On the robustness to misspecification of α-posteriors and their variational approximations*

Marco Avella Medina† José Luis Montiel Olea† Cynthia Rush†
Amilcar Velez‡

March 8, 2021

Abstract

α-posteriors and their variational approximations distort standard posterior inference by downweighting the likelihood and introducing variational approximation errors. We show that such distortions, if tuned appropriately, can help reducing the Kullback-Leibler (KL) divergence to the true, perhaps infeasible, posterior distribution when there is potential parametric model misspecification. To make this point, we derive a Bernstein-von Mises theorem in total variation distance for α-posteriors and their variational approximations. We use the limiting Gaussian distributions to evaluate the KL divergence between true and reported posteriors. We show this divergence is minimized by choosing α strictly smaller than one, assuming there is a vanishingly small probability of misspecification. The optimized value is smaller the worse the misspecification. The optimized KL divergence increases logarithmically in the degree of misspecifcation and not linearly as with the usual posterior.

Keywords: α-posterior, variational inference, model misspecification.

*We would like to thank Pierre Alquier, Stéphane Bonhomme, Debdeep Pati, and Yun Ju for extremely helpful comments and suggestions. All errors remain our own.
†Columbia University
‡Northwestern University.

1 Introduction

A recent body of work in Bayesian statistics and probabilistic machine learning argues that α-posterior distributions and their variational approximations are more robust to model misspecification than standard Bayesian inference.[1] The α-posteriors (also known as fractional, tempered, or power posteriors) differ from regular posterior distributions in that Bayes' rule is based on the product of the prior with the α-power of the likelihood.[2] Their variational approximations are defined as those distributions that minimize the Kullback-Leibler (KL) divergence to the α-posterior within some tractable subclass; see Definition 1.2 in Alquier and Ridgway (2020).

In this work, we will investigate what it means for α-posteriors and their variational approximations to be robust to model misspecification with a focus on parametric, low-dimensional models. Our suggestion—motivated by the seminal work of Gustafson (2001)—is based on a simple idea: if two different procedures both lead to incorrect a posteriori inference (either due to model misspecification or computational considerations), one procedure is more robust than the other if it is closer—in terms of KL divergence—to the *true* posterior. Thus, this paper analyzes the KL divergence between true posteriors and the distributions reported by either the α-posterior approach or their variational approximations.

To analyze our suggested measure of robustness, we rely on asymptotic approximations to α-posteriors and their variational approximations. We establish a Bernstein-von Mises theorem (BvM) in total variation distance for α-posteriors (Theorem 1) and for their (Gaussian mean-field) variational approximations (Theorem 2) that allows for both model misspecification and non i.i.d. data. The main assumptions are that the likelihood ratio of the presumed

[1]The robustness of α-posteriors to model misspecification is discussed in Bhattacharya, Pati, and Yang (2019); Miller and Dunson (2019); and in the seminal work on *safe-Bayes* by Grünwald (2012) and Grünwald and Van Ommen (2017). Also, the work on β-variational autoencoders by Higgins, Matthey, Pal, Burgess, Glorot, Botvinick, Mohamed, and Lerchner (2017) and Burgess, Higgins, Pal, Matthey, Watters, Desjardins, and Lerchner (2018) has provided some evidence on the benefits of introducing an additional hyperparameter to the variational optimization problem.

[2]A textbook definition of α-posteriors can be found in Chapter 8.6 in Ghosal and Van der Vaart (2017).

model is stochastically locally asymptotically normal (LAN) as in Kleijn and Van der Vaart (2012) and that the α-posterior concentrates around the (pseudo)-true parameter at rate $\sqrt{n}$.[3]

Using these approximations, our suggested measure of robustness can be reduced to finding the KL divergence between multivariate Gaussians having parameters that depend on the data, the sample size, and the 'curvature' of the likelihood. One interesting observation is that relative to the BvM theorem for the standard posterior or its variational approximation, the choice of α only re-scales the limiting variance. The new scaling is as if the observed sample size were $\alpha \cdot n$ instead of n, but the location for the Gaussian approximation continues to be the Maximum Likelihood estimator. Thus, the posterior mean of α-posteriors and their variational approximations has the same limit regardless of the value of α.

When computing our measures of robustness, we think of a researcher that, ex-ante, places some small exogenous probability ϵ_n of model misspecification and is thus interested in computing an *expected* KL. This forces us to analyze the KL divergence under both correct and incorrect specification. We assume that as the sample size n increases, the probability of misspecification decreases as $n\epsilon_n \to \varepsilon$ for constant $\varepsilon \in (0, \infty)$. We establish three main results (Theorem 3), which we think speak to the robustness of α-posteriors and their variational approximations.

