UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Scienze Statistiche
Corso di Dottorato di Ricerca in Scienze Statistiche
Ciclo XXXVI

Skew-symmetric approximations of posterior distributions

Coordinatore del Corso: Prof. Nicola Sartori

Supervisore: Prof. Bruno Scarpa

Co-supervisori: Prof. Daniele Durante and Prof. Botond Tibor Szabó

Dottorando: Francesco Pozza

3 Novembre 2023

Abstract

In Bayesian statistics, routinely implemented deterministic approximations of posterior distributions typically rely on symmetric densities, often taken to be Gaussian. Such a choice facilitates optimization and inference, but may compromise the quality of the overall approximation. In fact, even in simple parametric models, the posterior distribution can displays substantial asymmetries that yield major bias and reduced accuracy when considering symmetric approximations. Recent research has moved toward more flexible classes of approximating densities incorporating skewness. However, current solutions are model specific, lack general supporting theory and usually increase the computational challenges and complexity of the optimization problem.

This thesis aims to fill such a gap by developing a general, and theoretically supported, family of skew-symmetric approximations. To accomplish this goal, Chapter 1 demonstrates that in the idealized framework where the true data generating mechanism is known, the posterior distribution converges, in an appropriate sense, to a specific sequence of skew-symmetric distributions at a rate that is faster than the classical Gaussian one derived under the Bernstein-Von Mises theorem. In Chapter 2, these findings further motivate the development of practical plug-in versions that, besides enjoying the same theoretical guarantees, can approximate the posterior distribution in real-world scenarios. The approximations developed in the first two chapters are derived by exploiting asymptotic arguments. Chapter 3 offers a different perspective by introducing a general and provably optimal strategy to perturb any off-the-shelf symmetric approximation of a generic posterior distribution. Such a novel perturbation is derived without additional optimization steps and yields a similarly-tractable approximation within the class of skew-symmetric densities that provably improves the finite sample accuracy of the original symmetric approximation.

Sommario

Nella statistica bayesiana parametrica la distribuzione a posteriori viene spesso approssimata con densità simmetriche, in particolare gaussiane. Questa scelta è, in molti casi, computazionalmente conveniente ma può portare a risultati sub-ottimali. Infatti, anche semplici modelli parametrici possono dare vita a distribuzioni a posteriori sensibilmente asimmetriche che vengono mal descritte da approssimazioni che non tengono conto di tale caratteristica. Non è quindi un caso se, negli ultimi anni, si è osservato un crescente interesse nello sviluppo di approssimazioni deterministiche asimmetriche. Tuttavia, le soluzioni attuali sono sviluppate per modelli specifici, mancano di una teoria generale a supporto e, di solito, presentano una elevata complessità computazionale.

Questa tesi si propone di colmare tali lacune introducendo un'ampia e teoricamente giustificata famiglia di approssimazioni asimmetriche. Nel Capitolo 1 viene dimostrato come, sotto l'assunzione che il meccanismo generatore dei dati sia noto, sia possibile derivare una sequenza di distribuzioni asimmetriche alla quale la distribuzione a posteriori converge, in senso appropriato, più velocemente di quanto faccia verso il classico limite gaussiano derivato del teorema di Bernstein-Von Mises. Nel Capitolo 2, questo risultato teorico viene sfruttato per ottenere delle approssimazioni asimmetriche che, pur mantenendo le medesime garanzie teoriche, non si basano sulla conoscenza del meccanismo generatore dei dati e, quindi, possono essere utilizzate in pratica. I primi due capitoli della tesi si basano su giustificazioni e argomenti di carattere asintotico. Il Capitolo 3 offre, invece, una prospettiva diversa introducendo un metodo che permette di perturbare in modo ottimale ogni approssimazione simmetrica della distribuzione a posteriori. Per ogni numerosità campionaria, tale procedura fornisce una approssimazione asimmetrica che non è mai meno accurata di quella di partenza, pur essendo similmente trattabile. Rilevante è il fatto che la correzione sopra menzionata non necessiti di particolari processi di ottimizzazione ma è basata su semplici valutazioni della distribuzione a priori e della funzione di verosimiglianza.

Ai miei nonni

Acknowledgements

For me, these days mark the end of an eight-year journey at the University of Padova. During this time I have certainly had the opportunity to learn a lot and I am aware that this would not have been possible without the help and support of many people I have met along the way.

For this reason, I would like to start by thanking Bruno. Since my undergraduate days, you have instilled in me your passion for statistics as a way to solve interesting and complex problems and, throughout this PhD program, you have always made sure that I was in the best condition to learn and progress.

Thanks also to Daniele, from day one you bombarded me with a variety of difficult and challenging research ideas. Trying to keep up with your pace is a desperate task, but during these years I could always count on your support. You pushed me to the limit, but at the same time you gave me the chance to follow my intuition and, why not, to make mistakes. Today, when I look back on the work we have done, I am very proud of it.

It should also be acknowledged that I would not have been able to complete a large part of this thesis without the patient and constant guidance of Botond and of our endless afternoon meetings. I cannot say that I am now an expert in asymptotic statistics, but certainly what I know I owe mainly to you.

I am also very grateful to Alessandra, I learned a lot from you during these years, and to Anirban and Debdeep for giving me the opportunity to spend a research period at the Department of Statistics at Texas A&M University.

Finally, as I approach the end of this PhD program, I can say that I owe what I have accomplished most of all to my family: to my parents for the sacrifices they made to allow me to focus solely on my studies, to Aurora for always encouraging me to keep going during difficult times, to my grandparents for sharing their passion for science with me with great enthusiasm since I was a child, and to Davide, Elisabetta, my aunts, uncles, and cousins for their constant support and encouragement.

Contents

List of Figures

List of Tables

Introduction

Overview

Deterministic approximations of intractable posterior distributions provide a routinely implemented alternative to sampling-based methods in Bayesian inference (see e.g., Tierney and Kadane, 1986; Minka, 2001; Rue *et al.*, 2009; Blei *et al.*, 2017). Albeit derived under different arguments and optimization strategies, these solutions share a common trade-off between the need to avoid an overly simplified characterization of the posterior distribution, which may undermine inference accuracy, and the attempt to facilitate optimization and posterior inference via a sufficiently tractable approximation. For these purposes, Gaussian approximations represent a particularly convenient option as they allow to capture useful characteristics of the posterior distribution while maintaining a high level of analytical tractability.

From a theoretical perspective, the use of distributions within the Gaussian family is justified, in asymptotic regimes, by Bernstein-von Mises-type results. Indeed, in different frameworks ranging from standard parameteric models (Le Cam, 2012) to high-dimensional (Boucheron and Gassiat, 2009; Bontemps, 2011; Spokoiny and Panov, 2021) and semiparametric regimes (Bickel and Kleijn, 2012; Castillo and Rousseau, 2015), this class of theorems demonstrate that, under appropriate conditions, the posterior distribution converges in probability to a Gaussian, usually under the total variation distance.

Although these results are fundamental to our understanding of the theoretical properties of Bayesian methods, limiting Gaussianity may not be representative of the actual posterior behavior in non-asymptotic regimes. In fact, in situations where the sample size is limited, or when the log-posterior is a highly nonlinear function of the model parameters, it is not uncommon to observe posterior distributions that exhibit asymmetry and heavy tails, two characteristics that Gaussian approximations inherently fail to capture. As a consequence, solutions that either explicitly or implicitly take skewness into account (e.g., Rue *et al.*, 2009; Challis and Barber, 2012; Fasano *et al.*, 2022) tend to

perform better than their Gaussian counterparts. Unfortunately, these approximations are often model specific and a general justification similar to the Bernstein-von Mises theorem is not yet available. Indeed, current theory on skewed approximations is either lacking or tailored to the specific models and priors studied (e.g., Fasano *et al.*, 2022).

Main contributions of the thesis

This thesis aims to develop a broad theoretical and methodological framework that justifies the use of asymmetric approximations in parametric Bayesian inference. This is done by considering as approximating class the flexible family of skew-symmetric distributions (Azzalini and Capitanio, 2003; Ma and Genton, 2004). Particular emphasis is given on providing methods that not only produce satisfactory empirical performances but also give a clear quantification of the theoretical improvement provided over more classical solutions.

The skew Bernstein-von Mises theorem

Chapter 1 approaches the problem of approximating the posterior distribution in the idealized setting where the true generating mechanism of the data is known. In addition to provide interesting clues about the elements that play a role in the departure of the posterior from Gaussianity, this chapter introduces one of the key ideas of the thesis. More in detail, we show that, in parametric models admitting a refined version of the local asymptotic normality condition (Van der Vaart, 2000), it is possible to incorporate higher order terms, belonging to the Taylor expansion of both the likelihood and the prior, into the skewness inducing mechanism of a skew-symmetric density (Azzalini and Capitanio, 2003; Ma and Genton, 2004; Azzalini and Capitanio, 2014).

More specifically, we consider the setting where the data are modeled by a parametric family $\mathcal{P}_\Theta$, where each element is uniquely identified by a parameter $\theta \in \Theta \subseteq \mathbb{R}^d$. In full generality, we do not require the assumption that the true data generating process P_0 belongs to $\mathcal{P}_\Theta$, assuming that the final scope of the statistical analysis is to perform inference on θ_*, defined as the Kullback-Leibler projection of P_0 onto $\mathcal{P}_\Theta$. Let $\delta_n \to 0$ be a generic norming rate describing the rate at which the posterior distribution shrinks toward θ_*. For the reparametrization $h = \delta_n^{-1}(\theta - \theta_*)$, we show that, under mild technical conditions, it is possible to improve the rate of convergence, in total variation distance to the true posterior, by a multiplicative factor δ_n, by replacing the Gaussian distribution of the Bernstein-von Mises theorem with a skew-symmetric density of the

form $2\phi_d(h;\xi,\Omega)w(h-\xi)$. In the last expression, $\phi_d(h;\xi,\Omega)$ is the density function of a d-dimensional Gaussian distribution and $w(h-\xi)$ is a skewness-inducing factor depending on a cubic polynomial of the parameter with terms that disappear as $\delta_n \to 0$. In the limit, the mean parameter ξ and the covariance matrix Ω will also approach those predicted by the usual Gaussian Bernstein-von Mises theorems. As a consequence, as $\delta_n \to 0$, the skew-symmetric becomes increasingly close to a Gaussian.

This new approximation provides a series of theoretical and practical advancements relative to higher-order studies relying on Edgeworth or other types of techniques (see e.g., Johnson, 1970; Weng, 2010; Kolassa and Kuffner, 2020, and the references therein). Indeed, these contributions give supporting theory for arbitrarily truncated versions of infinite expansions which, however, do not necessarily correspond to closed-form valid densities, even after normalization - e.g., the density approximation is not guaranteed to be non-negative (see e.g., Kolassa and Kuffner, 2020, Remark 11). In contrast, the skewed Bernstein-von Mises results derived in Chapter 1, establish convergence to a valid and interpretable class of densities which are almost as tractable as the Gaussian counterpart. In fact, skew-symmetric random variables admit both closed-form normalizing constant and simple i.i.d. sampling schemes, which facilitate inference via Monte Carlo evaluation of any functional of the approximate posterior.

In the second part of the chapter, the skew-symmetric limiting law is specialized for the asymptotic regime where the parameter dimension d is fixed and $\delta_n = 1/\sqrt{n}$, with n being the sample size. In this case, the total variation distance between the skew-symmetric approximation and the posterior is shown to be asymptotically of order $1/n$ in probability, up to a logarithmic term, while under similar assumptions the rate for the Gaussian approximation is of order $1/\sqrt{n}$ in probability, again up to a logarithmic term. Under mild additional conditions, we show that the same improvement holds for the estimation of the posterior expectation of functions that are bounded by a polynomial.

Joint and marginal skew-modal approximations

The theoretical results derived in Chapter 1 describe the asymptotic behavior of the posterior distribution evaluated with respect to the unknown parameter θ_*. This perspective provides insightful indications about the factors influencing the asymmetry of the posterior distribution and, at the same time, allows us to demonstrate the validity of the skew-symmetric limiting law under mild regularity conditions. From a practical point of view, however, the dependence on θ_* prevents our method to be directly applied as an approximation technique in real-world scenarios. In addition, a second aspect to

consider is that the process of generating a sample from the skew-symmetric approximation requires multiple evaluations of the skewness-inducing factor $w(\cdot)$, an operation that is cubic in the parameter dimension d.

In Chapter 2, the above considerations naturally motivate the development of a novel practical class of skew-modal approximations. As a first step, we introduce a plug-in version of the joint skew-symmetric approximation which replace the unknown θ_* with a consistent estimator. Even though, in principle, many different solutions could be adopted, we specialize our results for the posterior mode, as it leads to a sensible simplification of many of the quantities involved in the evaluation of the approximation. Then, using arguments similar to those in Chapter 1, we derive a family of skew-marginal approximations that can be applied to any subset $\theta_{\mathcal{C}}$ of the model parameter θ.

Under mild regularity conditions, it is then proved that the newly derived "practical" versions approximate the posterior distribution with the same level of asymptotic accuracy as their theoretical counterpart based on θ_* and described in Chapter 1.

If the interest is in the posterior marginals, the possibility to rely on approximations with closed-form expressions provides important computational gains. In fact, beside removing the need of drawing samples from the joint approximation, the newly derived skew marginal approximations depend on skewness-inducing factors that, once their parameters are evaluated a first time, have a computational cost which is cubic in the dimension of $\theta_{\mathcal{C}}$ and, therefore, negligible for univariate and bivariate marginal distributions.

The applicability and the good empirical performances of the proposed methods is demonstrated both through a simulation study and an application on a binary regression model. The latter also shows how the proposed methodology is competitive with respect to other state of the art approximations such as expectation propagation (Minka, 2001) and mean field variational Bayes (see e.g., Durante and Rigon, 2019).

General skew-symmetric approximations of posterior distributions

The first two chapters of the thesis deal with skew approximations of the posterior distribution obtained by perturbing Gaussian densities via a scheme which is derived under asymptotic arguments. In principle, such a choice for the symmetric component is not the only possible option. In fact, routinely implemented approximation methods such as variational Bayes (Opper and Archambeau, 2009; Blei *et al.*, 2017) and expectation propagation (Minka, 2001; Chopin and Ridgway, 2017) provide alternative Gaussian approximations as the result of specific optimization processes. At the same time, from

a practical point of view, it is not clear in these cases which form the skewness-inducing factor should have and whether there is a general way to find it.

In Chapter 3, we provide an answer to the above challenging questions by deriving a broadly-applicable strategy to perturb, at no additional optimization costs, any off-the-shelf symmetric approximation, giving rise to a similarly-tractable, yet provably more accurate, skew-symmetric version. Such a novel solution arises by noticing that any symmetric approximation $f^*_{\hat{\theta}}(\theta)$ of a posterior distribution $\pi_n(\theta)$, is provably more accurate in approximating a suitably-symmetrized version $\bar{\pi}_{n,\hat{\theta}}(\theta) = (\pi_n(\theta) + \pi_n(2\hat{\theta} - \theta))/2$ of such a posterior rather than its original non-symmetrized form, under the total variation distance and any α-divergence. Albeit seemingly irrelevant, such a result is of fundamental importance. In fact, although the final target is the original posterior and not its symmetrized version, the two corresponding densities are crucially related in a way that suggests a direct strategy to perturb any off-the-shelf symmetric approximation of a generic posterior distribution. In particular, we demonstrate that the density of the original posterior can be expressed as the product of its symmetrized version and a skewness-inducing factor $w^*_{\hat{\theta}}(\theta) = \pi_n(\theta)/(\pi_n(\theta) + \pi_n(2\hat{\theta} - \theta))$ which is available in closed-form and does not depend on additional unknown parameters. Such a characterization relates to a fundamental existence and uniqueness result of skew-symmetric representations (Wang *et al.*, 2004) which has been never explored within the context of Bayesian inference and approximations, despite its unique potentials. In fact, such a parallel ensures that the perturbation of the original symmetric approximation $f^*_{\hat{\theta}}(\theta)$ by the newly-derived skewness-inducing factor yields a density $2f^*_{\hat{\theta}}(\theta)w^*_{\hat{\theta}}(\theta)$ that falls within the class of skew-symmetric distributions (Azzalini and Capitanio, 2003; Wang *et al.*, 2004; Ma and Genton, 2004; Genton and Loperfido, 2005; Azzalini and Capitanio, 2014), thus admitting a closed-form normalizing constant and straightforward i.i.d. sampling schemes. Such schemes only require simulation from the original symmetric approximation and evaluation of the analytically-available skewness-inducing factor. Importantly, such a factor depends only on ratios of posterior densities, thus allowing cancellation of the intractable normalizing constant and, hence, direct computation without additional optimization costs.

Using rigorous theoretical arguments we prove that the proposed strategy is not only computationally tractable, but also yields skewed approximating densities that are provably more accurate than the original symmetric approximation, for any sample size n and under several routinely-studied divergences. This fact is also supported with two real-world applications which further demonstrate the superior performance of our methodological proposal compared to its symmetric counterparts.

Chapter 1

A skewed Bernstein-von Mises theorem

1.1 Introduction

In Bayesian statistic, several commonly adopted deterministic approximations of the posterior distribution rely either explicitly or implicitly on Gaussian densities (e.g., Tierney and Kadane, 1986; Minka, 2001; Rue *et al.*, 2009; Blei *et al.*, 2017). From a theoretical point of view, the choice of the Gaussian family to approximate the posterior distribution is justified, in asymptotic regimes, by Bernstein - von Mises type results. In its original formulation (e.g. Laplace, 1810; Von Mises, 1931; Le Cam, 1953; Le Cam and Yang, 1990; Van der Vaart, 2000), the famous Bernstein-von Mises theorem states that, in sufficiently regular parametric models, the posterior distribution converges to a Gaussian in total variation distance with probability tending to one under the true law of the data. Both the mean and the covariance of such limiting Gaussian depends only on quantities related to the log-likelihood function, while the prior effect tends to disappear as the sample size increases. Extensions of the Bernstein-von Mises theorem to more complex settings such as misspecified (Kleijn and Van der Vaart, 2012), high-dimensional (Boucheron and Gassiat, 2009; Bontemps, 2011; Spokoiny and Panov, 2021) and semiparametric models (Bickel and Kleijn, 2012; Castillo and Rousseau, 2015; Ray and Van der Vaart, 2020) have also been made in recent years. In addition, Bernstein-von Mises type results with a special focus on scalable deterministic approximations such as expectation-propagation (EP), variational Bayes (VB) and Laplace methods have been obtained by Dehaene and Barthelmé (2018), Wang and Blei (2019) and Kasprzak *et al.* (2022), respectively.

Although considering different regimes, the above results share a common focus on

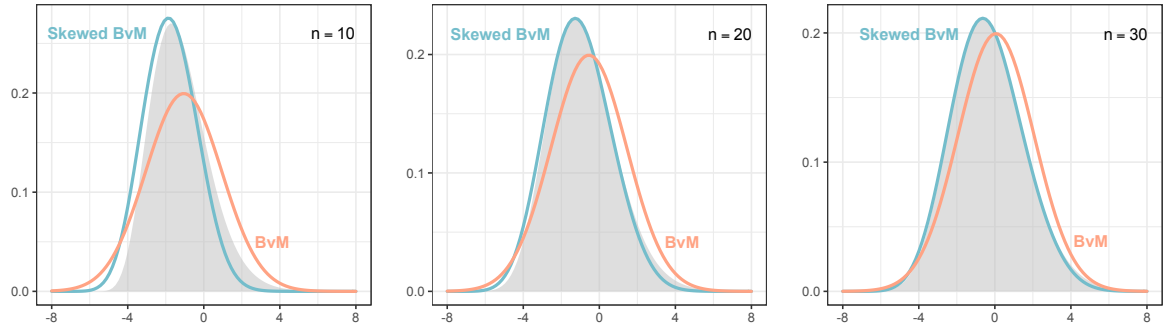

FIGURE 1.1: For varying sample size $n \in \{10, 20, 30\}$, graphical comparison between the approximations of the exact posterior density (grey area) provided by the classical Gaussian Bernstein-von Mises theorem (BvM) and the newly-developed skewed version relying on skew-symmetric approximating densities (Skewed BvM). Results refer to an exponential EXP(θ) model with EXP(1) prior on the rate parameter θ.

the asymptotic Gaussianity of the posterior distribution and on its frequentist properties, such as the coverage of credible intervals. While this perspective has led to fundamental advances in our understanding of Bayesian asymptotics, when the goal is to approximate the posterior distribution, Gaussian approximations may fail to capture relevant aspects such as asymmetry and heavy tails, as illustrated for a simple exponential model in Figure 1.1.

As a consequences, available extensions of Gaussian deterministic approximations that either explicitly or implicitly account for skewness (e.g., Rue *et al.*, 2009; Challis and Barber, 2012; Fasano *et al.*, 2022) have shown evidence of improved empirical accuracy relative to the Gaussian counterparts. Nevertheless, these approximations are often model specific, and a general justification similar to the Bernstein-von Mises theorem is not yet available. In fact, current theory on skewed approximations is either lacking or tailored to the specific models and priors studied (e.g., Fasano *et al.*, 2022). In this chapter, we take a step toward filling the aforementioned gaps by deriving a novel limiting law for posterior distributions which belongs to the class of skew-symmetric distributions (Azzalini and Capitanio, 2003; Ma and Genton, 2004; Azzalini and Capitanio, 2014) and yields noticeable improvements over the convergence rate of the classical Bernstein-von Mises theorem based on limiting Gaussian densities.

1.2 Notation

Let $\{X_i\}_{i=1}^n$, $n \in \mathbb{N}$, denote a sequence of random variables with unknown true distribution P_0^n. Let $\mathcal{P}_\Theta = \{P_\theta^n, \theta \in \Theta\}$, with $\Theta \subseteq \mathbb{R}^d$, be a parametric family of distributions. In the following, we assume that there exists a common σ-finite measure μ^n which dominates P_0^n as well as all measures P_θ^n and we denote by p_0^n and p_θ^n the

corresponding density functions. The Kullback-Leibler projection $P_{\theta_*}^n$ of P_0^n onto $\mathcal{P}_\Theta$ is defined as $P_{\theta_*}^n = \text{argmin}_{P_\theta^n \in \mathcal{P}_\Theta} \text{KL}[P_0^n \| P_\theta^n]$ where $\text{KL}[P_0^n \| P_\theta^n]$ denotes the Kullback-Leibler divergence between the two probability measures P_0^n and P_θ^n. For two generic probability densities $p_1(\theta)$ and $p_2(\theta)$, $\mathcal{D}_{\text{TV}}\big[p_1(\theta) \;||\; p_2(\theta)\big] = (1/2)\int |p_1(\theta) - p_2(\theta)| d\theta$ denotes their total variation distance (see e.g., Ghosal and Van der Vaart, 2017, Appendix B).

The, possibly misspecified, likelihood of the model is $L(\theta) = L(\theta; X^n) = p_\theta^n(X^n)$ while the log-likelihood is indicated with $\ell(\theta) = \ell(\theta; X^n) = \log L(\theta; X^n)$. Prior and posterior distributions are denoted by $\Pi(\cdot)$ and $\Pi_n(\cdot)$ while their densities are $\pi(\cdot)$ and $\pi_n(\cdot)$, respectively.

Our results rely on higher-order expansions and derivatives. To this end, we characterize operations among vectors, matrices and arrays in a compact manner by adopting the index notation along with the Einstein's summation convention (e.g., Pace and Salvan, 1997, pg. 335). More specifically, the inner product $Z^\intercal a$ between the generic random vector $Z \in \mathbb{R}^d$, with components Z_s for $s = 1, \ldots, d$, and the vector of coefficients $a \in \mathbb{R}^d$ having elements a_s for $s = 1, \ldots, d$, is expressed as $a_s Z_s$, with the sum being implicit in the repetition of the indexes. Similarly, if B is a $d \times d$ matrix with entries b_{st} for $s, t = 1, \ldots, d$, the quadratic form $Z^\intercal B Z$ is expressed as $b_{st} Z_s Z_t$. The generalization to operations involving arrays with higher dimensions is obtained under the same reasoning.

Leveraging the above notation, the score vector evaluated at θ_* is defined as

$$\ell_{\theta_*}^{(1)} = [\ell_s^{(1)}(\theta)]_{|\theta=\theta_*} = [(\partial/\partial\theta_s)\ell(\theta)]_{|\theta=\theta_*} \in \mathbb{R}^d,$$

whereas, the second, third and fourth order derivatives of $\ell(\theta)$, still evaluated at θ_*, are

$$\begin{aligned}
\ell_{\theta_*}^{(2)} &= [\ell_{st}^{(2)}(\theta)]_{|\theta=\theta_*} = [\partial/(\partial\theta_s\partial\theta_t)\ell(\theta)]_{|\theta=\theta_*} \in \mathbb{R}^{d\times d}, \\
\ell_{\theta_*}^{(3)} &= [\ell_{stl}^{(3)}(\theta)]_{|\theta=\theta_*} = [\partial/(\partial\theta_s\partial\theta_t\partial\theta_l)\ell(\theta)]_{|\theta=\theta_*} \in \mathbb{R}^{d\times d\times d}, \\
\ell_{\theta_*}^{(4)} &= [\ell_{stlk}^{(4)}(\theta)]_{|\theta=\theta_*} = [\partial/(\partial\theta_s\partial\theta_t\partial\theta_l\partial\theta_k)\ell(\theta)]_{|\theta=\theta_*} \in \mathbb{R}^{d\times d\times d\times d},
\end{aligned}$$

where all the indexes in the above definitions and in the subsequent ones go from 1 to d. The observed and expected Fisher information are denoted by $J_{\theta_*} = [j_{st}] = -[\ell_{\theta_*,st}^{(2)}] \in \mathbb{R}^{d\times d}$ and $I_{\theta_*} = [i_{st}] = [\mathbb{E}_0^n j_{st}] \in \mathbb{R}^{d\times d}$, where $\mathbb{E}_0^n$ is the expectation with respect to P_0^n. In addition,

$$\begin{aligned}
\log \pi_{\theta_*}^{(1)} &= [\log \pi(\theta)_s^{(1)}]_{|\theta=\theta_*} = [\partial/(\partial\theta_s)\log\pi(\theta)]_{|\theta=\theta_*} \in \mathbb{R}^d, \\
\log \pi_{\theta_*}^{(2)} &= [\log \pi(\theta)_{st}^{(2)}]_{|\theta=\theta_*} = [\partial/(\partial\theta_s\partial\theta_t)\log\pi(\theta)]_{|\theta=\theta_*} \in \mathbb{R}^{d\times d},
\end{aligned}$$

represent the first two derivatives of the log-prior density, evaluated at θ_*.

The Euclidean norm of a vector $a \in \mathbb{R}^d$ is denoted by $\|a\|$, whereas, for a generic $d \times d$ matrix B, the notation $|B|$ indicates its determinant and $\lambda_{\text{MIN}}(B)$ and $\lambda_{\text{MAX}}(B)$ its minimum and maximum eigenvalue, respectively. Furthermore, $u \wedge v$ and $u \vee v$ correspond to $\min\{u, v\}$ and $\max\{u, v\}$. For two positive sequences u_n, v_n we employ $u_n \lesssim v_n$ if there exists a universal positive constant C such that $u_n \leq C v_n$. When $u_n \lesssim v_n$ and $v_n \lesssim u_n$ are satisfied simultaneously, we write $u_n \asymp v_n$.

1.3 The skew-symmetric family of distributions

This thesis makes extensive use of approximating densities belonging to the skew-symmetric family of distributions (Azzalini and Capitanio, 2003; Ma and Genton, 2004; Azzalini and Capitanio, 2014). As illustrated in Definition 1.1, this class of random variables is characterized by a flexible mechanism which perturbs a generic symmetric density function with a skewness-inducing factor $w(\cdot)$. Such a simple mathematical structure gives rise to a wide variety of density functions that can encompass not only asymmetry but also multimodality (Ma and Genton, 2004).

Definition 1.1. A random variable θ taking values in $\Theta \subseteq \mathbb{R}^d$ is skew-symmetric (see e.g., Ma and Genton, 2004; Azzalini and Capitanio, 2014) if its density function can be written as

$$2p(\theta - \xi)w(\theta - \xi), \tag{1.1}$$

where $\xi \in \mathbb{R}^d$, $p(\cdot)$ is a symmetric density about zero and $w : \mathbb{R}^d \to [0, 1]$ is a skewness-inducing factor which satisfies $0 \leq w(x) \leq 1$ and $w(-x) = 1 - w(x)$.

Furthermore, $w(\cdot)$ can be equivalently expressed as the composition

$$w(\cdot) = F(\alpha(\cdot)), \tag{1.2}$$

where $F(\cdot)$ is the cumulative distribution function of a univarite random variable with density symmetric about zero and $\alpha : \mathbb{R}^d \to \mathbb{R}$ is an odd function (Ma and Genton, 2004).

A notable example in the skew-symmetric family is the skew-normal distribution (Azzalini, 1985; Azzalini and Dalla Valle, 1996), which is obtained by taking $p(\cdot)$ multivariate Gaussian, $F(\cdot)$ the cumulative distribution function of the standard normal distribution, and $\alpha(\theta - \xi) = a^\intercal(\theta - \xi)$ for $a \in \mathbb{R}^d$. A particularly attractive property of the skew-normal distribution it that its moments are available in closed form (Azzalini and Dalla Valle, 1996). This fact does not generally hold for any generic skew-symmetric

distributions, which therefore necessitate the use of numerical or Monte Carlo methods to evaluate quantities of interest in statistical inference. For this purpose, in Proposition 1.2 we recall a useful stochastic representation of skew-symmetric random variables. A direct consequence of Proposition 1.2 is that if simulating from $p(\cdot)$ is easy, then it is possible to obtain a sample from (1.1) in an exact manner and with little additional computational cost.

Proposition 1.2. *Let θ_{sym} be a realization of a random variable having density function $p(\cdot)$ as described in (1.1). Conditioned on θ_{sym} define $S_{\theta_{\text{sym}}}$ such that*

$$S_{\theta_{\text{sym}}} = \begin{cases} 1 & \textit{with probability } w(\theta_{\text{sym}}), \\ 0 & \textit{otherwise.} \end{cases} .$$

Then, the random variable $\theta = (2S_{\theta_{\text{sym}}} - 1)\theta_{\text{sym}} + \xi$ has density function (1.1) (Azzalini and Capitanio, 2014, pg. 6).

Finally, note that the class of distributions described by (1.1) is very broad. In our work, apart from some exceptions in Chapter 3, we will consider skew-symmetric distributions where the symmetric component $p(\cdot)$ belongs to the Gaussian family.

1.4 A skewed Bernstein-von Mises theorem

This section presents the first important contribution of the thesis. More specifically, we show how, in Bayesian models satisfying a refined version of the local asymptotic normality (LAN) condition (see e.g., Van der Vaart, 2000; Kleijn and Van der Vaart, 2012), a new treatment of higher-order terms can yield to a novel limiting law of densities within the skew-symmetric family (SKS) (Azzalini and Capitanio, 2003; Ma and Genton, 2004). Focusing on this alternative sequence of approximating densities, we then prove that, compared to the classical Gaussian approximation, the convergence rate of its total variation distance from the exact posterior distribution improves by at least one order of magnitude. In regular, possibly misspecified settings, this implies a gain of a factor of order $\sqrt{n}$ over the Gaussian limit predicted by the Bernstein-von Mises theorem.

Let $\delta_n \to 0$ be generic a norming rate governing the posterior contraction toward θ_*. Consistent with standard Bernstein-von Mises type theory (Van der Vaart, 2000; Kleijn and Van der Vaart, 2012), let us consider the re-parametrization $h = \delta_n^{-1}(\theta - \theta_*) \in \mathbb{R}^d$. Then, as illustrated below in Section 1.4.2, the newly-derived class of approximating

SKS densities p^n_{SKS} has the general form

$$p^n_{\text{SKS}}(h) = 2\phi_d(h;\xi,\Omega)\, w(h-\xi), \tag{1.3}$$

while $P^n_{\text{SKS}}(S) = \int_S p^n_{\text{SKS}}(h)dh$ is the corresponding cumulative distribution function. In Equation (1.3), $\phi_d(\cdot;\xi,\Omega)$ is the density of a d-variate Gaussian variable with mean vector ξ and covariance matrix Ω. The term $w(h-\xi) \in (0,1)$ is instead responsible for inducing skewness, and takes the form $w(h-\xi) = F(\alpha_\eta(h-\xi))$ where $F : \mathbb{R} \to [0,1]$ is any univariate cumulative distribution function which satisfies $F(-x) = 1-F(x)$ and $F(x) = 1/2 + \eta x + O(x^2)$, $x \to 0$, for some $\eta \in \mathbb{R}$, and $\alpha : \mathbb{R}^d \to \mathbb{R}$ is a third order polynomial of $(h-\xi)$.

In Section 1.5, we show that in the asymptotic regime where $n \to \infty$, d is fixed and $\delta_n^{-1} = \sqrt{n}$, the components of (1.3) take the form

$$\begin{aligned}
&\xi = [\xi_s] = [n(J^{-1}_{\theta_*})_{st} u_t] \in \mathbb{R}^d, \quad \text{with } u_t = (\ell^{(1)}_{\theta_*} + \log \pi^{(1)}_{\theta_*})_t/\sqrt{n}, \quad t = 1,\dots,d,\\
&\Omega^{-1} = [j_{st}/n - (\xi_l a^{(3),n}_{\theta_*,stl})/\sqrt{n}] \in \mathbb{R}^{d\times d},
\end{aligned}$$

and

$$\alpha_\eta(h-\xi) = a^{(3),n}_{\theta_*,stl}\{(h-\xi)_s(h-\xi)_t(h-\xi)_l + 3(h-\xi)_s\xi_t\xi_l\}/(12\eta\sqrt{n}) \in \mathbb{R},$$

with $a^{(3),n}_{\theta_*} = \ell^{(3)}_{\theta_*}/n$. Interestingly, in this last case, the first factor in the right hand side of (1.3) closely resembles the limiting Gaussian density with mean vector $\ell^{(1)}_{\theta_*}/\sqrt{n}$ and covariance matrix $(I_{\theta_*}/n)^{-1}$ from the classical Bernstein-von Mises theorem which, however, fails to incorporate skewness. For this reason, the symmetric density factor in (1.3) is further perturbed via a skewness-inducing mechanism regulated by $w(h-\xi)$ to obtain a valid and known asymmetric density with tractable normalizing constant. Indeed, since $\alpha_\eta(h-\xi)$ is an odd function of $(h-\xi)$, and $\phi_d(\cdot;\xi,\Omega)$ is symmetric about ξ, it follows that $p^n_{\text{SKS}}(h)$ in (1.3) is a density belonging to the skew-symmetric family introduced in Section 1.3.

The fact that this new approximation is a proper density function is particularly remarkable. Indeed, even though similar higher order approximations, based on Edgeworth or other expansions (e.g., Johnson, 1970; Weng, 2010; Kolassa and Kuffner, 2020), can be found in the literature, they generally suffer from several limitations such as the possibility of assuming negative values (Kolassa and Kuffner, 2020), a problem that does not affect (1.3). Moreover, the stochastic representation introduced in Proposition 1.2 allows to easily design Monte Carlo strategies to evaluate functionals of interest for statistical inference with a similar computational cost as simulating from a multivariate

Gaussian distribution.

The aforementioned discussion clarifies how our result points toward the possibility of employing a more refined family of valid approximating distributions arising from a simple perturbation of multivariate Gaussian densities that still allows tractable inference, while yielding provable, and noticeable, improvements in approximation accuracy.

1.4.1 Derivation of the skew-symmetric approximating distribution

Prior to state and prove the skewed Bernstein-von Mises theorem in Section 1.4.2, let us focus on providing a constructive derivation of the skew-symmetric approximating density in (1.3) via a third-order version of the Laplace method. To simplify the notation, we consider the simple univariate case $d = 1$ and $\delta_n^{-1} = \sqrt{n}$. The extension to $d > 1$ and different asymptotic frameworks yielding the general expression for $p^n_{\text{SKS}}(h)$ in (1.3) follows as a direct adaptation of this univariate case and are reported in Sections 1.4.2 and 1.5.

As a first step towards deriving the SKS density $p^n_{\text{SKS}}(h)$ in (1.3), notice that the posterior for $h = \sqrt{n}(\theta - \theta_*)$ can be expressed as

$$\pi(h \mid X^n) \propto \frac{p^n_{\theta_*+h/\sqrt{n}}}{p^n_{\theta_*}}(X^n)\frac{\pi(\theta_* + h/\sqrt{n})}{\pi(\theta_*)}, \tag{1.4}$$

since $p^n_{\theta_*}(X^n)$ and $\pi(\theta_*)$ do not depend on h, and $\theta = \theta_* + h/\sqrt{n}$.

Under suitable regularity conditions discussed in Section 1.4.2 and Section 1.5 below, the third-order Taylor's expansion for the logarithm of the likelihood ratio in Equation (1.4) is

$$\log \frac{p^n_{\theta_*+h/\sqrt{n}}}{p^n_{\theta_*}}(X^n) = \frac{\ell^{(1)}_{\theta_*}}{\sqrt{n}}h - \frac{1}{2}\frac{j_{\theta_*}}{n}h^2 + \frac{1}{6\sqrt{n}}\frac{\ell^{(3)}_{\theta_*}}{n}h^3 + O_{P^n_0}\big(n^{-1}\big), \tag{1.5}$$

whereas the first order Taylor's expansion of the log-prior ratio is

$$\log \frac{\pi(\theta_* + h/\sqrt{n})}{\pi(\theta_*)} = \frac{\log \pi^{(1)}_{\theta_*}}{\sqrt{n}}h + O\big(n^{-1}\big). \tag{1.6}$$

Combining (1.5) and (1.6) it is possible to reformulate the right-hand-side of Equation (1.4) as

$$\frac{p^n_{\theta_*+h/\sqrt{n}}}{p^n_{\theta_*}}(X^n)\frac{\pi(\theta_* + h/\sqrt{n})}{\pi(\theta_*)} = \exp\left(uh - \frac{1}{2}\frac{j_{\theta_*}}{n}h^2 + \frac{1}{6\sqrt{n}}\frac{\ell^{(3)}_{\theta_*}}{n}h^3\right) + O_{P^n_0}(n^{-1}), \tag{1.7}$$

where $u = (\ell^{(1)}_{\theta_*} + \log \pi^{(1)}_{\theta_*})/\sqrt{n}$.

Notice that, up to a multiplicative constant, the Gaussian density arising from the classical Bernstein-von Mises theorem can be obtained by neglecting all terms in (1.5)-(1.6) which converge to zero in probability. These correspond to the contribution of the prior, the difference between the observed and expected Fisher information, and the term associated to the third-order log-likelihood derivative. Maintaining these quantities would surely yield improved accuracy, but it is unclear whether a valid and similarly-tractable density can be still identified. In fact, current solutions (e.g., Johnson, 1970) consider approximations based on the sum among a Gaussian density and additional terms in the Taylor's expansion. However, as for related alternatives arising from Edgeworth-type expansions (e.g., Weng, 2010; Kolassa and Kuffner, 2020), there is no guarantee that such constructions provide valid densities.

As a first key contribution we prove below that a valid and tractable approximating density can be, in fact, derived from the above expansions and belongs to the SKS class. To this end, let $\omega = 1/v$ with $v = j_{\theta_*}/n - \xi\ell^{(3)}_{\theta_*}/n^{3/2}$ and $\xi = n(j_{\theta_*})^{-1}u$, and note that, by replacing h^3 in the right hand side of Equation (1.7) with $(h+\xi-\xi)^3$, the exponential term in (1.7) can be rewritten as proportional to

$$\phi(h;\xi,\omega)\exp(\{1/(6\sqrt{n})\}(\ell^{(3)}_{\theta_*}/n)\{(h-\xi)^3 + 3(h-\xi)\xi^2\}). \tag{1.8}$$

Next recall that, for $x \to 0$, we can write $\exp(x) = 1+x+O(x^2)$ and $2F(x) = 1+2\eta x+O(x^2)$, for some $\eta \in \mathbb{R}$, where $F(\cdot)$ is the univariate cumulative distribution function introduced in Section 1.4 above. Therefore, leveraging the similarity among these two expansions and the fact that the exponent in Equation (1.8) is an odd function of $(h-\xi)$ about 0, of order $O_{P^n_0}(n^{-1/2})$, it follows that (1.8) is equal to

$$2\phi(h;\xi,\omega)F(\alpha_\eta(h)) + O_{P^n_0}(n^{-1}),$$

with $\alpha_\eta(h)$ defined below Equation (1.3), thereby yielding for the univariate case the skew-symmetric density in (1.3), up to an additive $O_{P^n_0}(n^{-1})$ term. The direct extension of the above derivations to the multivariate case provides the general form of $p^n_{\text{SKS}}(h)$ in (1.3).

