix of psi evaluated at u
hess = function(u, params) { ... }

#### ----- Problem-specific initializations ----- ####

params = init( ... )     # initialize any hyperparameters
J = 5000                 # number of posterior samples to draw
samps = sample_post(J)   # (J x d) samples from the target distribution

# evaluate posterior samples using psi()
psi_df = hybrid::preprocess(samps, d, params)

# compute vanilla hybrid estimate for the log Z
hybrid::hybml_const(psi_df)$logz

# compute hybrid-ep estimate for the log Z
hybrid::hyb(psi_df, params, grad, hess)$logz
```

Figure B.1: Demonstration of how to use `hybrid` package in `R`. This package contains the implementation for both the vanilla Hybrid and Hybrid-EP algorithms for estimating the log normalizing constant of a target distribution.

computed using the built-in function, `graphml::getEdgeMat()`, which takes an adjacency matrix as input. We intentionally juxtapose these two approximation functions in Figure B.2 to demonstrate that the Hybrid-EP + JT method requires remarkably little user input and can be used just as easily as other state of the art methods. Recall that for the algorithms in the `hybrid` package, we require user-defined functions for the objective function, gradient, and Hessian, whereas in this package, all of these functions are already optimally implemented within the package. In the case of the G-Wishart density, these functions are quite cumbersome to implement because of their recursive structure, so removing this from the list of user responsibilities is especially convenient. The Github repository that contains the source code for this package can be found at `https://github.com/echuu/graphml`. Along with installation instructions, the repository also contains a working example of the `graphml` package that sets up the problem for computing the normalizing constants of a G-Wishart densities corresponding to graphs with 30 and 60 vertices and compares the runtime with that of Atay’s method. This example, along with the rest of the simulation results in Table 4.3, can easily be reproduced using the `graphml` package and following the approach outlined in the working example.

```r

#### ------------ GRAPHML / HYBRID-EP + JT ALGORITHM DEMO  ------------ ####

library(graphml)

#### ----- Initialize G-Wishart parameters ----- ####

G = ...                  # initialize adj. matrix representation of graph
b = ...                  # degrees of freedom, b > 2
V = ...                  # non-negative definite scale matrix
J = ...                  # number of samples to draw from target density

# Compute the edge matrix for the JT part of the algorithm
EdgeMat = graphml::getEdgeMat(G)

# ---- Compute the log normalizing constant of the GW(b, V) density ---- #

# Compute the log normalizing constant using Atay’s algorithm
BDgraph::gnorm(G, b, V, J)

# Compute the log normalziing constant using the Hybrid-EP + JT algorithm
graphml::hybridJT(G, EdgeMat, b, V, J)
```

Figure B.2: Demonstration of how to use `graphml` package in `R`. This package contains the implementation for Hybrid-EP + Junction Tree algorithm for estimating the log normalizing constant of G-Wishart densities. Atay’s estimator can similarly be computed using the `gnorm()` function from the `BDgraph` package.

B.3 Experiments

We elaborate on some of the posterior distribution and marginal likelihood calculations from the experiments sections. Code to reproduce the simulation results can be found in the `examples` sub-directories within the `hybrid` and `graphml` repositories.

B.3.1 Conjugate 2-d example

In the conjugate normal model in Section 3.4.1, the posterior distribution of (μ, σ^2) is well-known. In particular, $\mu \mid \sigma^2, y_{1:n} \sim \mathcal{N}\left(m_n, \sigma^2/w_n\right)$ and $\sigma^2 \mid y_{1:n} \sim \mathcal{IG}\left(r_n/2, s_n/2\right)$, with posterior parameters defined as follows,

$$\begin{aligned} m_n &= \frac{m\bar{y} + w_0 m_0}{n + w_0}, \quad w_n = w_0 + n, \quad r_n = r_0 + n, \\ s_n &= s_0 + \sum_{i=1}^{n} (y_i - \bar{y})^2 + \left(\frac{n w_0}{n + w_0}\right)(\bar{y} - m_0)^2. \end{aligned}$$

With this in place, the marginal likelihood can be computed in closed form,

$$\begin{aligned} p\left(y\right) &= \int \left[\prod_{i=1}^{n} \mathcal{N}\left(y_i \mid \mu, \sigma^2\right)\right] \mathcal{N}\left(\mu \mid m_0, \sigma^2/w_0\right) \mathcal{IG}\left(\sigma^2 \mid r_0/2, s_0/2\right) d\beta \, d\sigma^2 \\ &= \pi^{-n/2} \left(\frac{w_0}{w_n}\right)^{1/2} \frac{\Gamma\left(\frac{r_n}{2}\right)}{\Gamma\left(\frac{r_0}{2}\right)} \cdot \frac{s_0^{r_0/2}}{s_n^{r_n/2}}. \end{aligned} \tag{B.1}$$

For the experiments in Section 3.4.1, each of the $n = 100$ observations is drawn from a normal distribution with mean 30 and variance 4. The prior hyperparameters are $m_0 = 0, w_0 = 0.05, r_0 = 3, s_0 = 3$. Using these to compute the posterior and plugging these into Eq. (B.1), we calculate the true log marginal likelihood to be -113.143.

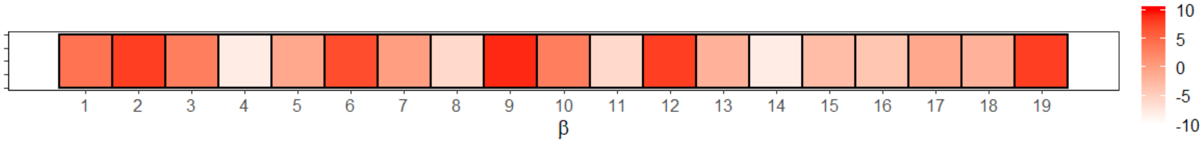


Figure B.3: True value of β; each component is represented as a tile and takes on value between -10 and 10. Values closer to 10 are red and values closer to -10 are white.