First, for a large enough sample size, the expected KL divergence between α-posteriors and the true posterior is minimized for some $\alpha_n^* \in (0, 1)$. This means that, for a properly tuned value of $\alpha \in (0, 1)$, inference based on the α-posterior is asymptotically more robust than regular posterior inference (corresponding to the case $\alpha = 1$). Our calculations suggest that α_n^* tends to be smaller as both the probability of misspecification ϵ_n and the difference between the true and pseudo-true parameter increase, where by pseudo-true parameter we mean the point in the parameter space that provides the best approximation (in terms of

[3]We thus provide a generalization of the results in Wang and Blei (2019a,b), who focus on the case in which $\alpha = 1$. We also extend the results of Li, Wang, and Yu (2019), who establish the BvM theorem for α-posteriors under a weaker norm, but under more primitive conditions.

KL divergence) to the true data generating process. In other words, this analysis makes the reasonable suggestion that as the probability of the likelihood function being wrong increases, one should put less emphasis on it when computing the posterior.

Second, we demonstrate that the Gaussian mean-field variational approximations of α-posteriors inherit some of the robustness properties of α-posteriors. In particular, it is shown that the expected KL divergence between the true posterior and the *mean-field* variational approximation to the α-posterior is also minimized at some $\tilde{\alpha}_n^* \in (0, 1)$. Our second result thus provides support to the claim that variational approximations to α-posteriors are more robust to model misspecification than regular variational approximations. Our result also provides some theoretical support for the recent work of Higgins et al. (2017) that suggests introducing an additional hyperparameter β, which tempers the posterior when implementing variational inference.[4]

Finally, we contrast the expected KL of the optimized α-posterior (and also of the optimized α-variational approximation) against the expected KL of the regular posterior. We find that the latter increases linearly in the magnitude of misspecification (measured by the difference between the true parameter θ_0 and the pseudo-true parameter θ^*), while the former does so logarithmically. This suggests that when the model misspecification is large, there will be significant gains in robustness from using α-posteriors and their variational approximations.

1.1 Related Work

A large part of the theoretical literature studying the robustness of α-posteriors and their variational approximations has focused on nonparametric or high-dimensional models, and in these works, the term robustness has been used to mean that, in large samples, α-posteriors and their variational approximations *concentrate* around the pseudo-true parameter even

[4]See Equation 8 below and the discussion that follows it.

when the standard posterior does not.[5] Although the comparison of the conditions required to verify concentration properties is natural in nonparametric or high-dimensional models, there are alternative suggestions in the literature that are relevant for the analysis of low-dimensional parametric models. For example, Wang and Blei (2019a) compare the posterior predictive distribution based on the regular posterior and also its variational approximation. They show that under likelihood misspecification, the difference between the two posterior predictive distributions converges to zero. Grünwald and Van Ommen (2017) also used (in-sample) predictive distributions to assess the performance of α-posteriors. They show that in a misspecified linear regression problem (where the researcher assumes homoskedasticity, but the data is heteorskedastic) α-posteriors perform better than regular posteriors.

Our work complements this recent body of work by demonstrating robustness in a different sense, but connecting to previous work on the subject. While it is true that one may just be content with studying 'first order' properties of α-posteriors (like contraction) whenever the BvM does not hold, it is not yet clear how to properly formalize the first order benefits of α-posteriors and their variational approximations under misspecification.[6] Our results exploit heavily the BvM theorem which, in a sense, is based on a 'second order' approximations. Perhaps our results can be helpful to study some more complicated models where BvM results are available, but the underlying model is not necessarily low-dimensional and/or parametric; for example, semiparametric models where the object of interest are smooth functionals as in Castillo and Rousseau (2015).

More importantly, our results are an attempt to provide a rationale for the use of variational inference methods based on robustness to model misspecification, as opposed to solely computational considerations. We hope that our results will help to encourage the use of

[5] Bhattacharya et al. (2019) illustrate this point by providing examples of heavy-tailed priors in both density estimation and regression analysis where the α-posterior can be guaranteed to concentrate at (near) minimax rate, but the regular posterior cannot. Alquier and Ridgway (2020) derive concentration rates for α-posteriors for high-dimensional and non-parametric models.

[6] Yang, Pati, and Bhattacharya (2020) show that the necessary conditions for optimal contraction in misspecified models can be relaxed, but there are no results yet that formally tease apart the benefits of α-posteriors compared to the usual posteriors.

variational methods in fields like applied statistics and economics, where variational inference remains somewhat underutilized despite its tremendous impact in machine learning.

The rest of this paper is organized as follows. Section 2 presents definitions, notations, and our general framework. Section 3 presents the Bernstein-von Mises theorem for α-posteriors and their variational approximations. Section 4 presents our suggested measure of robustness and its theoretical analysis. Section 5 presents an example concerning a Gaussian linear regression model with omitted variables. Section 6 presents some concluding remarks and discussion. The proofs of our main results are collected in Appendix A.