1.4.2 A general theorem

The take-home message of Section 1.4.1 is that, if the cubic term in the Taylor expansion of the log-posterior is sufficiently small, it is possible to incorporate it into

the skewing factor of a SKS distribution, improving the quality of the Gaussian approximation by an order of magnitude while avoiding classical problems of polynomial approximations such as regions with negative mass (see e.g McCullagh, 2018, pg. 154).

In this section, we show that this idea can be applied to obtain a skew-symmetric approximation in a wide variety of different settings, provided that the posterior contraction is governed by a generic norming rate $\delta_n \to 0$ and that some additional and reasonable regularity conditions are satisfied. More specifically, Theorem 1.3 requires Assumptions 1-4 stated below. For convenience, let us also introduce the notation $M_n = \sqrt{c_0 \log \delta_n^{-1}}$ where c_0 is a positive constant to be specified later.

Assumption 1. *The Kullback-Leibler projection $\theta_* \in \Theta$ is unique.*

Assumption 2. *There exists a sequence of d-dimensional random vectors $\Delta^n_{\theta_*} = O_{P_0^n}(1)$, a sequence of random matrices $V^n_{\theta_*} = [v^n_{st}]$ where $v^n_{st} = O_{P_0^n}(1)$ and a sequence of three dimensional arrays $a^{(3),n}_{\theta_*} = [a^{(3),n}_{\theta_*,stl}]$ with entries $a^{(3),n}_{\theta_*,stl} = O_{P_0^n}(1)$ so that*

$$\log \frac{p^n_{\theta_*+\delta_n h}}{p^n_{\theta_*}}(X^n) - h_s v^n_{st} \Delta^n_{\theta_*,t} + \frac{1}{2} h_s v^n_{st} h_t - \frac{\delta_n}{6} a^{(3),n}_{\theta_*,stl} h_s h_t h_l = r_{n,1}(h),$$

with $r_{n,1} := \sup_{h \in K_n} |r_{n,1}(h)| = O_{P_0^n}(\delta_n^2 M_n^{c_1})$, for some constant $c_1 > 0$, where $K_n = \{\|\theta - \theta_*\| \leq M_n \delta_n\}$.

In addition, there are two positive constants η_1^* and η_2^* such that the event $A_{n,0} = \{\lambda_{\text{MIN}}(V^n_{\theta_*}) > \eta_1^*\} \cap \{\lambda_{\text{MAX}}(V^n_{\theta_*}) < \eta_2^*\}$, holds with $P_0^n A_{n,0} = 1 - o(1)$.

Assumption 3. *There exist a d-dimensional vector $\log \pi^{(1)}$ for which $\log \pi(\theta_* + \delta_n h)/\pi(\theta_*) - \delta_n h_s \log \pi^{(1)}_s = r_{n,2}(h)$ with $\log \pi^{(1)} = O(1)$ and $r_{n,2} := \sup_{h \in K_n} |r_{n,2}(h)| = O(\delta_n^2 M_n^{c_2})$ for some constant $c_2 > 0$.*

Assumption 4. *It holds $\lim_{\delta_n \to 0} P_0^n \{\Pi_n(\|\theta - \theta_*\| > M_n \delta_n) < \delta_n^2\} = 1$.*

A brief discussion of the above assumptions is in order. Assumption 1 is mild and can be found, for example, in Kleijn and Van der Vaart (2012). Together with Assumption 4, it guarantees that asymptotically the posterior distribution is concentrated in the region where the two Taylor expansions described in Assumptions 2 and 3 hold with negligible remainder terms. Note also that, given Assumptions 2 and 3, the prior affects only the higher-order terms of the log-posterior expansion. This behavior is standard in the $n \to \infty$-fixed d asymptotic framework. On the contrary, recent studies in high-dimensional Bayesian statistics Spokoiny and Panov (2021); Spokoiny (2023) highlight that the prior effect often plays a critical role even through terms associated with its second-order derivatives. In particular, the prior should provide sufficient shrinkage to

control the behavior of the third and fourth-order components of the log-posterior. This framework is beyond the scope of the present section, but can be covered by directly imposing appropriate sufficient conditions on the behavior of the Taylor expansion of the log-posterior.

Theorem 1.3 below states the most important theoretical result of this section. It establishes in a general, misspecified setting, an improved rate of convergence for the total variation distance between the exact posterior and the skew-symmetric approximation introduced in (1.3), under the Assumptions 1-4. Our results clarify that a substantially more accurate representation for the asymptotic behavior of the exact posterior distribution, relative to that achieved under limiting multivariate Gaussians, can be obtained via a similarly tractable class of skew-symmetric densities. It is worth emphasizing that, since these results aim to better describe the behavior of the true posterior distribution, they necessarily inherit its frequentist asymptotic properties. This means that the improvement given by using (1.3) instead of the classical Gaussian approximation should be intended in terms of fidelity to the true posterior and does not, in general, affect quantities such as the frequentist coverage of the credible sets.

Theorem 1.3. *Let* $h = \delta_n^{-1}(\theta - \theta_*)$, *and* $K_n = \{h : \|h\| < M_n\}$. *Under Assumptions 1 to 4, if* θ_* *is an inner point of* Θ, *it holds*

$$\mathcal{D}_{\text{TV}}\big[\pi_n(h) \,||\, p^n_{\text{SKS}}(h)\big] = O_{P_0^n}(M_n^{c_3}\delta_n^2), \tag{1.9}$$

with $c_3 > 0$. *In (1.9) the skew-symmetric limiting density (1.3) has parameters* $\xi = \Delta^n_{\theta_*} + \delta_n(V^n_{\theta_*})^{-1}\log\pi^{(1)}$, $\Omega^{-1} = [v^n_{st} - \delta_n a^{(3),n}_{\theta_*,stl}\xi_l]$ *and* $w(h-\xi) = F(\alpha_\eta(h))$ *with* $F(\cdot)$ *be any univarite cdf satisfying* $F(-x) = 1 - F(x)$ *and* $F(x) = 1/2 + \eta x + O(x^2)$, *for some* $\eta \in \mathbb{R}$, *when* $x \to 0$ *and*

$$\alpha_\eta(h) = \frac{\delta_n}{12\eta}\{\Psi^{(3)}_{stl}(h-\xi)_s(h-\xi)_t(h-\xi)_l + 3\Psi^{(1)}_s(h-\xi)_s\},$$

for $\Psi^{(1)} = [a^{(3),n}_{\theta_*,stl}\xi_t\xi_l]$, $\Psi^{(3)} = [a^{(3),n}_{\theta_*,stl}]$.

Remark 1.4. Under related conditions and a simpler proof, it is possible to show that the order of convergence for the usual Bernstein-von Mises theorem based on Gaussian limiting distributions is $O_{P_0^n}(M_n^{c_4}\delta_n)$, for some $c_4 > 0$. Thus, Theorem 1.3 guarantees that by relying on SKS approximating distributions with the density defined in (1.3) and derived in Section 1.4.1, it is possible to improve the classical Bernstein-von Mises result by a multiplicative factor of δ_n, up a logarithmic term. This follows directly from the fact that the SKS approximation is able to include the terms of order δ_n that are present

in the Taylor expansion of the log-posterior but are neglected in the Gaussian limit. In addition, in order to work, (1.3) requires that the term δ_n is sufficiently small, a condition that is also necessary for the validity of the classical Gaussian approximation. The improvement provided by the use of the skew-symmetric limiting law is thus found when the posterior is sufficiently concentrated in a neighborhood of the location parameter, and it is due to a better repartitioning of the density of the approximation in the high posterior probability region. As illustrated in Sections 1.6-2.4, this simple correction is usually able to significantly improve the ability of the approximation to capture salient features of the posterior density.

Remark 1.5. The form of the univariate cumulative distribution function $F(\cdot)$, in Theorem 1.3, is not explicit since it is only required to satisfy $F(-x) = 1 - F(x)$ and $F(x) = 1/2+\eta x+O(x^2)$ for some $\eta \in \mathbb{R}$ when $x \to 0$. For practical purposes, it is useful to list some possible choices of $F(\cdot)$. Two good candidates are the cumulative distribution function of the standard Gaussian distribution, $\Phi(x)$, and the inverse logit function, $g(x) = \exp(x)/\{1+\exp(x)\}$. In fact, they both satisfy $F(-x) = 1-F(x)$ and their Taylor expansions take the form $\phi(x) = 1/2+x/\sqrt{2\pi}+O(x^3)$ and $g(x) = 1/2+x/4+O(x^3)$ for $x \to 0$. Moreover, in the first case, the combination of the standard normal cumulative distribution function with the Gaussian kernel in Theorem 1.3 yields a skew-symmetric approximation belonging to the family of generalized skew-normal distributions (Ma and Genton, 2004; Genton and Loperfido, 2005).

Before giving the proof of Theorem 1.3, it is worthwhile to discuss an interesting point regarding the interplay between skew-symmetric and Gaussian approximations. In general, a straightforward implication of Theorem 1.3 is that (1.3) provides the same asymptotic improvement over the Gaussian approximation in estimating the posterior expectation of any bounded function, with this fact usually extendable to functions bounded by a polynomial under mild additional conditions, as illustrated in Corollary 1.7. However, Lemma 1.6 below highlights how the skew-symmetric distributional invariance with respect to even functions (Wang *et al.*, 2004) implies that the skew approximation $2\phi_d(h;\xi,\Omega)\, w(h-\xi)$ and the corresponding Gaussian $\phi_d(h;\xi,\Omega)$, obtained by eliminating the skewing factor, provide the same level of accuracy in estimating the posterior expected value of functions that are symmetric with respect to the location parameter. Thus our result provides a new explanation on the phenomenon observed in Spokoiny and Panov (2021) and Spokoiny (2023), where the quality of the Gaussian approximation, in high-dimensional models, increases by an order of magnitude when evaluated on Borel sets which are centrally symmetric with respect to the location parameter (see e.g., Spokoiny, 2023, Thm 3.4).

Lemma 1.6. *Let $2\phi_d(\theta;\xi,\Omega)w(\theta-\xi)$ with $\xi\in\mathbb{R}^d$ and $\Omega\in\mathbb{R}^{d\times d}$, be a skew-symmetric approximation of $\pi_n(\theta)$ and let $G:\mathbb{R}^d\to\mathbb{R}$ be an even function. If both $\int G(\theta-\xi)\pi_n(\theta)d\theta$ and $\int G(\theta-\xi)2\phi_d(\theta;\xi,\Omega)w(\theta-\xi)d\theta$ are finite, it holds*

$$\int G(\theta-\xi)\{\pi_n(\theta)-2\phi_d(\theta;\xi,\Omega)w(\theta-\xi)\}d\theta=\int G(\theta-\xi)\{\pi_n(\theta)-\phi_d(\theta;\xi,\Omega)\}d\theta.$$

Proof. Lemma 1.6 is a direct consequence of Proposition 6 in Wang *et al.* (2004). □

The proof of Theorem 1.3 is reported below and it follows other general reasoning behind the classical Bernstein–von Mises type derivations (e.g., Kleijn and Van der Vaart, 2012), extended to SKS distributions. Nonetheless, as mentioned before, the need to derive a sharp rate which establishes a higher approximation accuracy, relative to Gaussian limiting distributions, requires a number of additional technical lemmas and refined arguments ensuring a tight control of the error terms in the expansions underlying Theorem 1.3.

Proof. By an application of triangle inequality the problem can be split in three parts

$$\begin{aligned}\int|\pi_n(h)-p^n_{\text{SKS}}(h)|dh\leq&\int|\pi_n(h)-\pi_n^{K_n}(h)|dh\\&+\int|\pi_n^{K_n}(h)-2\phi_d^{K_n}(h;\xi,\Omega)w(h-\xi)|dh\\&+\int|2\phi_d(h;\xi,\Omega)w(h-\xi)-2\phi_d^{K_n}(h;\xi,\Omega)w(h-\xi)|dh,\end{aligned}\tag{1.10}$$

where $\pi_n^{K_n}(h)=\pi_n(h)\mathbb{1}_{h\in K_n}/\int_{K_n}\pi_n(h)dh$ and

$$2\phi_d^{K_n}(h;\xi,\Omega)w(h-\xi)=2\phi_d(h;\xi,\Omega)w(h-\xi)\mathbb{1}_{h\in K_n}/\int_{K_n}2\phi_d(h;\xi,\Omega)w(h-\xi)dh,$$

are the constrained versions of the corresponding densities to K_n.

Assumption 4 and a standard inequality of the total variation norm give

$$\int|\pi_n(h)-\pi_n^{K_n}(h)|dh\leq 2\int_{h\,:\,\|h\|>M_n}\pi_n(h)dh=O_{P_0^n}(\delta_n^2).\tag{1.11}$$

We deal with the third term in a similar manner. Leveraging the same total variation inequality as above and $|w(x)|\leq 1$ we obtain

$$\int|2\phi_d(h;\xi,\Omega)w(h-\xi)-2\phi_d^{K_n}(h;\xi,\Omega)w(h-\xi)|dh\leq 4\int_{h\,:\,\|h\|>M_n}\phi_d(h;\xi,\Omega)dh.$$

Let $P_{\xi,\Omega}(\|h\| > M_n) := \int_{h\in K_n^c} \phi_d(h;\xi,\Omega)dh$ and $A_{n,1} = \{\lambda_{\text{MIN}}(\Omega) > \eta_1\} \cap \{\lambda_{\text{MAX}}(\Omega) < \eta_2\} \cap \{\|\xi\| < \tilde{M}_n\}$ for some sequence $\tilde{M}_n = o(M_n)$ going to infinity arbitrary slowly and some $\eta_1, \eta_2 > 0$. Moreover, note that $V_{\theta_*}^n - \Omega^{-1}$ has entries of order $O_{P_0^n}(\delta_n)$. As a consequence, in view of Assumptions 2-3 and Lemma B.2 we get $P_0^n A_{n,1} = 1 - o(1)$. Conditioned on $A_{n,1}$, the eigenvalues of Ω lay on a positive bounded range. This fact, together with Markov's inequality and $\tilde{M}_n/M_n \to 0$ imply, for every $\epsilon > 0$,

$$\begin{aligned} &P_0^n\Big(P_{\xi,\Omega}(\|h\| > M_n)/\delta_n^2 > \epsilon\Big) \;=\; P_0^n\Big(\{P_{\xi,\Omega}(\|h\| > M_n)/\delta_n^2 > \epsilon\} \cap A_{n,1}\Big) + o(1) \\ &\leq P_0^n\Big(e^{-\tilde{c}_1 M_n^2}/\delta_n^2 > \epsilon | A_{n,1}\Big) + o(1) = o(1), \end{aligned} \tag{1.12}$$

where $\tilde{c}_1$ is a sufficiently small positive constant and the last inequality follows from the tail behavior of the multivariate Gaussian for a sufficiently large choice of c_0 in $M_n = \sqrt{c_0 \log \delta_n^{-1}}$. This gives

$$\int |2\phi_d(h;\xi,\Omega)w(h-\xi) - 2\phi_d^{K_n}(h;\xi,\Omega)w(h-\xi)|dh = o_{P_0^n}(\delta_n^2). \tag{1.13}$$

We are left to deal with the term $\int |\pi_n^{K_n}(h) - 2\phi_d^{K_n}(h;\xi,\Omega)w(h-\xi)|$. Let us consider the event

$$A_{n,2} = A_{n,1} \cap \{\int_{K_n} \pi_n(h)dh > 0\} \cap \{\int_{K_n} 2\phi_d(h;\xi,\Omega)w(h-\xi)dh > 0\},$$

Note that $P_0^n\{\int_{K_n} \pi_n(h)dh > 0\} = 1 - o(1)$ by Assumption 4 and that, in view of (1.12), it follows

$$\begin{aligned} P_0^n\Big\{\int_{K_n} 2\phi_d(h;\xi,\Omega)w(h-\xi)dh > 0\Big\} =& P_0^n\Big\{1 - \int_{K_n^c} 2\phi_d(h;\xi,\Omega)w(h-\xi)dh > 0\Big\} \\ \geq& P_0^n\Big\{1 - 2P_{\xi,\Omega}(\|h\| > M_n) > 0\Big\} = 1 - o(1), \end{aligned}$$

implying, in turn, $P_0^n A_{n,2} = 1 - o(1)$. As a consequence, we can restrict our attention to

$$\begin{aligned} &\int |\pi_n^{K_n}(h) - 2\phi_d^{K_n}(h;\xi,\Omega)w(h-\xi)|dh\mathbb{1}_{A_{n,2}} \\ &= \int \Big[\Big|1 - \int_{K_n} \frac{2\phi_d^{K_n}(h;\xi,\Omega)w(h-\xi)}{2\phi_d^{K_n}(g;\xi,\Omega)w(g-\xi)} \frac{p_{\theta_*+\delta_n g}(X^n)\pi(\theta_*+\delta_n g)}{p_{\theta_*+\delta_n h}(X^n)\pi(\theta_*+\delta_n h)} 2\phi_d^{K_n}(g;\xi,\Omega)w(g-\xi)dg\Big| \\ &\quad \times \pi_n^{K_n}(h)dh\mathbb{1}_{A_{n,2}}\Big]. \end{aligned} \tag{1.14}$$

Note that the ratio $2\phi_d^{K_n}(h;\xi,\Omega)w(h-\xi)/2\phi_d^{K_n}(g;\xi,\Omega)w(g-\xi)$ corresponds to its unconditioned version $2\phi_d(h;\xi,\Omega)w(h-\xi)/2\phi_d(g;\xi,\Omega)w(g-\xi)$. This fact and an application

of Jensen inequality implies that the quantity in the last display is upper bounded by

$$\int_{K_n\times K_n}\left|1-\frac{2\phi_d(h;\xi,\Omega)w(h-\xi)}{2\phi_d(g;\xi,\Omega)w(g-\xi)}\frac{p_{\theta_*+\delta_n g}(X^n)\pi(\theta_*+\delta_n g)}{p_{\theta_*+\delta_n h}(X^n)\pi(\theta_*+\delta_n h)}\right|\pi_n^{K_n}(h)2\phi_d^{K_n}(g;\xi,\Omega)w(g-\xi)dhdg\mathbb{1}_{A_{n,2}}.$$

At this point, it is sufficient to recall Lemma B.1 and $e^x = 1 + x + e^{\beta x}x^2/2$, for some $\beta \in (0,1)$ to obtain

$$\begin{aligned}
&\int_{K_n}|\pi_n^{K_n}(h) - 2\phi_d^{K_n}(h;\xi,\Omega)w(h-\xi)|dh\mathbb{1}_{A_{n,2}}\\
&\leq \int_{K_n\times K_n}\left|1-e^{r_{n,4}(g)-r_{n,4}(h)}\right|\pi_n^{K_n}(h)2\phi_d^{K_n}(g;\xi,\Omega)w(g-\xi)dhdg\mathbb{1}_{A_{n,2}} \qquad (1.15)\\
&\leq 2|r_{n,4}| + 2\exp(2\beta|r_{n,4}|)r_{n,4}^2 = O_{P_0^n}(\delta_n^2 M_n^{c_3}),
\end{aligned}$$

where $r_{n,4} = \sup_{h\in K_n} r_{n,4}(h)$ and c_3 is some constant defined in Lemma B.1. Equation (1.9) of the theorem is proved by aggregating (1.11), (1.13) and (1.15). □

1.5 Skew-symmetric approximations in the standard asymptotic limit

The skew Bernstein-Von Mises theorem derived in Section 1.4.2 relies on general assumptions regarding the concentration of the posterior distribution and its Taylor expansion around θ_*. In particular, it holds in settings where the limiting behavior of the posterior is Gaussian, and the inclusion of the prior effect through a first-order Taylor expansion, as well as the third-order term in the Taylor expansion of the log-likelihood, provide an improvement of order δ_n over the Gaussian approximation.

In this section, we give a set of mild technical conditions which guarantee that the results presented in Section 1.4.2 holds, for a large class of parametric models, with the dimension d fixed and $\delta_n = n^{-1/2}$. Furthermore we show that the improvement over the Gaussian approximation can be extended to the expectation of general polynomially bounded functions with finite prior expectation.

Below, we state the assumptions needed, together with the fundamental Assumption 1, to adapt the general skew Bernstein-von Mises theorem developed in Section 1.4.2 to the classical fixed-d-increasing-n asymptotic setting.

Assumption 5. *The, possibly misspecified, log-likelihood of the model is four times differentiable at θ_* with*

$$\ell^{(1)}_{\theta_*,s} = O_{P_0^n}(n^{1/2}), \quad \ell^{(2)}_{\theta_*,st} = O_{P_0^n}(n), \quad \ell^{(3)}_{\theta_*,stl} = O_{P_0^n}(n), \quad \textit{for } \ s,t,l = 1,\dots,d,$$

and $\sup_{h\in K_n}|\ell^{(4)}_{\theta_*,stlk}(h)| = O_{P_0^n}(n)$, *for* $s,t,l,k = 1,\dots,d$, *with* $\ell^{(4)}_{\theta_*,stlk}(h) := \ell^{(4)}_{stlk}(\theta_* + h/\sqrt{n})$.

Assumption 6. *The entries of the Fisher information matrix satisfies* $i_{st} = O(n)$ *while* $j_{st}/n - i_{st}/n = O_{P_0^n}(n^{-1/2})$, *for* $s,t = 1,\dots,d$. *Moreover, there exist two positive constants* η_1 *and* η_2 *such that* $\lambda_{\text{MIN}}(I_{\theta_*}/n) > \eta_1$ *and* $\lambda_{\text{MAX}}(I_{\theta_*}/n) < \eta_2$.

Assumption 7. *The log-prior density* $\log\pi(\theta)$ *is two times continuously differentiable in a neighborhood of* θ_*, *and* $0 < \pi(\theta_*) < \infty$.

Assumption 8. *For every* $M_n \to \infty$ *there exists a constant* c_5 *such that*

$$\lim_{n\to\infty} P_0^n\Big\{\sup_{\|\theta-\theta_*\|>M_n/\sqrt{n}} \frac{1}{n}\{\ell(\theta)-\ell(\theta_*)\} < -c_5\frac{M_n^2}{n}\Big\} = 1.$$

Assumptions 5-6 are mild and usually considered standard in classical frequentist theory (see e.g. Pace and Salvan, 1997, pg. 347). In Lemma 1.9 we show that they allow to precisely control the error in the Taylor approximation of the log-likelihood. Assumption 7 is also satisfied by several priors that are commonly used in practice and it allows us to take a first order Taylor expansion of the form

$$\log\pi(\theta) = \log\pi(\theta_*) + \frac{\log\pi^{(1)}_{\theta_*,s}h_s}{\sqrt{n}} + r_{n,2}(h), \tag{1.16}$$

with $r_{n,2} := \sup_{h\in K_n} r_{n,2}(h) = O(M_n^2/n)$. Similarly, Assumption 8 is needed to control the the rate of contraction of the misspecified posterior distribution into K_n. In other modern versions of Berstein-Von Mises-type results, it is usually replaced by conditions on the existence of a suitable sequence of tests. Sufficient conditions for the well specified case can be found, for example, in Van der Vaart (2000). In the misspecified setting, assumptions ensuring the existence of such tests have been derived by Kleijn and Van der Vaart (2012). Another possible option is to assume, for every $\delta > 0$, the presence of a positive constant c_δ such that

$$\lim_{n\to\infty} P_0^n\{\sup_{\|\theta-\theta_*\|>\delta} \frac{1}{n}\{\ell(\theta)-\ell(\theta_*)\} < -c_\delta\} = 1. \tag{1.17}$$

In the misspecified setting this is done, for example, by Koers *et al.* (2023). Assumption 8 is just a slightly more restrictive version of (1.17). In fact, Lemma 1.11 below shows that it is implied by mild sufficient conditions.

Corollary 1.7 below states that, under Assumption 1 and Assumptions 5-8, Theorem 1.3 holds with the total variation distance between the true posterior and the skew

approximation of order $O_{P_0^n}(M_n^{c_6}/n)$, where c_6 is a fixed positive constant and $M_n = \sqrt{c_0 \log n}$. The result can be extended to polynomially bounded functionals provided that their expectation with respect to the prior distribution is finite.

Corollary 1.7. *Let $h = \sqrt{n}(\theta - \theta_*)$, and $K_n = \{h : \|h\| < M_n\}$. Under assumptions 1 and 5-8 it holds*

$$\mathcal{D}_{\text{TV}}\left[\pi_n(h) \,||\, p^n_{\text{SKS}}(h)\right] = O_{P_0^n}(M_n^{c_6}/n). \tag{1.18}$$

for some $c_6 > 0$. In (1.18) the skew-symmetric density (1.3) has parameters $\xi = \Delta^n_{\theta_*} + (1/\sqrt{n})(V^n_{\theta_*})^{-1} \log \pi^{(1)}_{\theta_*}$, $\Omega^{-1} = [v^n_{st} - (1/\sqrt{n}) a^{(3),n}_{\theta_*,stl}\xi_l]$ with $\Delta^n_{\theta_*,t} = j^{-1}_{st}\sqrt{n}\ell^{(1)}_{\theta_*,s}$, $V^n_{\theta_*} = J_{\theta_*}/n$ and $a^{(3),n} = \ell^{(3)}_{\theta_*}/n$. The skewness-inducing factor takes the form $w(h-\xi) = F(\alpha_\eta(h))$ with $F(\cdot)$ be any univarite cdf satisfying $F(-x) = 1 - F(x)$ and $F(x) = 1/2 + \eta x + O(x^2)$, for some $\eta \in \mathbb{R}$, when $x \to 0$ and

$$\alpha_\eta(h) = \frac{1}{12\eta\sqrt{n}}\{\Psi^{(3)}_{stl}(h-\xi)_s(h-\xi)_t(h-\xi)_l + 3\Psi^{(1)}_s(h-\xi)_s\},$$

for $\Psi^{(1)} = [a^{(3),n}_{\theta_*,stl}\xi_t\xi_l]$, $\Psi^{(3)} = [a^{(3),n}_{\theta_*,stl}]$.

In addition, let $G : \mathbb{R}^d \to R$ be a function satisfying $|G| \lesssim \|h\|^r$. If the prior is such that $\int \|h\|^r \pi(\theta + h/\sqrt{n})dh < \infty$ then

$$\int G(h)|\pi_n(h) - p^n_{\text{SKS}}(h)|dh = O_{P_0^n}(M_n^{c_6+r}/n). \tag{1.19}$$

Remark 1.8. As for Theorem 1.3, under conditions similar to those reported in Corollary 1.7, it is possible to show that the total variation distance between the true posterior and the Gaussian approximation predicted by the standard Bernstein-von Mises theorem is of order $O_{P_0^n}(M_n^{c_7}/\sqrt{n})$ for some fixed $c_7 > 0$. The improvement given by the use of the skew-symmetric approximation is maintained even when the interest is in the expectation of polynomially bounded functions.

For ease of reading, the demonstration of Corollary 1.7 is postponed to Appendix A.1.1.

1.5.1 Log-posterior asymptotic and posterior contraction

In order to move from the general theory of Theorem 1.3 to Corollary 1.7, two key points are the rate at which the posterior concentrates in $K_n = \{h : \|h\| < M_n\}$ and the behavior of the Taylor expansion of the log-likelihood in K_n. In this section, we show that Assumptions 1 and 5-8 are indeed sufficient for our purpose. In particular, Lemma 1.9 below establishes that, under Assumptions 5 and 6, the error given by replacing the

log-likelihood with its third-order Taylor approximation is uniformly of order M_n^4/n on K_n.

Finally, in Lemma 1.10, we show that under Assumptions 5 and 6 it is possible to choose c_0 and hence K_n such that the posterior distribution concentrates its mass in K_n, at any polynomial rate, with P_0^n probability tending to 1 as $n \to \infty$. This in turn implies that Assumption 4 of Section 1.4.2 is satisfied.

Lemma 1.9. *Under Assumptions 5 and 6, it holds in* $K_n = \{h : \|h\| \leq M_n\}$ *that*

$$\log \frac{p^n_{\theta_*+h/\sqrt{n}}}{p^n_{\theta_*}}(X^n) - h_s v^n_{st} \Delta^n_{\theta_*,t} + \frac{1}{2} v^n_{st} h_s h_t - \frac{1}{6\sqrt{n}} a^{(3),n}_{\theta_*,stl} h_s h_t h_l = r_{n,1}(h), \tag{1.20}$$

with $\Delta^n_{\theta_*,t} = j^{-1}_{st}\sqrt{n}\ell^{(1)}_{\theta_*,s} = O_{P_0^n}(1)$, $v^n_{st} = j_{st}/n = O_{P_0^n}(1)$ *and* $a^{(3),n}_{\theta_*,stl} = \ell^{(3)}_{stl}/n = O_{P_0^n}(1)$. *Moreover,*

$$r_{n,1} := \sup_{h \in K_n} |r_{n,1}(h)| = O_{P_0^n}(M_n^4/n). \tag{1.21}$$

Lemma 1.10 (Posterior contraction)**.** *Under Assumptions 5-8, there exists a choice of* $c_0 > 0$ *large enough in* $M_n = \sqrt{c_0 \log n}$, *such that for every* $D > 0$ *it holds*

$$\lim_{n\to\infty} P_0^n\{\Pi_n(K_n^c) < n^{-D}\} = 1,$$

where K_n^c *is the complement of* K_n.

Proof. For ease of reading, the proofs of both lemmas 1.9 and 1.10 are moved to Appendix A.1.1. □

1.5.2 Sufficient conditions for Assumption 8

The validity of the above Lemma 1.10 crucially depends on the fulfillment of Assumption 8 which allows precise control over how the log-likelihood ratio behaves outside the K_n set. In addition, Assumption 8 also plays an important role in the development of the skew-modal approximation, as discussed in Section 2.1. To make this assumption practical, we need a set of easily verifiable sufficient conditions that guarantee its validity. These conditions are given in detail in the Lemma 1.11.

Lemma 1.11. *Suppose that Assumptions 1 and 6 hold and that for every* $\delta > 0$ *there exist a positive constant* c_δ *such that*

$$\lim_{n\to\infty} P_0^n\{\sup_{\|\theta-\theta_*\|>\delta} \frac{1}{n}\{\ell(\theta) - \ell(\theta_*)\} < -c_\delta\} = 1. \tag{1.22}$$

If there exist $\bar{n} \in \mathbb{N}$ *and* $\delta_1 > 0$ *such that, for all* $n > \bar{n}$, *it holds*

R1) $\mathbb{E}_0^n\{\ell(\theta) - \ell(\theta_*)\}/n$ *is concave in* $\{\theta : \|\theta - \theta_*\| < \delta_1\}$*, two times differentiable at* θ_* *with negative Hessian equal to the Fisher information matrix* I_{θ_*}/n*,*

and

R2)

$$\sup_{0<\|\theta-\theta_*\|<\delta_1} \frac{1}{n\|\theta-\theta_*\|}[\{\ell(\theta) - \ell(\theta_*)\} - \mathbb{E}_0^n\{\ell(\theta) - \ell(\theta_*)\}] = O_{P_0^n}\left(n^{-1/2}\right), \quad (1.23)$$

then for every $M_n \to \infty$ *there exist a constant* c_5 *such that*

$$\lim_{n\to\infty} P_0^n\{\sup_{\|\theta-\theta_*\|>M_n/\sqrt{n}} \frac{1}{n}\{\ell(\theta) - \ell(\theta_*)\} < -c_5\frac{M_n^2}{n}\} = 1.$$

Proof. The proof of Lemma 1.11 is postponed to Appendix A.1.4. □

We conclude this section with a detailed discussion of the key assumptions made in Lemma 1.11. As already highlighted above, (1.22) is mild and can be found both in classical versions of the Bernstein-von Mises (Lehmann and Casella, 2006) as well as in modern misspecified results (Koers *et al.*, 2023). Condition R1 requires the expected log-likelihood to be sufficiently regular in a neighborhood of θ_* and it is closely related to standard assumptions on M-estimators (see e.g., Van der Vaart, 2000, Ch. 5). Finally, among the assumptions of Lemma 1.11, R2 is arguably the most specific. It requires that, for all θ such that $0 < \|\theta-\theta_*\| < \delta_1$, the quantity $[\{\ell(\theta)-\ell(\theta_*)\}-\mathbb{E}_0^n\{\ell(\theta)-\ell(\theta_*)\}]/(n\|\theta-\theta_*\|)$ converges uniformly to zero in probability with rate $n^{-1/2}$. This behavior is common in many routinely implemented statistical models, such as generalized linear models.

1.6 Empirical results

Sections 1.6.1-1.6.2 illustrate through simulation studies the validity of the asymptotic results developed in Section 1.5. More specifically, the focus is on providing empirical evidence of the improved accuracy achieved by the SKS limiting approximation in (1.3) relative to its Gaussian counterpart arising from the classical Bernstein-von Mises theorem both in well specified and misspecified settings.

Notice that, as for other versions of the classical Bernstein-von Mises theorem, also our results in Section 1.5 require knowledge of the Kullback-Leibler minimizer between the true data-generating model and the parametric family $\mathcal{P}_\Theta$, which is clearly unknown in practical implementations. In Chapter 2, we address this aspect via a plug-in

TABLE 1.1: In the exponential example, average, over 50 replicated studies, of the log-total-variation (TV) distances and first-moment-absolute-errors (FMAE) corresponding to the classical (BvM) and skewed (S-BvM) Bernstein-von Mises theorem in the well-specified simulation setting described in Section 1.6.1. Results are displayed for different sample sizes from $n = 10$ to $n = 1500$. The bold values indicate the best performance for each sample size.

	$n = 10$	$n = 50$	$n = 100$	$n = 500$	$n = 1000$	$n = 1500$
$\log \text{TV}^n_{\text{BvM}}$	-1.67	-2.50	-2.82	-3.59	-3.98	-4.18
$\log \text{TV}^n_{\text{S-BvM}}$	$\mathbf{-2.53}$	$\mathbf{-3.86}$	$\mathbf{-4.41}$	$\mathbf{-5.76}$	$\mathbf{-5.58}$	$\mathbf{-6.58}$
$\log \text{FMAE}^n_{\text{BvM}}$	-0.90	-1.77	-1.97	-2.85	-3.21	-3.33
$\log \text{FMAE}^n_{\text{S-BvM}}$	$\mathbf{-1.07}$	$\mathbf{-2.81}$	$\mathbf{-3.74}$	$\mathbf{-6.14}$	$\mathbf{-7.09}$	$\mathbf{-7.42}$

version of the skew-symmetric limiting approximation, which replaces θ_* with its maximum–a–posteriori estimate, to obtain similar theoretical and empirical support (see the additional simulations and real-data applications in Section 2.4).

1.6.1 Exponential model

Let $X_i \overset{iid}{\sim} \text{EXP}(\theta_0)$, for $i = 1, \dots, n$, where $\text{EXP}(\theta_0)$ denotes the exponential distribution with rate parameter $\theta_0 = 2$. In the following, we consider a well specified model having exponential likelihood and a $\text{EXP}(1)$ prior for θ. To obtain the skew-symmetric approximation for the posterior distribution induced by such a Bayesian model, let us first verify that all conditions of Corollary 1.7 hold.

To address this goal first notice that, since the model is well specified, Assumption 1 is met with $\theta_* = \theta_0$. The first four derivates of the log-likelihood at θ are $n/\theta - \sum_{i=1}^n x_i$, $-n/\theta^2$, $2n/\theta^3$ and $-6n/\theta^4$, respectively. Hence, Assumptions 5-6 are both satisfied, even around a small neighborhood of θ_0. Assumption 7 is met by a broad class of routinely-implemented priors. For instance, $\text{EXP}(1)$ can be considered in this case. Finally, we need to check Assumption 8. To this end, note that $\{\ell(\theta) - \ell(\theta_0)\}/n = \log \theta/\theta_0 + (\theta_0 - \theta) \sum_{i=1}^n x_i/n$ which, by the law of large number, converges in probability to a negative constant for every fixed θ implying (1.22). Additionally, $\mathbb{E}_0^n\{\ell(\theta) - \ell(\theta_0)\}/n = \log \theta/\theta_0 + (1 - \theta/\theta_0)$ is concave in θ and, therefore, it fulfills Assumption $R1$ of Lemma 1.11. Since $[\{\ell(\theta) - \ell(\theta_0)\} - \mathbb{E}_0^n\{\ell(\theta) - \ell(\theta_0)\}]/n = (\theta_0 - \theta)(\sum_{i=1}^n x_i/n - 1/\theta_0)$ also Assumption $R2$ in Lemma 1.11 is satisfied and, as a consequence, also Assumption 8.

The above derivations ensure that Corollary 1.7 holds. Hence, let us derive the parameters of the skew-symmetric approximating density in (1.3) under this exponential example. To this end, first notice that, since the prior is an $\text{EXP}(1)$, then $\log \pi^{(1)}_{\theta_0} = -1$. Therefore, $\xi = \theta_0^2(n/\theta_0 - \sum_{i=1}^n x_i - 1)/\sqrt{n}$ and $\Omega = 1/(\theta_0^{-2} - 2\theta_0^{-1}\{1/\theta_0 - (\sum_{i=1}^n x_i)/n - 1/n\})$. For what concerns the skewness-inducing factor, we choose $F(\cdot) = \Phi(\cdot)$ which implies a cubic function equal to $\alpha_\eta(h) = \{\sqrt{2\pi}/(6\sqrt{n}\theta_0^3)\}\{(h - \xi)^3 + 3(h - \xi)\xi^2\}$.

Table 1.1 compares the accuracy of the skew-symmetric (s-BvM) and the Gaussian (BvM) approximations corresponding to the newly-derived and classical Bernstein-von Mises theorems, respectively, under growing sample size and replicated experiments. More specifically, we consider 50 different simulated datasets with $\theta_0 = 2$ and sample size $n^* = 1500$. Then, within each of these 50 experiments, we derive the exact posterior under several subsets of data $x_1, \ldots, x_n$ with a growing sample size $n \in \{10, 50, 100, 500, 1000, 1500\}$. The difference between the true posterior and the two approximations is evaluated both in terms of total variation (TV) distances and absolute difference between the posterior mean and the mean of the two approximations. Since the exact posterior and the two approximating densities are available in closed-form, the total variation distances $\text{TV}^n_{\text{BvM}} = (1/2)\int_{\mathbb{R}} |\pi_n(h) - p^n_{\text{GAUSS}}(h)|dh$ and $\text{TV}^n_{\text{s-BvM}} = (1/2)\int_{\mathbb{R}} |\pi_n(h) - p^n_{\text{SKS}}(h)|dh$, as well as, the first moment absolute errors $\text{FMAE}^n_{\text{BvM}} = |\int_{\mathbb{R}} h\{\pi_n(h) - p^n_{\text{GAUSS}}(h)\}dh|$ and $\text{FMAE}^n_{\text{s-BvM}} = |\int_{\mathbb{R}} h\{\pi_n(h) - p^n_{\text{SKS}}(h)\}dh|$ can be evaluated numerically, for each n, via standard routines in `R`.

Table 1.1 displays, for each n, the logarithm of such summary statistics, computed across the 50 replicated experiments. It clarifies that the skew-symmetric approximation consistently yields substantial accuracy improvements relative to the Gaussian counterpart for any n. This empirical finding clarifies that our theoretical results for the limiting case are, in fact, visible also in finite, even small, sample size settings, thus making our theory of direct practical impact, while motivating the adoption of the skew-symmetric approximation in place of the Gaussian one.

1.6.2 Misspecified exponential model

The previous example deals with a well-specified case where the true generating mechanism is indeed included in the parametric family chosen to model the data. Since the results of Corollary 1.7 hold even when the model is misspecified, it is interesting to evaluate the differences between the accuracy of the skew-symmetric approximation and the Gaussian one also in this framework. To this end, let consider the case $X_i \overset{iid}{\sim}$ L-Norm$(-1.5, 1)$, for $i = 1, \ldots, n$, where L-Norm$(-1.5, 1)$ denotes the log-normal distribution with parameters $\mu = -1.5$ and $\sigma = 1$. As in Section 1.6.1, an exponential likelihood is assumed, parameterized by the rate parameter θ, and the prior is Exp(1).

With this specification, the minimizer of the Kullback-Leibler divergence between the log-normal distribution and the family of exponential distributions is unique and equal to $\theta_* \approx 2.71$. Similarly to Section 1.6.1 one can show that the conditions of Corollary 1.7 are satisfied, this time θ_* instead of θ_0 in the present setting.

TABLE 1.2: In the exponential example, average, over 50 replicated studies, of the log-total-variation (TV) distances and first-moment-absolute-errors (FMAE) corresponding to the classical (BvM) and skewed (S-BvM) Bernstein-von Mises theorem in the misspecified simulation setting described in Section 1.6.2. Results are displayed for different sample sizes from $n = 10$ to $n = 1500$. The bold values indicate the best performance for each sample size.