B.3.2 Multivariate normal inverse-gamma

Recall the linear regression setup given in Section 3.7.1 and 3.7.1.1. Because we are dealing with a conjugate prior, the posterior distribution can easily be shown to have the following form:

$$\beta \mid \sigma^2, y \sim \mathcal{N}\left(\mu_n, \sigma^2 V_n\right),$$
$$\sigma^2 \mid y \sim \mathcal{IG}\left(a_n, b_n\right),$$

with posterior parameters,

$$\begin{aligned}
\mu_n &= V_n(X'y + V_\beta^{-1}\mu_\beta),\\
V_n &= (X'X + V_\beta^{-1})^{-1},\\
a_n &= a_0 + n/2,\\
b_n &= b_0 + (y'y + \mu_\beta' V_\beta^{-1}\mu_\beta - \mu_n' V_n^{-1}\mu_n).
\end{aligned}$$

Then the marginal likelihood can be computed directly to be

$$p\left(y\right) = \frac{1}{(2\pi)^{n/2}} \frac{b_0^{a_0}}{b_n^{a_n}} \frac{\Gamma\left(a_n\right)}{\Gamma\left(a_0\right)} \frac{\det\left(V_n\right)^{1/2}}{\det\left(V_\beta\right)^{1/2}}. \tag{B.2}$$

Each of the 100 observations is drawn from a d-dimensional normal distribution according to the linear regression model presented in Section 3.7.1. In the experiments, we take $d = 19$, and the prior hyperparameters values are $\mu_\beta = 0_d, V_\beta = I_d, a_0 = 1, b_0 = 1$. The true value of β is shown as a heatmap in Figure B.3 and $\sigma^2 = 4$. Plugging these into Eq. (B.2), we compute the true log marginal likelihood to be -303.8482.

B.3.3 Truncated multivariate normal

With the truncated multivariate normal prior given in Section 3.7.1.2, we obtain the following form of the posterior distribution of β,

$$\beta \mid y \sim \mathcal{N}_d \left(\beta \mid Q^{-1}\eta, Q^{-1}\right) \cdot \mathbb{1}_{[0,\infty)^d},$$

with posterior parameters $Q = \frac{1}{\sigma^2}\left(X'X + \lambda I_d\right)$ and $\eta = \frac{1}{\sigma^2}X'y$. Each of the $n = 100$ observations is drawn from a d-dimensional normal distribution according to the linear regression model presented in Section 3.7.1. In the experiments, we take $d = 20$, and the prior hyperparameters values are $\sigma^2 = 4, \lambda = 0.25$. The true value of β is shown as a heat map in Figure B.4.

We provide the details for computing the baseline used for comparison in the experiments. Let $R = [0,\infty)^D$ and $\mu = Q^{-1}\eta$. Then, we integrate the product of the likelihood and prior to obtain the normalizing constant of the resulting posterior distribution.

$$\begin{aligned}
p(y) &= \int \mathcal{N}\left(y \mid X\beta, \sigma^2 I\right) \cdot \frac{1}{2^{-d}} \cdot \mathcal{N}\left(\beta \mid 0, \sigma^2\tau^{-1}I\right) \mathbb{1}_R\left(\beta\right)\, d\beta \\
&= 2^d \int_R \mathcal{N}\left(y \mid X\beta, \sigma^2 I\right) \cdot \mathcal{N}\left(\beta \mid 0, \sigma^2\tau^{-1}I\right) d\beta \\
&= 2^d \int_R \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left(y - X\beta\right)'\left(y - X\beta\right)\right\} \left(2\pi\sigma^2\right)^{-\frac{d}{2}} \tau^{\frac{d}{2}} \exp\left\{-\frac{\tau}{2\sigma^2}\beta'\beta\right\} d\beta \\
&= 2^d \cdot \left(2\pi\right)^{-\frac{n}{2}} \left(\sigma^2\right)^{-\frac{1}{2}(n+d)} \tau^{\frac{d}{2}} \cdot e^{-\frac{1}{2\sigma^2}y'y} \\
&\qquad \times \int_R \left(2\pi\right)^{-\frac{d}{2}} \exp\left\{-\frac{1}{2}\left(\beta'\frac{1}{\sigma^2}\left[X'X + \tau I\right]\beta - 2\beta'\left[\frac{1}{\sigma^2}X'y\right]\right)\right\} d\beta \\
&= 2^d \cdot \left(2\pi\right)^{-\frac{n}{2}} \left(\sigma^2\right)^{-\frac{1}{2}(n+d)} \tau^{\frac{d}{2}} \cdot e^{-\frac{1}{2\sigma^2}y'y} \cdot e^{\frac{1}{2}\eta'Q^{-1}\eta} \\
&\qquad \times \int_R \left(2\pi\right)^{-D/2} \exp\Big\{ -\frac{1}{2}\left[\beta'Q\beta + \eta'Q^{-1}\eta - 2\beta'\eta\right] \Big\} d\beta \\
&= 2^d \cdot \left(2\pi\right)^{-\frac{n}{2}} \left(\sigma^2\right)^{-\frac{1}{2}(n+d)} \tau^{\frac{d}{2}} \cdot e^{-\frac{1}{2\sigma^2}y'y} \cdot e^{\frac{1}{2}\eta'Q^{-1}\eta} \cdot |Q|^{-\frac{1}{2}} \\
&\qquad \times \int_R \left(2\pi\right)^{-\frac{d}{2}} |Q|^{\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(\beta - Q^{-1}\eta\right)' Q \left(\beta - Q^{-1}\eta\right)\right\} d\beta.
\end{aligned} \tag{B.3}$$

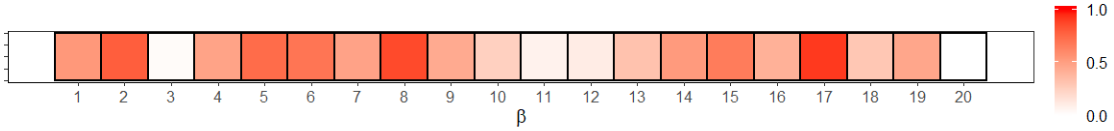


Figure B.4: True value of β; each component is represented as a tile and takes on value between 0 and 1. Values closer to 1 are red and values closer to 0 are white.

Since the integral in Eq. (B.3) is intractable, we use the `TruncatedNormal::pmvnorm()` function which implements Botev's minimax tilting method for estimating the normalizing constant of truncated multivariate normal distributions in high dimensions.