2 Notation and General Framework

2.1 Statistical Model

Let $\mathcal{F}_n \equiv \{f_n(\cdot \,|\, \theta) : \theta \in \Theta \subseteq \mathbb{R}^p\}$ be a parametric family of densities used as a statistical model for the random vector $X^n \equiv (X_1, \ldots, X_n)$. The statistical model may be *misspecified*, in the sense that if $f_{0,n}$ is the true density for the random vector X^n, then $f_{0,n}$ may not belong to $\mathcal{F}_n$. As usual, we define the Maximum Likelihood (ML) estimator—denoted by $\widehat{\theta}_{\text{ML-}\mathcal{F}_n}$—as

$$\widehat{\theta}_{\text{ML-}\mathcal{F}_n} \equiv \underset{\theta \in \Theta}{\arg\max}\, f_n(X^n \,|\, \theta). \tag{1}$$

For simplicity, we assume that the ML estimator is unique, and that there is a parameter value θ^* in the interior of Θ for which $\sqrt{n}\,(\widehat{\theta}_{\text{ML-}\mathcal{F}_n} - \theta^*)$ is asymptotically normal.[7] If the model is correctly specified, then θ^* is simply the true parameter, but if the model is misspecified, then θ^* provides the best approximation (in terms of KL divergence) to the true data generating process. In the latter misspecified case, it is then common to refer to θ^* as the *pseudo-true*

[7] Huber (1967); White (1982); Kleijn and Van der Vaart (2012) present sufficient conditions for the consistency and asymptotic normality of the ML estimator under model misspecification, with i.i.d. data. Examples of other papers establishing consistency and asymptotic normality under misspecification for certain types of non i.i.d. data are given in Pouzo, Psaradakis, and Sola (2016).

parameter. We focus on the case in which the ML estimator is asymptotically normal to highlight the fact that none of our results depend on atypical asymptotic distributions.

The main restriction we impose on $\mathcal{F}_n$ is that its corresponding likelihood ratio process approximates, asymptotically, that of a normal random variable. Formally, we impose the $\sqrt{n}$-stochastic local asymptotic normality (LAN) condition of (Kleijn and Van der Vaart, 2012), around θ^*.

Assumption 1 *Denote* $\Delta_{n,\theta^*} \equiv \sqrt{n}\,(\widehat{\theta}_{ML\text{-}\mathcal{F}_n} - \theta^*)$*. There exists a positive definite matrix* V_{θ^*} *such that*

$$R_n(h) \equiv \log\left(\frac{f_n\left(X^n \,|\, \theta^* + h/\sqrt{n}\right)}{f_n\left(X^n \,|\, \theta^*\right)}\right) - h^\top V_{\theta^*}\,\Delta_{n,\theta^*} + \frac{1}{2}\,h^\top V_{\theta^*}\,h \tag{2}$$

satisfies

$$\sup_{h\in K} |R_n(h)| \to 0, \tag{3}$$

in $f_{0,n}$*-probability, for any compact set* $K \subseteq \mathbb{R}^p$*.*[8]

Maybe add a short notation paragraph with definitions in footnotes 13 and 14?

2.2 α-posteriors and their Variational Approximations

We now present the definition of α-posteriors and their variational approximations. Starting from the statistical model $\mathcal{F}_n$, a prior density π for θ, and a scalar $\alpha > 0$, the α-posterior is defined as the distribution having density:

$$\pi_{n,\alpha}(\theta \,|\, X^n) \equiv \frac{[f_n(X^n \,|\, \theta)]^\alpha\, \pi(\theta)}{\int [f_n(X^n \,|\, \theta)]^\alpha\, \pi(\theta)\, d\theta}\,, \tag{5}$$

[8]Notice that

$$-\,h^\top V_{\theta^*}\,\Delta_{n,\theta^*} + \frac{1}{2}\,h^\top V_{\theta^*}\,h = \log\left(\frac{\phi(\theta^* + h/\sqrt{n}\,|\,\widehat{\theta}_{\text{ML-}\mathcal{F}_n},\,(nV_{\theta^*})^{-1})}{\phi(\theta^*\,|\,\widehat{\theta}_{\text{ML-}\mathcal{F}_n},(nV_{\theta^*})^{-1})}\right) \tag{4}$$

where $\phi(\cdot|\mu,\Sigma)$ is the density of a multivariate normal random vector with mean μ and covariance matrix Σ.

see Chapter 8.6 in (Ghosal and Van der Vaart, 2017) for a textbook definition.