	$n = 10$	$n = 50$	$n = 100$	$n = 500$	$n = 1000$	$n = 1500$
$\log \text{TV}^n_{\text{BvM}}$	-1.28	-2.16	-2.53	-3.28	-3.60	-3.86
$\log \text{TV}^n_{\text{S-BvM}}$	$\mathbf{-2.32}$	$\mathbf{-3.59}$	$\mathbf{-4.17}$	$\mathbf{-4.49}$	$\mathbf{-5.07}$	$\mathbf{-5.36}$
$\log \text{FMAE}^n_{\text{BvM}}$	0.15	-0.81	-1.27	-2.13	-2.18	-2.64
$\log \text{FMAE}^n_{\text{S-BvM}}$	$\mathbf{-0.56}$	$\mathbf{-2.35}$	$\mathbf{-3.30}$	$\mathbf{-5.05}$	$\mathbf{-6.15}$	$\mathbf{-6.80}$

Again, the quality of the two approximations is evaluated through a simulation study consisting of 50 different realizations of the data generation mechanism evaluated at different sample sizes, namely $n \in \{10, 50, 100, 500, 1000, 1500\}$. The results are reported in Table 1.2. As predicted by the theory, p^n_{SKS} is considerably more accurate than p^n_{GAUSS} in terms of both total variation distance and ability of approximating the posterior mean. More specifically, both summary statistics exhibit lower values and a quicker decrease for the skew approximation, even when the sample size is limited.

Chapter 2

Joint and marginal skew-modal approximations

2.1 Introduction

Besides refining classical Bernstein-von Mises type results and yielding improvements over available higher-order theoretical studies, the results of Chapter 1 naturally motivate the development of a novel practical class of skew-modal approximations. This class, which includes both approximations of the joint and marginal posterior distributions, relies mainly on the possibility of replacing, under mild conditions, the unknown quantities depending on the true data-generating mechanism by an estimate evaluated at the posterior mode. These novel plug-in versions are derived in Section 2.2, for the joint posterior approximation, and in Section 2.3, for the marginal ones. In these sections, we also verify that, under mild conditions, the use of the skew-modal approximations leads to an improvement of an order of magnitude compared to the classical Gaussian approximation, both in terms of the total variation distance and in the estimation of polynomially bounded posterior functionals.

2.2 Skew-modal approximation: derivation and theoretical guarantees

Consistent with the above discussion, we consider the plug-in version $\hat{p}^n_{\text{SKS}}$ of p^n_{SKS} in Equation (1.3), where the unknown θ_* is replaced by the MAP $\hat{\theta} = \text{argmax}_{\theta\in\Theta}\{\ell(\theta) + \log \pi(\theta)\}$. As a consequence, this yields the skew-symmetric density, for the rescaled parameter

$\hat{h} = \sqrt{n}(\theta - \hat{\theta}) \in \mathbb{R}^d$, defined as

$$\hat{p}^n_{\text{SKS}}(\hat{h}) \;=\; 2\phi_d(\hat{h}; 0, \hat{\Omega})\hat{w}(\hat{h}), \tag{2.1}$$

where $\hat{\Omega} = (\hat{V}^n)^{-1}$ with $\hat{V}^n = [\hat{v}^n_{st}] = [j_{\hat{\theta},st}/n] \in \mathbb{R}^{d\times d}$, while the skewness-inducing function $\hat{w}(\hat{h}) = F(\hat{\alpha}_\eta(\hat{h}))$ is the combination of $\hat{\alpha}_\eta(\hat{h}) = \{1/(12\eta\sqrt{n})\}a^{(3),n}_{\hat{\theta},stl}\hat{h}_s\hat{h}_t\hat{h}_l \in \mathbb{R}$, with $a^{(3),n}_{\hat{\theta}} = \ell^{(3)}_{\hat{\theta}}/n$, and a cumulative distribution function $F(\cdot)$ satisfying the same conditions of Theorem 1.3. Notice that, relative to the expression for $p^n_{\text{SKS}}(\cdot)$ in Equation (1.3), the location parameter $\hat{\xi}$ is zero in (2.1), since $\hat{\xi}$ is a function of the quantity $(\ell^{(1)}_{\hat{\theta}} + \log \pi^{(1)}_{\hat{\theta}})/\sqrt{n}$ which is zero by definition when $\hat{\theta}$ is the MAP. For the same reason, unlike its population version defined below (1.18) , in the expression for the precision matrix of the Gaussian density factor in (2.1) the additional term including the third order derivative disappears.

Equation (2.1) provides a practical skewed approximation of the exact posterior centered at the mode. As a consequence, such a solution is referred to as skew-modal approximation. In order to provide theoretical guarantees for this practical version, similar to those in Corollary 1.7, let us introduce two mild assumptions in addition to those outlined in Section 1.5.

M1 The MAP estimator $\hat{\theta} = \text{argmax}_{\theta\in\Theta}\{\ell(\theta) + \log \pi(\theta)\}$, satisfies $\mathbb{E}^n_0\|\hat{\theta} - \theta_*\|^2 = O(n^{-1})$.

M2 There exists two positive constants $\bar{\eta}_1, \bar{\eta}_2$ such that the event $\hat{A}_{n,0} = \{\lambda_{\text{MIN}}(\hat{\Omega}^{-1}) > \bar{\eta}_1\} \cap \{\lambda_{\text{MAX}}(\hat{\Omega}^{-1}) < \bar{\eta}_2\}$ holds with probability $P^n_0\hat{A}_{n,0} = 1 - o(1)$. Moreover, there exist two positive constants δ and L such that the inequalities $|\, \ell^{(3)}_{stl}(\theta)/n \,| < L$, $|\, \ell^{(4)}_{stlk}(\theta)/n \,| < L$, and $|\log \pi^{(2)}_{st}(\theta)| < L$ hold uniformly over $\theta \in B_\delta(\hat{\theta}) = \{\theta \in \Theta : \|\hat{\theta} - \theta\| < \delta\}$, with P^n_0-probability tending to one.

Condition M1 is mild and holds generally in regular parametric problems. This assumption ensures us that the MAP is in a small neighborhood of θ_*, where the centering took place in Corollary 1.7. Condition M2 is a weaker version of the analytical assumption for Laplace's method described in Kass *et al.* (1990). Note also that assumption M1 implies that M2 is a stronger version of Assumptions 5-6, requiring the upper bound to hold in a neighborhood of θ_*. These conditions ensure uniform control on the difference between the log-likelihood ratio and its third order Taylor's expansion.

Based on the above additional conditions we provide an asymptotic result for the skew-modal approximation in (2.1), similar to Corollary 1.7. The proof of the theorem follows closely the reasoning and derivations considered to prove Theorem 1.3 and Corollary 1.7.

Theorem 2.1. *Let* $\hat{h} = \sqrt{n}(\theta - \hat{\theta})$ *and* $\hat{K}_n = \{\hat{h} : \|\hat{h}\| < 2M_n\}$. *If Assumptions 1, 7-8, M1 and M2 are fulfilled, then the posterior for* $\hat{h}$ *satisfies*

$$\mathcal{D}_{\text{TV}}\big[\pi_n(\hat{h}) \,||\, \hat{p}^n_{\text{SKS}}(\hat{h})\big] = O_{P_0^n}\big(M_n^{c_8}/n\big), \tag{2.2}$$

for some fixed $c_8 > 0$ *with* $\hat{p}^n_{\text{SKS}}(\hat{h})$ *defined as in (2.1).*

In addition, let $G : \mathbb{R}^d \to R$ *be a function satisfying* $|G(\hat{h})| \lesssim \|\hat{h}\|^r$. *If the prior is such that* $\int \|\hat{h}\|^r \pi(\hat{\theta} + \hat{h}/\sqrt{n}) d\hat{h} < \infty$ *then*

$$\int G(\hat{h})|\pi_n(\hat{h}) - p^n_{\text{SKS}}(\hat{h})|d\hat{h} = O_{P_0^n}(M_n^{c_8+r}/n). \tag{2.3}$$

Proof. The proof of Theorem 2.1 is postponed to Appendix A.2.1 □

Remark 2.2. Since the total variation distance is invariant to scale and location transformations, the above result can be stated also for the original parametrization θ of interest. For the choice in which $F(\cdot) = \Phi(\cdot)$, this implies that the generalized skew-normal density

$$\hat{p}^n_{\text{SKS}}(\theta) = 2\phi_d(\theta; \hat{\theta}, J^{-1}_{\hat{\theta}})\Phi((\sqrt{2\pi}/12)\ell^{(3)}_{\hat{\theta},stl}(\theta - \hat{\theta})_s(\theta - \hat{\theta})_t(\theta - \hat{\theta})_l), \tag{2.4}$$

approximates the posterior density for θ with the rate derived in Theorem 2.1.

The skew-modal approximation in (2.4) provides, therefore, a similarly tractable, yet substantially more accurate, alternative to the classical Gaussian counterpart arising from the Laplace method. As discussed in Section 1.4, the closed-form density in (2.4) can be evaluated at a similar computational cost as the Gaussian density, when d is not too big, and further admits a straightforward i.i.d. sampling scheme that facilitates Monte Carlo estimation of any functional of interest. Such a scheme mainly relays on sampling from a d-variate Gaussian and, hence, can be implemented via standard `R` packages for simulating from these variables.

An open question is whether similar approximations can be obtained for the posterior marginal distributions. In the next section, we provide a positive answer by deriving an alternative approximation that allows us to focus directly on the marginal distributions of the posterior, providing for them skew-symmetric approximations that maintain the same level of asymptotic accuracy as the joint approximation described in (2.4).

2.3 Marginal skew-modal approximations

The skew-modal approximation developed in Section 2.1 targets the joint posterior distribution. In practical applications, marginal distributions and related functionals are often the final scope of the analysis. Usually, these quantities are not directly available from the joint approximation in (2.1). For this reason, it is necessary to rely on simulation methods by exploiting the stochastic representation of skew-symmetric random variables introduced in Section 1.3, which allows to draw i.i.d. observations from (2.1) simply by perturbing a sample obtained by a Gaussian distribution. This procedure requires multiple evaluations of the skewness-inducing factor $\hat{w}(\hat{h})$, an operation that is cubic in d. Even though the computational cost is in many case negligible, when the dimension of θ is big, it would be attractive to obtain directly a skewed, closed-form, expression for the marginal densities. In the following, we show that a precise use of asymptotic arguments leads to closed-form expressions for the marginal posterior distributions.

2.3.1 Derivation of a marginal skew-approximation

We now give a constructive derivation of the skew-symmetric approximation for marginal distributions. Let $\mathcal{C} \subseteq \{1, \ldots, d\}$ be a set containing the indexes for the elements of θ in which we are interested in, let $d_{\mathcal{C}}$ be its cardinality and use the notation $\bar{\mathcal{C}} = \mathcal{C}^c$ for the complement. We write $h = (h_{\mathcal{C}}, h_{\bar{\mathcal{C}}})$, accordingly, the matrix $\hat{\Omega} = (J_{\hat{\theta}}/n)^{-1}$ is partitioned as

$$\hat{\Omega} = \begin{bmatrix} \hat{\Omega}_{\mathcal{C}\mathcal{C}} & \hat{\Omega}_{\mathcal{C}\bar{\mathcal{C}}} \\ \hat{\Omega}_{\bar{\mathcal{C}}\mathcal{C}} & \hat{\Omega}_{\bar{\mathcal{C}}\bar{\mathcal{C}}} \end{bmatrix}.$$

Under the regularity conditions stated in Section 2.1, it is possible to write, for $n \to \infty$,

$$\pi_n(\hat{\theta} + \hat{h}/\sqrt{n}) \propto \exp\Big(- j_{\hat{\theta},st}\hat{h}_s\hat{h}_t/(2n) + a^{(3),n}_{\hat{\theta},stl}\hat{h}_s\hat{h}_t\hat{h}_l/(6\sqrt{n})\Big) + O_{P^n_0}(n^{-1}).$$

with $a^{(3),n}_{\hat{\theta}} = \ell^{(3)}_{\hat{\theta}}/n$. The second order term is proportional to a Gaussian kernel and it can be decomposed as follows: $\exp(-j_{\hat{\theta},st}\hat{h}_s\hat{h}_t/(2n)) \propto \phi_d(\hat{h}; 0, \hat{\Omega}) = \phi_{d_{\mathcal{C}}}(\hat{h}_{\mathcal{C}}; 0, \hat{\Omega}_{\mathcal{C}\mathcal{C}}) \times \phi_{d-d_{\mathcal{C}}}(\hat{h}_{\bar{\mathcal{C}}}; \Lambda_{\mathcal{C}}\hat{h}_{\mathcal{C}}, \bar{\Omega})$ where $\Lambda_{\mathcal{C}} = \hat{\Omega}_{\bar{\mathcal{C}}\mathcal{C}}\hat{\Omega}^{-1}_{\mathcal{C}\mathcal{C}}$ is a $(d - d_{\mathcal{C}}) \times d_{\mathcal{C}}$ matrix and $\bar{\Omega} = \hat{\Omega}_{\bar{\mathcal{C}}\bar{\mathcal{C}}} - \hat{\Omega}_{\bar{\mathcal{C}}\mathcal{C}}\hat{\Omega}^{-1}_{\mathcal{C}\mathcal{C}}\hat{\Omega}_{\mathcal{C}\bar{\mathcal{C}}}$ has dimension $(d - d_{\mathcal{C}}) \times (d - d_{\mathcal{C}})$.

To obtain a marginal skew-symmetric approximation, we use again that the third order term converges to zero in probability and that $e^x = 1 + x + O(x^2)$, for $x \to 0$. An

approximation on the marginal distribution is therefore proportional to

$$\int \phi_{d_{\mathcal{C}}}(h_{\mathcal{C}}; 0, \hat{\Omega}_{\mathcal{C}\mathcal{C}})\phi_{d-d_{\mathcal{C}}}(h_{\bar{\mathcal{C}}}; \Lambda_{\mathcal{C}} h_{\mathcal{C}}, \bar{\Omega})(1 + a^{(3),n}_{\hat{\theta},stl}\hat{h}_s\hat{h}_t\hat{h}_l/(6\sqrt{n}))dh_{\bar{\mathcal{C}}}. \tag{2.5}$$

To determine the value of the above integral, first note that the third order term in (2.5) can be decomposed as

$$a^{(3),n}_{\hat{\theta},stl}\hat{h}_s\hat{h}_t\hat{h}_l + 3a^{(3),n}_{\hat{\theta},str}\hat{h}_s\hat{h}_t\hat{h}_r + 3a^{(3),n}_{\hat{\theta},srv}\hat{h}_s\hat{h}_r\hat{h}_v + a^{(3),n}_{\hat{\theta},rvk}\hat{h}_r\hat{h}_v\hat{h}_k, \tag{2.6}$$

where $s, t, l \in \mathcal{C}$ and $r, v, k \in \bar{\mathcal{C}}$. The first term in (2.6) is not affected by the integral. We deal with the others one by one, separately. Simple calculations give

$$3\mathbb{E}_{\hat{h}_{\bar{\mathcal{C}}}|\hat{h}_{\mathcal{C}}}\left(a^{(3),n}_{\hat{\theta},str}\hat{h}_s\hat{h}_t\hat{h}_r\right) = \upsilon^{(1),n}_{3,stl}\hat{h}_s\hat{h}_t\hat{h}_l, \tag{2.7}$$

where $\mathbb{E}_{\hat{h}_{\bar{\mathcal{C}}}|\hat{h}_{\mathcal{C}}}$ denotes the expectation with respect to $\phi_{d-d_{\mathcal{C}}}(\hat{h}_{\bar{\mathcal{C}}}; \Lambda_{\mathcal{C}}\hat{h}_{\mathcal{C}}, \bar{\Omega})$ and $\upsilon^{(1)}_3$ is a $d^3_{\mathcal{C}}$-dimensional array with entries $\upsilon^{(1)}_{3,stl} = 3a^{(3)}_{\hat{\theta},str}\Lambda_{\mathcal{C},rl}$ with $s, t, l \in \mathcal{C}$ and $r \in \bar{\mathcal{C}}$. Similarly,

$$3\mathbb{E}_{\hat{h}_{\bar{\mathcal{C}}}|\hat{h}_{\mathcal{C}}}\left(a^{(3),n}_{\hat{\theta},srv}\hat{h}_s\hat{h}_r\hat{h}_v\right) = \upsilon^{(2),n}_{1,s}\hat{h}_s + \upsilon^{(2),n}_{3,stl}\hat{h}_s\hat{h}_t\hat{h}_l, \tag{2.8}$$

and

$$\mathbb{E}_{\hat{h}_{\bar{\mathcal{C}}}|\hat{h}_{\mathcal{C}}}\left(a^{(3),n}_{\hat{\theta},rvk}\hat{h}_r\hat{h}_v\hat{h}_k\right) = \upsilon^{(3),n}_{1,s}\hat{h}_s + \upsilon^{(3),n}_{3,stl}\hat{h}_s\hat{h}_t\hat{h}_l, \tag{2.9}$$

where $\upsilon^{(2),n}_{1,s} = 3a^{(3),n}_{\hat{\theta},srv}\bar{\Omega}_{rv}$ and $\upsilon^{(3),n}_{1,s} = 3a^{(3),n}_{\hat{\theta},rvk}\bar{\Omega}_{rv}\Lambda_{\mathcal{C},ks}$ are $d_{\mathcal{C}}$ dimensional vectors while $\upsilon^{(2),n}_{3,stl} = 3a^{(3),n}_{\hat{\theta},srv}\Lambda_{\mathcal{C},rt}\Lambda_{\mathcal{C},vl}$ and $\upsilon^{(3),n}_{3,stl} = a^{(3),n}_{\hat{\theta},rvk}\Lambda_{\mathcal{C},rs}\Lambda_{\mathcal{C},vt}\Lambda_{\mathcal{C},kl}$ are $d^3_{\mathcal{C}}$-dimensional arrays for $s, t, l \in \mathcal{C}$ and $r, v, k \in \bar{\mathcal{C}}$. Note that averaging over the conditional distribution removes all the components which depends on odds powers of the elements in $\bar{\mathcal{C}}$.

By combining (2.7), (2.8) and (2.9) we obtain the following expression for the integral in (2.5)

$$2\phi_{d_{\mathcal{C}}}(\hat{h}_{\mathcal{C}}; 0, \hat{\Omega}_{\mathcal{C}\mathcal{C}})(1/2 + \eta\alpha_{\mathcal{C}}(\hat{h}_{\mathcal{C}})), \tag{2.10}$$

where

$$\alpha_{\mathcal{C}}(\hat{h}_{\mathcal{C}}) = (\nu^n_{1,s}\hat{h}_s + \nu^n_{3,stl}\hat{h}_s\hat{h}_t\hat{h}_l)/(12\eta\sqrt{n}),$$

with $\nu^n_{1,s} = \upsilon^{(2),n}_{1,s} + \upsilon^{(3),n}_{1,s}$ and $\nu^n_{3,stl} = a^{(3),n}_{stl} + \upsilon^{(1),n}_{3,stl} + \upsilon^{(2),n}_{3,stl} + \upsilon^{(3),n}_{3,stl}$ for $s, t, l \in \mathcal{C}$. Note, also, that $\alpha_{\mathcal{C}}(\hat{h}_{\mathcal{C}})$ is an odd polynomial function of $\hat{h}_{\mathcal{C}}$ and that

$$\alpha_{\mathcal{C}}(\hat{h}_{\mathcal{C}}) = \mathbb{E}_{\hat{h}_{\bar{\mathcal{C}}}|\hat{h}_{\mathcal{C}}}\big(\hat{\alpha}_\eta(\hat{h})\big). \tag{2.11}$$

Then (2.10) can be reformulated as

$$2\phi_{d_{\mathcal{C}}}(\hat{h}_{\mathcal{C}};0,\hat{\Omega}_{\mathcal{C}\mathcal{C}})(1/2+\eta\alpha_{\mathcal{C}}(\hat{h})) = 2\phi_{d_{\mathcal{C}}}(\hat{h}_{\mathcal{C}};0,\hat{\Omega}_{\mathcal{C}\mathcal{C}})F(\alpha_{\mathcal{C}}(\hat{h}_{\mathcal{C}})) + O_{P_0^n}(n^{-1}),$$

where $\eta \in \mathbb{R}$ and $F(\cdot)$ is the cdf of a univariate random variable on $\mathbb{R}$ satisfying $F(-x) = 1 - F(x)$, $F(0) = 1/2$ and $F(x) = F(0) + \eta x + O(x^2)$. As a consequence, the posterior marginal density of $\hat{h}_{\mathcal{C}}$ can be approximated by the following skew-symmetric density

$$\hat{p}^n_{\text{SKS},\mathcal{C}} = 2\phi_{d_{\mathcal{C}}}(\hat{h}_{\mathcal{C}};0,\hat{\Omega}_{\mathcal{C}\mathcal{C}})w_{\mathcal{C}}(\hat{h}_{\mathcal{C}}), \tag{2.12}$$

where $w_{\mathcal{C}}(\hat{h}_{\mathcal{C}}) = F(\alpha_{\mathcal{C}}(\hat{h}_{\mathcal{C}})) = F(\mathbb{E}_{\hat{h}_{\bar{\mathcal{C}}}|\hat{h}_{\mathcal{C}}}\{\hat{\alpha}_\eta(\hat{h})\})$.

We briefly describe the salient features of (2.12) before formalizing its theoretical properties. The symmetric component is obtained by marginalizing the Gaussian approximation derived from the Laplace method and it is therefore very easy to obtain. The operation with the highest computational cost is the first evaluation of the parameters in $w_{\mathcal{C}}(\hat{h}_{\mathcal{C}})$, which involves summations over the three-dimensional array $a^{(3),n}_{\hat{\theta}}$. However, once this step is completed, each evaluation of (2.12) is cubic in $d_{\mathcal{C}}$. Since one is usually interested in the univariate or bivariate marginal distribution, the computational gain over the sampling from the joint approximation (2.1) can be substantial, also because functionals of interest, such as the posterior mean and variance, can be evaluated by standard low-dimensional numerical integration methods.

In the next section we prove that this marginal approximation has the same accuracy as the joint skew-symmetric approximation discussed in Theorem 2.1.

2.3.2 Theoretical guarantees for the marginal approximation

This section provides a rigorous justification for the use of the skew-symmetric marginal approximation derived above. To this end, we denote the marginal posterior density for $\hat{h}_{\mathcal{C}}$ with $\pi_{n,\mathcal{C}}(\hat{h}_{\mathcal{C}}) = \int \pi_n(\hat{h})d\hat{h}_{\bar{\mathcal{C}}}$. The following theorem shows that, under the same assumptions as in Theorem 2.1, the total variation distance between $\pi_{n,\mathcal{C}}(\hat{h}_{\mathcal{C}})$ and $\hat{p}^n_{\text{SKS},\mathcal{C}}$ is, up to a logarithmic factor, of order n^{-1} in P_0^n-probability.

Theorem 2.3. *Under the assumptions of Theorem 2.1*

$$\mathcal{D}_{\text{TV}}\left[\pi_{n,\mathcal{C}}(\hat{h}_{\mathcal{C}}) \,||\, \hat{p}^n_{\text{SKS},\mathcal{C}}(\hat{h}_{\mathcal{C}})\right] = O_{P_0^n}\left(M_n^{c_9}/n\right), \tag{2.13}$$

for some $c_9 > 0$ with $\hat{p}^n_{\text{SKS},\mathcal{C}}(\hat{h})$ defined as in (2.12).

Remark 2.4. Similarly to Remark 2.2, in view of the invariance to scale and location transformation of the total variation distance, the above results hold with the original

parametrization θ as well. This implies that by taking $F(\cdot) = \Phi(\cdot)$ the marginal posterior density is estimated with the marginal generalized skew–normal density

$$\hat{p}^n_{\text{SKS},\mathcal{C}}(\theta_{\mathcal{C}}) = 2\phi_{d_{\mathcal{C}}}(\theta_{\mathcal{C}}; \hat{\theta}_{\mathcal{C}}, J^{-1}_{\hat{\theta},\mathcal{C}\mathcal{C}})\Phi\Big(\frac{\sqrt{2\pi}}{12}\{\nu^n_{1,s}(\theta-\hat{\theta})_s + n\nu^n_{3,stl}(\theta-\hat{\theta})_s(\theta-\hat{\theta})_t(\theta-\hat{\theta})_l\}\Big), \quad (2.14)$$

with rate given in Theorem 2.3, $s, t, l \in \mathcal{C}$ and $\nu^n_{1,s}$,$\nu^n_{3,stl}$ defined in (2.10).

2.4 Empirical results joint and marginal approximations

Sections 2.4.1-2.4.2 demonstrate the practical applicability of the joint and marginal skew-modal approximations derived in Sections 2.1-2.3 on both synthetic datasets and a real-data application on Cushing's syndrome (Venables and Ripley, 2002). These empirical analyses provide consistent evidence of substantial accuracy improvements relative to the Gaussian approximation induced by the Laplace method. The comparison between the skew-modal approach and other deterministic approximations, such as mean-field variational Bayes (e.g., Blei *et al.*, 2017) and expectation-propagation (e.g., Minka, 2001; Vehtari *et al.*, 2020) is also discussed. In the following we take $F(\cdot) = \Phi(\cdot)$.

2.4.1 Exponential model revisited

Let us first replicate the simulation studies regarding the exponential model under both the well-specified and the misspecified setting as described in Sections 1.6.1-1.6.2. This time the focus is on the plug-in skew-modal approximation in (2.4), rather than its population version which assumes knowledge of θ_*. Consistent with this focus, the performance of the skew-modal approximations is compared against the Gaussian approximation $\text{N}(\hat{\theta}, J^{-1}_{\hat{\theta}})$ arising from the Laplace method (e.g., Gelman *et al.*, 2013, pg. 318). Note that both the well-specified and the misspecified models satisfy the additional assumptions M1-M2 required by Theorem 2.1 and Remark 2.2. In fact, $\hat{\theta}$ is asymptotically equivalent to the maximum likelihood estimator which implies that condition M1 is fulfilled. Moreover, in view of the expressions for the first three log-likelihood derivatives in Section 1.6.1 also M2 is satisfied.

Table 2.1 reports the same summaries as Tables 1.1-1.2, but now the two total variation distances from the exact posterior and the first-moment absolute errors are computed with respect to the skew-modal approximation in (2.4) and the Gaussian

TABLE 2.1: We consider both the well and misspecified exponential examples introduced in Sections 1.6.1- 1.6.2, respectively. In both cases the average (over 50 replicates) of the log–total–variation (TV) distances and first–moment–absolute–errors (FMAE) are reported, both for the Gaussian modal (GM) and the newly–developed skew–modal (SKEW–M) approximations. The sample size ranges from n=10 to n=1500. The best performing method in each experiment is highlighted in bold.

	$n = 10$	$n = 50$	$n = 100$	$n = 500$	$n = 1000$	$n = 1500$
WELL-SPECIFIED						
$\log \text{TV}^n_{\text{BvM}}$	-2.48	-3.28	-3.63	-4.43	-4.78	-4.98
$\log \text{TV}^n_{\text{s-BvM}}$	$\mathbf{-3.71}$	$\mathbf{-5.33}$	$\mathbf{-6.03}$	$\mathbf{-7.65}$	$\mathbf{-8.34}$	$\mathbf{-8.74}$
$\log \text{FMAE}^n_{\text{BvM}}$	-0.61	-1.30	-1.63	-2.41	-2.76	-2.96
$\log \text{FMAE}^n_{\text{s-BvM}}$	$\mathbf{-1.91}$	$\mathbf{-3.52}$	$\mathbf{-4.35}$	$\mathbf{-6.50}$	$\mathbf{-7.50}$	$\mathbf{-8.09}$
MISSPECIFIED						
$\log \text{TV}^n_{\text{BvM}}$	-2.48	-3.28	-3.63	-4.43	-4.78	-4.98
$\log \text{TV}^n_{\text{s-BvM}}$	$\mathbf{-3.71}$	$\mathbf{-5.33}$	$\mathbf{-6.03}$	$\mathbf{-7.65}$	$\mathbf{-8.35}$	$\mathbf{-8.75}$
$\log \text{FMAE}^n_{\text{BvM}}$	-0.41	-1.05	-1.36	-2.12	-2.46	-2.66
$\log \text{FMAE}^n_{\text{s-BvM}}$	$\mathbf{-1.71}$	$\mathbf{-3.28}$	$\mathbf{-4.08}$	$\mathbf{-6.21}$	$\mathbf{-7.21}$	$\mathbf{-7.79}$

$\text{N}(\hat{\theta}, J^{-1}_{\hat{\theta}})$, respectively. Similarly to the population versions assessed in Sections 1.6.1-1.6.2, also the practical skew-modal approximations provide higher accuracy compared to their Gaussian counterpart both in the well specified and misspecified frameworks. This confirms that, also in this context, the asymptotic theory in Theorem 2.1 and Remark 2.2 closely matches the empirical behavior observed in practice even for finite, possibly small, sample sizes.

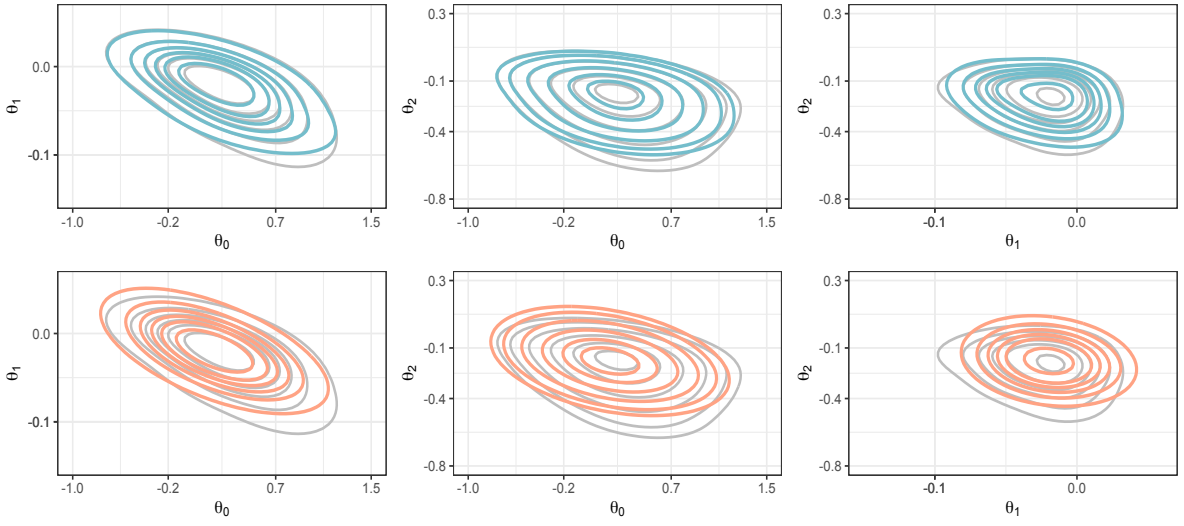

FIGURE 2.1: Visual comparison between skew-modal (blue) and Gaussian (orange) approximations of the exact bivariate posteriors (grey) for the three coefficients of the probit regression model in the Cushings application.

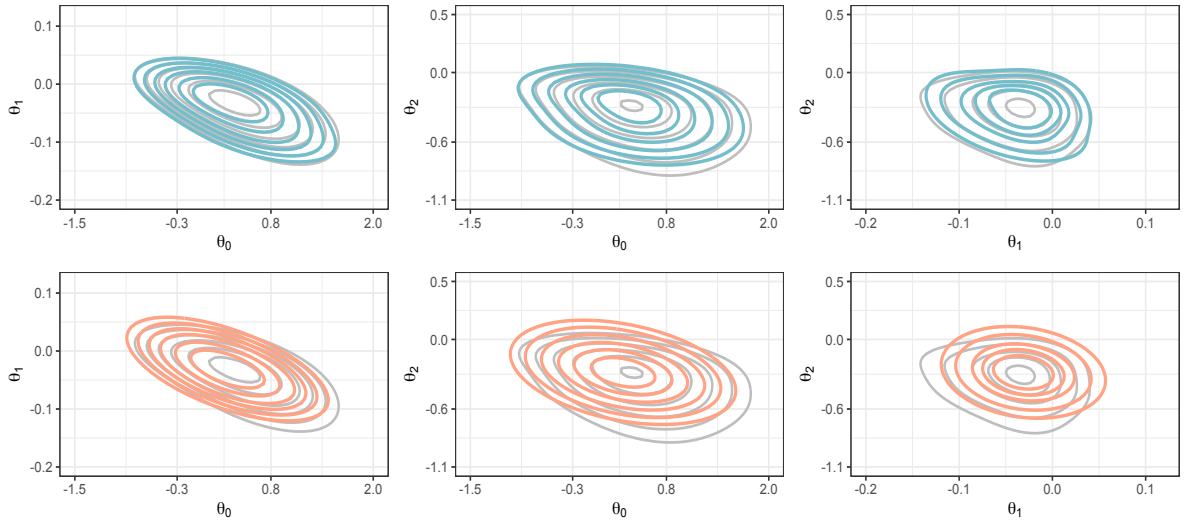

FIGURE 2.2: Visual comparison between skew-modal (blue) and Gaussian (orange) approximations of the exact bivariate posteriors (grey) for the three coefficients of the logistic regression model in the Cushings application.

2.4.2 Probit and logistic regression model

We now consider a real-data application on the Cushings dataset (Venables and Ripley, 2002), which is openly-available in the `R` library `Mass`. In this case the true data-generative model is not known and, therefore, this analysis is useful to evaluate performance in situations which are not guaranteed to meet the assumptions underlying our theory in Sections 2.1 and 2.3.

The data are obtained from a medical study on $n = 27$ subjects, aimed at investigating the relationship between four different sub-types of Cushing's syndrome and two steroid metabolites, *Tetrahydrocortisone* and *Pregnanetriol* respectively. To simplify the analysis, we consider the binary response $X_i \in \{0, 1\}$, $i = 1, \dots, n$ taking value 1 if the patient is affected by bilateral hyperplasia, and 0 otherwise, for $i = 1 \dots, n$. The observed covariates are $z_{i1} =$ *"urinary excretion rate (mg/24hr) of Tetrahydrocortisone for patient i"* and $z_{i2} =$ *"urinary excretion rate (mg/24hr) of Pregnanetriol for patient i"*. In the following, we focus on the two most widely-implemented regression models for binary data, namely:

Probit regression. $X_i \sim \text{Bern}(\Phi(\theta_0 + \theta_1 z_{i1} + \theta_2 z_{i2}))$, independently for $i = 1, \dots, n$, with regression coefficients $(\theta_0, \theta_1, \theta_2) = \theta \in \mathbb{R}^3$.

Logistic regression. $X_i \sim \text{Bern}(g(\theta_0 + \theta_1 z_{i1} + \theta_2 z_{i2}))$, independently for $i = 1, \dots, n$, with $(\theta_0, \theta_1, \theta_2) = \theta \in \mathbb{R}^3$, and $g(\cdot)$ the inverse logit function defined in Remark 1.5.

Under both models, Bayesian inference proceeds via routinely-implemented weakly

TABLE 2.2: For probit and logistic regression, estimated joint, bivariate and marginal total variation distances between the exact posterior and both the SKEW-M and the GM approximations under analysis in the Cushings application. The bold values indicate the best performance for each subset of parameters.

	TV_{θ}	$\text{TV}_{\theta_{01}}$	$\text{TV}_{\theta_{02}}$	$\text{TV}_{\theta_{12}}$	TV_{θ_0}	TV_{θ_1}	TV_{θ_2}
Probit							
SKEW-M	**0.11**	**0.05**	**0.06**	**0.09**	**0.03**	**0.04**	**0.05**
GM	0.19	0.10	0.13	0.18	0.09	0.08	0.11
Logit							
SKEW-M	**0.14**	**0.08**	**0.10**	**0.13**	**0.05**	**0.06**	**0.07**
GM	0.23	0.13	0.17	0.22	0.11	0.10	0.14

informative Gaussian prior $\text{N}(0, 25)$ for the three regression coefficients. Recalling Durante (2019), such a choice yields a closed-form unified skew-normal (Arellano-Valle and Azzalini, 2006) posterior for θ under probit regression, which admits tractable i.i.d. sampling schemes. Extending the results in Durante (2019), a recent contribution by Onorati and Liseo (2022) has proved that similar findings, for an extension of the unified skew-normal, can be obtained also under logistic models. Although yielding important advancements in the field of Bayesian inference for binary regression models, the derived closed-form posteriors require the evaluation of n-dimensional Gaussian cumulative distribution functions. As a consequence, inference under the exact posterior becomes intractable as n grows. This motivates the adoption of deterministic approximations in order to perform Bayesian inference while reducing the computational burden.

Both the joint and the marginal skew modal approximations (SKE-M), as well as, the classical Gaussian modal approximation (GM) can be readily derived from the closed-form derivatives of the log-likelihood and log-prior for both the probit and logistic regression models. Moreover, since the prior is Gaussian, the MAP under both models coincides with the ridge-regression estimator and hence can be computed via basic `R` functions. Table 2.2 displays the Monte Carlo estimates of the TV distance from the exact posteriors of θ for both GM and SKE-M approximations, under probit and logistic regression. Additionally, Table 2.3 reports the L_1-distance between the true and the approximated posterior means (BIAS), as well as, the average absolute error resulting in using GM and SKE-M, in place of π_n, to evaluate the expected value of the posterior probabilities of the model, in both probit and logistic regression, respectively. In the probit model, for a generic approximating density $\hat{p}^n_{\text{APP}}$, this last quantity takes the form AVE-PR $= \sum_{i=1}^n |\text{pr}_i - \hat{\text{pr}}_{\text{APP},i}|/n$ with $\text{pr}_i = \int \Phi(\theta_0 + \theta_1 z_{i1} + \theta_2 z_{i2})\pi_n(\theta)d\theta$, and $\hat{\text{pr}}_{\text{APP},i} = \int \Phi(\theta_0 + \theta_1 z_{i1} + \theta_2 z_{i2})\hat{p}^n_{\text{APP}}(\theta)d\theta$, for $i = 1, \ldots, n$. The logistic case follows by replacing $\Phi(\cdot)$ with $g(\cdot)$. In both models, the proposed SKEW-M solutions generally yield remarkable accuracy improvements relative to GM. More specifically, SKEW-M almost

TABLE 2.3: For probit and logistic regression, posterior mean bias and AVE-PR error obtained when replacing the exact posterior distribution with both SKEW-M and GM approximations under analysis in the Cushings application. The bold values indicate the best performance for each subset of parameters.

	BIAS_{θ_0}	BIAS_{θ_1}	BIAS_{θ_2}	AVE-PR
Probit				
SKEW-M	**0.004**	**0.002**	**0.015**	**0.006**
GM	−0.092	0.008	0.051	0.026
Logit				
SKEW-M	**0.069**	**−0.001**	**−0.008**	**0.009**
GM	−0.116	0.010	0.060	0.064

halves, on average, the TV distance associated with GM, while providing a much more accurate approximation for the posterior mean and posterior probabilities. This better fit is evident in Figures 2.1-2.2, which visually compare the approximation accuracy of the newly-developed bivariate marginal skew-modal approximations to the aforementioned GM.

We conclude this section by discussing the behavior of the SKEW-M approximation when compared to other, advanced, techniques within the class of deterministic approximation for binary regression models. State-of-the-art methods under this framework are mean-field variational Bayes (MF-VB) (Consonni and Marin, 2007; Durante and Rigon, 2019) and expectation-propagation (EP) (Chopin and Ridgway, 2017), while partially-factorized variational Bayes (PFM-VB) (Fasano *et al.*, 2022) is available only for probit regression. Table B.1 in the Appendix highlight how the use of SKEW-M yields noticeable improvements relative to MF-VB and PFM-VB. The gains over PFM-VB in the probit model are remarkable since also such a strategy leverages a skewed approximation of the exact posterior distribution. This yields higher accuracy than MF-VB, but the improvements are not as noticeable as SKEW-M. A reason for this result is that PFM-VB has been originally developed to provide high accuracy in high-dimensional $p > n$ settings (Fasano *et al.*, 2022), whereas in this study $p = 3$ and $n = 27$. Finally, SKEW-M also performs comparably well to state-of-the-art EP methods in terms of TV distances from the true posterior. This fact is noteworthy for at least two reasons. First, Gaussian EP methods are known to approximate the first two moments of the target density in a precise way, which, being a global characteristic of the density, usually leads to approximations that, even though symmetric, are difficult to improve. On the contrary, our SKEW-M is based on the simple GM, an approximation that only interpolates the local behavior of the posterior distribution in a neighborhood of its mode. It is therefore interesting to note that properly accounting for skewness can dramatically improve the global quality of an approximation, suggesting that a similar correction could also be

useful to refine more advanced methods. Second, EP techniques typically rely on a convenient factorization of the target density (see e.g., Gelman *et al.*, 2013, pg. 338), a condition that is not required for the adoption of SKEW-M, making the latter applicable to a wider range of approximation problems.