B.3.4 Approximate posterior samples

For the multivariate normal inverse-gamma example in Section 3.7.2 where we draw approximate posterior samples from the mean field approximation to the posterior distribution, each of the $n = 100$ observations is first drawn from a d-dimensional normal distribution according to the linear regression model presented in Section 3.7.1. In the experiments, we take $d = 9$, and the prior hyperparameters values are $\mu_\beta = 0_d, V_\beta = I_d, a_0 = 1, b_0 = 1$. The posterior distribution of β is approximated by a product of 3-dimensional normal distributions, each with mean and covariance components, $(\mu_n^{(i)}, V_n^{(i)})$ for $i = 1, 2, 3$. These are extracted from the true posterior parameters (μ_n, V_n), as defined in Section B.3.2, in the following way:

$$\mu_n = \begin{bmatrix} \mu_{n1} \\ \mu_{n2} \\ \vdots \\ \mu_{n9} \end{bmatrix} = \left[\begin{array}{c} \mu_n^{(1)} \\ \hline \mu_n^{(2)} \\ \hline \mu_n^{(3)} \end{array}\right], \quad V_n = \left[\begin{array}{c|c|c} V_n^{(1)} & & \\ \hline & V_n^{(2)} & \\ \hline & & V_n^{(3)} \end{array}\right].$$

Here, V_n is block diagonal, so the approximating distribution ignores some of the dependence structure that is present in the true posterior distribution. The true value of β is shown as a heat map in Figure B.5 and $\sigma^2 = 4$. Using the formula for the marginal likelihood derived in Eq. (B.2),

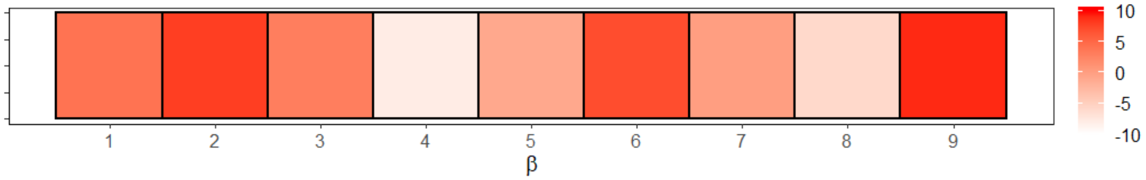


Figure B.5: True value of β; each component is represented as a tile and takes on value between -10 and 10. Values closer to 10 are red and values closer to -10 are white.

we compute the true log marginal likelihood to be -147.3245.

B.3.5 Unrestricted covariance matrices

We consider the inverse-Wishart prior on Σ, $\mathcal{W}^{-1}(\Lambda, \nu)$, where Λ is a positive definite $d \times d$ matrix, and $\nu > d - 1$. The prior density has the following form,

$$\pi\left(\Sigma\right) = C_{\Lambda,\nu}\, \det\left(\Sigma\right)^{-(\nu+d+1)/2} \exp\left\{- \mathrm{tr}\left(\Sigma^{-1}\Lambda\right)/2\right\},$$

where $C_{\Lambda,\nu} = \det\left(\Lambda\right)^{\nu/2} / \left(2^{\nu d/2}\,\Gamma_d\left(\nu/2\right)\right)$. Here, $\Gamma_d\left(\cdot\right)$ is the multivariate gamma function, given by

$$\Gamma_d\left(a\right) = \pi^{d(d-1)/4} \prod_{j=1}^{d} \Gamma\left(a + (1-j)/2\right),$$

where $\Gamma\left(\cdot\right)$ is the ordinary gamma function. Our choice of the prior admits the following closed form marginal likelihood:

$$\int L\left(\Sigma\right)\pi\left(\Sigma\right) d\Sigma = \frac{\Gamma_d\left((n+\nu)/2\right)}{\pi^{nd/2}\Gamma_d\left(\nu/2\right)} \frac{\det\left(\Lambda\right)^{\nu/2}}{\det\left(\Lambda + S\right)^{(n+\nu)/2}}. \tag{B.4}$$

In each of the 100 replications, we take $d = 4$ and draw $n = 100$ observations from a 4-dimensional normal distribution with mean vector 0_d and covariance matrix Σ, where

$$\Sigma = \begin{bmatrix} 1.662 & 1.640 & -1.985 & -0.007 \\ 1.640 & 7.163 & -4.146 & 5.654 \\ -1.985 & -4.146 & 4.906 & -1.237 \\ -0.007 & 5.654 & -1.237 & 6.779 \end{bmatrix}.$$

The prior hyperparameters are $\Lambda = I_4$ and $\nu = 5$. Plugging these into Eq. (B.4), we compute the true log marginal likelihood to be -673.7057.

B.3.6 Competing logistic regression models

The logistic regression models in Section 4.7.1 are used to predict diabetes for Pima Indians using a dataset originally provided by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). The dataset does not contain personally identifiable information or offensive content. To our knowledge, the NIDDK obtained consent from the subjects involved in the study. A copy of the data and the implementation of all competing methods shown in Figure 4.1 are available at `https://github.com/nbrosse/normalizingconstant`.

The predictors in question are: number of pregnancies (NP), plasma glucose concentration (PGC), diastolic blood pressure (BP), triceps skin fold thickness (TST), body mass index (BMI), diabetes pedigree function (DP), and age (AGE). Given these seven features, there are 2^7 potential models that we could consider, but in our experimental setup, we narrow our search to the two models that Friel and Wyse (2012) determined to have the highest posterior probability via a reversible jump MCMC algorithm (Green, 1995). The two models are: $\mathcal{M}_1 = \text{logit}\,(p) = 1 + \texttt{NP} + \texttt{PGC} + \texttt{BMI} + \texttt{DP}$ and $\mathcal{M}_2 = \text{logit}\,(p) = 1 + \texttt{NP} + \texttt{PGC} + \texttt{BMI} + \texttt{DP} + \texttt{AGE}$. The likelihood is defined given $\mathcal{M}_k$ for $k = 1, 2$ by

$$p\,(y \mid \theta, \mathcal{M}_k) = \exp\left(\sum_{i=1}^{n} y_i \theta' x_i^{(k)} - \log\left(1 + e^{\theta' x_i^{(k)}}\right)\right). \tag{B.5}$$

Here, $x_i^{(k)}$ denotes the i-th row of $X^{(k)}$, the $n \times d$ design matrix corresponding to $\mathcal{M}_k$. Using Eq. (B.5) together with the Gaussian prior defined on θ, we obtain the expression for the negative log posterior given in Eq. (4.11). In order to make use of the HYB-EP approximation, we also need the gradient and Hessian of Ψ, which can be written as

$$\nabla\Psi\left(\theta\right) = X'\mu + \tau\theta, \quad \nabla^2\Psi\left(\theta\right) = X'DX + \tau I_d,$$

where $\mu = (\sigma(z_1) - y_1, \ldots, \sigma(z_n) - y_n)'$, $\sigma(z_i) = \sigma(x_i'\theta) = \text{logistic}(x_i'\theta)$, and $D = \text{diag}(\sigma(z_1)(1 - \sigma(z_1)), \ldots, \sigma(z_n)(1 - \sigma(z_n)))$.