The projection of (5)—in Kullback-Leibler (KL) divergence—onto the space of probability distributions with independent marginals also referred to as the mean-field family and denoted $\mathcal{Q}_{\mathrm{MF}}$, provides the *mean-field variational approximation* to the α-posterior:

$$\tilde{\pi}_{n,\alpha}(\cdot \,|\, X^n) \in \underset{q \,\in\, \mathcal{Q}_{\mathrm{MF}}}{\arg\min}\, \mathcal{K}(q \,||\, \pi_{n,\alpha}(\cdot \,|\, X^n))\,, \tag{6}$$

where $\mathcal{K}(\cdot \,||\, \cdot)$ denotes the KL divergence, defined as

$$\mathcal{K}(q \,||\, \pi_{n,\alpha}(\cdot \,|\, X^n)) \equiv \int q(\theta) \log\left(\frac{q(\theta)}{\pi_{n,\alpha}(\theta \,|\, X^n)}\right) d\theta\,. \tag{7}$$

Equation (6) is a particular case of the more general variational approximations studied in the recent work of (Alquier and Ridgway, 2020), who allow for other sets of distributions over which the KL is minimized.

There is a trade-off between choosing a flexible and rich enough domain for the optimization in (6), so that q can be close to $\pi_{n,\alpha}(\theta \,|\, X^n)$, and choosing a domain that is also constrained enough such that the optimization is computationally feasible. A key insight of the variational framework is that minimizing the KL divergence in (6) is equivalent to solving the program

$$\tilde{\pi}_{n,\alpha}(\cdot \,|\, X^n) \equiv \underset{q \,\in\, \mathcal{Q}_{\mathrm{MF}}}{\arg\min} \left\{ \int q(\theta) \log\left(f_n(X^n \,|\, \theta)\right) \, d\theta - (1/\alpha)\, \mathcal{K}(q \,||\, \pi) \right\}\,. \tag{8}$$

The objective function in (8) is reminiscent of penalized estimation: it involves a data-fitting term (the average log-likelihood) and a regularization or penalization term that forces the distribution q to be close to a baseline prior π with regularization parameter $1/\alpha$.

The optimization problem in (8) has been the subject of recent work in the representation learning literature.[9] It has been argued that values of $\alpha < 1$ result in 'disentangled' latent

[9]See the definition of β-Variational Auto Encoder in (Burgess et al., 2018) and (Higgins et al., 2017).

representations.[10] The optimization problem has also been studied in axiomatic decision theory; see, for example, the *multiplier preferences* introduced in Hansen and Sargent (2001) and their axiomatization in Strzalecki (2011).[11]

3 Bernstein-von Mises Theorem for α-posteriors and their Variational Approximations

Before presenting a formal definition of the measure of robustness used in this paper, we show that α-posteriors and their variational approximations are asymptotically normal. This extends the Bernstein-von Mises (BvM) theorem for misspecified models in Kleijn and Van der Vaart (2012) (which shows that posteriors under misspecified models are asymptotically normal) and the recent Variational BvM of Wang and Blei (2019a) (which shows that variational approximations of true posteriors are asymptotically normal).

3.1 BvM for α-posteriors

We say that the α-posterior, $\pi_{n,\alpha}(\cdot \mid X^n)$ defined in (5), concentrates at rate $\sqrt{n}$ around θ^* if for every sequence of constants $r_n \to \infty$

$$\mathbb{E}_{f_{0,n}} \left[\int \mathbf{1} \left\{ \, \|\sqrt{n}\,(\theta - \theta^*)\| > r_n \, \right\} \, \pi_{n,\alpha}(\theta \mid X^n) \, d\theta \right] \to 0\,, \tag{9}$$

[10] Disentangled can be defined, loosely speaking, as distributions where 'single latent units are sensitive to changes in single generative factors, while being relatively invariant to changes in other factors'.

[11] More generally, the objective function in (8) with an arbitrary divergence function is analogous to the so-called *divergence preferences* studied in Maccheroni, Marinacci, and Rustichini (2006). We think this literature could be potentially useful in understanding the role of the different divergence functions in penalizations, as well as the multiplier parameter α.

where $\mathbf{1}\{A\}$ is the indicator function of event A, namely the function that takes the value 1 if event A occurs and 0 otherwise. Notice that

$$\mathbb{E}_{f_{0,n}}\left[\int \mathbf{1}\left\{ \|\sqrt{n}\,(\theta-\theta^*)\| > r_n \right\} \pi_{n,\alpha}(\theta \,|\, X^n)\, d\theta\right] \\ = \mathbb{E}_{f_{0,n}}\left[\mathbb{P}_{\pi_{n,\alpha}(\cdot \,|\, X^n)}\left(\|\sqrt{n}\,(\theta-\theta^*)\| > r_n\right)\right] .$$