2.4.3 High–dimensional logistic regression

We conclude our numerical analysis with another real-world binary regression problem that is particularly well suited for evaluating the performance of the marginal skew-modal approximation introduced in Section 2.3. The data, which are available in the `R` package `AppliedPredictiveModeling` (Kuhn and Johnson, 2018), are obtained from a clinical study on $n = 333$ subjects designed to investigate whether biological measurements from cerebrospinal fluid can be used to diagnose Alzheimer's disease (Craig-Schapiro *et al.*, 2011). Specifically, 130 explanatory variables are collected along with the response variable $X_i \in \{0, 1\}$, $i = 1, \dots, n$, which takes the value 1 if the patient is affected by Alzheimer's disease and 0 otherwise.

For the inference, we assume a logistic regression with independent Gaussian priors $\mathrm{N}(0, 4)$ on the coefficients. Note that the inclusion of the intercept and the presence of a categorical variable with 6 different levels imply that the number of parameters of the model is $d = 135$. As a consequence, although the sample size is not small in absolute terms, since $n/d \approx 2.46$, the behavior of the posterior distribution in this example is not necessarily well described by the asymptotic theory developed in Sections 2.2 and 2.3.2.

The posterior distribution is obtained via 7 chains of length 10 000 of Hamiltonian Monte Carlo using the `R` function `stan_glm` from the `rstanarm` package (Goodrich *et al.*, 2023). In the following, we focus on the marginal posterior distributions and compare the marginal univariate skew-modal approximation introduced in Section 2.3 with the corresponding Gaussian derived from the Laplace method (Gelman *et al.*, 2013, pg. 318). The comparison is made in terms of both the absolute difference between the exact posterior mean and its approximation (BIAS), and the total variation distances between each marginal posterior density and its approximation (TV). Table 2.4 reports both the mean and the median of these quantities for the skew-modal and Gaussian marginal approximations, respectively. It is clear that the skew-modal approximation outperforms the Gaussian for both quantities under investigation. Considering that the 95% of the posterior means are between -2.68 and 2.66, the improvement in terms of bias is particularly relevant.

TABLE 2.4: For the logistic regression described in Section 2.4.3, mean and median BIAS and total variation distance for both the marginal SKEW-M and the GM approximations. The bold values indicate the best performance.

	mean-BIAS	median-BIAS	mean-TV	median-TV
GM	0.425	0.347	0.145	0.120
SKEW-M	**0.139**	**0.068**	**0.104**	**0.078**

The results of this section provide further empirical evidence that our methodological proposal tends to perform well not only in settings where the parameter dimension is moderate and the departure from Gaussianity is mainly due to the small sample size, but also in frameworks where both the parameter dimension and the number of observations are not small.

Chapter 3

General skew-symmetric approximations

3.1 Introduction

The skew-symmetric approximations introduced in Chapters 1-2 are derived under asymptotic arguments. More specifically, both the perturbed Gaussian density and the skewness-inducing factor are obtained by manipulating higher-order Taylor expansions of the log-likelihood and of the log-prior. Interestingly, the symmetric component of the skew-modal approximation studied in Chapter 2, which is a Gaussian distribution with mean at the posterior mode and covariance matrix the inverse of the observed information, is closely related to the classical Gaussian distribution obtained from the Laplace method (e.g., Gelman *et al.*, 2013, Ch 13). In this regard, an important question is whether it is possible to replace such a Gaussian density by one of the many symmetric deterministic approximations available in the literature and, if so, what form the skewness-inducing factor should assume.

In this chapter, we give an answer to these questions by developing, in Section 3.3, a method which, starting from any approximation $f^*_{\tilde{\theta}}(\theta)$ of the posterior distribution that is symmetric about $\tilde{\theta} \in \mathbb{R}^d$, it provides a similarly tractable, yet provably more accurate, skew-symmetric approximating density $q^*_{\tilde{\theta}}(\theta) = 2f^*_{\tilde{\theta}}(\theta)w^*_{\tilde{\theta}}(\theta)$. Although this novel strategy is guaranteed to improve the accuracy of any symmetric approximation, in practice it is natural to perturb those obtained as an output of routinely-implemented Laplace, EP and Gaussian VB approximations which we briefly review in Section 3.2.

3.2 A brief overview of symmetric approximation of posterior distributions

When seeking a tractable symmetric approximation of a generic posterior distribution for the parameter $\theta \in \Theta \subseteq \mathbb{R}^d$, a simple option is to consider the classical Gaussian approximation arising from the Laplace method (e.g., Bishop, 2006, Ch. 4.4). Such a solution follows directly from a second-order Taylor expansion of the unnormalized log-posterior $\log[\pi(\theta)L(\theta; X^n)]$ at the maximum a posteriori (MAP), which yields a Gaussian approximating density centered at the MAP and with variance and covariance given by the inverse of the negative Hessian for $\log[\pi(\theta)L(\theta; X^n)]$, again evaluated at such MAP. Such a strategy provides, therefore, a simple approach which only requires estimation of the MAP via standard optimization schemes (see e.g., Gelman *et al.*, 2013, Ch. 13). This has stimulated broad applicability and several subsequent extensions, including, among others, integrated nested Laplace approximation (INLA) (Rue *et al.*, 2009) and approximate Laplace approximation (ALA) (Rossell *et al.*, 2021). INLA provides an effective framework which combines efficient numerical integration strategies and analytical approximations to yield accurate characterizations of posterior marginals for parameters of interest, under latent Gaussian models. Such a solution is inspired by Tierney and Kadane (1986) Laplace approximation of the marginal posterior density for a subset of parameters of interest which expresses such a marginal as proportional to the ratio between the joint posterior and the full conditional density of the remaining parameters, and then applies the Gaussian approximation from the Laplace method to such a latter density. Although this strategy can yield non-symmetric approximations of marginal posterior densities, it still relies on nested Gaussian approximations. Therefore, the closed-form skew-symmetric approximation we derive in Section 3.3 has potentials to further improve also INLA accuracy by replacing the commonly-used Gaussian approximation for the full conditionals, with the proposed skew-symmetric one. Our novel perturbation strategy can be also applied directly to improve the overall accuracy of the symmetric approximation underlying ALA. In fact, such a scalable procedure does not differ from the classical Laplace method in the shape of the approximating density, but rather in providing a computationally-cheaper and theoretically-supported strategy which avoids MAP estimation when locating the symmetric approximation.

While Laplace-type schemes provide simple strategies to approximate intractable posterior densities, recalling e.g. Bishop (2006, Ch. 4.4), these solutions arise from a Taylor expansion of the log-posterior at a given point, and thus fail to incorporate global properties beyond the local behavior around such a point. This problem has motivated

alternative approximation strategies, with an overarching focus on VB (e.g., Blei *et al.*, 2017) and EP (e.g., Vehtari *et al.*, 2020).

VB specifies a family $\mathcal{F}$ of tractable approximating densities and then identifies, within such a family, the one that is closest to the intractable posterior under the Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951) between the approximating density and such a posterior. Common practice in specifying $\mathcal{F}$ relies either on parametric families, often imposing a Gaussian approximation (see e.g., Opper and Archambeau, 2009; Challis and Barber, 2013; Tan and Nott, 2018), or on mean-field assumptions (e.g., Blei *et al.*, 2017), forcing the joint approximation density for θ to factorize as a product of marginals for suitably selected non-overlapping subsets of parameters. In the first case, the final output of the optimization problem is, by definition, a symmetric density, which can be readily improved under our proposed skew-symmetric approximation. Conversely, the second case may yield skewed solutions. In fact, as shown in e.g, Blei *et al.* (2017), the optimum for each factor, under mean-field VB, has the same shape as the actual-not necessarily symmetric-full conditional density for the subset of parameters in θ comprising that factor, and the associated variational coefficients can be estimated iteratively via coordinate ascent variational inference (CAVI) algorithms (e.g., Blei *et al.*, 2017). Nevertheless, several routinely used Bayesian models, such as logit (Durante and Rigon, 2019) and probit (Consonni and Marin, 2007) regression, allow Gaussian full conditional on the coefficients. Thus, the proposed perturbation may be useful even in mean-field VB to improve the accuracy of the symmetric density factors and thus the overall approximation for the entire posterior. These Gaussian density approximations also appear in Laplace variational inference, delta-method variational inference, and automatic differentiation variational inference (ADVI) (Wang and Blei, 2013; Kucukelbir *et al.*, 2017), which are often used to facilitate the implementation of mean-field VB, even in more complex models where CAVI does not allow simple closed-form updates.

Although VB is arguably the most widely studied and implemented deterministic approximation strategy, routine-use variational approximations often suffer from an underestimation of posterior uncertainty (e.g., Blei *et al.*, 2017; Giordano *et al.*, 2018). This is implicit in the expression of the KL minimized under VB that penalizes approximating densities placing mass where the one assigned by the actual posterior is low and does not enforces a similarly-strong penalty for the opposite case. As a consequence, minimizing such a KL yields more concentrated approximations avoiding regions where the actual posterior does not places substantial mass. While improved estimates of variances and covariances have been proposed in the context of VB (Giordano *et al.*, 2018),

another natural solution is to consider EP (e.g., Vehtari *et al.*, 2020), which addresses such a issue by minimizing a reverse form of the KL considered under VB. This implies penalizations in the opposite direction than those of VB and, therefore, a tendency to favor more global *zero-avoiding* approximations that match more closely the variability encoded in the target posterior. To obtain these approximations EP postulates that the target posterior density itself can be expressed as a product of factors-often arising as a direct consequence of the conditional independence structures in the likelihood-and then iteratively approximates each of these factors with an element of a tractable parametric family, almost always Gaussian. This results in a computational scheme updating each factor at-a-time via moment matching between the global approximating density and a hybrid one, more tractable than the target posterior, where the other factors are kept fixed at the most recent approximation (e.g., Vehtari *et al.*, 2020). Being often Gaussian, these EP approximating densities can be readily perturbed under our proposed strategy to further improve the quality of the approximation. In fact, although several empirical studies have highlighted remarkable accuracy of Gaussian EP (e.g., Chopin and Ridgway, 2017; Vehtari *et al.*, 2020; Anceschi *et al.*, 2023) relative to Laplace and VB, there is still a lack of tractable solutions capable of including skewness within these Gaussian densities to further improve the quality of EP approximations. As illustrated in the empirical studies in Section 3.6, incorporating these skewed behaviors yields further remarkable improvements over such an already-successful strategy.

Before moving to the proposed skew-symmetric perturbation in Section 3.3, it is important to emphasize that our strategy applies directly to any symmetric approximation, not necessarily Gaussian. This can be useful to perturb further extensions of the aforementioned methods aimed at capturing higher-order (e.g., tails) properties. For example, generalizations from Gaussian approximating densities to Student-t ones have been explored in the context of Laplace, VB and EP (see e.g., Ding *et al.*, 2011; Futami *et al.*, 2017; Liang *et al.*, 2022; Gelman *et al.*, 2013, Ch. 13), but there are no strategies, to date, for including skewness within such extensions. Our proposed solution applies also to these symmetric approximations and, as proved in Section 3.5, allows to progressively improve the asymptotic convergence rates to the exact posterior via perturbation of these increasingly accurate symmetric densities.

3.3 Skew-symmetric perturbation of symmetric approximations

As anticipated in Section 3.1, the newly-proposed skew-symmetric approximation arises from the perturbation of an already-available approximating density $f^*_{\tilde{\theta}}(\theta)$, which is symmetric about the point $\tilde{\theta}$, i.e., $f^*_{\tilde{\theta}}(\theta) = f^*_{\tilde{\theta}}(2\tilde{\theta} - \theta)$ for any $\theta \in \Theta$. Although $f^*_{\tilde{\theta}}(\theta)$ arises from the attempt to accurately approximate the target posterior density $\pi_n(\theta)$ via, e.g., one of the methods discussed in Section 3.2, the overall quality of $f^*_{\tilde{\theta}}(\theta)$ is clearly undermined by the fact that $\pi_n(\theta)$ is often skewed in practice, whereas $f^*_{\tilde{\theta}}(\theta)$ is symmetric. To this end, it is natural to ask whether the given density $f^*_{\tilde{\theta}}(\theta)$ would rather provide a more accurate approximation for a symmetrized version $\bar{\pi}_{n,\tilde{\theta}}(\theta)$ of $\pi_n(\theta)$ about $\tilde{\theta}$ than for the original posterior $\pi_n(\theta)$, under a suitable divergence. Lemma 3.1 states that this is the case when the symmetrized version $\bar{\pi}_{n,\tilde{\theta}}(\theta)$ is defined as

$$\bar{\pi}_{n,\tilde{\theta}}(\theta) = \frac{\pi_n(\theta) + \pi_n(2\tilde{\theta} - \theta)}{2}, \tag{3.1}$$

for every parameter value $\theta \in \Theta$ and known point of symmetry $\tilde{\theta} \in \Theta$.

Lemma 3.1. *Let $\pi_n(\theta)$ be a generic posterior density for the parameter $\theta \in \Theta$, and denote with $f^*_{\tilde{\theta}}(\theta)$ an already-available approximation of $\pi_n(\theta)$ which is symmetric about the point $\tilde{\theta} \in \Theta$. Moreover, define the symmetrized posterior density about $\tilde{\theta}$ as in (3.1). Then*

$$\mathcal{D}[\bar{\pi}_{n,\tilde{\theta}}(\theta) \,||\, f^*_{\tilde{\theta}}(\theta)] \leq \mathcal{D}[\pi_n(\theta) \,||\, f^*_{\tilde{\theta}}(\theta)],$$

for any $\tilde{\theta} \in \Theta$ and sample size n, where $\mathcal{D}$ is either the TV *distance ($\mathcal{D}_{\text{TV}}$) or any α-divergence ($\mathcal{D}_\alpha$).*

Proof. For ease of reading the proof of Lemma 3.1 is moved to Section A.3.1. □

Remark 3.2. Although other forms of symmetrization can be designed to define $\bar{\pi}_{n,\tilde{\theta}}(\theta)$, the one considered in (3.1) is arguably the most natural and simple. In fact, although it has been generally overlooked in the subsequent literature and, to the best of our knowledge, never explored in the context of Bayesian deterministic approximations, such a symmetrization has been successfully employed in classical frequentist literature to improve standard estimators of empirical distribution functions associated with underlying symmetric densities (Schuster, 1975; Hinkley, 1976; Lo, 1985; Meloche, 1991). Our contribution leverages such a symmetrization in an innovative manner and with a substantially different focus. Nonetheless, as clarified in the following, its distinctive form plays

a crucial role also within our context, in that it allows to design a theoretically-supported and similarly-tractable skew-symmetric perturbation of the known $f^*_{\tilde{\theta}}(\theta)$ without additional optimization steps.

Remark 3.3. Lemma 3.1 states a general result which applies to any symmetric approximation $f^*_{\tilde{\theta}}(\theta)$ of a generic posterior $\pi_n(\theta)$, and holds under both the TV distance $\mathcal{D}_{\text{TV}}[p(\theta) \;||\; q(\theta)] = \int_\Theta |p(\theta) - q(\theta)| d\theta/2$ and every α-divergence $\mathcal{D}_\alpha[p(\theta) \;||\; q(\theta)] = [1/(\alpha(1-\alpha))](\int_\Theta 1 - p(\theta)^\alpha q(\theta)^{1-\alpha} d\theta)$ for $\alpha \in \mathbb{R} \setminus \{0; 1\}$ (see e.g., Cichocki and Amari, 2010). Notice that, when $\alpha \to 0$ and $\alpha \to 1$, the distance $\mathcal{D}_\alpha$ reduces to the KL divergences $\text{KL}[q(\theta) \;||\; p(\theta)]$ and $\text{KL}[p(\theta) \;||\; q(\theta)]$, respectively (e.g., Cichocki and Amari, 2010). Hence, Lemma 3.1, and the subsequent theoretical results we derive, hold under the most widely used divergences within the context of deterministic approximations for posterior distributions. As discussed in Section 3.2, the two KL divergences $\text{KL}[q(\theta) \;||\; p(\theta)]$ and $\text{KL}[p(\theta) \;||\; q(\theta)]$ formally enter the formulation of the minimization problems underlying VB and EP (e.g., Blei *et al.*, 2017; Vehtari *et al.*, 2020), whereas TV is a reference distance in the study of the asymptotic accuracy of standard approximations (e.g., Kasprzak *et al.*, 2022; Wang and Blei, 2019). Recent literature has also explored variational approximations based on a generic α-divergence, beyond its limiting KL form, to obtain increased flexibility in the minimized loss function (Yang *et al.*, 2020). Hence, our methods and theory further extend to these strategies.

Although the result in Lemma 3.1 is interesting in its own right, the ultimate goal is to approximate the actual posterior $\pi_n(\theta)$, and not its symmetrized form $\bar{\pi}_{n,\tilde{\theta}}(\theta)$. With this objective in mind, Proposition 3.4 establishes an analytic relation among these two densities which is fundamental to derive our improved skew-symmetric approximation of $\pi_n(\theta)$ via a closed-form perturbation of the original $f^*_{\tilde{\theta}}(\theta)$.

Proposition 3.4 (Skew-symmetric representation of posterior densities)**.** *Consider a generic posterior density $\pi_n(\theta) = \pi(\theta)L(\theta; X^n)/c(X^n)$ for $\theta \in \Theta$, where $\pi(\theta)$ is the prior, $L(\theta; X^n)$ the likelihood, and $c(X^n)$ the normalizing constant. Moreover, denote with $\bar{\pi}_{n,\tilde{\theta}}(\theta) = [\pi_n(\theta) + \pi_n(2\tilde{\theta} - \theta)]/2$ the symmetrized form of such a posterior density about a point $\tilde{\theta} \in \Theta$, and let*

$$w^*_{\tilde{\theta}}(\theta) = \frac{\pi(\theta)L(\theta; X^n)}{\pi(\theta)L(\theta; X^n) + \pi(2\tilde{\theta} - \theta)L(2\tilde{\theta} - \theta; X^n)}. \tag{3.2}$$

Then, the posterior density $\pi_n(\theta)$ can be equivalently re-expressed as

$$\pi_n(\theta) = 2\bar{\pi}_{n,\tilde{\theta}}(\theta) w^*_{\tilde{\theta}}(\theta), \tag{3.3}$$

for any $\tilde{\theta} \in \Theta$ *and sample size* n, *where* $w^*_{\tilde{\theta}}(\theta) \in [0,1]$ *satisfies* $w^*_{\tilde{\theta}}(\theta) = 1 - w^*_{\tilde{\theta}}(2\tilde{\theta} - \theta)$.

The proof of Proposition 3.4 is direct and simply requires to note that the posterior density $\pi_n(\theta)$ can be equivalently re-written as $2\bar{\pi}_{n,\tilde{\theta}}(\theta)[\pi_n(\theta)/(2\bar{\pi}_{n,\tilde{\theta}}(\theta))]$. Therefore, replacing $\bar{\pi}_{n,\tilde{\theta}}(\theta)$ with its expression in (3.1), and $\pi_n(\theta)$ with $\pi(\theta)L(\theta; X^n)/c(X^n)$, yields the skewness-inducing factor $w^*_{\tilde{\theta}}(\theta)$ in (3.3), after noticing that the intractable normalizing constant $c(X^n)$ cancels out in the ratio between the target posterior density and its symmetrized form. The $[0,1]$ range for $w^*_{\tilde{\theta}}(\theta)$ and the property $w^*_{\tilde{\theta}}(\theta) = 1 - w^*_{\tilde{\theta}}(2\tilde{\theta} - \theta)$ follow directly from its definition. As clarified later, such properties for $w^*_{\tilde{\theta}}(\theta)$ are fundamental to ensure that the proposed perturbation for $f^*_{\tilde{\theta}}(\theta)$ belongs to a tractable class of skew–symmetric distributions (Azzalini and Capitanio, 2003; Wang *et al.*, 2004; Ma and Genton, 2004).

Proposition 3.4 crucially relates to a core result in Wang *et al.* (2004) which establishes the existence and uniqueness of skew–symmetric representations for generic densities. Such a parallel ensures that the equivalent expression derived for $\pi_n(\theta)$ in (3.3) is the one of a skew-symmetric density. Although the representation for $\pi_n(\theta)$ in Proposition 3.4 has been never explored in the context of Bayesian inference and deterministic approximations, as clarified in (3.3), the resulting skew-symmetric representation unveils a previously-overlooked form given by the product between a, possibly intractable, symmetrized posterior $\bar{\pi}_{n,\tilde{\theta}}(\theta)$ and a tractable skewness-inducing factor $w^*_{\tilde{\theta}}(\theta)$, where the normalizing constant $c(X^n)$ cancels out. Hence, such a representation, together with Lemma 3.1, suggest a natural strategy to obtain a skew-symmetric approximation via direct perturbation of $f^*_{\tilde{\theta}}(\theta)$. More specifically, since by Lemma 3.1, $f^*_{\tilde{\theta}}(\theta) \approx \bar{\pi}_{n,\tilde{\theta}}(\theta)$ more accurately than $\pi_n(\theta)$, and, as a direct consequence of (3.3), $\bar{\pi}_{n,\tilde{\theta}}(\theta) = \pi_n(\theta)/[2w^*_{\tilde{\theta}}(\theta)]$, it follows that

$$f^*_{\tilde{\theta}}(\theta) \approx \bar{\pi}_{n,\tilde{\theta}}(\theta) \quad \longrightarrow \quad f^*_{\tilde{\theta}}(\theta) \approx \pi_n(\theta)/[2w^*_{\tilde{\theta}}(\theta)] \quad \longrightarrow \quad 2f^*_{\tilde{\theta}}(\theta)w^*_{\tilde{\theta}}(\theta) \approx \pi_n(\theta). \quad (3.4)$$

Setting $q^*_{\tilde{\theta}}(\theta) = 2f^*_{\tilde{\theta}}(\theta)w^*_{\tilde{\theta}}(\theta)$ yields the novel skew-symmetric approximation in Definition 3.5.

Definition 3.5 (Skew-symmetric approximation of posterior densities)**.** Denote with $\pi_n(\theta)$ a generic posterior density for the parameter $\theta \in \Theta$, and let $f^*_{\tilde{\theta}}(\theta)$ be an already-known approximation of $\pi_n(\theta)$ which is symmetric about the point $\tilde{\theta} \in \Theta$. Moreover, denote with $w^*_{\tilde{\theta}}(\theta) \in [0,1]$ the skewness-inducing factor defined in equation (3.2). Then,

a skew-symmetric approximation of $\pi_n(\theta)$ arising from the perturbation of $f^*_{\tilde{\theta}}(\theta)$ is defined as

$$q^*_{\tilde{\theta}}(\theta) = 2 f^*_{\tilde{\theta}}(\theta) w^*_{\tilde{\theta}}(\theta), \tag{3.5}$$

for every known symmetry point $\tilde{\theta} \in \Theta$ and sample size n.

Remark 3.6. Note that, in (3.5), the skewness-inducing factor $w^*_{\tilde{\theta}}(\theta)$ admits a natural interpretation. Indeed it coincides with the relative proportion of posterior density assigned at $\theta \in \Theta$ with respect to the total given to such a θ and to $2\tilde{\theta} - \theta$. For every $\theta \in \Theta$, this yields a skewing function which quantifies differences in the posterior density at points $(\theta, 2\tilde{\theta} - \theta)$ that are symmetric with respect to the center of symmetry $\hat{\theta}$ of the original approximation $f^*_{\tilde{\theta}}(\theta)$. Hence, if the posterior is actually symmetric about $\tilde{\theta}$, then for all $\theta \in \Theta$, $w^*_{\tilde{\theta}}(\theta) = 1/2$ and $q^*_{\tilde{\theta}}(\theta)$ reduces to $f^*_{\tilde{\theta}}(\theta)$, as expected. Conversely, whenever there are asymmetries within $\pi_n(\theta)$, the original symmetric approximation $f^*_{\tilde{\theta}}(\theta)$ is re-weighted by $w^*_{\tilde{\theta}}(\theta)$ in order to properly re-distribute the total density at each pair $(\theta, 2\tilde{\theta} - \theta)$ according to the one assigned by the actual posterior to θ and $2\tilde{\theta} - \theta$. This yields an improved approximation $q^*_{\tilde{\theta}}(\theta)$ which incorporates the skewness in $\pi_n(\theta)$ with respect to the know symmetry point $\tilde{\theta}$. For instance, if $\pi_n(\theta) > \pi^{(n)}(2\tilde{\theta} - \theta)$, then, by definition, also $q^*_{\tilde{\theta}}(\theta) > q^*_{\tilde{\theta}}(2\tilde{\theta} - \theta)$, whereas $f^*_{\tilde{\theta}}(\theta) = f^*_{\tilde{\theta}}(2\tilde{\theta} - \theta)$ by construction.

As clarified in Definition 3.5 and in Remark 3.6, the proposed approximation $q^*_{\tilde{\theta}}(\theta)$ results from the re-weighting of the known $f^*_{\tilde{\theta}}(\theta)$ by a skewness-inducing factor $w^*_{\tilde{\theta}}(\theta)$ via a strategy which does not require additional optimization costs relative to those required for deriving the original $f^*_{\tilde{\theta}}(\theta)$. In fact, the expression for $w^*_{\tilde{\theta}}(\theta)$ in (3.2) does not depend on additional unknown parameters beyond $\tilde{\theta}$, which is in turn available as the output of the already-solved optimization problem that targeted the posterior $\pi_n(\theta)$ via the symmetric density $f^*_{\tilde{\theta}}(\theta)$. Proposition 3.7 below guarantees that, albeit more flexible and sophisticated than the original $f^*_{\tilde{\theta}}(\theta)$, the deterministic approximation $q^*_{\tilde{\theta}}(\theta)$ in Definition 3.5 still preserves similar tractability in inference, in that it belongs to the known class of skew-symmetric densities (Azzalini and Capitanio, 2003; Wang *et al.*, 2004; Ma and Genton, 2004; Genton and Loperfido, 2005).

Proposition 3.7. *The expression for $q^*_{\tilde{\theta}}(\theta)$ given in equation (3.5) coincides with the density of a skew-symmetric distribution with symmetric component $f^*_{\tilde{\theta}}(\theta)$ and skewing function $w^*_{\tilde{\theta}}(\theta)$.*

Proposition 3.7 follows directly from the general definition of skew-symmetric densities in, e.g., Wang *et al.* (2004), after noticing that, by construction, $f^*_{\tilde{\theta}}(\theta)$ is symmetric

Algorithm 1 i.i.d. sampling from the approximating distribution with density $q^*_{\tilde\theta}(\theta)$ in (3.5)

For $s = 1, \dots, N_{\text{SAMPL}}$ **do**:

1. Sample $\theta_0^{(s)}$ from the distribution having symmetric density $f_0^*(\theta_0)$,
2. Draw a sample from the Bernoulli random variable $u^{(s)} \sim \text{Be}(w^*_{\tilde\theta}(\tilde\theta + \theta_0^{(s)}))$,
3. Set $\theta^{(s)} = (2u^{(s)} - 1)\theta_0^{(s)} + \tilde\theta$.

Output: i.i.d. samples $\theta^{(1)}, \dots, \theta^{(N_{\text{SAMPL}})}$ from the approximating distribution with density $q^*_{\tilde\theta}(\theta)$.

about $\tilde\theta$ and, as proved in Proposition 3.4, $w^*_{\tilde\theta}(\theta)$ has support in $[0, 1]$ and satisfies $w^*_{\tilde\theta}(\theta) = 1 - w^*_{\tilde\theta}(2\tilde\theta - \theta)$.

The connection with skew-symmetric distributions established in Proposition 3.7 is crucial in facilitating inference also under $q^*_{\tilde\theta}(\theta)$, in that such a class admits a straightforward and rejection-free i.i.d. sampling scheme from the general stochastic representation of skew-symmetric random variables described in Chapter 1, Proposition 1.2. The algorithm to obtain a sample from $q^*_{\tilde\theta}(\theta)$ is outlined in Algorithm 1.

Note that Algorithm 1 only requires simulation from the original symmetric approximation $f^*_{\tilde\theta}(\theta)$ and computation of the skewness-inducing factor $w^*_{\tilde\theta}(\theta)$, which is analytically-available in Definition 3.5 and does not depend on intractable quantities. The first task is straightforward whenever the perturbed density $f^*_{\tilde\theta}(\theta)$ arises from one of the routinely-implemented approximation schemes discussed in Section 3.2, and, when the interest is in more complex functionals, it is often required also for inference under $f^*_{\tilde\theta}(\theta)$. The second, requires instead evaluation of the likelihood, which enters the definition of $w^*_{\tilde\theta}(\theta)$. Although this yields some increment in sampling costs relative to inference under $f^*_{\tilde\theta}(\theta)$, as illustrated in the empirical studies in Section 3.6, the noticeable gains we obtain in approximation accuracy justify this additional cost. Notice also that these likelihood evaluations are standard in state-of-the-art sampling-based inference schemes, such as, for example, importance sampling, Metropolis-Hastings, Hamiltonian Monte Carlo and sequential Monte Carlo, among others (e.g., Chopin and Ridgway, 2017; Chopin and Papaspiliopoulos, 2020). However, unlike for Algorithm 1, these schemes are characterized by additional complexities, which often require further tuning and do not ensure a rejection-free i.i.d. sampling strategy. To this end, Algorithm 1 achieves a sensibile balance between the inference tractability ensured by $f^*_{\tilde\theta}(\theta)$ and the increased accuracy associated with state-of-the-art sampling schemes targeting the exact posterior density.

Sections 3.4-3.5 clarify that the skew-symmetric approximation proposed in Definition 3.5 is not only tractable from both an optimization and inference perspective, but

also yields a provably more accurate characterization of the exact target posterior, both in finite sample settings and in asymptotic regimes.

3.4 Skew-symmetric correction: finite-sample properties and optimality

The original motivation for the skew-symmetric approximation $q^*_{\tilde{\theta}}(\theta)$ in (3.5) is to improve the accuracy of the original unperturbed $f^*_{\tilde{\theta}}(\theta)$. Theorem 3.8 provides theoretical support in finite samples to such an accuracy improvement and crucially clarifies that the overall quality of $q^*_{\tilde{\theta}}(\theta)$ solely depends on how accurate $f^*_{\tilde{\theta}}(\theta)$ is in approximating the symmetrized posterior $\bar{\pi}_{n,\tilde{\theta}}(\theta)$ in (3.1). These results are further deepened in Theorem 3.9 which proves that the skewness-inducing factor $w^*_{\tilde{\theta}}(\theta)$ in Definition 3.5 is optimal among all those yielding a skew-symmetric approximation for $\pi_n(\theta)$, with $f^*_{\tilde{\theta}}(\theta)$ as symmetric component.

Theorem 3.8 (Finite-sample accuracy)**.** *Let $\pi_n(\theta)$ be a generic posterior density for the parameter $\theta \in \Theta$, and denote with $f^*_{\tilde{\theta}}(\theta)$ an already-known approximation of $\pi_n(\theta)$ which is symmetric about the point $\tilde{\theta} \in \Theta$. Moreover, let $q^*_{\tilde{\theta}}(\theta) = 2f^*_{\tilde{\theta}}(\theta)w^*_{\tilde{\theta}}(\theta)$ with $w^*_{\tilde{\theta}}(\theta)$ as in equation (3.2). Then*

$$\mathcal{D}[\pi_n(\theta) \,||\, q^*_{\tilde{\theta}}(\theta)] \leq \mathcal{D}[\pi_n(\theta) \,||\, f^*_{\tilde{\theta}}(\theta)], \tag{3.6}$$

for any symmetry point $\tilde{\theta} \in \Theta$ and sample size n, where $\mathcal{D}$ is either the total variation distance ($\mathcal{D}_{\text{TV}}$) or any α-divergence ($\mathcal{D}_\alpha$). Moreover,

$$\mathcal{D}[\pi_n(\theta) \,||\, q^*_{\tilde{\theta}}(\theta)] = \mathcal{D}[\bar{\pi}_{n,\tilde{\theta}}(\theta) \,||\, f^*_{\tilde{\theta}}(\theta)], \tag{3.7}$$

for any $\tilde{\theta} \in \Theta$ and n, where $\bar{\pi}_{n,\tilde{\theta}}(\theta)$ is the symmetrized posterior density in equation (3.1).

Proof. For ease of reading the proof of Theorem 3.8 is postponed to Section A.3.2. □

Theorem 3.8 states two important results. First, according to (3.6), $q^*_{\tilde{\theta}}(\theta)$ is, provably, never less accurate than the original $f^*_{\tilde{\theta}}(\theta)$ in approximating the target posterior $\pi_n(\theta)$, irrespectively of the chosen $f^*_{\tilde{\theta}}(\theta)$, its symmetry point $\tilde{\theta}$, and the sample size n. Second, as clarified in (3.7), the overall quality of $q^*_{\tilde{\theta}}(\theta)$ actually coincides with the one achieved by the original unperturbed $f^*_{\tilde{\theta}}(\theta)$ in approximating the symmetrized posterior $\bar{\pi}_{n,\tilde{\theta}}(\theta)$ in (3.1). Such a latter result is of direct practical interest in that, among the approximating densities with the same symmetry point, it suggests to prioritize those

yielding a more accurate approximation of the symmetrized posterior in (3.1), rather than of the original $\pi_n(\theta)$. Therefore, although this objective goes beyond our original scope of perturbing an already-available $f^*_{\tilde{\theta}}(\theta)$, equations (3.6)-(3.7) stimulate the development of novel symmetric approximations explicitly targeting $\bar{\pi}_{n,\tilde{\theta}}(\theta)$ rather than the original posterior $\pi_n(\theta)$. In fact, as a consequence of (3.6)-(3.7), the perturbation of these approximations under the proposed strategy can yield increasingly accurate characterizations of $\pi_n(\theta)$.

In addition, the results in Theorem 3.8 follow from the specific form of the skewness-inducing factor $w^*_{\tilde{\theta}}(\theta)$ in equation (3.2). Theorem 3.9 proves the optimality of such a choice.

Theorem 3.9 (Optimality of the skewness-inducing factor)**.** *Let $\pi_n(\theta)$ denote a generic posterior density for the parameter $\theta \in \Theta$, and let $f^*_{\tilde{\theta}}(\theta)$ be an already-known approximation of $\pi_n(\theta)$ which is symmetric about $\tilde{\theta} \in \Theta$. Moreover, denote with $q^*_{\tilde{\theta}}(\theta) = 2f^*_{\tilde{\theta}}(\theta)w^*_{\tilde{\theta}}(\theta)$ the proposed approximation in Definition 3.5, and define with $q_{\tilde{\theta}}(\theta) = 2f^*_{\tilde{\theta}}(\theta)w_{\tilde{\theta}}(\theta)$ an alternative skew-symmetric perturbation of $f^*_{\tilde{\theta}}(\theta)$, where $w_{\tilde{\theta}}(\theta)$ correspond to a generic skewing function such that $w_{\tilde{\theta}}(\theta) \in [0,1]$ and $w_{\tilde{\theta}}(\theta) = 1 - w_{\tilde{\theta}}(2\tilde{\theta} - \theta)$. Then, for every $w_{\tilde{\theta}}(\theta)$, it holds that*

$$\mathcal{D}[\pi_n(\theta) \,||\, q^*_{\tilde{\theta}}(\theta)] \leq \mathcal{D}[\pi_n(\theta) \,||\, q_{\tilde{\theta}}(\theta)], \tag{3.8}$$

for any $\tilde{\theta} \in \Theta$ and sample size n, where $\mathcal{D}$ is either the TV *distance ($\mathcal{D}_{\text{TV}}$) or any α-divergence ($\mathcal{D}_\alpha$).*

Proof. The proof of Theorem 3.9 is postponed to Section A.3.3 □

According to Theorem 3.9, the skewness-inducing factor $w^*_{\tilde{\theta}}(\theta) = \pi(\theta)L(\theta; X^n)/[\pi(\theta)L(\theta; X^n) + \pi(2\tilde{\theta} - \theta)L(2\tilde{\theta} - \theta; X^n)]$ we derive is guaranteed to provide a perturbation $q^*_{\tilde{\theta}}(\theta)$ of $f^*_{\tilde{\theta}}(\theta)$ which is never less accurate in approximating the target posterior $\pi_n(\theta)$ when compared to those arising from any other skew-symmetric density $q_{\tilde{\theta}}(\theta)$ with symmetric component $f^*_{\tilde{\theta}}(\theta)$ and generic skewing function $w_{\tilde{\theta}}(\theta)$. Notice that to ensure that $q_{\tilde{\theta}}(\theta) = 2f^*_{\tilde{\theta}}(\theta)w_{\tilde{\theta}}(\theta)$ is a skew-symmetric density it suffices that the skewing function satisfies $w_{\tilde{\theta}}(\theta) \in [0,1]$ and $w_{\tilde{\theta}}(\theta) = 1 - w_{\tilde{\theta}}(2\tilde{\theta} - \theta)$. Hence, in principle, there are infinitely-many options to perturb the original symmetric approximation so that the resulting density falls in the skew-symmetric class. Some interesting examples of skewing functions have been derived in Azzalini and Capitanio (2003); Ma and Genton (2004) and Genton and Loperfido (2005) with a focus on generalizations of skew-normal and skew-elliptical densities which belong to the broader skew-symmetric family. According to Theorem 3.9, all these options would be suboptimal relative to the proposed skewing

factor $w^*_{\tilde{\theta}}(\theta)$. Intuitively, this is because, unlike other possible choices of skewing functions, $w^*_{\tilde{\theta}}(\theta)$ exactly matches the skewness factor of the actual target posterior, when expressed in skew-symmetric form as in Proposition 3.4.

Besides proving the optimality of the skewness-inducing factor $w^*_{\tilde{\theta}}(\theta)$, Theorem 3.9 also allows to formalize the proposed skew-symmetric approximation in Definition 3.5 as the solution of a well-defined optimization problem. Such a result is stated in Corollary 3.10 and is useful to establish connections with state-of-the-art approximation strategies arising from optimization of specific divergences, e.g., VB (Blei *et al.*, 2017) and EP (Vehtari *et al.*, 2020). In addition, as discussed in Remark 3.11, it provides the premises to further expand the scope of the perspective considered in this thesis.

Corollary 3.10. *Consider the generic posterior density $\pi_n(\theta)$ for the parameter of interest $\theta \in \Theta$, and let $f^*_{\tilde{\theta}}(\theta)$ be an already-available approximation of $\pi_n(\theta)$ which is symmetric about the point $\tilde{\theta} \in \Theta$. Moreover, denote with $q^*_{\tilde{\theta}}(\theta) = 2f^*_{\tilde{\theta}}(\theta)w^*_{\tilde{\theta}}(\theta)$ the proposed skewed approximation in Definition 3.5, and define with*

$$\mathcal{Q} = \{q_{\tilde{\theta}}(\theta) = 2f^*_{\tilde{\theta}}(\theta)w_{\tilde{\theta}}(\theta) : w_{\tilde{\theta}}(\theta) \in [0,1], w_{\tilde{\theta}}(\theta) = 1 - w_{\tilde{\theta}}(2\tilde{\theta} - \theta)\},$$

the family of skew-symmetric densities that arise from the perturbation of $f^*_{\tilde{\theta}}(\theta)$ via a generic skewing function $w_{\tilde{\theta}}(\theta)$. Then

$$q^*_{\tilde{\theta}}(\theta) = \text{argmin}_{q_{\tilde{\theta}}(\theta) \in \mathcal{Q}} \mathcal{D}[\pi_n(\theta) \,||\, q_{\tilde{\theta}}(\theta)],$$

for any $\tilde{\theta} \in \Theta$ and sample size n, where $\mathcal{D}$ is either the TV distance ($\mathcal{D}_{\text{TV}}$) or any α-divergence ($\mathcal{D}_\alpha$).

Although the skew-symmetric approximation derived in Section 3.3 is not explicitly obtained as the solution of a suitably-defined optimization problem, Corollary 3.10 clarifies that, in fact, the proposed $q^*_{\tilde{\theta}}(\theta)$ can be formalized also under such a perspective. In particular, the skew-symmetric density $q^*_{\tilde{\theta}}(\theta)$ in Definition 3.5 actually coincides with the solution of the constrained minimization for a suitable divergence $\mathcal{D}$ between the target posterior $\pi_n(\theta)$ and a given approximating density $q_{\tilde{\theta}}(\theta)$ within the family $\mathcal{Q}$ of skew-symmetric densities with symmetric component fixed at the original approximating density $f^*_{\tilde{\theta}}(\theta)$ to be perturbed. This interpretation allows to establish connections with the optimization-based perspectives underlying VB (Blei *et al.*, 2017) and EP (Vehtari *et al.*, 2020) solutions. However, unlike these methods, Corollary 3.10 holds under a broader class of divergences, rather than a specific one, and, when compared to routinely-implemented VB and EP schemes yielding symmetric approximations, it considers an expanded family which ensures improvements in accuracy. Notice that,

consistent with the focus of this chapter, $f^*_{\tilde{\theta}}(\theta)$ and, as a consequence, $\tilde{\theta} \in \Theta$ are known and fixed in Corollary 3.10 and, hence, the only quantity to be derived is the skewness-inducing factor. Crucially, as clarified Theorem 3.9, the solution $w^*_{\tilde{\theta}}(\theta)$ to such a minimization with respect to $w_{\tilde{\theta}}(\theta)$ does not require optimization of additional parameters beyond the already-available $\tilde{\theta} \in \Theta$. Although extending the optimization problem in Corollary 3.10 to the case in which also $f^*_{\tilde{\theta}}(\theta)$ is unknown goes beyond the scope of our contribution, as clarified in Remark 3.11, such a direction can be of substantial interest to further improve the accuracy of $q^*_{\tilde{\theta}}(\theta)$, and our results open several avenues to stimulate future advancements along these lines.