APPENDIX C

GRAPHICAL MODELS

C.1 Basic graph theory

Definition 1 (Graph)**.** *A graph is a pair $G = (V, E)$, where V is a finite set of vertices and $E = \{(u, v \mid u \in V, v \in V, u \neq v\}$ is the edge set that link the vertices so that E is a subset of the set of ordered pairs of distinct vertices $V \times V$. For all examples in this dissertation, we assume that G is simple, i.e., does not contain loops.*

Definition 2 (Undirected)**.** *Edges $(u, v) \in E$ with both (u, v) and (v, u) in E are called undirected, whereas an edge (u, v) with its opposite (v, u) not in E is called directed. If the graph has only undirected edges, it is an undirected graph and if all edges are directed, the graph is said to be directed. For all examples in this dissertation, we assume that G is an undirected graph, for which $(u, v) \in E$ implies $(v, u) \in E$.*

Definition 3 (Induced Subgraph)**.** *For any vertex set $A \subseteq V$, we define the edge set associated with it as $E_A := \{(u, v) \in E \mid u, v \in A\} = E \cap (A \times A)$. Let $G_A = (A, E_A)$ denote the subgraph of G induced by A, where E_A is obtained from G by keeping edges with both endpoints in A.*

Definition 4 (Adjacent)**.** *Two vertices $u, v \in V$ are adjacent (neighbors) in an undirected graph if $(u, v) \in E$. We write $u \sim v$ in G. Hence $E = \{(i, j) \mid i, j \in V, i \sim j\}$.*

Definition 5 (Neighbor)**.** *The set of neighbors of a vertex i in G is the set of all vertices that are adjacent to i in G. This is denoted as $\mathrm{ne}\,(i)$, where $\mathrm{ne}\,(i) = \{j \in V \mid i \sim j\} \setminus \{i\}$. The relation $i \sim j$ is symmetric and equivalent to each of $j \in \mathrm{ne}\,(i)$ and $i \in \mathrm{ne}\,(j)$. We write $\mathrm{ne}\,(A)$ to denote the neighbor of the vertex set A, which can be written explicitly as $\mathrm{ne}\,(A) = \cup_{\alpha \in A} \mathrm{ne}\,(\alpha) \setminus \{A\}$.*

Definition 6 (Boundary)**.** *The boundary of a subset of vertices $A \subseteq V$, denoted $bd\,(A)$, is the set* *of vertices in $V \setminus A$ adjacent to at least one vertex in A,*

$$bd\,(A) := \{v \in V \mid v \notin A \textit{ and } (u, v) \in E \textit{ for some vertex } u \in A\}.$$

Definition 7 (Complete)**.** *An induced subgraph G_A is complete if the vertices in A are pairwise adjacent (every pair of vertices are adjacent) in G. We also say that A is complete in G. A graph is complete if all vertices are joined by an edge and is said to be fully connected. Every node in a complete graph A is a neighbor of every other such node. A subset is complete if it induces a complete subgraph.*

Definition 8 (Clique)**.** *A complete vertex set A in G that is maximal (with respect to $\subseteq$) is a clique, i.e., a maximally complete subgraph. By maximal, we mean that a clique is not contained in a larger complete subgraph. That is, A is complete and we cannot add a further node that shares an edge with each node of A.*

Definition 9 (Proper Subgraph)**.** *Proper subgraphs of a clique$A$ (all subgraphs apart from A itself) are complete but not maximal.*

Definition 10 (Path)**.** *Let $u, v \in V$. A path (or chain) of length k in G from u to v is an alternating sequence of its $k - 1$ vertices and k edges of the form $u = v_0, e_1, v_1, e_2, \ldots, e_k, v_k = v$ of distinct vertices such that $(v_{i-1}, v_i) \in E$ for all $i = 1, \ldots, n$ and vertices $v_{i=1}$ and v_i are endpoints of edge e_i for each i. Since an edge is uniquely characterized by its endpoints, we may denote paths by the sequence of vertices only.*

Definition 11 (Adjacency Matrix)**.** *Let G be a graph having p vertices labelled $v_1, \ldots, v_p$. Then the adjacency matrix A of G is the $p \times p$ matrix whose ij-th entry $A_{ij} = 1$ if $(v_i, v_j) \in E$, and $A_{ij} = 0$ otherwise. Since the graph is simple, diagonal elements A_{ii} are all zero.*

Definition 12 (Cycle)**.** *The path is a n-cycle if the end points are allowed to be the same, $u = v$.*

Definition 13 (Connected)**.** *If there is a path from u to v we say that u and v are connected.*

A graph G is connected if all the pairs of vertices are connected. Otherwise G will consist of connected components which are maximal connected subgraphs of G.