Theorem 1 *Let $\phi(\cdot \,|\, \mu, \Sigma)$ denote the density of a multivariate normal random vector with mean μ and covariance matrix Σ. Suppose that the prior density π is continuous and positive on a neighborhood around the (pseudo-) true parameter θ^*, and that $\pi_{n,\alpha}(\cdot \,|\, X^n)$ concentrates at rate $\sqrt{n}$ around θ^* as in (9). If Assumption 1 holds, then*

$$d_{TV}\left(\pi_{n,\alpha}(\cdot \,|\, X^n),\, \phi(\cdot \,|\, \widehat{\theta}_{ML\text{-}\mathcal{F}_n},\, V_{\theta^*}^{-1}/(\alpha n))\right) \to 0\,, \tag{10}$$

in $f_{0,n}$-probability, where $d_{TV}(p,q)$ denotes the total variation distance between two densities p and q and V_{θ^*} is the positive definite matrix satisfying Assumption 1.[12]

In a nutshell, the theorem states that the α-posterior distribution behaves asymptotically as a multivariate normal distribution, centered at the ML estimator, $\widehat{\theta}_{\text{ML-}\mathcal{F}_n}$, which is based on the potentially misspecified model $\mathcal{F}_n$. Thus, the theorem shows that the choice of α does not asymptotically affect the location of the α-posterior distribution.

However, Theorem 1 shows that the asymptotic covariance matrix of the α-posterior is given by $V_{\theta^*}^{-1}/(\alpha n)$, hence, the parameter α inflates the asymptotic variance when $\alpha < 1$, and deflates it otherwise. The matrix V_{θ^*} is the second-order term in the stochastic LAN approximation in Assumption 1, and it is the usual variance in the BvM theorem for correctly or incorrectly specified models. Intuitively, V_{θ^*} can be thought of as measuring the curvature of the likelihood.

[12] If p and q are two densities with respect to Lebesgue measure in $\mathbb{R}^p$, we define the total variation distance between them as

$$d_{\text{TV}}(p,\, q) \equiv \frac{1}{2}\int_{\mathbb{R}^p} |\, p(u) - q(u)\,|\, du. \tag{11}$$

See Section 2.9 in (Van der Vaart, 2000).

The proof and its details are presented in Appendix A.1. For the sake of exposition, we present a brief intuitive argument for why the result should hold. By assumption, the α-posterior concentrates around θ^* at rate $\sqrt{n}$, in the sense of (9). Consider the log-likelihood ratio, for some vector $h \in \mathbb{R}^d$,

$$\log\left(\frac{\pi_{n,\alpha}(\theta^* + h/\sqrt{n} \mid X^n)}{\pi_{n,\alpha}(\theta^* \mid X^n)}\right) = \log\left(\left[\frac{f(X^n \mid \theta^* + h/\sqrt{n})}{f(X^n \mid \theta^*)}\right]^{\alpha} \frac{\pi(\theta^* + h/\sqrt{n})}{\pi(\theta^*)}\right). \tag{12}$$

If $\pi_{n,\alpha}(\cdot \mid X^n)$ were exactly a multivariate normal with mean $\widehat{\theta}_{\text{ML-}\mathcal{F}_n}$ and covariance matrix $V_{\theta^*}^{-1}/(\alpha n)$ the log-likelihood ratio in (12) would equal

$$h^\top(\alpha V_{\theta^*})\sqrt{n}\,(\widehat{\theta}_{\text{ML-}\mathcal{F}_n} - \theta^*) - \frac{1}{2}h^\top(\alpha V_{\theta^*})h. \tag{13}$$

The log-likelihood ratios in (12) and (13) are not equal, but the continuity of π at θ^* and the stochastic LAN property of Assumption 1 makes them equal up to an $o_{f_{0,n}}(1)$ term.[13]

Most of the work in the proof of Theorem 1 consists of relating the closeness in log-likelihood ratios to closeness in total variation distance. The arguments we use to make this connection follow verbatim the arguments used by Kleijn and Van der Vaart (2012). To the best of our knowledge the result in Theorem 1 has not appeared previously in the literature, although Section 4.1 in Li et al. (2019) present a different version of this result in a weaker metric (convergence in distribution, as opposed to total variation), but using lower-level conditions (as opposed to our high-level assumptions). Also, the result for $\alpha = 1$, i.e. the usual posterior under model misspecification, appears in Wang and Blei (2019a).

3.2 BvM for Variational Approximations of α-posteriors

Given that the $\pi_{n,\alpha}$ is close to a multivariate normal distribution, it is natural to conjecture that the variational approximation $\tilde{\pi}_{n,\alpha}$ will converge to the projection of such multivariate

[13] $X_n = o_{f_{0,n}}(1)$ if $\lim_n \mathbb{P}_{f_{0,n}}(\|X_n\| > \epsilon) = 0$ for every $\epsilon > 0$.

normal distribution onto the mean-field family $\mathcal{Q}_{\mathrm{MF}}$.

To formalize this argument, let $\mathcal{Q}_{\mathrm{GMF}-p}$ denote the family of multivariate normal distributions of dimension p with independent marginals. An element in this family is parameterized by a vector $\mu \in \mathbb{R}^p$ and a positive semi-definite diagonal covariance matrix $\Sigma \in \mathbb{R}^{p\times p}$. When convenient, we denote such an element as $q(\cdot|\mu, \Sigma)$ and will always implicitly assume that Σ is diagonal.