Remark 3.11. As already discussed, the overarching focus of this chapter is to improve the accuracy of state-of-the-art symmetric approximations of posterior distributions via a broadly-applicable perturbation scheme which can be derived at no additional optimization costs and applied directly to the output $f^*_{\tilde{\theta}}(\theta)$ of standard implementations. To this end, $f^*_{\tilde{\theta}}(\theta)$ is kept fixed and known in our derivations. However, although the optimization of such a symmetric component goes beyond the scope of our contribution, combining the results in Theorem 3.8 and Corollary 3.10 with the skew-symmetric representation of posterior densities in Proposition 3.4, opens promising directions to further improve approximation accuracy via the additional optimization of the symmetric component. In particular, notice that as a consequence of Corollary 3.10 and equation (3.7), $\min_{q_{\tilde{\theta}}(\theta)\in\mathcal{Q}} \mathcal{D}[\pi_n(\theta) \,||\, q_{\tilde{\theta}}(\theta)] = \mathcal{D}[\pi_n(\theta) \,||\, q^*_{\tilde{\theta}}(\theta)] = \mathcal{D}[\bar{\pi}_{n,\tilde{\theta}}(\theta) \,||\, f^*_{\tilde{\theta}}(\theta)]$, for any, possibly unknown, $f^*_{\tilde{\theta}}(\theta)$. Hence, when solving the minimization problem in Corollary 3.10 also with respect to $f^*_{\tilde{\theta}}(\theta)$, it suffices to find the closest symmetric density to the symmetrized posterior in equation (3.1) and then perturb such a density via the already-derived optimal skewness-inducing factor $w^*_{\tilde{\theta}}(\theta)$. This is expected to further improve accuracy relative to perturbations of currently-implemented symmetric approximations that target the actual posterior instead of its symmetrized version. In fact, to the best of our knowledge, such a different target has been never considered and, hence, our innovative perspective and results can open unexplored avenues to derive improved classes of tractable deterministic approximations, along with novel computational methods to obtain these approximations. This is because the minimization of $\mathcal{D}[\bar{\pi}_{n,\tilde{\theta}}(\theta) \,||\, f^*_{\tilde{\theta}}(\theta)]$ implies a peculiar reasoning which first requires finding that location $\tilde{\theta}$ inducing the symmetrized posterior in (3.1) more accurately approximated by a density within the assumed symmetric class and then, among the densities symmetric about such a location $\tilde{\theta}$, it seeks the one closest to the posterior symmetrized about $\tilde{\theta}$. Therefore, intuitively, such schemes implicitly require searching for both an optimal symmetrized posterior and also for an accurate approximation to such a symmetrized

posterior. When specializing $\mathcal{D}$ to the KL minimized under VB, which is a limiting case of α-divergence, a promising direction could be to solve such an optimization problem via automatic differentiation variational inference schemes (Kucukelbir *et al.*, 2017).

Theorems 3.8 and 3.9 provide finite-sample support on the accuracy improvements of the proposed skew-symmetric approximation $q^*_{\tilde{\theta}}(\theta)$. The theory in Section 3.5 further strengthens these results by quantifying the magnitude of such improvements in asymptotic regimes.

3.5 Asymptotic properties

Theorems 3.8-3.9 provide finite-sample theoretical guarantees for the better approximation quality of the newly developed skew-symmetric solution compared to its symmetric counterpart. At the same time, it is not clear whether such an improvement can have an effect also from an asymptotic point of view. Indeed, a general answer to this question is difficult to give, as a straightforward implication of Theorem (3.8) is that the behavior of the skew-symmetric approximation $q^*_{\tilde{\theta}}(\theta)$ only depends on how the symmetric component $f^*_{\tilde{\theta}}(\theta)$ approximates the symmetrized posterior $\bar{\pi}_{n,\tilde{\theta}}(\theta)$.

Building upon the results of Chapter 2, in this section, we provide an answer to such a challenging question for two particular symmetric approximations, in the asymptotic regime where d is fixed and $n \to \infty$. More specifically, we consider the cases in which $f^*_{\tilde{\theta}}(\theta)$ is either the Gaussian approximation centered at the posterior mode $\hat{\theta}$, and derived from the Laplace method (see Section 3.2),

$$f_{\hat{\theta},1}(\theta) = \phi_d(\theta; \hat{\theta}; \tilde{J}^{-1}_{\hat{\theta}}),$$

with $\tilde{J}_{\hat{\theta}} = -(\ell^{(2)}_{\hat{\theta}} + \log \pi^{(2)}_{\hat{\theta}})$, or a newly derived higher-order extension of it belonging to the family of semi-nonparametric distributions (Gallant and Nychka, 1987). This approximation is obtained by exploiting a fourth-order expansion of the symmetrized log-posterior distribution at $\hat{\theta}$ which, compared to the Gaussian density, is able to capture more accurately the behavior of the posterior distribution at the tails. In this case, the probability density function takes the form

$$f_{\hat{\theta},2}(\theta) = \phi_d(\theta; \hat{\theta}; \tilde{J}^{-1}_{\hat{\theta}}) P(\theta - \hat{\theta}) / \mathbb{E}_{f_{\hat{\theta},1}}\{P(\theta - \hat{\theta})\}, \tag{3.9}$$

where, $\mathbb{E}_{f_{\hat{\theta},1}}\{P(\theta-\hat{\theta})\} = \int P(\theta-\hat{\theta})\phi_d(\theta;\hat{\theta},\tilde{J}_{\hat{\theta}}^{-1})d\theta$ and $P(\theta-\hat{\theta})$ is a non-negative polynomial taking the form

$$\begin{aligned}P(\theta-\hat{\theta}) =& 1+\frac{1}{24}\ell^{(4)}_{\hat{\theta},stlk}(\theta-\hat{\theta})_s(\theta-\hat{\theta})_t(\theta-\hat{\theta})_l(\theta-\hat{\theta})_k \\ &+\frac{1}{2}\Big(\frac{1}{24}\ell^{(4)}_{\hat{\theta},stlk}(\theta-\hat{\theta})_s(\theta-\hat{\theta})_t(\theta-\hat{\theta})_l(\theta-\hat{\theta})_k\Big)^2+\frac{1}{2}\Big(\frac{1}{6}\ell^{(3)}_{\hat{\theta},stl}(\theta-\hat{\theta})_s(\theta-\hat{\theta})_t(\theta-\hat{\theta})_l\Big)^2.\end{aligned}$$

Note that, since $P(\theta-\hat{\theta})$ is always non negative, $f_{\hat{\theta},2}(\theta)$ is a proper density function which is symmetric about $\hat{\theta}$. As a consequence, these two particular choices of the symmetric component give rise to different skew-symmetric approximations $q_{\tilde{\theta},1}(\theta) = 2f_{\hat{\theta},1}(\theta)w_{\hat{\theta}}(\theta)$, $q_{\tilde{\theta},2}(\theta) = 2f_{\hat{\theta},2}(\theta)w_{\hat{\theta}}(\theta)$ which share the same skewness-inducing factor $w_{\hat{\theta}}(\theta)$.

Theorem 3.12 below, demonstrates that the total variation distance between the posterior and $q_{\tilde{\theta},1}(\theta)$ converges to zero in probability with rate $1/n$, up to a logarithmic term, while for $q_{\tilde{\theta},2}(\theta)$ the rate improves to $1/n^2$ again up to a logarithmic term. Informally, the difference between the performance of the two methods is explained by the fact that $f_{\hat{\theta},2}(\theta)$ is, asymptotically, a better approximation of the symmetrized posterior $\bar{\pi}_{n,\hat{\theta}}(\theta)$ and that, in view of Theorem 3.8, the same level of accuracy if maintained by $q_{\tilde{\theta},2}(\theta)$ in approximating the original posterior. To prove the validity of Theorem 3.12 we require the same assumptions of Theorem 2.1 with Assumption M2 replaced by Assumption M3 below.

M3 There exists two positive constants $\bar{\eta}_1, \bar{\eta}_2$ such that the event $\tilde{A}_{n,0} = \{\lambda_{\text{MIN}}(\tilde{J}_{\hat{\theta}}/n) > \bar{\eta}_1\}\cap\{\lambda_{\text{MAX}}(\tilde{J}_{\hat{\theta}}/n) < \bar{\eta}_2\}$ holds with probability $P_0^n\hat{A}_{n,0} = 1-o(1)$. Moreover, there exist two positive constants δ and L such that the inequalities $|\,\ell^{(3)}_{stl}(\theta)/n\,| < L$, $|\,\ell^{(4)}_{stlk}(\theta)/n\,| < L$, and $|\log\pi^{(2)}_{st}(\theta)| < L$ hold uniformly over $\theta\in B_\delta(\hat{\theta}) = \{\theta\in\Theta : \|\hat{\theta}-\theta\| < \delta\}$, with P_0^n-probability tending to one.

When $f_{\hat{\theta},2}(\theta)$ is considered, for the same δ and L as above, the inequalities $|\,\ell^{(5)}_{stl}(\theta)/n\,| < L$, $|\,\ell^{(6)}_{stlk}(\theta)/n\,| < L$, and $|\log\pi^{(3)}_{st}(\theta)| < L$ hold uniformly over $\theta\in B_\delta(\hat{\theta})$, with P_0^n-probability tending to one.

The first part of Assumption M3 is essentially equivalent to Assumption M2 described in Chapter 2. The additional conditions required when $f_{\hat{\theta},2}(\theta)$ is adopted, as symmetric component, are clearly slightly more restrictive since they concern the behavior of additional log-likelihood and log-prior higher-order derivatives. At the same time, they are the price to pay if one wants to reach a higher level of asymptotic accuracy. Note also that they are still fulfilled in many, commonly adopted, regular statistical models such Poisson regression, logistic regression and other generalized linear models.

The combination of Assumptions 1, 7-8, M1, described in Chapters 1-2, with Assumptions M3 allows us to demonstrate Theorem 3.12. The proof of the theorem is based on the one already developed for Theorem 2.1 but adapted to exploit the results derived in this chapter.

Theorem 3.12 (Asymptotic accuracy)**.** *Under Assumptions 1, 7-8, M1 and M3 it holds*

$$\mathcal{D}_{\text{TV}}[\pi_n(\cdot) \,||\, q_{\tilde{\theta},1}(\theta)] = O_{P_0^n}\big(M_n^{c_{13}}/n\big), \tag{3.10}$$

and

$$\mathcal{D}_{\text{TV}}[\pi_n(\cdot) \,||\, q_{\tilde{\theta},2}(\theta)] = O_{P_0^n}\big(M_n^{c_{14}}/n^2\big), \tag{3.11}$$

with $M_n = \sqrt{c_0 \log n}$ *and some fixed constants* $c_0, c_{13}, c_{14} > 0$.

Remark 3.13. Under the same assumptions, it is possible to prove that the rate of convergence of the symmetric components $f_{\hat{\theta},1}(\theta)$ and $f_{\hat{\theta},2}(\theta)$ to the posterior distribution is $O_{P_0^n}(M_n^{c_{15}}/\sqrt{n})$, for some $c_{15} > 0$. The different asymptotic performances observed in (3.10) and (3.11) are thus entirely due to how each symmetric component approximates $\bar{\pi}_{n,\hat{\theta}}(\theta)$.

The Gaussian density $f_{\hat{\theta},1}(\theta)$ is closely related to the symmetric component of the skew-modal approximation introduced in (2.4). In fact, the only difference between the two is the presence of the second log-prior derivative in $\tilde{J}_{\hat{\theta}}$, since the covariance matrix in (2.4) takes the form $J_{\hat{\theta}}^{-1} = -(\ell_{\hat{\theta}}^{(2)})^{-1}$. If d is fixed and $n \to \infty$, such a difference affects only terms of order $1/n$, and as a consequence, from an asymptotic point of view, using $\phi_d(\theta; \hat{\theta}; J_{\hat{\theta}})$ instead of $f_{\hat{\theta},1}(\theta)$ leads to skew-symmetric approximations with the same asymptotic accuracy.

3.6 Empirical studies

This section provides empirical evidence for the superior performance of the skew-symmetric correction introduced in Section 3.3 when compared (and applied) to three common symmetric approximations: the Gaussian approximation derived from the Laplace method (LA) (see e.g. Gelman *et al.*, 2013, Chapter 13), Gaussian variational Bayes with full covariance matrix (GVB) (Opper and Archambeau, 2009), and Gaussian expectation propagation (GEP) Minka (2001).

Two generalized linear models are considered, a logistic regression, and a Poisson regression with a logarithmic link function. In both cases, draws from the target posterior distribution are obtained via 5 chains of length 10000 under Hamiltonian Markov Chain Monte Carlo using the function `stan_glm`, available in the `R` package `rstanarm`

TABLE 3.1: For the logistic regression applied to the Glioma Grading Clinical and Mutation Features dataset, estimated values of the 7 summary statistics introduced in Section 3.6 for the Gaussian approximations LA, GVB, and GEP, and their skew-symmetric counterparts SKEW-LA, SKEW-GVB, and SKEW-GEP.

	Q1.ERR	MEDIAN.ERR	Q3.ERR	MEAN.ERR	WASS	KS	TV
LA	0.205	0.158	0.138	0.188	0.191	0.088	0.088
SKEW-LA	0.162	0.117	0.093	0.138	0.144	0.067	0.067
GVB	0.170	0.135	0.127	0.164	0.166	0.072	0.071
SKEW-GVB	0.117	0.087	0.071	0.105	0.109	0.048	0.047
GEP	0.017	0.036	0.032	0.009	0.037	0.021	0.028
SKEW-GEP	0.016	0.008	0.015	0.009	0.020	0.012	0.016

(Goodrich *et al.*, 2023). The parameters of LA and GVB approximations are obtained from the same function `stan_glm` with the specification `algorithm="optimizing"` and `algorithm="fullrank"`, respectively. GEP is computed with `Julia` using the function `ep_glm` (Barthelmé, 2023).

The quality of the resulting approximations is evaluated by considering the average, over all the marginal distributions, of seven different summary statistics. These quantities are estimated via Monte Carlo by drawing a sample of dimension 10000 from each approximation. Note that, given the Gaussianity of LA, GVB and GEP, this task can be easily done by exploiting Algorithm 1. The first 4 summary statistics are: mean absolute difference between the exact posterior first quartile and its approximation (Q1.ERR), mean absolute difference between the exact posterior median and its approximation (MEDIAN.ERR), mean absolute difference between the exact posterior third quartile and its approximation (Q3.ERR), and mean absolute difference between the exact posterior mean and its approximation (MEAN.ERR). In addition, the average of the empirical Wasserstein distances (WASS), the Kolmogorov-Smirnov statistics (KS), and the total variation distances (TV) between each marginal posterior density and its approximation is considered.

3.6.1 Example 1: logistic regression

We consider the Glioma Grading Clinical and Mutation Features dataset (Tasci *et al.*, 2022). These data, which are freely available in the UCI Machine Learning Repository, were obtained in a clinical study aimed at investigating, on $n = 839$ subjects, the ability of $d = 23$ explanatory variables to discriminate between patients with low-grade glioma or glioblastoma multiforme, two different forms of gliomas, the most common primary tumors of the brain (Tasci *et al.*, 2022). The 23 predictors consist of 3 demographic

TABLE 3.2: For the Poisson regression in Section 3.6.2, estimated values of the 7 summary statistics introduced in Section 3.6 for the Gaussian approximations LA, GVB, and GEP, and their skew-symmetric counterparts SKEW-LA, SKEW-GVB, and SKEW-GEP.

	Q1.ERR	MEDIAN.ERR	Q3.ERR	MEAN.ERR	WASS	KS	TV
LA	0.009	0.007	0.007	0.008	0.011	0.022	0.023
SKEW-LA	0.005	0.003	0.005	0.004	0.007	0.014	0.018
GVB	0.013	0.012	0.014	0.016	0.016	0.025	0.025
SKEW-GVB	0.007	0.006	0.007	0.008	0.009	0.016	0.017
GEP	0.004	0.004	0.004	0.003	0.006	0.012	0.017
SKEW-GEP	0.002	0.002	0.003	0.003	0.004	0.010	0.013

variables, gender, age at diagnosis and race; and the expression of 20 genes known to be associated with the two conditions.

To perform statistical inference, we define the response variable $x_i \in \{0, 1\}$, where $i = 1, \dots, n$, takes the value 1 if the patient is affected by low-grade glioma and 0 otherwise. A logistic regression model is assumed, meaning that each x_i is an independent realization of a Bernoulli random variable $X_i \sim \text{Be}(\exp(z_i^\intercal\theta)/\{1 + \exp(z_i^\intercal\theta)\})$, where $z_i \in \mathbb{R}^{24}$ is a vector containing the intercept and the explanatory variables, while $\theta \in \mathbb{R}^{24}$ is the parameter of interest. Finally, for each element of θ we assume a weakly informative Gaussian prior $N(0, \sigma^2)$, $\sigma = 5$.

Table 3.1 compares, for the model under study, the accuracy of LA, GVB, and GEP approximations with that of their skew-symmetric counterparts, SKEW-LA, SKEW-GVB, and SKEW-GEP, respectively. In this case, the improvement provided by our newly derived skew-symmetric solution is clear and uniform across all summary statistics. This highlights how in this case our skewness-inducing correction leads to a much more accurate representation of the true posterior shape.

3.6.2 Example 2: Poisson regression

As a second example, we consider a survey study in which 2276 high school students were asked whether they had ever used alcohol, cigarettes, or marijuana (see e.g. Agresti, 2015, Ex. 7.2.6). The dataset, which can also be retrieved using the `R` package `MLGdata` (Sartori *et al.*, 2020), consists of $n = 32$ entries representing all possible combinations of 5 explanatory variables: alcohol, cigarette and marijuana use together with 2 additional binary predictors: gender and race (classified as white or other). The variable of interest is $x_i =$ "number of students in the i-th combination of the predictors", for $i = 1, \dots, n$.

For these data, we assume a Poisson regression model, where each x_i is an independent realization of $X_i \sim \text{Pois}(\exp(z_i^\intercal\theta))$, with $z_i \in \mathbb{R}^{16}$ being a vector containing the

intercept, the 5 binary explanatory variables and all their possible pairwise interactions, while $\theta \in \mathbb{R}^{16}$ is the parameter of the model.

The comparison between the accuracy of LA, GVB and GEP and their corresponding skew-symmetric versions is reported in Table 3.2. As in the previous example, correcting for asymmetry leads to approximations of the posterior distribution that are significantly more accurate than the symmetric ones. This phenomenon is clear for all summary statistics considered.

Conclusions

Discussion

This thesis proves that substantial theoretical and practical improvements in accuracy can be achieved by replacing Gaussian, and more generally symmetric, deterministic approximations of the posterior distribution with suitably defined skew-symmetric densities.

Such a result is achieved first by adopting an asymptotic perspective in Chapters 1-2, where different limiting skew-symmetric approximations are constructively derived from a third-order version of the Laplace method. Beside demonstrating that, to approximate the posterior distribution, the same level of accuracy obtained by the adoption of higher-order polynomial approximations (Johnson, 1970; Weng, 2010; Kolassa and Kuffner, 2020) can be reached with a family of proper density functions, our contribution has also a remarkable methodological impact. Indeed, the joint and marginal skew-modal approximations derived in Chapter 2 can be directly applied to a wide class of statistical problems, both as a stand-alone method and as a building block of more sophisticated strategies, such as INLA (Rue *et al.*, 2009), which currently rely on Gaussian approximations in some of their parts.

The methodological contribution of the thesis is further broadened in Chapter 3 by introducing a novel and widely-applicable strategy to perturb any given symmetric approximation of a generic posterior density for obtaining an improved, yet tractable, skew-symmetric approximation with rigorous theoretical guarantees of improved accuracy and remarkable quality in empirical studies. Unlike recently-developed deterministic approximations based on generalizations of skew-normal distributions, the proposed solution [i] applies to generic posterior densities and to any symmetric approximation of such densities, [ii] does not imply additional optimization costs relative to those required for obtaining the original symmetric approximation to be improved, [iii] is simple and can be applied directly to any output of state-of-the-art software yielding symmetric approximations of posterior densities, [iv] has strong theoretical guarantees in terms of accuracy, both in finite-sample settings and in asymptotic regimes, and, finally, [v]

it arises from a yet-unexplored skew-symmetric representation of posterior densities in Proposition 3.4, which opens interesting avenues for future research in the context of deterministic approximations for Bayesian inference.

Future directions of research

The results of the thesis stimulate a number of interesting future research directions which we now briefly discuss. First of all, the methods we have derived are designed for Bayesian parametric models. Therefore, the extension of both theory and methods derived in Chapters 1-2 to semiparametric (Bickel and Kleijn, 2012; Castillo and Rousseau, 2015) problems opens an interesting and challenging area of research. Regarding the results of Chapter 3, another promising line of study is discussed in Remark 3.11 and relates to the case where the symmetric approximation density $f^*_{\hat{\theta}}(\theta)$ is also unknown and part of the novel optimization problem formalized in Corollary 3.10. Such an extension, combined with our results in Section 3.4, interestingly supports a change of perspective that suggests focusing on the symmetrized posterior densities (3.1), rather than on the original posterior, as the target of symmetric approximations. To the best of our knowledge, such a perspective has never been considered before, and is therefore expected to stimulate active research motivated by the novel questions associated with this task. For example, as mentioned in Remark 3.11, such a novel perspective implicitly requires extracting an appropriate symmetric component from the target posterior density which can be accurately approximated by a tractable symmetric density whose perturbation yields the final skew-symmetric approximation of $\pi_n(\theta)$.

Another important question is whether the perturbation strategy developed in Chapter 3 can be used to obtain a non-asymptotic counterpart of the marginal skew-modal approximations described in Section 2.3. Indeed, as for that class of approximations, such a solution would allow one to obtain quantities of interest for inference, such as posterior means and variances, by using of standard numerical integration methods and thereofore completely bypassing the need to sample from the approximation. From a practical point of view, the main challenge in this case is that the skewness-inducing factor of each marginal component depends on quantities that are not directly available in the un-normalized posterior and therefore need to be estimated. Since the error introduced by this additional approximation step is generally not easy to quantify, the quality of the resulting marginal skew-symmetric approximation is also difficult to assess.

Finally, we shall emphasize that, although our focus is on improving the accuracy of deterministic approximations for posterior densities, the results of Chapter 3 can have

important consequences also in refining sampling-based inference schemes. In fact, several of these schemes, including, e.g., importance sampling, Metropolis-Hastings, Hamiltonian Monte Carlo and sequential Monte Carlo methods (e.g., Chopin and Ridgway, 2017; Chopin and Papaspiliopoulos, 2020), rely on tractable proposal densities, often symmetric, whose closeness to the target posterior is a key in driving the convergence, mixing and degeneracy properties of the resulting sampling scheme. Perturbing such densities via our proposed strategy yields a skew-symmetric proposal which is provably-closer to the target posterior and, as discussed in Section 2.4, similarly-tractable from a sampling perspective. Hence, our contribution is expected to stimulate advancements also in the context of sampling-based Bayesian inference aimed at improving convergence, mixing and degeneracy issues encountered by state-of-the-art schemes.

Appendix A

Proofs main results of Chapters 1, 2 and 3

A.1 Proofs main results of Chapter 1

In this section we report the proofs of the theoretical results introduced in Chapter 1. The notation is the same as the main part of the thesis.

A.1.1 Proof of Corollary 1.7

Proof. To prove (1.18) note that the general Assumptions 1-2 and 4, introduced in Section 1.4.2, are implied by Assumptions 1-5-6-8 together with Lemma 1.9 and Lemma 1.10. In addition, Assumption 7 implies that Assumption 3 is verified with $\log \pi^{(1)} = (\partial/\partial\theta) \log \pi(\theta)_{|\theta=\theta_*}$. Hence all the conditions Theorem 1.3 are satisfied with $\delta_n = 1/\sqrt{n}$, proving the validity of the first part of corollary.

To deal with (1.19), recall that $K_n = \{h : \|h\| < M_n\} = \{\theta : \|\theta - \theta_*\| < M_n/\sqrt{n}\}$ and note that it is sufficient to prove the statement for $\|h\|^r$. Leveraging triangle inequality we split the problem in three parts

$$\begin{aligned}\int \|h\|^r |\pi_n(h) - p^n_{\text{SKS}}(h)| dh \leq & \int_{K_n^c} \|h\|^r \pi_n(h) dh + \int_{K_n^c} \|h\|^r p^n_{\text{SKS}}(h) dh \\ & + \int_{K_n} \|h\|^r |\pi_n(h) - p^n_{\text{SKS}}(h)| dh.\end{aligned} \tag{A.1}$$

Recall that, $A_{n,0} = \{\lambda_{\text{MIN}}(J_{\theta_*}/n) > \eta_1^*\} \cap \{\lambda_{\text{MAX}}(J_{\theta_*}/n) < \eta_2^*\}$, for some positive constants η_1^*, η_2^*, and $A_{n,1} = A_{n,0} \cap \{\|\xi\| < \tilde{M}_n\}$ for some $\tilde{M}_n$ going to infinity arbitrary slow. Note that from Assumptions 5-6, Lemma B.2 and Lemma 1.9 it follows $P_0^n A_{n,1} = 1 - o(1)$.

To bound the first element on the right-hand-side of the inequality we use that, from Assumptions 7-8 and equation (A.7) of Lemma 1.10 the event

$$A_{n,3} = A_{n,1} \cap \{ \sup_{\|\theta-\theta_*\|>M_n/\sqrt{n}} \{\ell(\theta)-\ell(\theta_*)\} < -c_5 M_n^2 \} \cap \{ \int_{K_n} e^{\ell(\theta_*+h/\sqrt{n})-\ell(\theta_*)} \pi(\theta_*+h/\sqrt{n}) dh > \tilde{c}_1 \},$$

where $\tilde{c}_1$ is an arbitrary small and fixed positive constant, satisfies $P_0^n A_{n,3} = 1 - o(1)$. By combining this fact with $\int \|h\|^r \pi(\theta_* + h/\sqrt{n}) dh < \infty$ and Jensen's inequality we obtain

$$\begin{aligned} \int_{K_n^c} \|h\|^r \pi_n(h) dh \mathbb{1}_{A_{n,3}} &\leq \int_{K_n^c} \|h\|^r \frac{e^{\ell(\theta_*+h/\sqrt{n})-\ell(\theta_*)} \pi(\theta_* + h/\sqrt{n})}{\int_{K_n} e^{\ell(\theta_*+g/\sqrt{n})-\ell(\theta_*)} \pi(\theta_* + g/\sqrt{n}) dg} dh \mathbb{1}_{A_{n,3}} \\ &\lesssim \frac{1}{n^{c_0 c_5}} \int \|h\|^r \pi(\theta_* + h/\sqrt{n}) = O(n^{-1}), \end{aligned}$$

for a sufficiently large choice of c_0 in M_n. Since $P_0^n A_{n,3} = 1 - o(1)$, this implies

$$\int_{K_n^c} \|h\|^r \pi_n(h) dh = O_{P_0^n}(n^{-1}). \tag{A.2}$$

Similarly the boundedness of $w(h-\xi)$ and the tail behavior of the Gaussian distribution give

$$\int_{K_n^c} \|h\|^r p^n_{\text{SKS}}(h) dh \mathbb{1}_{A_{n,1}} \leq 2 \int_{K_n^c} \|h\|^r \phi_d(h; \xi, \Omega) dh \mathbb{1}_{A_{n,1}} = O(n^{-1}),$$

for a sufficiently large choice of c_0. In turn, this implies

$$\int_{K_n^c} \|h\|^r p^n_{\text{SKS}}(h) dh = O_{P_0^n}(n^{-1}). \tag{A.3}$$

Finally, equation (1.18) gives

$$\int_{K_n} \|h\|^r \left| \pi_n(h) - p^n_{\text{SKS}}(h) \right| dh \leq M_n^r \int |\pi_n(h) - p^n_{\text{SKS}}(h)| dh = O_{P_0^n}(M_n^{c_6+r}/n). \tag{A.4}$$

Equation (1.19) follows by the combination of (A.2), (A.3) and (A.4). □

A.1.2 Proof of Lemma 1.9

Proof. The statement of the lemma easily follows from Assumptions 5-6. In fact, they allow to take, in $K_n = \{h : \|h\| \leq M_n\}$, the following Taylor expansion

$$\log \frac{p^n_{\theta_*+h/\sqrt{n}}}{p^n_{\theta_*}}(X^n) = h_s \frac{\ell^{(1)}_{\theta_*,s}}{\sqrt{n}} - \frac{1}{2} \frac{j_{st}}{n} h_s h_t + \frac{1}{6\sqrt{n}} \frac{\ell^{(3)}_{\theta_*,stl}}{n} h_s h_t h_l + r_{n,1}(h),$$

with $\ell^{(1)}_{\theta_*,s}/\sqrt{n} = O_{P_0^n}(1)$, $j_{st}/n = O_{P_0^n}(1)$, $\ell^{(3)}_{\theta_*,stl}/n = O_{P_0^n}(1)$ and

$$\sup_{h\in K_n} r_{n,1}(h) = \sup_{h\in K_n} \frac{1}{24n}\frac{\ell^{(4)}_{\theta_*,stlk}(\beta h)}{n} h_s h_t h_l h_k = O_{P_0^n}(M_n^4/n),$$

for some $\beta \in (0,1)$. To conclude, we need to check that the first term can be written as $h_s(j_{st}/n)\Delta^n_{\theta_*,t}$ with $\Delta^n_{\theta_*,t} = j_{st}^{-1}\sqrt{n}\ell^{(1)}_{\theta_*,s} = O_{P_0^n}(1)$. To this end, note that, in view of Assumption 6, Lemma B.2 implies that $\lambda_{\text{MAX}}(J_{\theta_*}/n)$ and $\lambda_{\text{MIN}}(J_{\theta_*}/n)$ are bounded from above and below, respectively, with probability tending to 1 as $n \to \infty$. Since, by the eigendecomposition (we assume the eigenvectors to be normalized) follows that the entries of $(J_{\theta_*}/n)^{-1}$ are bounded, in absolute value, by $d/\lambda_{\text{MIN}}(J_{\theta_*}/n)$, we get $nj_{st}^{-1} = O_{P_0^n}(1)$, which implies, in turn, $\Delta^n_{\theta_*,t} = O_{P_0^n}(1)$.

□

A.1.3 Proof of Lemma 1.10

Proof. To prove Lemma 1.10 we start by writing

$$\pi_n(K_n^c) \leq \frac{\int_{K_n^c} p^n_{\theta_*+h/\sqrt{n}}(X^n)\pi(\theta_*+h/\sqrt{n})dh}{\int_{K_n} p^n_{\theta_*+h/\sqrt{n}}(X^n)\pi(\theta_*+h/\sqrt{n})dh}. \tag{A.5}$$

Recall that under Assumption 8 it holds

$$\lim_{n\to\infty} P_0^n\{\sup_{\|h\|>M_n}\{\ell(\theta_*+h/\sqrt{n})-\ell(\theta_*)\} < -c_5 M_n^2\} = 1.$$

As a consequence, for every $D > 1$,

$$\int_{K_n^c} (p^n_{\theta_*+h/\sqrt{n}}/p^n_{\theta_*})(X^n)\pi(\theta_*+h/\sqrt{n})/\pi(\theta_*)dh = O_{P_0^n}(n^{-D}), \tag{A.6}$$

given a sufficiently large constant c_0 in M_n and the boundedness condition in Assumption 7. For the denominator of the right-hand-side of (A.5), we use Assumptions 5-6-7 to take the Taylor expansions reported in (1.16) and (1.20). Recall that from Assumption 6 and Lemma B.2 there exist two positive constants η_1^* and η_2^* such that $A_{n,0} = \{\lambda_{\text{MIN}}(V^n_{\theta_*}) > \eta_1^*\} \cap \{\lambda_{\text{MAX}}(V^n_{\theta_*}) < \eta_2^*\}$, holds with probability $P_0^n A_{n,0} = 1 - o(1)$. As a consequence, if we include the third order term in (1.20) and the prior effect in the remainder, we get

$$(p^n_{\theta_*+h/\sqrt{n}}/p^n_{\theta_*})(X^n)\pi(\theta_*+h/\sqrt{n})/\pi(\theta_*)/\mathbb{1}_{A_{n,0}} = \exp\left\{h_s v^n_{st}\Delta^n_{\theta_*,t} - \frac{1}{2}v^n_{st}h_s h_t + r_{n,5}(h)\right\}/\mathbb{1}_{A_{n,0}}$$
$$= \exp\left\{-\frac{1}{2}v^n_{st}(h-\Delta^n_{\theta_*})_s(h-\Delta^n_{\theta_*})_t + \gamma_n + r_{n,5}(h)\right\}/\mathbb{1}_{A_{n,0}},$$

with $\gamma_n = v^n_{st}\Delta^n_{\theta_*,s}\Delta^n_{\theta_*,t}/2 > 0$, since $V^n_{\theta_*}$ is positive definite when conditioned on $A_{n,0}$,

$$r_{n,5}(h) = r_{n,1}(h) + r_{n,2}(h) + \frac{1}{6\sqrt{n}}a^{(3),n}_{\theta_*,stl}h_s h_t h_l + \frac{1}{\sqrt{n}}\log \pi^{(1)}_{\theta_*,s}h_s,$$

and $r_{n,5} = \sup_{h\in K_n} r_{n,5}(h) = O_{P^n_0}(M^3_n/\sqrt{n})$.

Define the event $A_{n,4} = A_{n,0} \cap \{\|\Delta^n_{\theta_*}\| < \tilde{M}_n\} \cap \{|r_{n,5}| < \gamma_1\}$ for some $\tilde{M}_n$ going to infinity arbitrary slow and γ_1 a fixed positive constant. Since $P^n_0(A_{n,4}) = 1 - o(1)$ and $\gamma_n > 0$, we can equivalently study the asymptotic behavior of the following lower bound

$$\begin{aligned}
&\int_{K_n} (p^n_{\theta_*+h/\sqrt{n}}/p^n_{\theta_*})(X^n)\pi(\theta_* + h/\sqrt{n})/\pi(\theta_*)dh/\mathbb{1}_{A_{n,4}} \\
\geq &\exp(-\gamma_1)\int_{K_n} \exp\left\{-v^n_{st}(h-\Delta^n_{\theta_*})_s(h-\Delta^n_{\theta_*})_t/2\right\} dh/\mathbb{1}_{A_{n,4}} \\
= &\exp(-\gamma_1)(2\pi)^{d/2}|V^n_{\theta_*}|^{-1/2}\int_{K_n}(2\pi)^{-d/2}|V^n_{\theta_*}|^{1/2}\exp\left\{-v^n_{st}(h-\Delta^n_{\theta_*})_s(h-\Delta^n_{\theta_*})_t/2\right\} dh/\mathbb{1}_{A_{n,4}}.
\end{aligned} \tag{A.7}$$

Due to the fact that, in K_n, $M_n \to \infty$, $\Delta^n_{\theta_*} = O_{P^n_0}(1)$ and that, in $\mathbb{1}_{A_{n,4}}$, the eigenvalues of $V^n_{\theta_*}$ lay on a bounded and positive range, the quantity in the last display is positive and bounded away from zero. This, together with (A.6), gives the statement of Lemma 1.10. □

A.1.4 Proof of Lemma 1.11

Proof. To prove the statement of Lemma 1.11 we deal with the cases $\|\theta - \theta_*\| > \delta_1$ and $M_n/\sqrt{n} < \|\theta - \theta_*\| < \delta_1$ separately. For $\|\theta - \theta_*\| > \delta_1$ the claim trivially follows from (1.22). We are left to deal with the case $M_n/\sqrt{n} < \|\theta - \theta_*\| < \delta_1$. We write

$$\frac{1}{n}\{\ell(\theta) - \ell(\theta_*)\} = D_n(\theta, \theta_*) + \bar{D}_n(\theta, \theta_*).$$

where $D_n(\theta,\theta_*) = [\{\ell(\theta) - \ell(\theta_*)\} - \mathbb{E}^n_0\{\ell(\theta) - \ell(\theta_*)\}]/n$ and $\bar{D}_n(\theta,\theta_*) = \mathbb{E}^n_0\{\ell(\theta) - \ell(\theta_*)\}/n$. Note that Assumption R1 implies that $\bar{D}_n(\theta,\theta_*)$ is concave in $\|\theta - \theta_*\| < \delta_1$ with Hessian $-I_{\theta_*}/n$. Since I_{θ_*}/n is positive definite from Assumption 6, for sufficiently small choice of $\rho > 0$ it follows

$$\begin{aligned}
\bar{D}_n(\theta,\theta_*) \leq &-\rho i_{st}(\theta-\theta_*)_s(\theta-\theta_*)_t/n \\
\leq &-\rho\lambda_{\text{MIN}}(I_{\theta_*}/n)(\theta-\theta_*)_s(\theta-\theta_*)_s \leq -\rho\eta_1\|\theta-\theta_*\|^2.
\end{aligned}$$

In addition, define the event $A_{n,5} = \{\sup_{0<\|\theta-\theta_*\|<\delta_1} D_n(\theta,\theta_*)/\|\theta-\theta_*\| < \tilde{c}_1\tilde{M}_n/\sqrt{n}\}$ for a sufficiently large constant $\tilde{c}_1$ and a sequence $\tilde{M}_n$ which goes to infinity arbitrary slow. Note that $P^n_0(A_{n,5}) = 1 - o(1)$ from Assumption R2. As a consequence, conditioned on

$A_{n,5}$, we have for every θ such that $0 < \|\theta - \theta_*\| < \delta_1$ that

$$\begin{aligned} D_n(\theta, \theta_*) + \bar{D}_n(\theta, \theta_*) \leq & \tilde{c}_1 \|\theta - \theta_*\| \tilde{M}_n / \sqrt{n} - \rho \eta_1 \|\theta - \theta_*\|^2 \\ = & \{\tilde{c}_1 \tilde{M}_n / (\|\theta - \theta_*\| \sqrt{n}) - \rho \eta_1\} \|\theta - \theta_*\|^2. \end{aligned}$$

Since $\tilde{M}_n$ can be chosen such that $\tilde{M}_n / M_n \to 0$, the first component in the right-hand-side of the last display becomes always negative for n large enough and the whole expression is asymptotically maximized when $\|\theta - \theta_*\|$ is at its minimum. This implies

$$\sup_{M_n/\sqrt{n} < \|\theta - \theta_*\| < \delta_1} D_n(\theta, \theta_*) + \bar{D}_n(\theta, \theta_*) \leq -c_5 M_n^2 / n,$$

for some $c_5 > 0$, with P_0^n-probability tending to one. This concludes the proof of the lemma. □

A.2 Proofs main results of Chapter 2

This section collects the proof for Theorem 2.1 and Theorem 2.3 which were given in Chapter 2.The notation is the same as the main part of the thesis.

A.2.1 Proof of Theorem 2.1

Proof. The proof of Theorem 2.1 is closely related to the ones of Theorem 1.3 and Corollary 1.7. The main additional step is the need to condition on the event $B_n = \{\|\hat{\theta} - \theta_*\| \leq M_n / \sqrt{n}\}$ in order to deal with the set $\hat{K}_n = \{\theta : \sqrt{n}\|\theta - \hat{\theta}\| < 2M_n\}$ using what has been proved in Corollary 1.7 for $K_n = \{\theta : \sqrt{n}\|\theta - \theta_*\| < M_n\}$.

We start by splitting the problem in three parts

$$\begin{aligned} \int |\pi_n(\hat{h}) - 2\phi_d(\hat{h}; 0, \hat{\Omega})\hat{w}(\hat{h})| dh \leq & \int |\pi_n(\hat{h}) - \pi_n^{\hat{K}_n}(\hat{h})| d\hat{h} \\ & + \int |\pi_n^{\hat{K}_n}(\hat{h}) - 2\phi_d^{\hat{K}_n}(\hat{h}; 0, \hat{\Omega})\hat{w}(\hat{h})| d\hat{h} \\ & + \int |2\phi_d(\hat{h}; 0, \hat{\Omega})\hat{w}(\hat{h}) - 2\phi_d^{\hat{K}_n}(\hat{h}; 0, \hat{\Omega})\hat{w}(\hat{h})| d\hat{h}, \end{aligned} \tag{A.8}$$

where

$$\begin{aligned} \pi_n^{\hat{K}_n}(\hat{h}) &= \pi_n(\hat{h}) \mathbb{1}_{\hat{h} \in \hat{K}_n} / \int_{\hat{K}_n} \pi_n(\hat{h}) d\hat{h}, \\ 2\phi_d^{\hat{K}_n}(\hat{h}; 0, \hat{\Omega})\hat{w}(\hat{h}) &= 2\phi_d(\hat{h}; 0, \hat{\Omega})\hat{w}(\hat{h}) \mathbb{1}_{\hat{h} \in \hat{K}_n} / \int_{\hat{K}_n} 2\phi_d(\hat{h}; 0, \hat{\Omega})\hat{w}(\hat{h}) d\hat{h}, \end{aligned}$$

are versions of $\pi_n(\hat{h})$ and $2\phi_d(\hat{h};0,\hat{\Omega})\hat{w}(\hat{h})$ conditioned to $\hat{K}_n$.