Definition 14 (Separator). *A subset $C \subseteq V$is said to be an uv-separator if all paths from u to v intersect C (or every path from from u to v includes at least one vertex in C).*

Definition 15 (Decomposition). *A triple (A, B, C) of disjoint subsets of the vertex set V of an undirected graph G is said to form a decomposition of G if $V = A \cup B \cup C$ and (i) C separates A from B; (ii) C is a complete subset of V. In this case, (A, B, C) decomposes G into the induced component subgraphs $G_{A \cup C}$ and $G_{B \cup C}$.*

Definition 16 (Decomposable). *An undirected graph is said to be decomposable if it is complete, or if there exists a proper decomposition (A, B, C) into decomposable subgraphs $G_{A \cup C}$ and $G_{B \cup C}$. Note that this definition is recursive.*

Definition 17 (Simplicial). *A vertex $v \in V$ is simplicial in $G = (V, E)$ if $\mathrm{bd}(v)$ is a clique. A subset B is a simplical subset if $\mathrm{bd}\,(B)$ is complete.*

Definition 18 (Ordering). *An ordering of G is a bijection from the vertex set V to a set of labels* $\{1, 2, \ldots, n\}$.

Definition 19 (Perfect Elimination Order). *The ordering $v_1, \ldots, v_n$ is a perfect elimination ordering if v_i is simplificial in the graph $G\{v_i, v_{i+1}, \ldots, v_n\}$ for $i = 1, \ldots, n$.*

Definition 20 (Leaves). *Let $G = (V, E)$ be a connected graph having a clique separator C, and let $V_1, \ldots V_s$ be the vertex set of the connected components of $G \backslash C$. The subgraphs $G_{V_1 \cup C}, \ldots, G_{V_s \cup C}$ are the leaves of G produced by C.*

Definition 21 (Tree). *A tree is a connected, undirected graph with no cycles. In a tree, there is a unique path between any two vertices.*

Definition 22 (Junction Graph). *The junction graph of a decomposable graph has nodes, where every pair of nodes is connected. Each link is associated with the intersection of the two cliques that it connects, and has a weight (possibly zero) equal to the cardinality of the intersection.*

Definition 23 (Spanning Tree)**.** *A spanning tree of a graph is a subgraph that contains all the vertices and is a tree. A graph may have many spanning trees.*

Definition 24 (Junction Tree)**.** *Let J be any spanning tree of the junction graph. J is a junction tree if for any two cliques C and D of G, every node on the unique path between C and D in J contains $C \setminus D$.*

C.2 Hyper-inverse Wishart clique and separator densities

We first introduce some notation to help us obtain a closed form for the marginal likelihood in the case of a decomposable graph G. For an $n \times d$ matrix X, X_C is defined as the sub-matrix of X consisting of columns with indices in the clique C. Let $(X_1, X_2, \ldots, X_d) = (x_1, x_2, \ldots, x_n)'$, where X_i is the ith column of $X_{n\times d}$. If $C = \{i_1, i_2, \ldots, i_{|C|}\}$, where $1 \leq i_1 < i_2 < \ldots < i_{|C|} \leq d$, then $X_C = (X_{i_1}, X_{i_2}, \ldots, X_{i_{|C|}})$. For any square matrix $A = (a_{ij})_{d\times d}$, define $A_C = (a_{ij})_{|C|\times|C|}$ where $i, j \in C$, and the order of entries carries into the new sub-matrix A_C. Therefore, $X_C'X_C = (X'X)_C$.

Decomposable graphs correspond to a special kind of sparsity pattern in Σ, henceforth denoted Σ_G. Suppose we have a $\text{HIW}_G\,(b, D)$ distribution on the cone of $d \times d$ positive definite matrices with $b > 2$ degrees of freedom and a fixed $d \times d$ positive definite matrix D such that the joint density factorizes on the junction tree of the given decomposable graph G as

$$p\,(\Sigma_G \mid b, D) = \frac{\prod_{C\in\mathcal{C}} p\,(\Sigma_C \mid b, D_C)}{\prod_{S\in\mathcal{S}} p\,(\Sigma_S \mid b, D_S)}, \tag{C.1}$$

where for each $C \in \mathcal{C}$, $\Sigma_C \sim \mathcal{W}^{-1}_{|C|}\,(b, D_C)$ has the density

$$p\,(\Sigma_C \mid b, D_C) \propto |\Sigma_C|^{-(b+2|C|)/2}\,\text{etr}\Big\{-\frac{1}{2}\Sigma_C^{-1}D_C\Big\}, \tag{C.2}$$

where $|C|$ is the cardinality of the clique C and $\text{etr}\,(\cdot) = \exp\big\{\text{tr}\,(\cdot)\big\}$. Here, $\mathcal{W}_d^{-1}\,(b, D)$ denotes the inverse-Wishart distribution with degrees of freedom b and a fixed $d\times d$ positive definite matrix

D with normalizing constant

$$\left|\frac{1}{2}D\right|^{(b+d-1)/2}\Gamma_d^{-1}\left(\frac{b+d-1}{2}\right).$$