We focus on the *Gaussian mean-field approximation to the α-posterior.* Given a sample of size n, this can be defined as the Gaussian distribution with parameters $\tilde{\mu}_n, \tilde{\Sigma}_n$ satisfying

$$\tilde{\pi}_{n,\alpha}(\cdot \mid X^n) = q(\cdot \mid \tilde{\mu}_n,\, \tilde{\Sigma}_n) \in \underset{q \,\in\, \mathcal{Q}_{\text{GMF-p}}}{\arg\min}\; \mathcal{K}(q(\cdot) \,||\, \pi_{n,\alpha}(\cdot \mid X^n)). \tag{14}$$

Theorem 1 has shown that $\pi_{n,\alpha}(\cdot \mid X^n)$ is close to a multivariate normal distribution with mean $\widehat{\theta}_{\text{ML-}\mathcal{F}_n}$ and variance $V_{\theta^*}^{-1}/(\alpha n)$. Let $q(\cdot \mid \mu_n^*, \Sigma_n^*)$ denote the multivariate normal distribution in the Gaussian mean-field family closest to such limit. That is,

$$q(\cdot \mid \mu_n^*, \Sigma_n^*) \;=\; \underset{q \,\in\, \mathcal{Q}_{\text{GMF-p}}}{\arg\min}\; \mathcal{K}(q(\cdot) \,||\, \phi(\cdot \mid \widehat{\theta}_{\text{ML-}\mathcal{F}_n},\, V_{\theta^*}^{-1}/(\alpha n))). \tag{15}$$

Algebra shows that the optimization problem (15) has a simple (and unique) closed-form solution when V_{θ^*} is positive definite; namely:

$$\mu_n^* = \widehat{\theta}_{\text{ML-}\mathcal{F}_n}, \quad \Sigma_n^* = \operatorname{diag}(V_{\theta^*})^{-1}/(\alpha n). \tag{16}$$

In words, $q(\cdot \mid \mu_n^*, \Sigma_n^*)$ is the distribution in the Gaussian mean-field family having the same marginal distributions as the limiting distribution of $\pi_{n,\alpha}(\cdot \mid X^n)$. We would like to show that the total variation distance between the distributions $\tilde{\pi}_{n,\alpha}(\cdot \mid X^n;\, \tilde{\mu}_n,\, \tilde{\Sigma}_n)$ and $q(\cdot|\mu_n^*, \Sigma_n^*)$ converges in probability to zero, provided the prior and the likelihood satisfy some regularity conditions. To do this, let θ^* and $R_n(h)$ be defined as in Assumption 1.

Assumption 2 *The prior π and the residual $R_n(h)$ in Assumption 1 are such that for any sequence (μ_n, Σ_n) for which $(\sqrt{n}(\mu_n - \theta^*), n\Sigma_n)$ is bounded in $f_{0,n}$-probability*

$$\int \phi(h \,|\, \sqrt{n}(\mu_n - \theta^*), n\Sigma_n) \log\left(\frac{\pi(\theta^* + h/\sqrt{n})}{\pi(\theta^*)}\right) dh \;\to 0, \tag{17}$$

and

$$\int \phi(h \,|\, \sqrt{n}(\mu_n - \theta^*), n\Sigma_n) \, R_n(h) \, dh \to 0. \tag{18}$$

In both cases the convergence is in $f_{0,n}$-probability.[14]

Theorem 2 *Suppose that the sequence $(\tilde{\mu}_n, \tilde{\Sigma}_n)$ defining $\tilde{\pi}_{n,\alpha}(\cdot|X^n) = q(\cdot|\tilde{\mu}_n, \tilde{\Sigma}_n)$ is such that $(\sqrt{n}(\tilde{\mu}_n - \theta^*), n\tilde{\Sigma}_n))$ is bounded in $f_{0,n}$-probability. If Assumptions 1 and 2 hold, then*

$$d_{TV}\left(\tilde{\pi}_{n,\alpha}(\cdot \mid X^n),\, q(\cdot \mid \mu_n^*, \Sigma_n^*)\right) \to 0\,, \tag{19}$$

in $f_{0,n}$-probability, where μ_n^* and Σ_n^* are defined in (16).

In words, Theorem 2 shows that the Gaussian mean-field approximation to the α-posterior converges to the Gaussian mean-field approximation of asymptotic distribution of the α-posterior. Indeed, the mean and variance parameter of this normal distribution are obtained by *projecting* the limiting distribution obtained in Theorem 1 onto the Gaussian mean-field family.