A standard inequality of the total variation norm gives $\int |\pi_n(\hat{h}) - \pi_n^{\hat{K}_n}(\hat{h})|d\hat{h} \leq 2\Pi_n(\hat{K}_n^c)$. We show below that $\Pi_n(\hat{K}_n^c)$ is $O_{P_0^n}(n^{-1})$. To this end, note that, from assumption M1 and Markov's inequality it follows $P_0^n B_n = 1 - o(1)$. From triangle inequality and the definition of B_n we get

$$\Pi_n(\hat{K}_n^c)\mathbb{1}_{B_n} \leq \Pi_n(\{\theta : \|\theta - \theta_*\| > M_n/\sqrt{n}\})\mathbb{1}_{B_n}.$$

In view of Lemma 1.10 the right-hand-side of the previous display is of order $O_{P_0^n}(n^{-1})$ and, as a consequence, the same is true for $\Pi_n(\hat{K}_n^c)$. This implies

$$\int |\pi_n(\hat{h}) - \pi_n^{\hat{K}_n}(\hat{h})|d\hat{h} = O_{P_0^n}(n^{-1}), \tag{A.9}$$

for a sufficiently large choice of c_0 in M_n.

We deal with the third term in a similar manner. The same total variation inequality and the skew-symmetric invariance with respect to even functions (see e.g. Azzalini and Capitanio, 2014, Prop. 1.4) give

$$\begin{aligned}\int |2\phi_d(\hat{h};0,\hat{\Omega})\hat{w}(\hat{h}) - 2\phi_d^{\hat{K}_n}(\hat{h};0,\hat{\Omega})\hat{w}(\hat{h})|d\hat{h} \leq & 2\int_{\hat{h}\,:\,\|\hat{h}\|>2M_n} 2\phi_d(\hat{h};0,\hat{\Omega})\hat{w}(\hat{h})d\hat{h} \\ = & 2\int_{\hat{h}\,:\,\|\hat{h}\|>2M_n} \phi_d(\hat{h};0,\hat{\Omega})d\hat{h}.\end{aligned}$$

Let $P_{0,\hat{\Omega}}(\|\hat{h}\| > 2M_n) := \int_{\hat{h}\,:\,\|\hat{h}\|>2M_n} \phi_d(\hat{h};0,\hat{\Omega})dh$, and recall $\hat{A}_{n,0} = \{\lambda_{\text{MIN}}(\hat{\Omega}^{-1}) > \bar{\eta}_1\} \cap \{\lambda_{\text{MAX}}(\hat{\Omega}^{-1}) < \bar{\eta}_2\}$. For every $\epsilon > 0$, Assumption M2 implies

$$\begin{aligned}&P_0^n\Big(nP_{0,\hat{\Omega}}(\|\hat{h}\| > 2M_n) > \epsilon\Big) = P_0^n\Big(\{nP_{0,\hat{\Omega}}(\|\hat{h}\| > 2M_n) > \epsilon\} \cap \hat{A}_{n,0}\Big) + o(1) \\ &\leq P_0^n\Big(ne^{-\tilde{c}_1 M_n^2} > \epsilon | \hat{A}_{n,0}\Big) + o(1) = o(1),\end{aligned} \tag{A.10}$$

where $\tilde{c}_1$ is a sufficiently small positive constant and the last inequality follows from the tail behavior of the multivariate Gaussian for a sufficiently large choice of c_0 in $M_n = \sqrt{c_0 \log n}$. This gives

$$\int |2\phi_d(\hat{h};0,\hat{\Omega})\hat{w}(\hat{h}) - 2\phi_d^{\hat{K}_n}(\hat{h};0,\hat{\Omega})\hat{w}(\hat{h})|d\hat{h} = o_{P_0^n}(n^{-1}). \tag{A.11}$$

We are left to deal with $\int |\pi_n^{\hat{K}_n}(\hat{h}) - 2\phi_d^{\hat{K}_n}(\hat{h};0,\hat{\Omega})\hat{w}(\hat{h})|d\hat{h}$. Define

$$\hat{A}_{n,1} = \hat{A}_{n,0} \cap \{\int_{\hat{K}_n} \pi_n(\hat{h})d\hat{h} > 0\} \cap \{\int_{\hat{K}_n} 2\phi_d(\hat{h};0,\hat{\Omega})\hat{w}(\hat{h})d\hat{h} > 0\}.$$

Note that $P_0^n\{\int_{\hat{K}_n}\pi_n(\hat{h})d\hat{h}>0\}=1-o(1)$, by the order of $\Pi_n(\hat{K}_n^c)$ derived above (A.9), and $P_0^n\{\int_{\hat{K}_n}2\phi_d(\hat{h};0,\hat{\Omega})\hat{w}(\hat{h})d\hat{h}>0\}=1-o(1)$ by (A.10), which implies $P_0^n\hat{A}_{n,1}=1-o(1)$. As a consequence we can work similarly as in the proof of Theorem 1.3, use Lemma B.3 and $e^x=1+x+e^{\beta x}x^2/2$, for some $\beta\in(0,1)$, to obtain

$$\begin{aligned}&\int|\pi_n^{\hat{K}_n}(\hat{h})-2\phi_d^{\hat{K}_n}(\hat{h};0,\hat{\Omega})\hat{w}(\hat{h})|d\hat{h}\mathbb{1}_{\hat{A}_{n,1}}\\&\leq\int_{\hat{K}_n\times\hat{K}_n}\left|1-e^{\hat{r}_{n,4}(\hat{g})-\hat{r}_{n,4}(\hat{h})}\right|\pi_n^{\hat{K}_n}(\hat{h})2\phi_d^{\hat{K}_n}(\hat{g};0,\hat{\Omega})w(\hat{g})d\hat{h}d\hat{g}\mathbb{1}_{\hat{A}_{n,1}}\\&\leq 2|\hat{r}_{n,4}|+2\exp(2\beta|\hat{r}_{n,4}|)(\hat{r}_{n,4})^2=O_{P_0^n}(M_n^{c_8}/n),\end{aligned}\tag{A.12}$$

for $\hat{r}_{n,4}:=\sup_{\hat{h},\hat{g}\in\hat{K}_n}\hat{r}_{n,4}(h)$ and some constant c_8 defined in Lemma B.3. The result reported in Equation (2.2) is obtained by aggregating (A.9), (A.11) and (A.12).

It remains to deal with (2.3). Note that it is sufficient to prove the statement for $\|\hat{h}\|^r$. Using triangle inequality it is possible to split the problem in three parts

$$\begin{aligned}\int\|\hat{h}\|^r|\pi_n(\hat{h})-\hat{p}^n_{\text{SKS}}(\hat{h})|d\hat{h}\leq&\int_{\hat{K}_n^c}\|\hat{h}\|^r\pi_n(\hat{h})d\hat{h}+\int_{\hat{K}_n^c}\|\hat{h}\|^r\hat{p}^n_{\text{SKS}}(\hat{h})d\hat{h}\\&+\int_{\hat{K}_n}\|\hat{h}\|^r|\pi_n(\hat{h})-\hat{p}^n_{\text{SKS}}(\hat{h})|d\hat{h}.\end{aligned}\tag{A.13}$$

To bound the first element on the right-hand-side of the inequality we note that, from Assumption 8 and the fact that $P_0^nB_n^c=o(1)$ where $B_n=\{\|\hat{\theta}-\theta_*\|\leq M_n/\sqrt{n}\}$, it follows

$$\lim_{n\to\infty}P_0^n\{\sup_{\|\theta-\hat{\theta}\|>2M_n/\sqrt{n}}\frac{1}{n}\{\ell(\theta)-\ell(\hat{\theta})\}<-\frac{\tilde{c}_2M_n^2}{n}\}=1,$$

for some $\tilde{c}_2>0$. Moreover, conditioned on B_n, $K_n\subset\hat{K}_n$ which, combined with equation (A.7) of Lemma 1.10, implies that $\int e^{\ell(\hat{\theta}+\hat{h}/\sqrt{n})-\ell(\hat{\theta})}\pi(\hat{\theta}+\hat{h}/\sqrt{n})d\hat{h}>\int_{\hat{K}_n}e^{\ell(\hat{\theta}+\hat{h}/\sqrt{n})-\ell(\hat{\theta})}\pi(\hat{\theta}+\hat{h}/\sqrt{n})d\hat{h}$ is positive and bounded away from zero with P_0^n-probability tending to one. Let

$$\begin{aligned}\tilde{A}_{n,2}=&\tilde{A}_{n,1}\cap\{\int e^{\ell(\hat{\theta}+\hat{h}/\sqrt{n})-\ell(\hat{\theta})}\pi(\hat{\theta}+\hat{h}/\sqrt{n})d\hat{h}>\tilde{c}_2\}\\&\cap\{\sup_{\|\theta-\hat{\theta}\|>2M_n/\sqrt{n}}\{\ell(\theta)-\ell(\hat{\theta})\}/n<-\tilde{c}_2M_n^2/n\},\end{aligned}$$

and note that from Assumptions M1 and 8, together with $K_n^c\supset\hat{K}_n^c$, imply $P_0^n\tilde{A}_{n,2}=1-o(1)$, for some fixed small positive constant $\tilde{c}_2$. By combining these two facts with

$\int \|\hat{h}\|^r \pi(\hat{\theta} + \hat{h}/\sqrt{n}) d\hat{h} < \infty$ and Jensen's inequality we obtain

$$\begin{aligned}\int_{\hat{K}_n^c} \|\hat{h}\|^r \pi_n(\hat{h}) d\hat{h} \mathbb{1}_{\tilde{A}_{n,2}} &\leq \int_{\hat{K}_n^c} \|\hat{h}\|^r \frac{e^{\ell(\hat{\theta}+\hat{h}/\sqrt{n})-\ell(\hat{\theta})}\pi(\hat{\theta}+\hat{h}/\sqrt{n})}{\int_{\hat{K}_n} e^{\ell(\hat{\theta}+\hat{g}/\sqrt{n})-\ell(\hat{\theta})}\pi(\hat{\theta}+\hat{g}/\sqrt{n}) d\hat{g}} d\hat{h} \mathbb{1}_{\tilde{A}_{n,2}} \\ &\lesssim \frac{1}{n^{c_0 \tilde{c}_2}} \int \|\hat{h}\|^r \pi(\hat{\theta} + \hat{h}/\sqrt{n}) d\hat{h} \mathbb{1}_{\tilde{A}_{n,2}} = o(n^{-1}),\end{aligned}$$

for a sufficiently large choice of c_0. Since $P_0^n \tilde{A}_{n,2} = 1 - o(1)$, this implies

$$\int_{\hat{K}_n^c} \|\hat{h}\|^r \pi_n(\hat{h}) d\hat{h} = o_{P_0^n}(n^{-1}). \tag{A.14}$$

Similarly, conditioned on $\hat{A}_{n,0}$, the boundedness of $\hat{w}(\hat{h})$, together with the tail behavior of the Gaussian distribution implies

$$\int_{\hat{K}_n^c} \|\hat{h}\|^r \hat{p}^n_{\text{SKS}}(\hat{h}) d\hat{h} \mathbb{1}_{A_{n,0}} \leq 2 \int_{\hat{K}_n^c} \|\hat{h}\|^r \phi_d(\hat{h}; 0, \hat{\Omega}) d\hat{h} \mathbb{1}_{A_{n,0}} = o(n^{-1}),$$

for a sufficiently large choice of c_0. In turn, this implies

$$\int_{\hat{K}_n^c} \|\hat{h}\|^r \hat{p}^n_{\text{SKS}}(\hat{h}) d\hat{h} = o_{P_0^n}(n^{-1}). \tag{A.15}$$

Finally, (2.2) implies

$$\int_{\hat{K}_n} \|\hat{h}\|^r \left|\pi_n(\hat{h}) - \hat{p}^n_{\text{SKS}}(\hat{h})\right| d\hat{h} \leq (2M_n)^r \int |\pi_n(\hat{h}) - \hat{p}^n_{\text{SKS}}(\hat{h})| d\hat{h} = O_{P_0^n}(M_n^{c_8+r}/n), \tag{A.16}$$

where c_8 is defined in (2.2). Equation (2.3) follows by the combination of (A.14), (A.15) and (A.16). □

A.2.2 Proof of Theorem 2.3

Proof. Let $\hat{K}_{n,\mathcal{C}} = \{\hat{h}_{\mathcal{C}} : \|\hat{h}_{\mathcal{C}}\| < 2M_n\}$. The total variation distance between $\pi_{n,\mathcal{C}}(\hat{h}_{\mathcal{C}})$ and $\hat{p}^n_{\text{SKS},\mathcal{C}}$ is given by $(1/2)\int |\pi_{n,\mathcal{C}}(\hat{h}_{\mathcal{C}}) - p^n_{\text{SKS},\mathcal{C}}(\hat{h}_{\mathcal{C}})| dh_{\mathcal{C}}$. By adding and subtracting $\int p^n_{\text{SKS}}(\hat{h}) d\hat{h}_{\bar{\mathcal{C}}}$ and by exploiting Jensen's and triangle inequality, we obtain the following upper bound

$$\begin{aligned}\int |\pi_{n,\mathcal{C}}(\hat{h}_{\mathcal{C}}) - p^n_{\text{SKS},\mathcal{C}}(\hat{h}_{\mathcal{C}})| d\hat{h}_{\mathcal{C}} \leq &\int |\pi_n(\hat{h}) - p^n_{\text{SKS}}(\hat{h})| d\hat{h} + \\ &\int \Big| \int p^n_{\text{SKS}}(\hat{h}) d\hat{h}_{\bar{\mathcal{C}}} - p^n_{\text{SKS},\mathcal{C}}(\hat{h}_{\mathcal{C}}) \Big| d\hat{h}_{\mathcal{C}}.\end{aligned}$$

It follows from Theorem 2.1 that

$$\int |\pi_n(\hat{h}) - p^n_{\text{SKS}}(\hat{h})| d\hat{h} = O_{P_0^n}(M_n^{c_8}/n), \tag{A.17}$$

for some $c_8 > 0$. Therefore, it is sufficient to study $\int |\int p^n_{\text{SKS}}(\hat{h}) d\hat{h}_{\bar{\mathcal{C}}} - p^n_{\text{SKS},\mathcal{C}}(\hat{h}_{\mathcal{C}})| d\hat{h}_{\mathcal{C}}$. Note that, from (2.11), it follows that

$$\int p^n_{\text{SKS}}(\hat{h}) d\hat{h}_{\bar{\mathcal{C}}} - p^n_{\text{SKS},\mathcal{C}}(\hat{h}_{\mathcal{C}}) = 2\phi_{d_{\mathcal{C}}}(\hat{h}_{\mathcal{C}}; 0, \hat{\Omega}_{\mathcal{C}\mathcal{C}}) \mathbb{E}_{\hat{h}_{\bar{\mathcal{C}}}|\hat{h}_{\mathcal{C}}} \Big[F(\hat{\alpha}_\eta(\hat{h})) - F(\mathbb{E}_{\hat{h}_{\bar{\mathcal{C}}}|\hat{h}_{\mathcal{C}}}\{\hat{\alpha}_\eta(\hat{h})\}) \Big].$$

Let $C_{n,0} = \{\lambda_{\text{MIN}}(\hat{\Omega}_{\mathcal{C}\mathcal{C}}) > \eta_{1,\mathcal{C}}\} \cap \{\lambda_{\text{MAX}}(\hat{\Omega}_{\mathcal{C}\mathcal{C}}) < \eta_{2,\mathcal{C}}\}$ for some fixed $\eta_{1,\mathcal{C}}, \eta_{2,\mathcal{C}} > 0$. From Assumption M2 and Lemma B.4 it follows $P_0^n C_{n,0} = 1 - o(1)$. We condition on $C_{n,0}$, and we split the integral

$$\int \Big| 2\phi_{d_{\mathcal{C}}}(\hat{h}_{\mathcal{C}}; 0, \hat{\Omega}_{\mathcal{C}\mathcal{C}}) \mathbb{E}_{\hat{h}_{\bar{\mathcal{C}}}|\hat{h}_{\mathcal{C}}} \Big[F(\hat{\alpha}_\eta(\hat{h})) - F(\mathbb{E}_{\hat{h}_{\bar{\mathcal{C}}}|\hat{h}_{\mathcal{C}}}\{\hat{\alpha}_\eta(\hat{h})\}) \Big] \Big| d\hat{h}_{\mathcal{C}} \mathbb{1}_{C_{n,0}},$$

between $\hat{K}_{n,\mathcal{C}}$ and $\hat{K}^c_{n,\mathcal{C}}$. It follows from the boundedness of $\mathbb{E}_{\hat{h}_{\bar{\mathcal{C}}}|\hat{h}_{\mathcal{C}}} \Big[F(\hat{\alpha}_\eta(\hat{h})) - F(\mathbb{E}_{\hat{h}_{\bar{\mathcal{C}}}|\hat{h}_{\mathcal{C}}}\{\hat{\alpha}_\eta(\hat{h})\}) \Big]$ and from the tail behavior of the Gaussian distribution that

$$\int_{\hat{h}_{\mathcal{C}} \in \hat{K}^c_{n,\mathcal{C}}} \Big| 2\phi_{d_{\mathcal{C}}}(\hat{h}_{\mathcal{C}}; 0, \hat{\Omega}_{\mathcal{C}\mathcal{C}}) \mathbb{E}_{\hat{h}_{\bar{\mathcal{C}}}|\hat{h}_{\mathcal{C}}} \Big[F(\hat{\alpha}_\eta(\hat{h})) - F(\mathbb{E}_{\hat{h}_{\bar{\mathcal{C}}}|\hat{h}_{\mathcal{C}}}\{\hat{\alpha}_\eta(\hat{h})\}) \Big] \Big| d\hat{h}_{\mathcal{C}} \mathbb{1}_{C_{n,0}} \leq 4e^{-\tilde{c}_1 M_n^2},$$

for some constant $\tilde{c}_1 > 0$, which in turn implies that

$$\int_{\hat{h}_{\mathcal{C}} \in \hat{K}^c_{n,\mathcal{C}}} \Big| 2\phi_{d_{\mathcal{C}}}(\hat{h}_{\mathcal{C}}; 0, \hat{\Omega}_{\mathcal{C}\mathcal{C}}) \mathbb{E}_{\hat{h}_{\bar{\mathcal{C}}}|\hat{h}_{\mathcal{C}}} \Big[F(\hat{\alpha}_\eta(\hat{h})) - F(\mathbb{E}_{\hat{h}_{\bar{\mathcal{C}}}|\hat{h}_{\mathcal{C}}}\{\hat{\alpha}_\eta(\hat{h})\}) \Big] \Big| d\hat{h}_{\mathcal{C}} = O_{P_0^n}(n^{-1}), \tag{A.18}$$

for a sufficiently large constant c_0 in M_n. In addition, from Lemma B.5 it follows

$$\sup_{\hat{h}_{\mathcal{C}} \in \hat{K}_{n,\mathcal{C}}} \Big| \mathbb{E}_{\hat{h}_{\bar{\mathcal{C}}}|\hat{h}_{\mathcal{C}}} \Big[F(\hat{\alpha}_\eta(\hat{h})) - F(\mathbb{E}_{\hat{h}_{\bar{\mathcal{C}}}|\hat{h}_{\mathcal{C}}}\{\hat{\alpha}_\eta(\hat{h})\}) \Big] \Big| = O_{P_0^n}(M_n^{c_{10}}/n),$$

from some $c_{10} > 0$, which implies

$$\int_{\hat{h}_{\mathcal{C}} \in \hat{K}_{n,\mathcal{C}}} \Big| 2\phi_{d_{\mathcal{C}}}(\hat{h}_{\mathcal{C}}; 0, \hat{\Omega}_{\mathcal{C}\mathcal{C}}) \mathbb{E}_{\hat{h}_{\bar{\mathcal{C}}}|\hat{h}_{\mathcal{C}}} \Big[F(\alpha_\eta(\hat{h})) - F(\mathbb{E}_{\hat{h}_{\bar{\mathcal{C}}}|\hat{h}_{\mathcal{C}}}\{\alpha_\eta(\hat{h})\}) \Big] \Big| d\hat{h}_{\mathcal{C}} \mathbb{1}_{C_{n,0}} = O_{P_0^n}(M_n^{c_{10}}/n). \tag{A.19}$$

The combination of (A.17),(A.18) and (A.19) concludes the proof of the theorem with $c_9 = c_8 \vee c_{10}$. □

A.3 Proofs main results of Chapter 3

This section reports the proofs of the theoretical results introduced in Chapter 3. The notation is the same as the main part of the thesis.

A.3.1 Proof of Lemma 3.1

Proof. Without loss of generality consider the case $\tilde{\theta} = 0$. Moreover, to lighten the notation we drop $\tilde{\theta}$ from $\bar{\pi}_{n,\tilde{\theta}}(\theta)$,$f^*_{\tilde{\theta}}(\theta)$ and $f^*_{\tilde{\theta}}(\theta)$ and we write $\Theta = \Theta_+ \cup \Theta_-$ where for every $\theta \in \Theta_+$ it holds $-\theta \in \Theta_-$.

First, we deal with the case in which $\mathcal{D}[\cdot||\cdot] = \mathcal{D}_\alpha[\cdot||\cdot]$ is an α-divergence. Assume that α is fixed and $\alpha \notin \{0, 1\}$. By exploiting the skew-symmetric representation of the posterior distribution (3.3), $w^*(-\theta) = 1 - w^*(\theta)$, $0 \leq w^*(\theta) \leq 1$ and the symmetry of $\bar{\pi}_n(\theta)$ and $f^*(\theta)$ we obtain

$$\begin{aligned} &\mathcal{D}_\alpha[\pi_n(\theta) \,||\, f^*(\theta)] - \mathcal{D}_\alpha[\bar{\pi}_n(\theta) \,||\, f^*(\theta)] \\ =&\frac{1}{\alpha(1-\alpha)} \int \left[\bar{\pi}_n(\theta)^\alpha f^*(\theta)^{1-\alpha} - \{2\bar{\pi}_n(\theta) w^*(\theta)\}^\alpha f^*(\theta)^{1-\alpha}\right] d\theta \\ =&\int_{\Theta_+} \left[2 - \{2w^*(\theta)\}^\alpha - \{2(1-w^*(\theta))\}^\alpha\right] \bar{\pi}_n(\theta)^\alpha f^*(\theta)^{1-\alpha} d\theta. \end{aligned} \tag{A.20}$$

Now, note that the function $\kappa(k) = \left[2 - (2k)^\alpha - \{2(1-k)\}^\alpha\right] / \{\alpha(1-\alpha)\}$ has first and second derivative, with respect to k, equal to $2\{-(2k)^{\alpha-1} + \{2(1-k)\}^{\alpha-1}\}/(1-\alpha)$ and $4\{(2k)^{\alpha-1} + \{2(1-k)\}^{\alpha-1}\}$, respectively. For $k \in \left[0, 1\right]$, the fist derivative is equal to zero if and only if $k = 1 - k = 0.5$ while the second derivative is always positive. This implies that $k = 0.5$ is a point of minimum. Since $\kappa(0.5) = 0$, it follows from $\bar{\pi}_n(\theta)^\alpha > 0$ $f^*(\theta)^{1-\alpha} > 0$ that the last integral in (A.20) is always grater than or equal to zero, implying, in turn,

$$\mathcal{D}_\alpha[\bar{\pi}_n(\theta) \,||\, f^*(\theta)] \leq \mathcal{D}_\alpha[\pi_n(\theta) \,||\, f^*(\theta)], \tag{A.21}$$

for any $\alpha \notin 0, 1$. We are left to the cases $\alpha \to 1$ and $\alpha \to 0$.

In this regard, a useful fact is that, for two generic densities $p(\theta)$ and $q(\theta)$, $\lim_{\alpha\to 1} \mathcal{D}_\alpha[p(\theta) \,||\, q(\theta)] = \mathrm{KL}[p(\theta) \,||\, q(\theta)]$ while $\lim_{\alpha\to 0} \mathcal{D}_\alpha[p(\theta) \,||\, q(\theta)] = \mathrm{KL}[q(\theta) \,||\, p(\theta)]$ (see e.g., Cichocki and Amari, 2010). As a consequence, by exploiting again $w^*(-\theta) = 1 - w^*(\theta)$,

$0 \leq w^*(\theta) \leq 1$ and the symmetry of $\bar{\pi}_n(\theta)$ and $f^*(\theta)$ we get

$$\begin{aligned}
&\lim_{\alpha\to 1}\mathcal{D}_\alpha[\pi_n(\theta) \,||\, f^*(\theta)] - \mathcal{D}_\alpha[\bar{\pi}_n(\theta) \,||\, f^*(\theta)] \\
&\quad = \text{KL}[\pi_n(\theta) \,||\, f^*(\theta)] - \text{KL}[\bar{\pi}_n(\theta) \,||\, f^*(\theta)] \\
&\quad = \int \Big[2w^*(\theta)\log\big(2w^*(\theta)\big)\Big]\bar{\pi}_n(\theta)d\theta \\
&\quad = \int_{\Theta_+} \big[2w^*(\theta)\log\big(2w^*(\theta)\big) + 2(1-w^*(\theta))\log\big(2\{1-w^*(\theta)\}\big)\big]\bar{\pi}_n(\theta)d\theta.
\end{aligned} \tag{A.22}$$

The function $\kappa_1(k) = 2k\log\big(2k\big) + 2(1-k)\log\big(2\{1-k\}\big)$ has first and second derivative, with respect to k, equal to $2\{\log(k) - \log(\{1-k\})\}$ and $2(1/k + 1/(1-k))$, respectively. For $k \in \big[0,1\big]$, the first derivative is equal to zero when $k = 1-k = 0.5$ while the second derivative is always positive. As a consequence, $k = 0.5$ is a point of minimum for $\kappa_1(\cdot)$. Since $\kappa_1(0.5) = 0$ and $\bar{\pi}_n(\theta)^\alpha > 0$ the last line in (A.22) is always grater or equal then zero. As a result,

$$\text{KL}[\pi_n(\theta) \,||\, f^*(\theta)] - \text{KL}[\bar{\pi}_n(\theta) \,||\, f^*(\theta)] \geq 0. \tag{A.23}$$

Finally, for the case $\alpha \to 0$ we use again $w^*(-\theta) = 1 - w^*(\theta)$, $0 \leq w^*(\theta) \leq 1$ and the symmetry of $f^*(\theta)$ to obtain

$$\begin{aligned}
&\lim_{\alpha\to 0}\mathcal{D}_\alpha[\pi_n(\theta) \,||\, f^*(\theta)] - \mathcal{D}_\alpha[\bar{\pi}_n(\theta) \,||\, f^*(\theta)] \\
&\quad = \text{KL}[f^*(\theta) \,||\, \pi_n(\theta)] - \text{KL}[f^*(\theta) \,||\, \bar{\pi}_n(\theta)] \\
&\quad = -\int \log(2w^*(\theta))f^*(\theta)d\theta = -\int \big[\log(2w^*(\theta)) + \log(2\{1-w^*(\theta)\})\big]f^*(\theta)d\theta.
\end{aligned} \tag{A.24}$$

The function $\kappa_2(k) = -\big[\log(2k) + \log(2\{1-k\})\big]$, with $k \in [0,1]$, has first derivative equal to $-\{1/k - 1/(1-k)\}$ while the second derivative takes the form $\{1/k^2 + 1/(1-k)^2\}$. As in the previous cases, the first derivatives cancel-out at $k = 0.5$ while the second one is always positive, indicating that $\kappa_2(0.5) = 0$ is a point of minimum. This implies,

$$\text{KL}[f^*(\theta) \,||\, \pi_n(\theta)] - \text{KL}[f^*(\theta) \,||\, \bar{\pi}_n(\theta)] \geq 0. \tag{A.25}$$

The statement of the lemma regarding the α-divergences follows by aggregating (A.21),(A.23) and (A.25).

We are left to deal with the total variation. To this end it is sufficient to note that by the skew-symmetric representation of the posterior distribution (3.3), $w^*(-\theta) = 1 - w^*(\theta)$, $0 \leq w^*(\theta) \leq 1$ and the symmetry of $\bar{\pi}_n(\theta)$ and $f^*(\theta)$ it follows from the

triangle inequality

$$
\begin{aligned}
&2\Big(\mathcal{D}_{\mathrm{TV}}[\pi_n(\theta) \,||\, f^*(\theta)] - \mathcal{D}_{\mathrm{TV}}[\bar{\pi}_n(\theta) \,||\, f^*(\theta)]\Big) \\
=&\int \big|2\bar{\pi}_n(\theta)w^*(\theta) - f^*(\theta)\big|d\theta - \int \big|\bar{\pi}_n(\theta) - f^*(\theta)\big|d\theta \\
=&\int_{\Theta_+} \Big\{\big|2\bar{\pi}_n(\theta)w^*(\theta) - f^*(\theta)\big| + \big|2\bar{\pi}_n(\theta)(1-w^*(\theta)) - f^*(\theta)\big|\Big\}d\theta - \int \big|\bar{\pi}_n(\theta) - f^*(\theta)\big|d\theta \\
\geq& 2\int_{\Theta_+} \big|\bar{\pi}_n(\theta) - f^*(\theta)\big|d\theta - \int \big|\bar{\pi}_n(\theta) - f^*(\theta)\big|d\theta = 0.
\end{aligned}
$$

This concludes the proof of the lemma. □

A.3.2 Proof of Theorem 3.8

Proof. To prove Theorem 3.8, note that, in view of Lemma 3.1, it is sufficent to prove (3.7), as (3.6) directly follows from the combination of Lemma 3.1 and (3.7).

Without loss of generality consider the case $\tilde{\theta} = 0$. Moreover, to lighten the notation we drop $\tilde{\theta}$ from $\bar{\pi}_{n,\tilde{\theta}}(\theta)$,$f^*_{\tilde{\theta}}(\theta)$,$q^*_{\tilde{\theta}}(\theta)$ and $f^*_{\tilde{\theta}}(\theta)$. Let also, $\Theta = \Theta_+ \cup \Theta_-$ where for every $\theta \in \Theta_+$ it holds $-\theta \in \Theta_-$.

Consider the case in which $\mathcal{D}[\cdot||\cdot] = \mathcal{D}_{\mathrm{TV}}[\cdot||\cdot]$ is the total variation distance. By exploiting the skew-symmetric representation of the posterior distribution (3.3), $w^*(-\theta) = 1 - w^*(\theta)$, $0 \leq w^*(\theta) \leq 1$, the symmetry of $\bar{\pi}_n(\theta)$ and $f^*(\theta)$ we obtain

$$
\begin{aligned}
&\mathcal{D}_{\mathrm{TV}}[\pi_n(\theta) \,||\, q^*(\theta)] = \frac{1}{2}\int \big|2\bar{\pi}_n(\theta)w^*(\theta) - 2f^*(\theta)w^*(\theta)\big|d\theta \\
=&\frac{1}{2}\int_{\Theta_+} \Big\{2w^*(\theta)\big|\bar{\pi}_n(\theta) - f^*(\theta)\big| + 2(1-w^*(\theta))\big|\bar{\pi}_n(\theta) - f^*(\theta)\big|\Big\}d\theta \\
=&\int_{\Theta_+} \big|\bar{\pi}_n(\theta) - f^*(\theta)\big|d\theta = \mathcal{D}_{\mathrm{TV}}[\bar{\pi}_n(\theta) \,||\, f^*(\theta)].
\end{aligned}
\tag{A.26}
$$

This proves the validity of (3.7) for the total variation distance.

Now consider the case in which $\mathcal{D}[\cdot||\cdot] = \mathcal{D}_\alpha[\cdot||\cdot]$ is an α-divergence. From the definition of α-divergence and the same arguments as above it easily follows that

$$
\begin{aligned}
&\mathcal{D}_\alpha[\pi_n(\theta) \,||\, q^*(\theta)] = \frac{1}{\alpha(1-\alpha)}\Big\{1 - \int 2w^*(\theta)\bar{\pi}_n(\theta)^\alpha f^*(\theta)^{1-\alpha}d\theta\Big\} \\
=&\frac{1}{\alpha(1-\alpha)}\Big\{1 - 2\int_{\Theta_+} \bar{\pi}_n(\theta)^\alpha f^*(\theta)^{1-\alpha}d\theta\Big\} = \mathcal{D}_\alpha[\bar{\pi}_n(\theta) \,||\, f^*(\theta)].
\end{aligned}
\tag{A.27}
$$

Note that (A.27) holds also in the cases when $\alpha \to 0$ and $\alpha \to 1$. This because $\{2w^*(\theta)^\alpha\}\{2w^*(\theta)^{1-\alpha}\} = 2w^*(\theta)$ for every $\alpha \in [0,1]$. The combination of (A.26) and

(A.27) concludes the proof of the theorem. □

A.3.3 Proof of Theorem 3.9

Proof. Without loss of generality consider the case $\tilde{\theta} = 0$ and drop $\tilde{\theta}$ from $\bar{\pi}_{n,\tilde{\theta}}(\theta)$,$f^*_{\tilde{\theta}}(\theta)$, $q_{\tilde{\theta}}(\theta)$, $q^*_{\tilde{\theta}}(\theta)$ and $f^*_{\tilde{\theta}}(\theta)$. Recall also that $\Theta = \Theta_+ \cup \Theta_-$ where for every $\theta \in \Theta_+$ it holds $-\theta \in \Theta_-$.

In the following, we exploit the skew-symmetric representation of the posterior distribution $2\bar{\pi}_n(\theta)w^*(\theta)$ introduced in Proposition 3.4, as well as, the properties of the skewing factors $w(\theta)$ and $w^*(\theta)$.

For the total variation distance case, triangle inequality and equation (3.7) in Theorem 3.8 imply

$$\begin{aligned}
&\mathcal{D}_{\text{TV}}[\pi_n(\theta) \,||\, q(\theta)] = \frac{1}{2}\int \big|2\bar{\pi}_n(\theta)w^*(\theta) - 2f^*(\theta)w(\theta)\big|d\theta \\
=&\frac{1}{2}\int_{\Theta_+} \big|2\bar{\pi}_n(\theta)w^*(\theta) - 2f^*(\theta)w(\theta)\big| + \big|2\bar{\pi}_n(\theta)(1-w^*(\theta)) - 2f^*(\theta)(1-w(\theta))\big|d\theta \\
\geq&\frac{1}{2}\int \big|\bar{\pi}_n(\theta) - f^*(\theta)\big|d\theta = \mathcal{D}_{\text{TV}}[\pi_n(\theta) \,||\, q^*(\theta)].
\end{aligned} \tag{A.28}$$

For what concerns the α-divergences, consider first the case with fixed $\alpha \notin \{0,1\}$. From equation (3.7) of Theorem 3.8 it follows

$$\begin{aligned}
&\mathcal{D}_\alpha[\pi_n(\theta) \,||\, q(\theta)] - \mathcal{D}_\alpha[\pi_n(\theta) \,||\, q^*(\theta)] \\
=&\frac{1}{\alpha(1-\alpha)}\Big[\int \bar{\pi}_n(\theta)^\alpha f^*(\theta)^{1-\alpha}d\theta - \int \{2\bar{\pi}_n(\theta)w^*(\theta)\}^\alpha\{2f^*(\theta)w(\theta)\}^{1-\alpha}d\theta\Big] \\
=&\frac{2}{\alpha(1-\alpha)}\int_{\Theta_+}\big[1 - w^*(\theta)^\alpha w(\theta)^{1-\alpha} - \{1-w^*(\theta)\}^\alpha\{1-w(\theta)\}^{1-\alpha}\big]\bar{\pi}_n(\theta)^\alpha f^*(\theta)^{1-\alpha}d\theta.
\end{aligned}$$

We study the behavior of the function $\kappa_3(k,v) = \big[1-k^\alpha v^{1-\alpha} - \{1-k\}^\alpha\{1-v\}^{1-\alpha}\big]/\{\alpha(1-\alpha)\}$. To do it, fix $k \in (0,1)$ and denote with $\kappa_{3,k}(v)$ the function obtained from $\kappa_3(k,v)$ when maintaining k fixed. Note now that, for every $\alpha \notin \{0,1\}$

$$\lim_{v\to 0+}\kappa_{3,k}(v) \geq 0, \quad \text{and} \quad \lim_{v\to 1-}\kappa_{3,k}(v) \geq 0.$$

Moreover, $\kappa_{3,k}(v)$ is continuous in $(0,1)$ with first derivative $\big[-k^\alpha v^{-\alpha} + \{1-k\}^\alpha\{1-v\}^{-\alpha}\big]/\alpha$ which cancels out if and only if $k = v$ and second derivative $\big[k^\alpha v^{-(\alpha+1)} + \{1-k\}^\alpha\{1-v\}^{-(\alpha+1)}\big]$ which is always positive for $v \in (0,1)$. Since $\kappa_{3,k}(k) = 0$, this implies $\kappa_{3,k}(v) \geq 0$. To show that $\kappa_3(k,v) \geq 0$ for every fixed $\alpha \notin \{0,1\}$ we need to check also the behavior of $\kappa_3(k,v)$ at the boundary of its domain. To this end, note that the cases

in which both k and v are zero should be discarded as they represent points where all π_n, $q(\theta)$ and $q^*(\theta)$ have null mass. If $k \to 0+$ and v is fixed and bigger than 0, we get

$$\lim_{k\to 0+} \kappa_3(k,v) = \begin{cases} \geq 0 & \text{if} \quad \alpha \in (0,1), \\ +\infty & \text{if} \quad \alpha < 0, \\ \geq 0 & \text{if} \quad \alpha > 1, \end{cases}$$

and the same limits hold for $k \to 1-$ and $v \in (0,1)$, fixed. In addition, if $k \to 1-$ and $v \to 0+$

$$\lim_{k\to 1-, v\to 0+} \kappa_3(k,v) = \begin{cases} 1 & \text{if} \quad \alpha \in (0,1), \\ +\infty & \text{if} \quad \alpha < 0 \text{ or } \alpha > 1. \end{cases}$$

Finally, for $k \to 1-$, $v \to 1-$ and $\alpha \in (0,1)$ we get $\lim_{k\to 1-, v\to 1-} \kappa_3(k,v) = 0$. On the contrary, for the case $\alpha < 0$ or $\alpha > 1$ the limit does not necessary exists but $-(1-k)^\alpha(1-v)^{1-\alpha}/\{\alpha(1-\alpha)\}$ is always a non negative quantity. As a consequence, $\kappa_3(k,v) \geq (1-k^\alpha v^{1-\alpha})/\{\alpha(1-\alpha)\}$ with $\lim_{k\to 1-, v\to 1-}(1-k^\alpha v^{1-\alpha})/\{\alpha(1-\alpha)\} = 0$.