Note that we can establish equivalence to the parametrization used in Section 4.7.2.1 by taking $\delta = b+d-1$. Since the joint density in Eq. (3.9) factorizes over cliques and separators in the same way as in Eq. (C.1) and (C.2),

$$f\left(X \mid \Sigma_G\right) = (2\pi)^{-\frac{np}{2}}\frac{\prod_{C\in\mathcal{C}}\left|\Sigma_C\right|^{-\frac{n}{2}}\text{etr}\Big(-\frac{1}{2}\Sigma_C^{-1}X_C'X_C\Big)}{\prod_{S\in\mathcal{S}}\left|\Sigma_S\right|^{-\frac{n}{2}}\text{etr}\Big(-\frac{1}{2}\Sigma_S^{-1}X_S'X_S\Big)}. \tag{C.3}$$

The HIW (b, D) density can be written as

$$\begin{aligned} f\left(\Sigma_G \mid G\right) &= \frac{\prod_{C\in\mathcal{C}} p\left(\Sigma_C \mid b, D_C\right)}{\prod_{S\in\mathcal{S}} p\left(\Sigma_S \mid b, D_S\right)} \\ &= \frac{\prod_{C\in\mathcal{C}}\left|\frac{1}{2}D_C\right|^{\frac{b+|C|-1}{2}}\Gamma_{|C|}^{-1}\left(\frac{b+|C|-1}{2}\right)\left|\Sigma_C\right|^{-\frac{b+2|C|}{2}}\text{etr}\left(-\frac{1}{2}\Sigma_C^{-1}D_C\right)}{\prod_{S\in\mathcal{S}}\left|\frac{1}{2}D_S\right|^{\frac{b+|S|-1}{2}}\Gamma_{|S|}^{-1}\left(\frac{b+|S|-1}{2}\right)\left|\Sigma_S\right|^{-\frac{b+2|S|}{2}}\text{etr}\left(-\frac{1}{2}\Sigma_S^{-1}D_S\right)}. \end{aligned}$$

Then, it is straightforward to obtain the marginal likelihood of the decomposable graph G,

$$f\left(X \mid G\right) = (2\pi)^{-\frac{np}{2}}\frac{h\left(G, b, D\right)}{h\left(G, b+n, D+S\right)} = (2\pi)^{-\frac{np}{2}}\frac{\prod_{C\in\mathcal{C}} w\left(C\right)}{\prod_{S\in\mathcal{S}} w\left(S\right)}, \tag{C.4}$$

where

$$h\left(G, b, D\right) = \frac{\prod_{C\in\mathcal{C}}\left|\frac{1}{2}D_C\right|^{\frac{b+|C|-1}{2}}\Gamma_{|C|}^{-1}\left(\frac{b+|C|-1}{2}\right)}{\prod_{S\in\mathcal{S}}\left|\frac{1}{2}D_S\right|^{\frac{b+|S|-1}{2}}\Gamma_{|S|}^{-1}\left(\frac{b+|S|-1}{2}\right)}, \tag{C.5}$$

$$w\left(C\right) = \frac{\left|D_C\right|^{\frac{b+|C|-1}{2}}\left|D_C + X_C'X_C\right|^{-\frac{b+n+|C|-1}{2}}}{2^{-\frac{n|C|}{2}}\Gamma_{|C|}\left(\frac{b+|C|-1}{2}\right)\Gamma_{|C|}^{-1}\left(\frac{b+n+|C|-1}{2}\right)}. \tag{C.6}$$

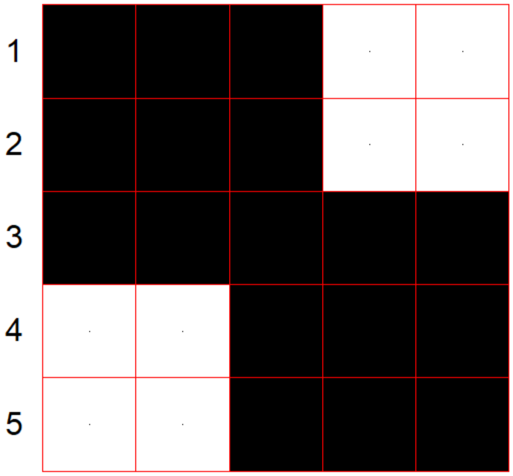


Figure C.1: In the undirected graph G, with vertex set $V = \{1, 2, 3, 4, 5\}$, the (i, j)-th box is black if the corresponding edge is present in G and white otherwise.

C.3 Hyper-inverse Wishart objective function

Recall that we take the Cholesky decomposition of $\Omega = \phi'\phi$, where ϕ is upper triangular. Using Eq. (4.12) and (3.13), we define $\Psi(\phi) = -\log L(\phi) - \log \pi(\phi)$. Even though Ψ is expressed as a function of the upper Cholesky factor ϕ, it is inherently a function of only the free elements of ϕ. As a result, the gradient of Ψ should only be taken with respect to the free elements of the upper Cholesky factor, $(i, j) \in \mathscr{V}$. This calculation can be done element-wise,

$$\frac{\partial \Psi}{\partial \phi_{ij}} = \begin{cases} -\frac{1}{\phi_{ii}}(\eta_i + n) + \phi_{ii} + \sum_{m=i}^{p} \phi_{im} s_{mi} & i = j, \\ \phi_{ij} + \sum_{m=i}^{p} \phi_{im} s_{mj} & i \neq j. \end{cases}$$

Using the above expression for the gradient, elements on and above the diagonal of the Hessian matrix can also be computed element-wise,

$$\frac{\partial^2 \Psi}{\partial \phi_{ij} \partial \phi_{kl}} = \begin{cases} 0 & i \neq k, \\ \frac{1}{\phi_{ii}^2}\left(n + \eta_i\right) + s_{ii} + 1 & i = j = k = l, \\ s_{li} & i = j, i = k, l > j, \\ 1 + s_{lj} & i \neq j, i = k, l = j, \\ s_{lj} & i \neq j, i = k, l > j, \end{cases}$$

where $(i, j), (k, l) \in \mathscr{V}$, $\eta_i = \delta + \nu_i - 1$, and $S = (s_{ij})_{1 \leq i,j \leq p}$.

Next, we discuss the experimental setup used in Section 3.7.4 to obtain the results in Table 3.2. Conditional on the graph G, which is represented in Figure C.1, we consider a hyper-inverse Wishart prior on $\Sigma = \Omega^{-1}$, $\text{HIW}_G\left(\delta, B\right)$, where the prior hyperparameters are $B = I_5$ and $\delta = 3$. We then draw $n = 100$ observations from a 5-dimensional normal distribution with mean vector 0 and a sparse inverse covariance matrix Ω, where the dependence structure in Ω is dictated by the graph G. Using the formula for the marginal likelihood derived in Eq. (C.4), we compute the true log marginal likelihood to be -506.3061.