A detailed proof of Theorem 2 can be found in Appendix A.2. A similar result was obtained by Wang and Blei (2019a) for the case of $\alpha = 1$. Thus, Theorem 2 can be viewed as a generalization of their variational BvM Theorem, applicable to the variational approximations of α-posteriors. We note however that our proof technique is quite different from theirs. Indeed, we require a simpler set of assumptions because we restrict ourselves to Gaussian

[14] A sequence of random variables X_n is bounded in $f_{0,n}$-probability if for every $\epsilon > 0$ there exists $M_\epsilon > 0$ such that $\mathbb{P}_{f_{0,n}}(\|X_n\| < M_\epsilon) \geq 1 - \epsilon$

mean-field variational approximations to the α-posterior. This enables us to work out a simplified argument that explicitly leverages formulas obtained by computing the KL divergence between two Gaussians. The key intermediate step in our proof is the asymptotic representation result stated in Lemma 3 in the Appendix. It shows that, under Assumption 1 and 2, for any sequence (μ_n, Σ_n) such that $(\sqrt{n}(\mu_n - \theta^*), n\Sigma_n)$ is bounded in $f_{0,n}$-probability we have that:

$$\begin{aligned}
&\mathcal{K}(q(\cdot \,|\, \mu_n,\, \Sigma_n) \,||\, \pi_{n,\alpha}(\cdot \,|\, X^n)) = \\
&\qquad \mathcal{K}\left(q(\cdot \,|\, \mu_n,\, \Sigma_n) \,||\, \phi(\cdot \,|\, \widehat{\theta}_{\text{ML-}\mathcal{F}_n},\, V_{\theta^*}^{-1}/(\alpha n))\right) + o_{f_{0,n}}(1).
\end{aligned}$$

We use this lemma to show that projecting the α-posterior onto the space of Gaussian mean-field distributions is approximately equal to projecting the α-posterior's total variation limit in Theorem 1. In particular, Theorem 2 shows that

$$\mathcal{K}(\tilde{\pi}_{n,\alpha}(\cdot \mid X^n; \tilde{\mu}_n, \tilde{\Sigma}_n) \,||\, q(\cdot \,|\, \mu_n^*, \Sigma_n^*)) \to 0, \tag{20}$$

in $f_{0,n}$-probability. The statement in (19) follows from the above limit by Pinsker's inequality i.e. $d_{TV}(P, Q) \leq \sqrt{2\mathcal{K}(P \,||\, Q)}$ for any two probability distributions P and Q.[15]

4 Misspecification Robustness Analysis

This section introduces the measure of robustness suggested in this paper and the main results. As we will explain below, the main idea is to measure closeness—in terms of KL divergence—of the α-posterior or their variational approximations to the correct posterior.

To formalize this discussion, we introduce a bit more of notation. In order to analyze model misspecification, following Gustafson (2001), we posit a correctly specified parametric model that takes the form $\mathcal{G}_n \equiv \{g_n(\cdot \,|\, \theta, \gamma) : \theta \in \Theta, \gamma \in \Gamma\}$ (notice that under model

[15]See part iii) of Lemma B.1 in Ghosal and Van der Vaart (2017) for a textbook reference on Pinsker's inequality

misspecification, $\mathcal{G}_n$ will differ from $\mathcal{F}_n$). Correct specification of $\mathcal{G}_n$ here simply means that, for any sample size n, there exist parameters $(\theta_n^*, \gamma_n^*) \in \Theta \times \Gamma$ for which $g_n(\cdot \,|\, \theta_n^*, \gamma_n^*)$ equals $f_{0,n}$. Note that the parameter θ is well-defined in both $\mathcal{F}_n$ and $\mathcal{G}_n$.

Let π^* be a prior over $\Theta \times \Gamma$. We let $\pi_n^*(\theta \,|\, X^n)$ denote the posterior for θ based on $\mathcal{G}_n$ and π^*. When $\mathcal{F}_n$ is misspecified, we refer to $\pi_n^*(\theta \,|\, X^n)$ as the *true* posterior, and when $\mathcal{F}_n$ is correctly specified, the true posterior is simply given by $\pi_{n,\alpha}$ evaluated at $\alpha = 1$. In a slight abuse of notation, we denote the ML estimator of θ based on $\mathcal{G}_n$ simply as $\widehat{\theta}_{\mathrm{ML}}$.

Our goal is to compute the KL divergence between α-posteriors (or their variational approximations) and the correct posterior. To do so, we need to consider two cases: one in which $\mathcal{F}_n$ is misspecified (in the sense that $f_{0,n} \notin \mathcal{F}_n$) and another in which $\mathcal{F}_n$ is correctly specified. We assume that the decision maker does not know ex-ante whether the model is correctly specified or not, and we denote by ϵ_n the probability of being misspecified (consequently, $1 - \epsilon_n$ denotes the probability of correct specification).