All these considerations imply, for fixed $\alpha \notin \{0,1\}$, that

$$\mathcal{D}_\alpha[\pi_n(\theta) \,||\, q(\theta)] \geq \mathcal{D}_\alpha[\pi_n(\theta) \,||\, q^*(\theta)]. \tag{A.29}$$

To conclude the proof of the theorem we need to study the limits $\alpha \to 0$ and $\alpha \to 1$. As in the proof of Theorem 3.8 we use the fact that this cases correspond to Kullback-Leibler divergences (see e.g. Cichocki and Amari, 2010). When $\alpha \to 1$

$$\begin{aligned} &\lim_{\alpha\to 1} \mathcal{D}_\alpha[\pi_n(\theta) \,||\, q(\theta)] = \textsc{kl}[\pi_n(\theta) \,||\, q(\theta)] \\ &= \int \log\Big(\frac{2\bar{\pi}_n(\theta)w^*(\theta)}{2f^*(\theta)w(\theta)}\Big) 2\bar{\pi}_n(\theta)w^*(\theta)d\theta \\ &= \int \log\Big(\frac{2\bar{\pi}_n(\theta)w^*(\theta)}{2\bar{\pi}_n(\theta)w(\theta)}\Big) 2\bar{\pi}_n(\theta)w^*(\theta)d\theta + \int \log\Big(\frac{2\bar{\pi}_n(\theta)w(\theta)}{2f^*(\theta)w(\theta)}\Big) 2\bar{\pi}_n(\theta)w^*(\theta)d\theta \\ &= \textsc{kl}[\pi_n(\theta) \,||\, 2\bar{\pi}_n(\theta)w(\theta)] + \textsc{kl}[\bar{\pi}_n(\theta) \,||\, f^*(\theta)] \\ &\geq \textsc{kl}[\bar{\pi}_n(\theta) \,||\, f^*(\theta)] = \textsc{kl}[\pi_n(\theta) \,||\, q^*(\theta)], \end{aligned} \tag{A.30}$$

with the second equality which follows by adding and subtracting $\int \log(2\bar{\pi}_n(\theta)w(\theta))$ $2\bar{\pi}_n(\theta)w^*(\theta)d\theta$, the third by the definition of KL-divergence, as well as, $w^*(-\theta) = 1 - w^*(\theta)$, the inequality from the positive definiteness of the KL-divergence and the last

equality form (3.7) of Theorem 3.8. The statement

$$\begin{aligned}&\lim_{\alpha\to 0}\mathcal{D}_\alpha[\pi_n(\theta)\,||\,q(\theta)] = \text{KL}[q(\theta)\,||\,\pi_n(\theta)]\\&\geq \text{KL}[f^*(\theta)\,||\,\bar{\pi}_n(\theta)] = \text{KL}[q^*(\theta)\,||\,\pi_n(\theta)],\end{aligned} \tag{A.31}$$

follows as in (A.30) with the role of $\pi_n(\theta)$ and $q(\theta)$ reversed. The combination of (A.28), (A.29),(A.30) and (A.31) completes the proof of the theorem. □

A.3.4 Proof of Theorem 3.12

Proof. To prove Theorem 3.12, we demonstrate the validity of (3.11), (3.10) follows in a similar manner but with less technical steps. For practical convenience, in view of the invariance of the total variation distance with respect to scale and location transformations, we proceed by reparameterizing with respect to $\hat{h} = \sqrt{n}(\theta - \hat{\theta})$. Note that, in this parametrization, the approximating density $q_{\tilde{\theta},2}(\hat{h})$ transforms to $2\phi_d(\hat{h};0,\tilde{\Omega})P(\hat{h})w_0(\hat{h})/\mathbb{E}_{0,\tilde{\Omega}}\{P(\hat{h})\}$ where $\tilde{\Omega}^{-1} = -(\ell^{(2)}_{\hat{\theta}} + \log\pi^{(2)}_{\hat{\theta}})/n$, $w_0(\hat{h}) = w_{\hat{\theta}}(\hat{\theta} + \hat{h}/\sqrt{n})$, $\mathbb{E}_{0,\tilde{\Omega}}\{P(\hat{h})\} = \int P(\hat{h})\phi_d(\hat{h};0,\tilde{\Omega})d\hat{h}$ and

$$P(\hat{h}) = 1 + \frac{\hat{a}^{(4)}_{\hat{\theta},stlk}\hat{h}_s\hat{h}_t\hat{h}_l\hat{h}_k}{24n} + \frac{1}{2}\Big(\frac{\hat{a}^{(4)}_{\hat{\theta},stlk}\hat{h}_s\hat{h}_t\hat{h}_l\hat{h}_k}{24n}\Big)^2 + \frac{1}{2}\Big(\frac{\hat{a}^{(3)}_{\hat{\theta},stl}\hat{h}_s\hat{h}_t\hat{h}_l}{6\sqrt{n}}\Big)^2,$$

with $\hat{a}^{(3)}_{\hat{\theta}} = \ell^{(3)}_{\hat{\theta}}/n$ and $\hat{a}^{(4)}_{\hat{\theta}} = \ell^{(4)}_{\hat{\theta}}/n$. In addition, define also the event $B_n = \{\|\hat{\theta} - \theta_*\| \leq M_n/\sqrt{n}\}$ and the sets $\hat{K}_n = \{\theta : \|\sqrt{n}(\theta - \hat{\theta})\| < 2M_n\}$ and $K_n = \{\theta : \|\sqrt{n}(\theta - \theta_*)\| < M_n\}$.

As a first step we split the problem in three parts

$$\begin{aligned}\int|\pi_n(\hat{h}) - 2\phi_d(\hat{h};0,\tilde{\Omega})P(\hat{h})w_0(\hat{h})|dh \leq &\int|\pi_n(\hat{h}) - \pi_n^{\hat{K}_n}(\hat{h})|d\hat{h}\\&+\int|\pi_n^{\hat{K}_n}(\hat{h}) - 2\phi_d^{\hat{K}_n}(\hat{h};0,\tilde{\Omega})P(\hat{h})w_0(\hat{h})|d\hat{h}\\&+\int|2\phi_d(\hat{h};0,\tilde{\Omega})P(\hat{h})w_0(\hat{h})/\mathbb{E}_{0,\tilde{\Omega}}\{P(\hat{h})\} - 2\phi_d^{\hat{K}_n}(\hat{h};0,\tilde{\Omega})P(\hat{h})w_0(\hat{h})|d\hat{h},\end{aligned} \tag{A.32}$$

where

$$\pi_n^{\hat{K}_n}(\hat{h}) = \pi_n(\hat{h})\mathbb{1}_{\hat{h}\in\hat{K}_n}/\int_{\hat{K}_n}\pi_n(\hat{h})d\hat{h},$$

$$2\phi_d^{\hat{K}_n}(\hat{h};0,\tilde{\Omega})P(\hat{h})w_0(\hat{h}) = 2\phi_d(\hat{h};0,\tilde{\Omega})P(\hat{h})w_0(\hat{h})\mathbb{1}_{\hat{h}\in\hat{K}_n}/\int_{\hat{K}_n}2\phi_d(\hat{h};0,\tilde{\Omega})P(\hat{h})w_0(\hat{h})d\hat{h},$$

are versions of $\pi_n(\hat{h})$ and $2\phi_d(\hat{h};0,\tilde{\Omega})P(\hat{h})w_0(\hat{h})$ conditioned to $\hat{K}_n$. Notice that, since

$w_0(-\hat{h}) = 1 - w_0(\hat{h})$ and that $\hat{K}_n$ is symmetric about 0, the normalizing constant of the skew-symmetric approximation can be equivalently rewritten as $\int_{\hat{K}_n} 2\phi_d(\hat{h}; 0, \tilde{\Omega})P(\hat{h})w_0(\hat{h})$ $d\hat{h} = \int_{\hat{K}_n} \phi_d(\hat{h}; 0, \tilde{\Omega})P(\hat{h})d\hat{h}$.

In view of a standard inequality of the total variation distance, we get $\int |\pi_n(\hat{h}) - \pi_n^{\hat{K}_n}(\hat{h})|d\hat{h} \leq 2\Pi_n(\hat{K}_n^c)$. Moreover note that, from assumption M1 and Markov's inequality it follows $P_0^n B_n = 1 - o(1)$. As a consequence, from triangle inequality and the definitions of B_n, K_n and $\hat{K}_n$ we get

$$\Pi_n(\hat{K}_n^c)\mathbb{1}_{B_n} \leq \Pi_n(K_n^c)\mathbb{1}_{B_n}.$$

From Lemma 1.10, with a sufficiently large choice of c_0 in M_n, the right-hand-side of the previous display is of order $O_{P_0^n}(n^{-2})$ and the same is true for $\Pi_n(\hat{K}_n^c)$, implying in turn,

$$\int |\pi_n(\hat{h}) - \pi_n^{\hat{K}_n}(\hat{h})|d\hat{h} = O_{P_0^n}(n^{-2}). \tag{A.33}$$

In order to deal with the third term in the right-hand-side of (A.32), we exploit the same total variation inequality. This, the symmetry of the set $\hat{K}_n^c$ with respect to 0 and the skew-symmetric invariance with respect to even functions (see e.g. Azzalini and Capitanio, 2014, Prop. 1.4) give

$$\begin{aligned}
&\int |2\phi_d(\hat{h}; 0, \tilde{\Omega})P(\hat{h})w_0(\hat{h})/\mathbb{E}_{0,\tilde{\Omega}}\{P(\hat{h})\} - 2\phi_d^{\hat{K}_n}(\hat{h}; 0, \tilde{\Omega})P(\hat{h})w_0(\hat{h})|d\hat{h} \\
\leq & 2\int_{\hat{h}\,:\,\|\hat{h}\|>2M_n} 2\phi_d(\hat{h}; 0, \tilde{\Omega})P(\hat{h})w_0(\hat{h})/\mathbb{E}_{0,\tilde{\Omega}}\{P(\hat{h})\}d\hat{h} \\
= & 2\int_{\hat{h}\,:\,\|\hat{h}\|>2M_n} \phi_d(\hat{h}; 0, \tilde{\Omega})P(\hat{h})/\mathbb{E}_{0,\tilde{\Omega}}\{P(\hat{h})\}d\hat{h}.
\end{aligned}$$

Recall that $\tilde{A}_{n,0} = \{\lambda_{\text{MIN}}(\tilde{\Omega}^{-1}) > \bar{\eta}_1\} \cap \{\lambda_{\text{MAX}}(\tilde{\Omega}^{-1}) < \bar{\eta}_2\}$. Conditioned on $\tilde{A}_{n,0}$, $\mathbb{E}_{0,\tilde{\Omega}}\{P(\hat{h})\}$ lies on a bounded positive range and, for n sufficiently large,

$$1 - \log\{P(\hat{h})\}/(\hat{h}^\intercal\tilde{\Omega}^{-1}\hat{h}/2) > 0.5,$$

uniformly in $\hat{h}$. As a consequence, for large n,

$$2\int_{\hat{h}\,:\,\|\hat{h}\|>2M_n} \phi_d(\hat{h}; 0, \tilde{\Omega})P(\hat{h})/\mathbb{E}_{0,\tilde{\Omega}}\{P(\hat{h})\}d\hat{h}\mathbb{1}_{\tilde{A}_{n,0}} \lesssim 2\int_{\hat{h}\,:\,\|\hat{h}\|>2M_n} \phi_d(\hat{h}; 0, 2\tilde{\Omega})d\hat{h}\mathbb{1}_{\tilde{A}_{n,0}}. \tag{A.34}$$

For every $\epsilon > 0$, Assumption M3, the tail behavior of the Gaussian distribution and sufficiently large choice of c_0 in $M_n = \sqrt{c_0 \log n}$ imply

$$\begin{aligned} &P_0^n\Big(n^2 P_{0,2\tilde{\Omega}}(\|\hat{h}\| > 2M_n) > \epsilon\Big) \;=\; P_0^n\Big(\{n^2 P_{0,2\tilde{\Omega}}(\|\hat{h}\| > 2M_n) > \epsilon\} \cap \tilde{A}_{n,0}\Big) + o(1) \\ &\leq P_0^n\Big(n^2 e^{-\tilde{c}_1 M_n^2} > \epsilon | \tilde{A}_{n,0}\Big) + o(1) = o(1), \end{aligned} \tag{A.35}$$

where $\tilde{c}_1$ is a sufficiently small positive constant. This gives

$$\int |2\phi_d(\hat{h}; 0, \tilde{\Omega}) P(\hat{h}) w_0(\hat{h}) / \mathbb{E}_{0,\tilde{\Omega}}\{P(\hat{h})\} - 2\phi_d^{\hat{K}_n}(\hat{h}; 0, \tilde{\Omega}) P(\hat{h}) w_0(\hat{h})| d\hat{h} = O_{P_0^n}(n^{-2}). \tag{A.36}$$

We are left to deal with $\int |\pi_n^{\hat{K}_n}(\hat{h}) - 2\phi_d^{\hat{K}_n}(\hat{h}; 0, \tilde{\Omega}) P(\hat{h}) w_0(\hat{h})| d\hat{h}$. To this end, define the event

$$\tilde{A}_{n,1} = \tilde{A}_{n,0} \cap \{\int_{\hat{K}_n} \pi_n(\hat{h}) d\hat{h} > 0\} \cap \{\int_{\hat{K}_n} 2\phi_d(\hat{h}; 0, \tilde{\Omega}) P(\hat{h}) w_0(\hat{h}) d\hat{h} > 0\}.$$

Note that $P_0^n\{\int_{\hat{K}_n} \pi_n(\hat{h}) d\hat{h} > 0\} = 1 - o(1)$, by the order of $\Pi_n(\hat{K}_n^c)$ derived above (A.33), and $P_0^n\{\int_{\hat{K}_n} 2\phi_d(\hat{h}; 0, \tilde{\Omega}) P(\hat{h}) w_0(\hat{h}) d\hat{h} > 0\} = 1 - o(1)$ by (A.34) and (A.35), implying $P_0^n \tilde{A}_{n,1} = 1 - o(1)$. To conclude the proof of the theorem, we express the posterior distribution in its skew-symmetric form $\pi_n(\hat{h}) = 2\bar{\pi}_{n,\hat{\theta}}(\hat{h}) w_0(\hat{h})$. In view of Theorem 3.8, equation (3.7), the problem can be rewritten only in terms of symmetric densities as

$$\begin{aligned} &\int \Big|2\bar{\pi}_{n,\hat{\theta}}^{\hat{K}_n}(\hat{h}) w_0(\hat{h}) - 2\phi_d^{\hat{K}_n}(\hat{h}; 0, \tilde{\Omega}) P(\hat{h}) w_0(\hat{h})\Big| d\hat{h} \\ &= \int \Big|\bar{\pi}_{n,\hat{\theta}}^{\hat{K}_n}(\hat{h}) - \phi_d^{\hat{K}_n}(\hat{h}; 0, \tilde{\Omega}) P(\hat{h})\Big| d\hat{h}, \end{aligned}$$

with

$$2\bar{\pi}_{n,\hat{\theta}}^{\hat{K}_n}(\hat{h}) w_0(\hat{h}) = 2\bar{\pi}_{n,\hat{\theta}}(\hat{h}) w_0(\hat{h}) \mathbb{1}_{\hat{h} \in \hat{K}_n} / \int_{\hat{K}_n} 2\bar{\pi}_{n,\hat{\theta}}(\hat{h}) w_0(\hat{h}) d\hat{h},$$

where, in view of the properties of $w_0(\cdot)$ and of the symmetry of $\hat{K}_n$, the normalizing constant can be rewritten as $\int_{\hat{K}_n} 2\bar{\pi}_{n,\hat{\theta}}^{\hat{K}_n}(\hat{h}) w_0(\hat{h}) d\hat{h} = \int_{\hat{K}_n} \bar{\pi}_{n,\hat{\theta}}^{\hat{K}_n}(\hat{h}) d\hat{h}$.

At this point, it is sufficient to proceed as in Theorem 1.3. We restrict our attention to

$$\begin{aligned} &\int_{\hat{K}_n} |\bar{\pi}_{n,\hat{\theta}}^{\hat{K}_n}(\hat{h}) - \phi_d^{\hat{K}_n}(\hat{h}; 0, \tilde{\Omega}) P(\hat{h})| d\hat{h} \mathbb{1}_{\tilde{A}_{n,1}} \\ &= \int_{\hat{K}_n} \Big|1 - \int_{\hat{K}_n} \frac{\phi_d^{\hat{K}_n}(\hat{h}; 0, \tilde{\Omega}) P(\hat{h})}{\phi_d^{\hat{K}_n}(\hat{g}; 0, \tilde{\Omega}) P(\hat{g})} \frac{\bar{\pi}_{n,\hat{\theta}}^{\hat{K}_n}(\hat{g})}{\bar{\pi}_{n,\hat{\theta}}^{\hat{K}_n}(\hat{h})} \phi_d^{\hat{K}_n}(\hat{g}; 0, \tilde{\Omega}) P(\hat{g}) d\hat{g}\Big| \bar{\pi}_{n,\hat{\theta}}^{\hat{K}_n}(\hat{h}) d\hat{h} \mathbb{1}_{\tilde{A}_{n,1}}. \end{aligned} \tag{A.37}$$

where the ratios $\phi_d^{\hat{K}_n}(\hat{h}; 0, \tilde{\Omega}) P(\hat{h}) / \{\phi_d^{\hat{K}_n}(\hat{g}; 0, \tilde{\Omega}) P(\hat{g})\}$ and $\bar{\pi}_{n,\hat{\theta}}^{\hat{K}_n}(\hat{h}) / \bar{\pi}_{n,\hat{\theta}}^{\hat{K}_n}(\hat{g})$ correspond

to their unconditioned version $\phi_d(\hat{h};0,\tilde{\Omega})/\phi_d(\hat{g};0,\tilde{\Omega})$ and $\bar{\pi}_{n,\hat{\theta}}(\hat{h})/\bar{\pi}_{n,\hat{\theta}}(\hat{g})$, for $\hat{h},\hat{g}\in\hat{K}_n$, respectively. This fact and an application of Jensen inequality implies that the quantity in the last display is upper bounded by

$$\int_{\hat{K}_n\times\hat{K}_n}\Big|1-\frac{\phi_d(\hat{h};0,\tilde{\Omega})P(\hat{h})}{\phi_d(\hat{g};0,\tilde{\Omega})P(\hat{g})}\frac{\bar{\pi}_{n,\hat{\theta}}(\hat{g})}{\bar{\pi}_{n,\hat{\theta}}(\hat{h})}\Big|\bar{\pi}_{n,\hat{\theta}}^{\hat{K}_n}(\hat{h})\phi_d^{K_n}(\hat{g};0,\tilde{\Omega})P(\hat{g})d\hat{h}d\hat{g}\mathbb{1}_{\tilde{A}_{n,1}}.$$

Now, recall that $\bar{\pi}_{n,\hat{\theta}}(\hat{h})$ is proportional to

$$\frac{\pi(\hat{\theta}+\hat{h}/\sqrt{n})L(\hat{\theta}+\hat{h}/\sqrt{n};X^n)+\pi(\hat{\theta}-\hat{h}/\sqrt{n})L(\hat{\theta}-\hat{h}/\sqrt{n};X^n)}{2\pi(\hat{\theta})L(\hat{\theta};X^n)}.$$

Therefore, in view of Lemma B.6, the definitions of $\tilde{\Omega}$ and $P(\hat{h})$, and $e^x=1+x+e^{\beta x}x^2/2$, for some $\beta\in(0,1)$ we obtain

$$\begin{aligned}
&\int_{\hat{K}_n}|\bar{\pi}_{n,\hat{\theta}}^{\hat{K}_n}(\hat{h})-\phi_d^{\hat{K}_n}(\hat{h};0,\tilde{\Omega})P(\hat{h})|d\hat{h}\mathbb{1}_{\tilde{A}_{n,1}}\\
&\leq\int_{\hat{K}_n\times\hat{K}_n}\Big|1-e^{\hat{r}_{n,6}(\hat{h})-\hat{r}_{n,6}(\hat{g})}\Big|\,\bar{\pi}_{n,\hat{\theta}}^{\hat{K}_n}(\hat{h})\phi_d^{K_n}(\hat{g};0,\tilde{\Omega})P(\hat{g})d\hat{h}d\hat{g}\mathbb{1}_{\tilde{A}_{n,1}} \qquad (A.38)\\
&\leq 2\hat{r}_{n,6}+2\exp(2\beta\hat{r}_{n,6})\hat{r}_{n,6}^2=O_{P_0^n}(M_n^{c_{12}}/n^2),
\end{aligned}$$

where $\hat{r}_{n,6}=\sup_{\hat{h}\in\hat{K}_n}|\hat{r}_{n,6}(\hat{h})|$ and c_{12} is some constant defined in Lemma B.6.

Equation (3.11) of Theorem 3.12 is proved by combining (A.33), (A.36) and (A.38)

□

Appendix B

Additional results for Chapters 1 - 2 and 3

B.1 Technical lemmas

We collect in the section the technical lemmas applied in the proofs of our main results.

Lemma B.1. *Let F be the cdf of a univariate random variable on $\mathbb{R}$ such that $F(-x) = 1-F(x)$, $F(0) = 1/2$ and $F(x) = F(0)+\eta x+O(x^2)$ for some $\eta \in \mathbb{R}$. Under Assumptions 2 and 3 it follows that*

$$\begin{aligned} &\log \frac{p^n_{\theta_*+\delta_n h}}{p^n_{\theta_*}}(X^n)\frac{\pi(\theta_* + \delta_n h)}{\pi(\theta_*)} + (h-\xi)_s\Omega^{-1}_{st}(h-\xi)_t/2 \\ &- \log 2w(h) + \delta = r_{n,4}(h), \end{aligned} \tag{B.1}$$

where δ is a constant not depending on h, $\xi = \Delta^n_{\theta_*} + \delta_n(V^n_{\theta_*})^{-1}\log\pi^{(1)}$, $\Omega^{-1} = [v^n_{st} - a^{(3),n}_{\theta_*,stl}\xi_l]$. As a consequence

$$\alpha_\eta(h) = \frac{\delta_n}{12\eta}\{\Psi^{(3)}_{stl}(h-\xi)_s(h-\xi)_t(h-\xi)_l + 3\Psi^{(1)}_s(h-\xi)_s\},$$

with $\Psi^{(1)} = [a^{(3),n}_{\theta_*,stl}\xi_t\xi_l]$, $\Psi^{(3)} = [a^{(3),n}_{\theta_*,stl}]$.

Moreover,

$$r_{n,4} := \sup_{h\in K_n} r_{n,4}(h) = O_{P^n_0}(\delta^2_n M^{c_3}_n), \tag{B.2}$$

for some constant $c_3 > 0$.

Proof. We start by noting that Assumptions 2 and 3 imply

$$\begin{aligned}&\log\frac{p^n_{\theta_*+\delta_n h}}{p^n_{\theta_*}}(X^n)\frac{\pi(\theta_*+\delta_n h)}{\pi(\theta_*)} - h_s v^n_{st}\Delta^n_{\theta_*,t} - \delta_n h_s \log\pi^{(1)}_s \\ &+ \frac{1}{2}h_s v^n_{st}h_t - \frac{\delta_n}{6}a^{(3),n}_{\theta_*,stl}h_s h_t h_l = r_{n,1}(h) + r_{n,2}(h).\end{aligned} \tag{B.3}$$

Furthermore, note that

$$h_s v^n_{st}\Delta^n_{\theta_*,t} + \delta_n h_s \log\pi^{(1)}_s - \frac{1}{2}h_s v^n_{st}h_t = -(h-\xi)_s v^n_{st}(h-\xi)_t/2 + \delta_1, \tag{B.4}$$

where $\xi = \Delta^n_{\theta_*} + \delta_n (V^n_{\theta_*})^{-1}\log\pi^{(1)}$ and δ_1 is a quantity not depending on h.

Second, we add and subtract ξ from h in the three dimensional array part, obtaining

$$\begin{aligned}\frac{\delta_n}{6}a^{(3),n}_{\theta_*,stl}h_s h_t h_l &= \frac{\delta_n}{6}a^{(3),n}_{\theta_*,stl}(h-\xi)_s(h-\xi)_t(h-\xi)_l + \frac{3\delta_n}{6}a^{(3),n}_{stl}(h-\xi)_s\xi_t\xi_l \\ &+ \frac{3}{6}\delta_n a^{(3),n}_{\theta_*,stl}(h-\xi)_s(h-\xi)_t\xi_l + \delta_2,\end{aligned} \tag{B.5}$$

where δ_2 does not depend on h.

By combining (B.4) and (B.5) it is possible to rewrite (B.3) as

$$\begin{aligned}&\log\frac{p^n_{\theta_*+\delta_n h}}{p^n_{\theta_*}}(X^n)\frac{\pi(\theta_*+\delta_n h)}{\pi(\theta_*)} + (h-\xi)_s\Omega^{-1}_{st}(h-\xi)_t/2 \\ &- \frac{\delta_n}{6}\Psi^{(3)}_{stl}(h-\xi)_s(h-\xi)_t(h-\xi)_l - \frac{3\delta_n}{6}\Psi^{(1)}_s(h-\xi)_s + \delta \\ &= r_{n,1}(h) + r_{n,2}(h),\end{aligned} \tag{B.6}$$

where $\Omega^{-1} = V^n_{\theta_*} - \delta_n\Psi^{(2)}$, $\Psi^{(3)} = [a^{(3),n}_{\theta_*,stl}]$, $\Psi^{(2)} = [a^{(3),n}_{\theta_*,stl}\xi_l]$, $\Psi^{(1)} = [a^{(3),n}_{\theta_*,stl}\xi_t\xi_l]$ and $\delta = -(\delta_1+\delta_2)$.

To conclude the proof of the lemma note that, Assumption 2, the fact that the parameter dimension d is fixed and Cauchy-Schwarz inequality imply

$$\frac{\delta_n}{6}\Psi^{(3)}_{stl}(h-\xi)_s(h-\xi)_t(h-\xi)_l + \frac{3\delta_n}{6}\Psi^{(1)}_s(h-\xi)_s = O_{P^n_0}(\delta_n\{\|h\|^3\vee 1\}).$$

By exploiting the conditions imposed on $F(\cdot)$ we write,

$$\begin{aligned}&F\left(\frac{\delta_n}{6}\left\{\Psi^{(3)}_{stl}(h-\xi)_s(h-\xi)_t(h-\xi)_l + 3\Psi^{(1)}_s(h-\xi)_s\right\}\right) \\ =&\frac{1}{2}\left[1 + 2\eta\frac{\delta_n}{6}\left\{\Psi^{(3)}_{stl}(h-\xi)_s(h-\xi)_t(h-\xi)_l + 3\Psi^{(1)}_s(h-\xi)_s\right\} + O_{P^n_0}(\delta^2_n\{\|h\|^6\vee 1\})\right],\end{aligned}$$

since the argument of $F(\cdot)$ converges to zero in probability. An additional Taylor expansion, this time $\log(1+x) = x + O(x^2)$ for $x \to 0$, gives

$$\begin{aligned}
&\log 2F\left(\frac{\delta_n}{12\eta}\{\Psi^{(3)}_{stl}(h-\xi)_s(h-\xi)_t(h-\xi)_l + 3a^{(3)}_{stl}(h-\xi)_s\xi_t\xi_l\}\right)\\
&= \log\left(1 + \frac{\delta_n}{6}\{\Psi^{(3)}_{stl}(h-\xi)_s(h-\xi)_t(h-\xi)_l + 3\Psi^{(1)}_{s}(h-\xi)_s\} + r_1^{**}(h)\right)\\
&= \frac{\delta_n}{6}\{\Psi^{(3)}_{stl}(h-\xi)_s(h-\xi)_t(h-\xi)_l + 3\Psi^{(1)}_{s}(h-\xi)_s\} + \tilde{r}_{n,1}(h),
\end{aligned} \tag{B.7}$$

where the remainder term $\tilde{r}_{n,1}(h)$ is $O_{P_0^n}(\delta_n^2\{\|h\|^6 \vee 1\})$. Note that, when restricted on K_n, $\tilde{r}_{n,1} = \sup_{h\in K_n}\tilde{r}_{n,1}(h) = O_{P_0^n}(\delta_n^2 M_n^6)$. This fact combined with (B.6) and (B.7) gives

$$\begin{aligned}
&\log\frac{p_{\theta_*+\delta_n h}}{p_{\theta_*}}(X^n)\frac{\pi(\theta_*+\delta_n h)}{\pi(\theta_*)} + (h-\xi)_s\Omega^{-1}_{st}(h-\xi)_t/2\\
&-\log 2F\Big(\frac{\delta_n}{12\eta}\{\Psi^{(1)}_{s}(h-\xi)_s + \Psi^{(3)}_{stl}(h-\xi)_s(h-\xi)_t(h-\xi)_l\}\Big) + \delta = r_{n,4}(h),
\end{aligned}$$

where $r_{n,4}(h) = r_{n,1}(h) + r_{n,2}(h) - \tilde{r}_{n,1}(h)$. Then, Assumptions 2, 3 imply that

$$\sup_{h\in K_n} r_{n,4}(h) = O_{P_0^n}(\delta_n^2 M_n^{c_3}), \tag{B.8}$$

for some constant $c_3 > 0$, concluding the proof. □

Lemma B.2. *Let A and $\hat{A}$ be two $d \times d$ real symmetric matrices. Suppose that the entries of $\hat{A}$ are random and satisfy $a_{st} = O(1)$, $\hat{a}_{st} = O_{P_0^n}(1)$ and*

$$a_{st} - \hat{a}_{st} = O_{P_0^n}(\delta_n), \tag{B.9}$$

for some norming rate $\delta_n \to 0$ and $s,t \in \{1,\dots,d\}$. If there exist two positive constants η_1 and η_2 such that $\lambda_{\text{MIN}}(A) > \eta_1$ and $\lambda_{\text{MAX}}(A) < \eta_2$ then, with P_0^n-probability tending to 1, there exist two positive constants η_1^* and η_2^* such that $\lambda_{\text{MIN}}(\hat{A}) > \eta_1^*$ and $\lambda_{\text{MAX}}(\hat{A}) < \eta_2^*$.

Proof. Leveraging (B.9), let us first notice that $\hat{A} = A + R$ with R having entries of order $\mathcal{O}_{P_0^n}(\delta_n)$. As a consequence, there exist constants $\tilde{c}_1 > 0$ and $\tilde{c}_2 > 1$ such that

$$P_0^n(|R_{st}| > \tilde{c}_1\delta_n^{\tilde{c}_2}) = o(1),$$

for each $s,t = 1,\dots,d$. Define now the matrix M with entries $M_{st} = |R_{st}| \wedge \tilde{c}_1\delta_n^{\tilde{c}_2}$ for $s,t = 1,\dots,d$. From Wielandt's theorem (Zwillinger and Jeffrey, 2007), with probability $1 - o(1)$, the spectral radius of M is an upper bound of the spectral radius of R.

Moreover, since M is a non-negative matrix, the Perron-Frobenius theorem (Perron, 1907; Frobenius, 1912) implies that the largest eigenvalue in absolute value is bounded by constant times $\delta_n^{\tilde{c}_2}$. Since both A and R are real symmetric matrices, in view of the Weyl's inequalities (e.g., Tao, 2011, equation (1.54)), the eigenvalues of $\hat{A}$ and A can differ at most by constant times $\delta_n^{\tilde{c}_2}$ with probability $1 - o(1)$. As a consequence, since the lemma assumes the existence of two positive constants η_1 and η_2 such that $\lambda_{\text{MIN}}(A) > \eta_1$ and $\lambda_{\text{MAX}}(A) < \eta_2$, it follows that there exist $\eta_1^*, \eta_2^* > 0$ such that, with probability $1 - o(1)$, $\lambda_{\text{MIN}}(\hat{A}) > \eta_1^*$ and $\lambda_{\text{MAX}}(\hat{A}) < \eta_2^*$. □

Lemma B.3. *Let $F(\cdot)$ be any cdf of a univariate random variable on $\mathbb{R}$ such that $F(-x) = 1 - F(x)$, $F(0) = 1/2$ and $F(x) = F(0) + \eta x + O(x^2)$.*

Under assumption M2 it follows that

$$\log \frac{p^n_{\hat{\theta}+\hat{h}/\sqrt{n}}}{p^n_{\hat{\theta}}}(X^n)\frac{\pi(\hat{\theta}+\hat{h}/\sqrt{n})}{\pi(\hat{\theta})} + \hat{\Omega}^{-1}_{st}\hat{h}_s\hat{h}_t/2 - \log 2\hat{w}(\hat{h}) = \hat{r}_{n,4}(\hat{h}), \tag{B.10}$$

where $\hat{\Omega}^{-1} = [V^n_{\hat{\theta},st}]$, the skewing factor $\hat{w}(\hat{h})$ depends on $\hat{\alpha}_\eta(\hat{h}) = \hat{a}^{(3),n}_{\hat{\theta},stl}\hat{h}_s\hat{h}_t\hat{h}_l/(12\eta\sqrt{n})$ and $\hat{a}^{(3),n}_{\hat{\theta},stl} = \ell^{(3)}_{\hat{\theta}}/n$.

Moreover,

$$\hat{r}_{n,4} := \sup_{\hat{h}\in\hat{K}_n} \hat{r}_{n,4}(\hat{h}) = O_{P_0^n}(M_n^{c_8}/n), \tag{B.11}$$

with $\hat{K}_n = \{\hat{h} : \|\hat{h}\| < 2M_n\}$, for some constant $c_8 > 0$.

Proof. Let $\hat{h}$ be fixed. Note that at $\hat{\theta}$ the first log-posterior derivative is null by definition. As a consequence, from Assumption M2 it follows that the third order Taylor expansion of the log-posterior takes the form

$$\log \frac{p_{\hat{\theta}+\hat{h}/\sqrt{n}}}{p_{\hat{\theta}}}(X^n)\frac{\pi(\hat{\theta}+\hat{h}/\sqrt{n})}{\pi(\hat{\theta})} = -\hat{h}_s\hat{\Omega}^{-1}_{st}\hat{h}_t/2 + \frac{1}{6\sqrt{n}}\frac{\ell^{(3)}_{\hat{\theta},stl}}{n}\hat{h}_s\hat{h}_t\hat{h}_l + O_{P_0^n}(\{\|\hat{h}\|^4 \vee 1\}/n),$$

with

$$\frac{1}{6\sqrt{n}}\frac{\ell^{(3)}_{\hat{\theta},stl}}{n}\hat{h}_s\hat{h}_t\hat{h}_l = O_{P_0^n}(\{\|h\|^3 \vee 1\}/\sqrt{n}).$$

This fact, together with $F(x) = 1/2 + \eta x + O(x^2)$ and $\log(1+x) = x + O(x^2)$ implies

$$\begin{aligned}\log\{2F(\hat{\alpha}_\eta(\hat{h}))\} =& \log\left\{1 + \frac{1}{6\sqrt{n}}\frac{\ell^{(3)}_{\hat{\theta},stl}}{n}\hat{h}_s\hat{h}_t\hat{h}_l + O_{P_0^n}(\{\|\hat{h}\|^6 \vee 1\}/n)\right\}\\ =& \frac{1}{6\sqrt{n}}\frac{\ell^{(3)}_{\hat{\theta},stl}}{n}\hat{h}_s\hat{h}_t\hat{h}_l + O_{P_0^n}(\{\|\hat{h}\|^6 \vee 1\}/n),\end{aligned}$$

By combining all the pieces together we obtain

$$\hat{r}_{n,4}(\hat{h}) = \log \frac{p_{\hat{\theta}+\hat{h}/\sqrt{n}}}{p_{\hat{\theta}}}(X^n)\frac{\pi(\hat{\theta}+\hat{h}/\sqrt{n})}{\pi(\hat{\theta})} + \hat{h}_s\hat{\Omega}_{st}^{-1}\hat{h}_t/2 - \log\{2\hat{w}(\hat{h})\}, \tag{B.12}$$

where $\hat{r}_{n,4}(\hat{h}) = O_{P_0^n}(\{\|\hat{h}\|^6 \vee 1\}/n)$ and hence $\hat{r}_{n,4} = O_{P_0^n}(M_n^{c_8}/n)$, with $c_8 = 6$. □

Lemma B.4. *Let A be a $d \times d$ symmetric positive definite matrix satisfying $\lambda_{\text{MIN}}(A) \geq \eta_{1A}$ and $\lambda_{\text{MAX}}(A) \leq \eta_{2A}$. Let $\mathcal{S} \subseteq \{1, \dots d\}$ be a set of indexes having cardinality d_*, and B be the $d_* \times d_*$ submatrix obtained by keeping only rows and columns of A whose position is in $\mathcal{S}$. Then it holds that $\lambda_{\text{MIN}}(B) \geq \eta_{1A}$ and $\lambda_{\text{MAX}}(B) \leq \eta_{2A}$.*

Proof. Without loss of generality assume that the elements of $\mathcal{S}$ are increasing order. Note, also, that $B = SAS^\intercal$ where S is a $d_* \times d$ matrix having entries $s_{ij} = 1$ if $j = \mathcal{S}_i$, where $\mathcal{S}_i$ denotes the i-th element of $\mathcal{S}$, and $s_{ij} = 0$ otherwise. Recall that, from the relation between minimum and maximum eigenvalues and the Rayleigh quotient it follows that

$$\lambda_{\text{MIN}}(A) = \min_{x \in \mathbb{R}^d} \frac{x^\intercal A x}{x^\intercal x}, \quad \text{and} \quad \lambda_{\text{MAX}}(A) = \max_{x \in \mathbb{R}^d} \frac{x^\intercal A x}{x^\intercal x}.$$

Similarly, leveraging $SS^\intercal = I_{d_*}$, it holds for B that

$$\lambda_{\text{MIN}}(B) = \min_{x_* \in \mathbb{R}^{d_*}} \frac{x_*^\intercal B x_*}{x_*^\intercal x_*} = \min_{x_* \in \mathbb{R}^{d_*}} \frac{x_*^\intercal SAS^\intercal x_*}{x_*^\intercal SS^\intercal x_*} \geq \min_{x \in \mathbb{R}^d} \frac{x^\intercal A x}{x^\intercal x} = \lambda_{\text{MIN}}(A),$$

with the inequality that follows from the fact that $\{x \in \mathbb{R}^d : x = x_*^\intercal S \text{ for } x_* \in \mathbb{R}^{d_*}\} \subseteq R^d$. Following the same line of reasoning it is possible to prove $\lambda_{\text{MAX}}(A) \geq \lambda_{\text{MAX}}(B)$ which concludes the proof of the lemma.

□

Lemma B.5. *Under the assumptions stated in Theorem 2.3, for every univariate cdf $F(\cdot)$ satisfying $F(-x) = 1 - F(x)$, $F(0) = 1/2$ and $F(x) = F(0) + \eta x + O(x^2)$, it holds*

$$\sup_{\hat{h}_{\mathcal{C}} \in \hat{K}_{n,\mathcal{C}}} \left| \mathbb{E}_{\hat{h}_{\bar{\mathcal{C}}}|\hat{h}_{\mathcal{C}}} F(\hat{\alpha}_\eta(\hat{h})) - F(\mathbb{E}_{\hat{h}_{\bar{\mathcal{C}}}|\hat{h}_{\mathcal{C}}}\{\hat{\alpha}_\eta(\hat{h})\}) \right| = O_{P_0^n}(M_n^{c_{10}}/n),$$

where $\hat{K}_{n,\mathcal{C}} = \{\hat{h}_{\mathcal{C}} : \|\hat{h}_{\mathcal{C}}\| < 2M_n\}$, c_{10} denotes a positive constant and $\hat{\alpha}_\eta(\hat{h})$ is as defined in Section 2.2.

Proof. Recall that the covariance matrix $\bar{\Omega}$, associated to the Gaussian measure $P_{\hat{h}_{\bar{\mathcal{C}}}|\hat{h}_{\mathcal{C}}}$, is the Schur complement of the block $\hat{\Omega}_{\mathcal{C}\mathcal{C}}$ of the matrix $\hat{\Omega}$. This implies that, in view of Assumption M2, Lemma B.4 and the properties of the Schur complement, there exist constants $0 < \tilde{c}_1 < \tilde{c}_2 < \infty$ such that the eigenvalues of $\bar{\Omega}$ are bounded from below by $\tilde{c}_1$ and above by $\tilde{c}_2$ with probability tending to one.