C.4 G-Wishart prior for general graphs

Here, we outline the GNORM algorithm which uses the results from Atay-Kayis and Massam Atay-Kayis and Massam (2005). Keeping consistent with the notation used in Section 4.7.2.4, we revisit the formulation of the normalizing constant in Eq. (4.17),

$$C_G(\delta, \Lambda) = 2^p \prod_{i=1}^{p} (t_{ii}^2)^{(\delta + b_i - 1)/2} \int \exp\left(-\frac{1}{2} \sum_{(i,j) \in \bar{\mathscr{V}}} \zeta_{ij}^2\right) \prod_{i=1}^{p} (\zeta_{ii}^2)^{(\delta + \nu_i - 1)/2} \exp\left(-\frac{1}{2} \sum_{i=1}^{p} \zeta_{ii}^2\right)$$
$$\times \exp\left(-\frac{1}{2} \sum_{(i,j) \in \mathscr{V}, i \neq j} \zeta_{ij}^2\right) \prod_{i=1}^{p} d\zeta_{ii} \prod_{(i,j) \in \mathscr{V}, i \neq j} d\zeta_{ij},$$

and we note that since $d\zeta_{ii} = \frac{1}{2}\zeta_{ii}^{-1} d\left(\zeta_{ii}^2\right)$, we can write the normalizing constant as the following integral,

$$\begin{aligned} C_G(\delta,\Lambda) = \prod_{i=1}^{p} (t_{ii}^2)^{(\delta+\nu_i)/2} \, (2\pi)^{\nu_i/2} \, \Gamma\left(\frac{\delta+\nu_i}{2}\right) \prod_{i=1}^{p} (t_{ii}^2)^{(\delta+b_i-1)/2} \\ \times \int \exp\left(-\frac{1}{2}\sum_{(i,j)\in\bar{\mathscr{V}}} \zeta_{ij}^2\right) \prod_{i=1}^{p} \frac{1}{\Gamma\left(\left(\delta+\nu_i\right)/2\right)} \left(\frac{\zeta_{ii}^2}{2}\right)^{(\delta+\nu_i)/2-1} \exp\left(-\frac{1}{2}\zeta_{ii}^2\right) \\ \times \prod_{(i,j)\in\mathscr{V}, i\neq j} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\zeta_{ij}^2\right) \prod_{i=1}^{p} d\left(\zeta_{ii}^2\right) \prod_{(i,j)\in\mathscr{V}, i\neq j} d\zeta_{ij}. \end{aligned}$$

This conveniently allows us to interpret the above integral as an expectation

$$C_G(\delta,\Lambda) = \prod_{i=1}^{p} (t_{ii}^2)^{(\delta+\nu_i)/2} \, (2\pi)^{\nu_i/2} \, \Gamma\left(\frac{\delta+\nu_i}{2}\right) \prod_{i=1}^{p} (t_{ii}^2)^{(\delta+b_i-1)/2} \, \mathbb{E}\left[f\left(\zeta^{\mathscr{V}}\right)\right], \tag{C.7}$$

where

$$f\left(\zeta^{\mathscr{V}}\right) = \exp\left(-\frac{1}{2}\sum_{(i,j)\in\bar{\mathscr{V}}} \zeta_{ij}^2\right).$$

Note, that the expectation in Eq. (C.7) is taken with respect to the distribution with density proportional to the product of independent chi-squared distributions and standard normal distributions. In particular, $\zeta_{ii} \sim \sqrt{\chi^2_{\delta+\nu_i}}$ for $i = 1, \dots, p$, and $\zeta_{ij} \sim \mathcal{N}(0,1)$ for $(i,j) \in \mathscr{V}$. With this in place, we can estimate the expectation in Eq. (C.7) using the following Monte Carlo average,

$$\frac{1}{N}\sum_{i=1}^{N} f\left(\zeta_i^{\mathscr{V}}\right),$$

where each $\zeta_i^{\mathscr{V}}$ is formed by random draws from the distribution defined above on ζ_{ij}, for $i = j$ and $(i,j) \in \mathscr{V}$. Therefore, by forming the Monte Carlo estimate and keeping track of the remaining constants in Eq. (C.7), we arrive at the GNORM estimator for the normalizing constant of the G-Wishart density.

Next, we provide the calculation details for the derivation of the gradient and Hessian of $\Psi(\zeta)$, as defined in Eq. (4.18). First, we can compute the terms of the gradient element-wise by taking the derivative of Ψ with respect to the free elements of ζ,

$$\frac{\partial \Psi(\zeta)}{\partial \zeta_{ij}} = \begin{cases} \displaystyle\sum_{(r,s)\in\bar{\mathscr{V}}} \zeta_{rs} \frac{\partial \zeta_{rs}}{\partial \zeta_{ii}} - \frac{(\delta + \nu_i - 1)}{\zeta_{ii}} + \zeta_{ii} & i = j, \\ \displaystyle\sum_{(r,s)\in\bar{\mathscr{V}}} \zeta_{rs} \frac{\partial \zeta_{rs}}{\partial \zeta_{ij}} + \zeta_{ij} & i \neq j, (i,j) \in \mathscr{V}. \end{cases} \tag{C.8}$$

Note that because the non-free elements are functions of the free elements, each gradient term involves additional recursive derivative calculations. As given in Eq. (4.19), for $(r,s) \in \bar{\mathscr{V}}$ and $r < s$,

$$\zeta_{rs} = \sum_{j=r}^{s-1} \left(-\zeta_{rj} \frac{\lambda_{js}}{\lambda_{ss}} \right) - \sum_{i=1}^{r-1} \left(\frac{\zeta_{ir} + \sum_{j=i}^{r-1} \zeta_{ij} \frac{\lambda_{jr}}{\lambda_{rr}}}{\zeta_{rr}} \right) \left(\zeta_{is} + \sum_{j=i}^{s-1} \zeta_{ij} \frac{\lambda_{js}}{\lambda_{ss}} \right). \tag{C.9}$$

Since we have taken $\Lambda = (\lambda_{ij}) = I_p$, then $\lambda_{ij} = 0$ for $i \neq j$, so Eq. (4.19) can be simplified significantly and written as

$$\zeta_{rs} = -\frac{1}{\zeta_{rr}} \sum_{k=1}^{r-1} \zeta_{kr} \zeta_{ks}. \tag{C.10}$$

After an application of the product rule, the partial derivative terms corresponding to the non-free elements in the gradient can be calculated using the following recursive definition:

$$\frac{\partial \zeta_{rs}}{\partial \zeta_{ij}} = -\frac{1}{\zeta_{rr}} \sum_{k=1}^{r-1} \left[\zeta_{ks} \frac{\partial \zeta_{kr}}{\partial \zeta_{ij}} + \zeta_{kr} \frac{\partial \zeta_{ks}}{\partial \zeta_{ij}} \right], \qquad (r,s) \in \bar{\mathscr{V}},\ r < s. \tag{C.11}$$

Finally, using the expression for the gradient in Eq. (C.8), we can perform a similar calculation for the Hessian. The elements on and above the diagonal of the Hessian matrix can be computed as follows,

$$\frac{\partial^2\Psi(\zeta)}{\partial\zeta_{ij}\partial\zeta_{kl}} = \begin{cases} \displaystyle\sum_{(r,s)\in\bar{\mathscr{V}}} \left[\left(\frac{\partial\zeta_{rs}}{\partial\zeta_{ii}}\right)^2 + \zeta_{rs}\frac{\partial^2\zeta_{rs}}{\partial\zeta_{ii}^2} \right] + \frac{(\delta+\nu_i-1)}{\zeta_{ii}^2} + 1, & i=j=k=l, \\ \displaystyle\sum_{(r,s)\in\bar{\mathscr{V}}} \left[\frac{\partial\zeta_{rs}}{\partial\zeta_{kl}}\frac{\partial\zeta_{rs}}{\partial\zeta_{ij}} + \zeta_{rs}\frac{\partial^2\zeta_{rs}}{\partial\zeta_{ij}\partial\zeta_{kl}} \right] + \frac{\partial\zeta_{ij}}{\partial\zeta_{kl}}, & i\neq j, (i,j),(k,l)\in\mathscr{V}, \end{cases} \tag{C.12}$$

where any subsequent derivatives can be computed using Eq. (C.11).