As a consequence, there exists a large enough constant $c_{0,\bar{\mathcal{C}}} > 0$ such that the complement of the set $\hat{K}_{n,\bar{\mathcal{C}}} = \{\hat{h}_{\bar{\mathcal{C}}} : \|\hat{h}_{\bar{\mathcal{C}}} - \mathbb{E}_{\hat{h}_{\bar{\mathcal{C}}}|\hat{h}_{\mathcal{C}}}(\hat{h}_{\bar{\mathcal{C}}})\| < 2\sqrt{c_{0,\bar{\mathcal{C}}} \log n}\}$, has negligible mass, i.e. the event $L_{n,0} = \big\{P_{\hat{h}_{\bar{\mathcal{C}}}|\hat{h}_{\mathcal{C}}}(\hat{h}_{\bar{\mathcal{C}}} \in \hat{K}^c_{n,\bar{\mathcal{C}}}) < \tilde{c}_3/n, \text{ for } \hat{h}_{\mathcal{C}} \in \hat{K}_{n,\mathcal{C}}\big\}$ has probability $P_0^n(L_{n,0}) = 1 - o(1)$. This fact and the boundedness of $F(\cdot)$ imply

$$\begin{aligned}
&\sup_{\hat{h}_{\mathcal{C}} \in \hat{K}_{n,\mathcal{C}}} \big|\mathbb{E}_{\hat{h}_{\bar{\mathcal{C}}}|\hat{h}_{\mathcal{C}}}[F(\hat{\alpha}_\eta(\hat{h})) - F(\mathbb{E}_{\hat{h}_{\bar{\mathcal{C}}}|\hat{h}_{\mathcal{C}}}\{\hat{\alpha}_\eta(\hat{h})\})\big| \mathbb{1}_{L_{n,0}} \\
&\leq \sup_{\hat{h}_{\mathcal{C}} \in \hat{K}_{n,\mathcal{C}}} \big|\mathbb{E}_{\hat{h}_{\bar{\mathcal{C}}}|\hat{h}_{\mathcal{C}}}\big\{F(\hat{\alpha}_\eta(\hat{h})) - F(\mathbb{E}_{\hat{h}_{\bar{\mathcal{C}}}|\hat{h}_{\mathcal{C}}}\hat{\alpha}_\eta(\hat{h}))\big\}\mathbb{1}_{\hat{h}_{\bar{\mathcal{C}}} \in \hat{K}_{n,\bar{\mathcal{C}}}}\big|\mathbb{1}_{L_{n,0}} + O(n^{-1}).
\end{aligned} \tag{B.13}$$

We further decompose the first term on the right–hand–side of the last display by adding and subtracting the first order Taylor expansion of $F(\hat{\alpha}_\eta(\hat{h}))$. Moreover, recall that under the assumptions of Theorem 2.3, there exists $\tilde{c}_3 > 0$ such that $L_{n,1} = L_{n,0} \cap \{\max_{s,t,l} |a^{(3),n}_{\hat{\theta},stl}| < \tilde{c}_3\}$ holds with probability $P_0^n L_{n,1} = 1 - o(1)$. Since $\|\hat{h}\| \leq \|\hat{h}_{\mathcal{C}}\| + \|\hat{h}_{\bar{\mathcal{C}}}\|$ it follows also that, conditioned on $L_{n,1}$, both $\sup_{\hat{h} \in \hat{K}_{n,\mathcal{C}} \cap \hat{K}_{n,\bar{\mathcal{C}}}} |\hat{\alpha}_\eta(\hat{h})|$ and $\sup_{\hat{h}_{\mathcal{C}} \in \hat{K}_{n,\mathcal{C}}} |\mathbb{E}_{\hat{h}_{\bar{\mathcal{C}}}|\hat{h}_{\mathcal{C}}}\{\hat{\alpha}_\eta(\hat{h})\}|$ are of order $O(M_n^{\tilde{c}_4}/\sqrt{n})$, while $\sup_{\hat{h}_{\mathcal{C}} \in \hat{K}_{n,\mathcal{C}}} \mathbb{E}_{\hat{h}_{\bar{\mathcal{C}}}|\hat{h}_{\mathcal{C}}}\{\hat{\alpha}_\eta(\hat{h})^2\} = O(M_n^{\tilde{c}_5}/n)$ for some $\tilde{c}_4, \tilde{c}_5 > 0$. This, together with $F(x) = 1/2 + \eta x + O(x^2)$, $x \to 0$, implies for n sufficiently large

$$\begin{aligned}
&\sup_{\hat{h}_{\mathcal{C}} \in \hat{K}_{n,\mathcal{C}}} \big|\mathbb{E}_{\hat{h}_{\bar{\mathcal{C}}}|\hat{h}_{\mathcal{C}}}\big\{F(\hat{\alpha}_\eta(\hat{h})) - 1/2 - \eta\hat{\alpha}_\eta(\hat{h})\big\}\mathbb{1}_{\hat{h}_{\bar{\mathcal{C}}} \in \hat{K}_{n,\bar{\mathcal{C}}}}\big|\mathbb{1}_{L_{n,1}} \\
&\lesssim \sup_{\hat{h}_{\mathcal{C}} \in \hat{K}_{n,\mathcal{C}}} \mathbb{E}_{\hat{h}_{\bar{\mathcal{C}}}|\hat{h}_{\mathcal{C}}}\{\hat{\alpha}_\eta(\hat{h})^2\}\mathbb{1}_{L_{n,1}} = O(M_n^{\tilde{c}_5}/n).
\end{aligned}$$

As a consequence,

$$\sup_{\hat{h}_{\mathcal{C}} \in \hat{K}_{n,\mathcal{C}}} \big|\mathbb{E}_{\hat{h}_{\bar{\mathcal{C}}}|\hat{h}_{\mathcal{C}}}\big\{F(\hat{\alpha}_\eta(\hat{h})) - 1/2 - \eta\hat{\alpha}_\eta(\hat{h})\big\}\mathbb{1}_{\hat{h}_{\bar{\mathcal{C}}} \in \hat{K}_{n,\bar{\mathcal{C}}}}\big| = O_{P_0^n}(M_n^{\tilde{c}_5}/n). \tag{B.14}$$

To conclude, note that for n sufficiently large,

$$\begin{aligned}
&\sup_{\hat{h}_{\mathcal{C}} \in \hat{K}_{n,\mathcal{C}}} |\mathbb{E}_{\hat{h}_{\bar{\mathcal{C}}}|\hat{h}_{\mathcal{C}}}\big\{F(\mathbb{E}_{\hat{h}_{\bar{\mathcal{C}}}|\hat{h}_{\mathcal{C}}}\hat{\alpha}_\eta(\hat{h})) - 1/2 - \eta\hat{\alpha}_\eta(\hat{h})\big\}\mathbb{1}_{\hat{h}_{\bar{\mathcal{C}}} \in \hat{K}_{n,\bar{\mathcal{C}}}}\mathbb{1}_{L_{n,1}}| \\
&\leq \sup_{\hat{h}_{\mathcal{C}} \in \hat{K}_{n,\mathcal{C}}} |\mathbb{E}_{\hat{h}_{\bar{\mathcal{C}}}|\hat{h}_{\mathcal{C}}}[\eta\big\{\mathbb{E}_{\hat{h}_{\bar{\mathcal{C}}}|\hat{h}_{\mathcal{C}}}(\hat{\alpha}_\eta(\hat{h})) - \hat{\alpha}_\eta(\hat{h})\big\}\mathbb{1}_{\hat{h}_{\bar{\mathcal{C}}} \in \hat{K}_{n,\bar{\mathcal{C}}}}]\mathbb{1}_{L_{n,1}}| \\
&\qquad\qquad + O(\{\mathbb{E}_{\hat{h}_{\bar{\mathcal{C}}}|\hat{h}_{\mathcal{C}}}\hat{\alpha}_\eta(\hat{h})\}^2\mathbb{1}_{L_{n,1}}) \\
&= \sup_{\hat{h}_{\mathcal{C}} \in \hat{K}_{n,\mathcal{C}}} |\mathbb{E}_{\hat{h}_{\bar{\mathcal{C}}}|\hat{h}_{\mathcal{C}}}[\eta\big\{\mathbb{E}_{\hat{h}_{\bar{\mathcal{C}}}|\hat{h}_{\mathcal{C}}}(\hat{\alpha}_\eta(\hat{h})) - \hat{\alpha}_\eta(\hat{h})\big\}\mathbb{1}_{\hat{h}_{\bar{\mathcal{C}}} \in \hat{K}^c_{n,\bar{\mathcal{C}}}}]\mathbb{1}_{L_{n,1}}| + O(M_n^{2\tilde{c}_4}/n) \\
&\leq \sup_{\hat{h}_{\mathcal{C}} \in \hat{K}_{n,\mathcal{C}}} \eta[\mathbb{E}_{\hat{h}_{\bar{\mathcal{C}}}|\hat{h}_{\mathcal{C}}}\big\{\mathbb{E}_{\hat{h}_{\bar{\mathcal{C}}}|\hat{h}_{\mathcal{C}}}(\hat{\alpha}_\eta(\hat{h})) - \hat{\alpha}_\eta(\hat{h})\big\}^2]^{\frac{1}{2}}[\mathbb{E}_{\hat{h}_{\bar{\mathcal{C}}}|\hat{h}_{\mathcal{C}}}\mathbb{1}_{\hat{h}_{\bar{\mathcal{C}}} \in \hat{K}^c_{n,\bar{\mathcal{C}}}}]^{\frac{1}{2}}\mathbb{1}_{L_{n,1}} + O(M_n^{2\tilde{c}_4}/n),
\end{aligned}$$

with the first inequality that follows from the Taylor expansion of $F(\cdot)$ at 0, the equality from $\mathbb{1}_{\hat{h}_{\bar{\mathcal{C}}} \in \hat{K}_{n,\bar{\mathcal{C}}}} = 1 - \mathbb{1}_{\hat{h}_{\bar{\mathcal{C}}} \in \hat{K}^c_{n,\bar{\mathcal{C}}}}$ and the last line from Cauchy-Schwarz inequality. Finally,

note that $[\mathbb{E}_{\hat{h}_{\bar{\mathcal{C}}}|\hat{h}_{\mathcal{C}}} \mathbb{1}_{\hat{h}_{\bar{\mathcal{C}}} \in \hat{K}^c_{n,\bar{\mathcal{C}}}}]^{\frac{1}{2}} \mathbb{1}_{L_{n,1}} = O(n^{-1/2})$ while

$$\sup_{\hat{h}_{\mathcal{C}} \in \hat{K}_{n,\mathcal{C}}} [\mathbb{E}_{\hat{h}_{\bar{\mathcal{C}}}|\hat{h}_{\mathcal{C}}} \{\mathbb{E}_{\hat{h}_{\bar{\mathcal{C}}}|\hat{h}_{\mathcal{C}}}(\hat{\alpha}_\eta(\hat{h})) - \hat{\alpha}_\eta(\hat{h})\}^2]^{\frac{1}{2}} \mathbb{1}_{L_{n,1}} = O(M_n^{\tilde{c}_6}/\sqrt{n}),$$

for some $\tilde{c}_6 > 0$ large enough. From the previous display it follows

$$\sup_{\hat{h}_{\mathcal{C}} \in \hat{K}_{n,\mathcal{C}}} \left| \mathbb{E}_{\hat{h}_{\bar{\mathcal{C}}}|\hat{h}_{\mathcal{C}}} [\{F(\mathbb{E}_{\hat{h}_{\bar{\mathcal{C}}}|\hat{h}_{\mathcal{C}}} \hat{\alpha}_\eta(\hat{h})) - 1/2 - \eta \hat{\alpha}_\eta(\hat{h})\} \mathbb{1}_{\hat{h}_{\bar{\mathcal{C}}} \in \hat{K}_{n,\bar{\mathcal{C}}}} \right| = O_{P_0^n}(M_n^{\tilde{c}_7}/n). \tag{B.15}$$

for some $\tilde{c}_7$ large enough. The combination of (B.13), (B.14) and (B.15) concludes the proof. □

Lemma B.6. *Let $\hat{h} = \sqrt{n}(\theta - \hat{\theta})$, $\hat{K}_n = \{h : \|h\| < 2M_n\}$ and $M_n = \sqrt{c_0 \log n}$ for some $c_0 > 0$. Moreover, let $\tilde{\ell}(\theta) = \log \pi(\theta) L(\theta; X^n)$ be the un-normalized log-posterior. Under Assumption M3 it follows that*

$$\begin{aligned} &\log\Big(\big[\exp\{\tilde{\ell}(\hat{\theta} + \hat{h}/\sqrt{n}) - \tilde{\ell}(\hat{\theta})\} + \exp\{\tilde{\ell}(\hat{\theta} - \hat{h}/\sqrt{n}) - \tilde{\ell}(\hat{\theta})\} \big]/2 \Big) \\ &+ \tilde{\Omega}^{-1}_{st} \hat{h}_s \hat{h}_t/2 = \hat{r}_{n,5}(\hat{h}), \end{aligned} \tag{B.16}$$

$$\begin{aligned} &\log\Big(\big[\exp\{\tilde{\ell}(\hat{\theta} + \hat{h}/\sqrt{n}) - \tilde{\ell}(\hat{\theta})\} + \exp\{\tilde{\ell}(\hat{\theta} - \hat{h}/\sqrt{n}) - \tilde{\ell}(\hat{\theta})\} \big]/2 \Big) \\ &+ \tilde{\Omega}^{-1}_{st} \hat{h}_s \hat{h}_t/2 - \log P(\hat{h}) = \hat{r}_{n,6}(\hat{h}), \end{aligned} \tag{B.17}$$

with $\tilde{\Omega}^{-1} = -(\ell^{(2)}_{\hat{\theta}} + \log \pi^{(2)}_{\hat{\theta}})/n$ and

$$P(\hat{h}) = 1 + \frac{\hat{a}^{(4)}_{\hat{\theta},stlk} \hat{h}_s \hat{h}_t \hat{h}_l \hat{h}_k}{24n} + \frac{1}{2}\Big(\frac{\hat{a}^{(4)}_{\hat{\theta},stlk} \hat{h}_s \hat{h}_t \hat{h}_l \hat{h}_k}{24n} \Big)^2 + \frac{1}{2}\Big(\frac{\hat{a}^{(3)}_{\hat{\theta},stl} \hat{h}_s \hat{h}_t \hat{h}_l}{6\sqrt{n}} \Big)^2,$$

where $P(\hat{h})$ is a non negative polynomial, $\hat{a}^{(3)}_{\hat{\theta}} = \ell^{(3)}_{\hat{\theta}}/n$ and $\hat{a}^{(4)}_{\hat{\theta}} = \ell^{(4)}_{\hat{\theta}}/n$.

In addition,

$$\hat{r}_{n,5} := \sup_{\hat{h} \in \hat{K}_n} |\hat{r}_{n,5}(\hat{h})| = O_{P_0^n}(M_n^{c_{11}}/n), \tag{B.18}$$

and

$$\hat{r}_{n,6} := \sup_{\hat{h} \in \hat{K}_n} |\hat{r}_{n,6}(\hat{h})| = O_{P_0^n}(M_n^{c_{12}}/n^2), \tag{B.19}$$

for some constants $c_{11}, c_{12} > 0$.

Proof. We demonstrate the validity of (B.17) and (B.19). (B.16) and (B.18) follows in a similar way but with less tedious algebra. For reasons of compactness we adopt the notation

$$\frac{\langle \ell^{(k)}_{\hat{\theta}}, \hat{h}^{\otimes k} \rangle}{k! n^{k/2}} = \frac{\ell^{(k)}_{\hat{\theta},s_1 \cdots s_k} \hat{h}_{s_1} \times \cdots \times \hat{h}_{s_k}}{k! n^{k/2}},$$

to denote the k-the element of the log-likelihood Taylor expansion. The same notation is adopted for the Taylor expansion of the log-prior with $\ell^{(k)}_{\hat\theta}$ replaced by $\log \pi^{(k)}_{\hat\theta}$.

Recall that under condition M1 the event $\{\|\hat\theta - \theta_*\| \leq \delta\}$ for any $\delta > 0$ has probability tending to 1. As a consequence from Assumption M3 it is possible to expand the log-likelihood ratio around $\hat\theta$ as

$$\log \frac{p_{\hat\theta+\hat h/\sqrt{n}}}{p_{\hat\theta}}(X^n) \;=\; \sum_{k=1}^{5} \frac{\langle \ell^{(k)}_{\hat\theta}, \hat h^{\otimes k}\rangle}{k! n^{k/2}} + O_{P_0^n}\Big(\frac{\{\|\hat h\|^6 \vee 1\}}{n^2}\Big).$$

Similarly, the Taylor expansion of the log-prior takes the form

$$\log \frac{\pi_{\hat\theta+\hat h/\sqrt{n}}}{\pi_{\hat\theta}} \;=\; \sum_{k=1}^{2} \frac{\langle \log \pi^{(k)}_{\hat\theta}, \hat h^{\otimes k}\rangle}{k! n^{k/2}} + O_{P_0^n}\Big(\frac{\{\|\hat h\|^3 \vee 1\}}{n^2}\Big).$$

Combining these facts, we can write the log-posterior ratio as

$$\tilde\ell(\theta + \hat h/\sqrt{n}) - \tilde\ell(\theta) \;=\; \sum_{k=1}^{5} \frac{\langle \hat a^{(k)}_{\hat\theta}, \hat h^{\otimes k}\rangle}{k! n^{(k-2)/2}} + O_{P_0^n}\Big(\frac{\{\|\hat h\|^6 \vee 1\}}{n^2}\Big), \tag{B.20}$$

where $\hat a^{(k)}_{\hat\theta} = (\ell^{(k)}_{\hat\theta} + \log \pi^{(k)}_{\hat\theta})/n$, if $k \leq 2$, and $\hat a^{(k)}_{\hat\theta} = \ell^{(k)}_{\hat\theta}/n$ otherwise. Moreover, since $\hat\theta$ is the posterior mode, by definition, $\hat a^{(1)}_{\hat\theta} = 0$.

By exploiting (B.20) and $e^x = 1 + O(x)$, $x \to 0$ we obtain

$$\begin{aligned}
&\frac{1}{2}\exp\{\tilde\ell(\hat\theta + \hat h/\sqrt{n}) - \tilde\ell(\hat\theta)\} + \frac{1}{2}\exp\{\tilde\ell(\hat\theta - \hat h/\sqrt{n}) - \tilde\ell(\hat\theta)\} \\
&= \frac{1}{2}\exp\Big(\sum_{k=2}^{5} \frac{\langle \hat a^{(k)}_{\hat\theta}, \hat h^{\otimes k}\rangle}{k! n^{(k-2)/2}}\Big) + \frac{1}{2}\exp\Big(\sum_{k=2}^{5} \frac{\langle \hat a^{(k)}_{\hat\theta}, -\hat h^{\otimes k}\rangle}{k! n^{(k-2)/2}}\Big) + O_{P_0^n}\Big(\frac{\{\|\hat h\|^6 \vee 1\}}{n^2}\Big), \\
&= \frac{1}{2}\exp\Big(\sum_{k=1}^{2} \frac{\langle \hat a^{(2k)}_{\hat\theta}, \hat h^{\otimes 2k}\rangle}{(2k)! n^{(k-2)/2}}\Big)\left\{\exp\Big(\sum_{k=1}^{2} \frac{\langle \hat a^{(2k+1)}_{\hat\theta}, \hat h^{\otimes (2k+1)}\rangle}{(2k+1)! n^{(2k-1)/2}}\Big) + \exp\Big(\sum_{k=1}^{2} \frac{\langle \hat a^{(2k+1)}_{\hat\theta}, -\hat h^{\otimes (2k+1)}\rangle}{(2k+1)! n^{(2k-1)/2}}\Big)\right\} \\
&+ O_{P_0^n}\Big(\frac{\{\|\hat h\|^6 \vee 1\}}{n^2}\Big).
\end{aligned} \tag{B.21}$$

Note that from Assumption M3 it follows that, for $k \geq 2$, $\tilde a^{(k)}_{\hat\theta} = O_{P_0^n}(1)$. By exploiting $e^x = 1 + x + x^2/2 + x^3/6 + O(x^4)$, we can manipulate the the second multiplicative term on the right-hand-side of the last equality of (B.21) as double the

$$\begin{aligned}
\exp\Big(\sum_{k=1}^{2} \frac{\langle \hat a^{(2k+1)}_{\hat\theta}, \hat h^{\otimes (2k+1)}\rangle}{(2k+1)! n^{(2k-1)/2}}\Big) =& 1 + \sum_{k=1}^{2} \frac{\langle \hat a^{(2k+1)}_{\hat\theta}, \hat h^{\otimes (2k+1)}\rangle}{(2k+1)! n^{(2k-1)/2}} + \frac{1}{2}\left(\sum_{k=1}^{2} \frac{\langle \hat a^{(2k+1)}_{\hat\theta}, \hat h^{\otimes (2k+1)}\rangle}{(2k+1)! n^{(2k-1)/2}}\right)^2 \\
&+ \frac{1}{6}\left(\sum_{k=1}^{2} \frac{\langle \hat a^{(2k+1)}_{\hat\theta}, \hat h^{\otimes (2k+1)}\rangle}{(2k+1)! n^{(2k-1)/2}}\right)^3 + O_{P_0^n}\Big(\frac{\{\|\hat h\|^{12} \vee 1\}}{n^2}\Big),
\end{aligned}$$

and

$$\exp\Big(\sum_{k=1}^{2}\frac{\langle\hat{a}_{\hat{\theta}}^{(2k+1)},-\hat{h}^{\otimes(2k+1)}\rangle}{(2k+1)!n^{(2k-1)/2}}\Big)=1-\sum_{k=1}^{2}\frac{\langle\hat{a}_{\hat{\theta}}^{(2k+1)},\hat{h}^{\otimes(2k+1)}\rangle}{(2k+1)!n^{(2k-1)/2}}+\frac{1}{2}\left(\sum_{k=1}^{2}\frac{\langle\hat{a}_{\hat{\theta}}^{(2k+1)},\hat{h}^{\otimes(2k+1)}\rangle}{(2k+1)!n^{(2k-1)/2}}\right)^{2}$$
$$-\frac{1}{6}\left(\sum_{k=1}^{2}\frac{\langle\hat{a}_{\hat{\theta}}^{(2k+1)},\hat{h}^{\otimes(2k+1)}\rangle}{(2k+1)!n^{(2k-1)/2}}\right)^{3}+O_{P_0^n}\Big(\frac{\{\|\hat{h}\|^{12}\vee 1\}}{n^2}\Big).$$

As a consequence, recalling the asymptotic order of $\tilde{a}_{\hat{\theta}}^{(r)}$, we can rewrite the second multiplicative term in the right-hand-side of the last equality of (B.21) as

$$\begin{aligned}&\frac{1}{2}\left\{\exp\Big(\sum_{k=1}^{2}\frac{\langle\hat{a}_{\hat{\theta}}^{(2k+1)},\hat{h}^{\otimes(2k+1)}\rangle}{(2k+1)!n^{(2k-1)/2}}\Big)+\exp\Big(\sum_{k=1}^{2}\frac{\langle\hat{a}_{\hat{\theta}}^{(2k+1)},-\hat{h}^{\otimes(2k+1)}\rangle}{(2k+1)!n^{(2k-1)/2}}\Big)\right\}\\&=1+\frac{1}{2}\left(\sum_{k=1}^{2}\frac{\langle\hat{a}_{\hat{\theta}}^{(2k+1)},\hat{h}^{\otimes(2k+1)}\rangle}{(2k+1)!n^{(2k-1)/2}}\right)^{2}+O_{P_0^n}\Big(\frac{\{\|\hat{h}\|^{12}\vee 1\}}{n^2}\Big)\\&=1+\frac{1}{2}\left(\frac{\langle\hat{a}_{\hat{\theta}}^{(3)},\hat{h}^{\otimes 3}\rangle}{6n}\right)^{2}+O_{P_0^n}\Big(\frac{\{\|\hat{h}\|^{12}\vee 1\}}{n^2}\Big).\end{aligned}\tag{B.22}$$

To conclude, we exploit again $e^x=1+x+x^2/2+O(x^3)$, and the fact that even, partial, sums of the exponential series are always positive (Zemyan, 2005). This leads to the following higher order, non-negative, approximation of the un-normalized, symmetrized, posterior density

$$\begin{aligned}&\frac{1}{2}\exp\{\tilde{\ell}(\hat{\theta}+\hat{h}/\sqrt{n})-\tilde{\ell}(\hat{\theta})\}+\frac{1}{2}\exp\{\tilde{\ell}(\hat{\theta}-\hat{h}/\sqrt{n})-\tilde{\ell}(\hat{\theta})\}\\&=\exp\Big(\frac{\langle\hat{a}_{\hat{\theta}}^{(2)},\hat{h}^{\otimes 2}\rangle}{2}\Big)\Big(1+\frac{\langle\hat{a}_{\hat{\theta}}^{(4)},\hat{h}^{\otimes 4}\rangle}{24n}+\frac{1}{2}\Big(\frac{\langle\hat{a}_{\hat{\theta}}^{(4)},\hat{h}^{\otimes 4}\rangle}{24n}\Big)^{2}+O_{P_0^n}\Big(\frac{\{\|\hat{h}\|^{12}\vee 1\}}{n^3}\Big)\Big)\\&\quad\times\Big(1+\frac{1}{2}\Big(\frac{\langle\hat{a}_{\hat{\theta}}^{(3)},\hat{h}^{\otimes 3}\rangle}{6n^{1/2}}\Big)^{2}+O_{P_0^n}\Big(\frac{\{\|\hat{h}\|^{12}\vee 1\}}{n^2}\Big)\Big)+O_{P_0^n}\Big(\frac{\{\|\hat{h}\|^{6}\vee 1\}}{n^2}\Big)\\&=\exp\Big(\frac{\langle\hat{a}_{\hat{\theta}}^{(2)},\hat{h}^{\otimes 2}\rangle}{2}\Big)\Big(1+\frac{\langle\hat{a}_{\hat{\theta}}^{(4)},\hat{h}^{\otimes 4}\rangle}{24n}+\frac{1}{2}\Big(\frac{\langle\hat{a}_{\hat{\theta}}^{(4)},\hat{h}^{\otimes 4}\rangle}{24n}\Big)^{2}+\frac{1}{2}\Big(\frac{\langle\hat{a}_{\hat{\theta}}^{(3)},\hat{h}^{\otimes 3}\rangle}{6n^{1/2}}\Big)^{2}\Big)+O_{P_0^n}\Big(\frac{\{\|\hat{h}\|^{12}\vee 1\}}{n^2}\Big)\\&=\exp\Big(-\frac{1}{2}\tilde{\Omega}_{st}^{-1}\hat{h}_s\hat{h}_t\Big)P(\hat{h})+O_{P_0^n}\Big(\frac{\{\|\hat{h}\|^{12}\vee 1\}}{n^2}\Big),\end{aligned}$$

where the last lines follows from $\tilde{\Omega}=-(\hat{a}_{\hat{\theta}}^{(2)})^{-1}$ and from the definition of $P(\hat{h})$. To conclude, it suffices to note that $\log(1+x)=O(x)$ for $x\to 0$ and therefore,

$$\begin{aligned}\hat{r}_{n,6}(\hat{h})=&\log\Big(\big[\exp\{\tilde{\ell}(\hat{\theta}+\hat{h}/\sqrt{n})-\tilde{\ell}(\hat{\theta})\}+\exp\{\tilde{\ell}(\hat{\theta}-\hat{h}/\sqrt{n})-\tilde{\ell}(\hat{\theta})\}\big]/2\Big)\\&+\tilde{\Omega}_{st}^{-1}\hat{h}_s\hat{h}_t/2-\log P(\hat{h}),\end{aligned}$$

TABLE B.1: For probit and logistic regression, estimated joint, bivariate and marginal total variation distances between the exact posterior and the deterministic approximations under analysis in the Cushings application. The bold values indicate the best performance for each subset of parameters.

	TV_{θ}	$\text{TV}_{\theta_{01}}$	$\text{TV}_{\theta_{02}}$	$\text{TV}_{\theta_{12}}$	TV_{θ_0}	TV_{θ_1}	TV_{θ_2}
Probit							
SKEW-M	**0.11**	**0.05**	**0.06**	**0.09**	0.03	**0.04**	**0.05**
GM	0.19	0.10	0.13	0.18	0.09	0.08	0.11
EP	0.13	0.07	0.09	0.11	**0.01**	0.07	0.09
MF-VB	0.50	0.32	0.41	0.47	0.18	0.28	0.35
PFM-VB	0.25	0.12	0.22	0.23	0.06	0.09	0.19
Logit							
SKEW-M	**0.14**	0.08	**0.10**	0.13	0.05	**0.06**	**0.07**
GM	0.23	0.13	0.17	0.22	0.11	0.10	0.14
EP	**0.14**	**0.07**	0.11	**0.12**	**0.01**	0.07	0.10
MF-VB	0.25	0.13	0.21	0.24	0.07	0.10	0.19

with $\hat{r}_{n,6}(\hat{h}) = O_{P_0^n}\left(\frac{\{\|\hat{h}\|^{12} \vee 1\}}{n^2}\right)$ and

$$\hat{r}_{n,6} := \sup_{\hat{h} \in \hat{K}_n} |\hat{r}_{n,6}(\hat{h})| = O_{P_0^n}(M_n^{c_{12}}/n^2),$$

for $c_{12} = 12$. This concludes the proof of the lemma. □

B.2 Cushings dataset

This section reports some additional details regarding the real data analysis described in Section 2.4.2. In particular, the comparison with the performance of the joint and marginal skew-modal approximations is extended to include additional state-of-the-art deterministic Gaussian approximation methods, such as mean-field variational Bayes (MF-VB) (Consonni and Marin, 2007; Durante and Rigon, 2019), expectation propagation (Chopin and Ridgway, 2017) (EP) and, in the case of the probit model, partially factorized variational Bayes (PFM-VB). MF-VB and PFM-VB for probit regression use the implementation in the GitHub repository `Probit-PFMVB` (Fasano *et al.*, 2022), while in the logistic setting we rely on the codes in the repository `logisticVB` (Durante and Rigon, 2019). Note that PFM-VB is designed for probit regression only. Finally, EP is implemented under both models using the `R` library `EPGLM`. (Chopin and Ridgway, 2017).

Table B.1 reports the total variation distance from the true posterior for each of the 5 approximations considered. It highlights how the skew-modal approximations generally tend to outperform not only GM, but also MF-VB and PFM-VB. In addition,

SKEW-M tends to work comparably well as EP. As discussed in Section 2.4.2, the similar performance between SKEW-M and EP is probably due to the better quality of the EP approximation compared to GM, which is a key component in our SKEW-M method. At the same time, one of the main features that makes SKEW-M attractive over EP is the fact that, besides its very good performance, it can in principle be applied to a wider range of problems, since it only assumes the ability to compute the Laplace approximation and the third derivatives of the log-posterior, and does not require any special factorization of the log-posterior to be used.

Bibliography

Agresti, A. (2015) *Foundations of Linear and Generalized Linear Models*. John Wiley & Sons.

Anceschi, N., Fasano, A., Durante, D. and Zanella, G. (2023) Bayesian conjugacy in probit, tobit, multinomial probit and extensions: A review and new results. *Journal of the American Statistical Association* **118**(542), 1451–1469.

Arellano-Valle, R. B. and Azzalini, A. (2006) On the unification of families of skew-normal distributions. *Scandinavian Journal of Statistics* **33**(3), 561–574.

Azzalini, A. (1985) A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics* **12**(1), 171–178.

Azzalini, A. and Capitanio, A. (2003) Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. *Journal of the Royal Statistical Society: Series B* **65**(2), 367–389.

Azzalini, A. and Capitanio, A. (2014) *The Skew-Normal and Related Families*. Cambridge University Press.

Azzalini, A. and Dalla Valle, A. (1996) The multivariate skew-normal distribution. *Biometrika* **83**(4), 715–726.

Barthelmé, S. (2023) GaussianEP. https://github.com/dahtah/GaussianEP.jl.

Bickel, P. J. and Kleijn, B. J. (2012) The semiparametric Bernstein–von Mises theorem. *The Annals of Statistics* **40**(1), 206–237.

Bishop, C. (2006) *Pattern Recognition and Machine Learning*. Springer.

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

Bontemps, D. (2011) Bernstein–von Mises theorems for Gaussian regression with increasing number of regressors. *The Annals of Statistics* **39**(5), 2557–2584.

Boucheron, S. and Gassiat, E. (2009) A Bernstein-von Mises theorem for discrete probability distributions. *Electronic Journal of Statistics* **3**, 114–148.

Castillo, I. and Rousseau, J. (2015) A Bernstein–von Mises theorem for smooth functionals in semiparametric models. *The Annals of Statistics* **43**(6), 2353–2383.

Challis, E. and Barber, D. (2012) Affine independent variational inference. *Advances in Neural Information Processing Systems* **25**, 1–9.

Challis, E. and Barber, D. (2013) Gaussian Kullback–Leibler approximate inference. *Journal of Machine Learning Research* **14**, 2239–2286.

Chopin, N. and Papaspiliopoulos, O. (2020) *An Introduction to Sequential Monte Carlo*. Springer.

Chopin, N. and Ridgway, J. (2017) Leave pima indians alone: Binary regression as a benchmark for Bayesian computation. *Statistical Science* **32**(1), 64–87.

Cichocki, A. and Amari, S.-I. (2010) Families of alpha-beta-and gamma-divergences: Flexible and robust measures of similarities. *Entropy* **12**(6), 1532–1568.

Consonni, G. and Marin, J.-M. (2007) Mean-field variational approximate Bayesian inference for latent variable models. *Computational Statistics & Data Analysis* **52**(2), 790–798.

Craig-Schapiro, R., Kuhn, M., Xiong, C., Pickering, E. H., Liu, J., Misko, T. P., Perrin, R. J., Bales, K. R., Soares, H., Fagan, A. M. *et al.* (2011) Multiplexed immunoassay panel identifies novel csf biomarkers for alzheimer's disease diagnosis and prognosis. *Plos One* **6**(4), e18850.

Dehaene, G. and Barthelmé, S. (2018) Expectation propagation in the large data limit. *Journal of the Royal Statistical Society: Series B* **80**(1), 199–217.

Ding, N., Qi, Y. and Vishwanathan, S. (2011) t-divergence based approximate inference. *Advances in Neural Information Processing Systems* **24**, 1–9.

Durante, D. (2019) Conjugate Bayes for probit regression via unified skew-normal distributions. *Biometrika* **106**(4), 765–779.

Durante, D. and Rigon, T. (2019) Conditionally conjugate mean-field variational Bayes for logistic models. *Statistical Science* **34**(3), 472–485.

Fasano, A., Durante, D. and Zanella, G. (2022) Scalable and accurate variational Bayes for high-dimensional binary regression models. *Biometrika* **109**(4), 901–919.

Frobenius, G. (1912) Über matrizen aus nicht negativen elementen. *Sitzungsberichte der Königlich Preussischen Akademie der Wissenschaften* **23**, 456–477.

Futami, F., Sato, I. and Sugiyama, M. (2017) Expectation propagation for t-exponential family using q-algebra. *Advances in Neural Information Processing Systems* **30**, 1–10.

Gallant, A. R. and Nychka, D. W. (1987) Semi-nonparametric maximum likelihood estimation. *Econometrica: Journal of the Econometric Society* **55**(2), 363–390.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. and Rubin, D. B. (2013) *Bayesian Data Analysis*. Chapman and Hall/CRC.

Genton, M. G. and Loperfido, N. M. (2005) Generalized skew-elliptical distributions and their quadratic forms. *Annals of the Institute of Statistical Mathematics* **57**(2), 389–401.

Ghosal, S. and Van der Vaart, A. (2017) *Fundamentals of Nonparametric Bayesian Inference*. Cambridge University Press.

Giordano, R., Broderick, T. and Jordan, M. I. (2018) Covariances, robustness and variational Bayes. *Journal of Machine Learning Research* **19**, 1–49.

Goodrich, B., Gabry, J., Ali, I. and Brilleman, S. (2023) rstanarm: Bayesian applied regression modeling via Stan. R package version 2.21.4.

Hinkley, D. (1976) On estimating a symmetric distribution. *Biometrika* **63**(3), 680–681.

Johnson, R. A. (1970) Asymptotic expansions associated with posterior distributions. *The Annals of Mathematical Statistics* **41**(3), 851–864.

Kasprzak, M. J., Giordano, R. and Broderick, T. (2022) How good is your Gaussian approximation of the posterior? finite-sample computable error bounds for a variety of useful divergences. *arXiv:2209.14992* .

Kass, R. E., Tierney, L. and Kadane, J. B. (1990) The validity of posterior expansions based on Laplace's method. *Bayesian and Likelihood Methods in Statistics and Econometrics: Essays in Honor of George A. Barnard* pp. 473–487.

Kleijn, B. J. and Van der Vaart, A. W. (2012) The Bernstein-von-Mises theorem under misspecification. *Electronic Journal of Statistics* **6**, 354–381.

Koers, G., Szabo, B. and Van der Vaart, A. (2023) Misspecified Bernstein-von Mises theorem for hierarchical models. *arXiv:2308.07803* .

Kolassa, J. E. and Kuffner, T. A. (2020) On the validity of the formal Edgeworth expansion for posterior densities. *The Annals of Statistics* **48**(4), 1940–1958.

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A. and Blei, D. M. (2017) Automatic differentiation variational inference. *Journal of Machine Learning Research* **18**, 1–45.

Kuhn, M. and Johnson, K. (2018) Appliedpredictivemodeling: Functions and data sets for 'applied predictive modeling'. R package version 1.1-7.

Kullback, S. and Leibler, R. A. (1951) On information and sufficiency. *The Annals of Mathematical Statistics* **22**(1), 79–86.

Laplace, P. S. (1810) *Théorie Analytique des Probabilités*. Courcier.

Le Cam, L. (1953) On some asymptotic properties of maximum likelihood estimates and related Bayes' estimates. *Univ. Calif. Publ. in Statist.* **1**, 277–330.

Le Cam, L. (2012) *Asymptotic Methods in Statistical Decision Theory*. Springer Science & Business Media.

Le Cam, L. and Yang, G. L. (1990) *Asymptotics in Statistics*. Springer.

Lehmann, E. L. and Casella, G. (2006) *Theory of Point Estimation*. Springer Science & Business Media.

Liang, F., Mahoney, M. and Hodgkinson, L. (2022) Fat–tailed variational inference with anisotropic tail adaptive flows. *Proceedings of the International Conference on Machine Learning* **162**, 13257–13270.

Lo, S. H. (1985) Estimation of a symmetric distribution. *The Annals of Statistics* **13**(3), 1097–1113.

Ma, Y. and Genton, M. G. (2004) Flexible class of skew-symmetric distributions. *Scandinavian Journal of Statistics* **31**(3), 459–468.

McCullagh, P. (2018) *Tensor Methods in Statistics*. Courier Dover Publications.

Meloche, J. (1991) Estimation of a symmetric density. *The Canadian Journal of Statistics* **19**(2), 151–164.

Minka, T. P. (2001) Expectation propagation for approximate Bayesian inference. *Proceedings of Uncertainty in Artificial Intelligence* **17**, 362–369.

Onorati, P. and Liseo, B. (2022) An extension of the unified skew-normal family of distributions and application to Bayesian binary regression. *arXiv:2209.03474* .

Opper, M. and Archambeau, C. (2009) The variational Gaussian approximation revisited. *Neural Computation* **21**(3), 786–792.

Pace, L. and Salvan, A. (1997) *Principles of Statistical inference: From a Neo-Fisherian Perspective*. World scientific.

Perron, O. (1907) Zur theorie der matrices. *Mathematische Annalen* **64**(2), 248–263.

Ray, K. and Van der Vaart, A. (2020) Semiparametric Bayesian causal inference. *The Annals of Statistics* **48**(5).

Rossell, D., Abril, O. and Bhattacharya, A. (2021) Approximate Laplace approximations for scalable model selection. *Journal of the Royal Statistical Society: Series B* **83**(4), 853–879.

Rue, H., Martino, S. and Chopin, N. (2009) Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B* **71**(2), 319–392.

Sartori, N., Salvan, A. and Pace, L. (2020) *MLGdata: Datasets for Use with Salvan, Sartori and Pace (2020)*. R package version 0.1.0.

Schuster, E. F. (1975) Estimating the distribution function of a symmetric distribution. *Biometrika* **62**(3), 631–635.

Spokoiny, V. (2023) Inexact Laplace approximation and the use of posterior mean in Bayesian inference. *Bayesian Analysis* **in print**.

Spokoiny, V. and Panov, M. (2021) Accuracy of Gaussian approximation for high-dimensional posterior distribution. *Bernoulli* **in print**.

Tan, L. S. and Nott, D. J. (2018) Gaussian variational approximation with sparse precision matrices. *Statistics and Computing* **28**, 259–275.

Tao, T. (2011) *Topics in Random Matrix Theory*. American Mathematical Society.

Tasci, E., Zhuge, Y., Kaur, H., Camphausen, K. and Krauze, A. V. (2022) Hierarchical voting-based feature selection and ensemble learning model scheme for glioma grading with clinical and molecular characteristics. *International Journal of Molecular Sciences* **23**(22), 141155.

Tierney, L. and Kadane, J. B. (1986) Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association* **81**(393), 82–86.

Van der Vaart, A. W. (2000) *Asymptotic Statistics*. Cambridge University Press.

Vehtari, A., Gelman, A., Sivula, T., Jylänki, P., Tran, D., Sahai, S., Blomstedt, P., Cunningham, J. P., Schiminovich, D. and Robert, C. P. (2020) Expectation propagation as a way of life: A framework for Bayesian inference on partitioned data. *Journal of Machine Learning Research* **21**, 1–53.

Venables, W. N. and Ripley, B. D. (2002) *Modern Applied Statistics with S*. Fourth edition. Springer.

Von Mises, R. (1931) *Wahrscheinlichkeitsrechnung*. Springer Verlagr.

Wang, C. and Blei, D. M. (2013) Variational inference in nonconjugate models. *Journal of Machine Learning Research* **14**, 1005–1031.

Wang, J., Boyer, J. and Genton, M. G. (2004) A skew-symmetric representation of multivariate distributions. *Statistica Sinica* **14**(4), 1259–1270.

Wang, Y. and Blei, D. M. (2019) Frequentist consistency of variational Bayes. *Journal of the American Statistical Association* **114**(527), 1147–1161.

Weng, R. C. (2010) A Bayesian Edgeworth expansion by Stein's identity. *Bayesian Analysis* **5**(4), 741–763.

Yang, Y., Pati, D. and Bhattacharya, A. (2020) α-variational inference with statistical guarantees. *The Annals of Statistics* **48**(2), 886–905.

Zemyan, S. M. (2005) On the zeroes of the nth partial sum of the exponential series. *The American Mathematical Monthly* **112**(10), 891–909.

Zwillinger, D. and Jeffrey, A. (2007) *Table of Integrals, Series, and Products*. Elsevier.