C.5 G-Wishart prior for general graphs with non-diagonal scale matrix

The setup for the case where Λ is non-diagonal is similar to the diagonal case, but the recursive formulation of ζ_{rs} does not enjoy the simplification given in Eq. (C.10) and (C.11). Subsequently, the gradient and Hessian calculations become more involved. The piecewise definitions of the gradient and Hessian functions in Eq. (C.8) and (C.12) remain the same, but the individual elements in the summations of both functions are different. Below, we provide the derivation for the partial derivative terms of the non-free elements taken with respect to the free elements by again starting with the recursive definition of ζ_{rs} in Eq. (4.19) and making repeated use of the product rule. In the following calculations, we define $\xi_{ks} = \lambda_{ks}/\lambda_{ss}$, where $\Lambda = (\lambda_{ij})$, where $1 \leq i, j \leq p$. We consider two cases.

Case 1: for $(r,s) \in \bar{\mathscr{V}}$, $r < s$, $i \neq j$, and (r,s) coming after (i,j), we have

$$\begin{aligned} \frac{\partial\zeta_{rs}}{\partial\zeta_{ij}} &= -\sum_{k=r}^{s-1}\xi_{ks}\frac{\partial\zeta_{rk}}{\partial\zeta_{ij}} \\ &\quad - \frac{1}{\zeta_{rr}}\frac{\partial}{\partial\zeta_{ij}}\sum_{k=1}^{r-1}\left[\zeta_{ks}\zeta_{kr} + \zeta_{kr}\sum_{l=k}^{s-1}\zeta_{kl}\xi_{ls} + \zeta_{ks}\sum_{l=k}^{r-1}\zeta_{kl}\xi_{lr} + \left(\sum_{l=k}^{k-1}\zeta_{kl}\xi_{kr}\right)\left(\sum_{l=k}^{s-1}\zeta_{kl}\xi_{ks}\right)\right]. \end{aligned} \tag{C.13}$$

Case 2: for $r = i = j$, and $s > r$, we obtain the derivative:

$$
\begin{aligned}
\frac{\partial \zeta_{rs}}{\partial \zeta_{rr}} = &-\sum_{k=r}^{s-1} \xi_{ks} \frac{\partial \zeta_{rk}}{\partial \zeta_{rr}} \\
&+ \frac{1}{\zeta_{rr}^2} \sum_{k=1}^{r-1} \left[\zeta_{ks}\zeta_{kr} + \zeta_{kr} \sum_{l=k}^{s-1} \zeta_{kl}\xi_{ls} + \zeta_{ks} \sum_{l=k}^{r-1} \zeta_{kl}\xi_{lr} + \left(\sum_{l=k}^{k-1} \zeta_{kl}\xi_{kr} \right) \left(\sum_{l=k}^{s-1} \zeta_{kl}\xi_{ks} \right) \right].
\end{aligned}
$$

In case 1, each term in the outer summation in Eq. (C.13) can be computed separately as follows,

$$
\frac{\partial}{\partial \zeta_{ij}} \left[\zeta_{ks}\zeta_{kr} \right] = \frac{\partial \zeta_{kr}}{\partial \zeta_{ij}} \zeta_{ks} + \zeta_{kr} \frac{\partial \zeta_{ks}}{\partial \zeta_{ij}},
$$

$$
\frac{\partial}{\partial \zeta_{ij}} \left[\zeta_{kr} \sum_{l=k}^{s-1} \zeta_{kl}\xi_{ls} \right] = \frac{\partial \zeta_{kr}}{\partial \zeta_{ij}} \sum_{l=k}^{s-1} \zeta_{kl}\xi_{ls} + \zeta_{kr} \sum_{l=k}^{s-1} \frac{\partial \zeta_{kl}}{\partial \zeta_{ij}} \xi_{ls},
$$

$$
\frac{\partial}{\partial \zeta_{ij}} \left[\zeta_{ks} \sum_{l=k}^{r-1} \zeta_{kl}\xi_{lr} \right] = \frac{\partial \zeta_{ks}}{\partial \zeta_{ij}} \sum_{l=k}^{r-1} \zeta_{kl}\xi_{lr} + \zeta_{ks} \sum_{l=k}^{r-1} \frac{\partial \zeta_{kl}}{\partial \zeta_{ij}} \xi_{lr},
$$

$$
\frac{\partial}{\partial \zeta_{ij}} \left[\left(\sum_{l=k}^{r-1} \zeta_{kl}\xi_{kr} \right) \left(\sum_{l=k}^{s-1} \zeta_{kl}\xi_{ks} \right) \right] = \left[\sum_{l=k}^{r-1} \xi_{kr} \frac{\partial \zeta_{kl}}{\zeta_{ij}} \right] \left[\sum_{l=k}^{s-1} \zeta_{kl}\xi_{ks} \right] + \left[\sum_{l=k}^{r-1} \zeta_{kl}\xi_{kr} \right] \left[\sum_{l=k}^{s-1} \xi_{ks} \frac{\partial \zeta_{kl}}{\partial \zeta_{ij}} \right].
$$

These scalar, element-wise quantities can be substituted back into the piecewise definitions of the gradient and Hessian matrices, as given in Eq. (C.8) and (C.12), respectively.