To assess the robustness of α-posteriors our interest is to compute the expected value of KL divergence:

$$r_n(\alpha) \equiv \epsilon_n \mathcal{K}(\pi_n^*(\theta|X^n) \,||\, \pi_{n,\alpha}(\theta|X^n)) + (1 - \epsilon_n)\mathcal{K}(\pi_{n,1}(\theta|X^n) \,||\, \pi_{n,\alpha}(\theta|X^n)). \tag{21}$$

The first term in (21) measures how difficult is to distinguish between the *true* posterior under model misspecification and the α-posterior. The second term is the KL divergence between the regular posterior ($\alpha = 1$) and the α-posterior. These terms are weighted by the probability of misspecification. Values of α that lead to smaller values of $r_n(\alpha)$ are said to be more robust to parametric specification, as the corresponding α-posterior is 'closer' to the true posterior. To the best of our knolewdge, the idea of using the KL divergence between reported posteriors and true posteriors was first introduced by Gustafson (2001).

Likewise, we could analyze the robustness of variational α-posteriors by studying:

$$\tilde{r}_n(\alpha) \equiv \epsilon_n \mathcal{K}(\pi_n^*(\theta|X^n) \,||\, \tilde{\pi}_{n,\alpha}(\theta|X^n)) + (1-\epsilon_n)\mathcal{K}(\pi_{n,1}(\theta|X^n) \,||\, \tilde{\pi}_{n,\alpha}(\theta|X^n)), \tag{22}$$

where (22) is the same as (21), except we have replaced the α-posterior $\pi_{n,\alpha}(\theta|X^n)$ in (21) with its variational approximation $\tilde{\pi}_{n,\alpha}(\theta|X^n)$ in (22). Both $r_n(\alpha)$ and $\tilde{r}_n(\alpha)$ are random variables (where the randomness comes from the sampled data X^n used to construct the posterior distributions) and their magnitudes depend on the structure of the correctly and incorrectly specified models, on the priors, and on the sample size. We are able to make progress on the analysis of (21) and (22) by relying on asymptotic approximations to the infeasible posterior, the regular posterior, and the α-posterior and its variational approximation.

It is well-known that the Bernstein-von Mises theorem for correctly specified models—e.g., (DasGupta, 2008), p. 291—implies that under some regularity conditions on the statistical model $\mathcal{G}_n$ and the prior π^* (analogous to Assumption 1 and (9)), the true, infeasible posterior $\pi_n^*(\theta|X^n)$ is close—in total variation distance—to the p.d.f. of a $\mathcal{N}(\widehat{\theta}_{\mathrm{ML}}, \Omega/n)$ random variable. Theorems 1 and 2, naturally suggests surrogates for (21) and (22) where we replace the densities by their asymptotically normal approximations. We denote the surrogate measures by $r_n^*(\alpha)$ and $\tilde{r}_n^*(\alpha)$. Define

$$\alpha_n^* \equiv \arg\min_{\alpha \geq 0} r_n^*(\alpha), \qquad \tilde{\alpha}_n^* \equiv \arg\min_{\alpha \geq 0} \tilde{r}_n^*(\alpha), \tag{23}$$

where

$$\begin{aligned} r_n^*(\alpha) = \epsilon_n \mathcal{K}\left(\phi(\cdot \mid \widehat{\theta}_{\mathrm{ML}}, \Omega/n) \,||\, \phi(\cdot \mid \widehat{\theta}_{\text{ML-}\mathcal{F}_n}, V_{\theta^*}^{-1}/(\alpha n))\right) \\ + (1-\epsilon_n)\mathcal{K}\left(\phi(\cdot \mid \widehat{\theta}_{\text{ML-}\mathcal{F}_n}, V_{\theta^*}^{-1}/n) \,||\, \phi(\cdot \mid \widehat{\theta}_{\text{ML-}\mathcal{F}_n}, V_{\theta^*}^{-1}/(\alpha n))\right). \end{aligned}$$

And

$$\tilde{r}_n^*(\alpha) = \epsilon_n \mathcal{K}(\phi(\cdot \mid \widehat{\theta}_{\mathrm{ML}}, \Omega/n) \,||\, \phi(\cdot \mid \widehat{\theta}_{\mathrm{ML}\text{-}\mathcal{F}_n}, \mathrm{diag}(V_{\theta^*})^{-1}/(\alpha n)))$$
$$+ (1-\epsilon_n)\mathcal{K}(\phi(\cdot \mid \widehat{\theta}_{\mathrm{ML}\text{-}\mathcal{F}_n}, V_{\theta^*}^{-1}/n) \,||\, \phi(\cdot \mid \widehat{\theta}_{\mathrm{ML}\text{-}\mathcal{F}_n}, \mathrm{diag}(V_{\theta^*})^{-1}/(\alpha n))).$$

Theorem 3 *Let $p \equiv dim(\theta)$ and let V_{θ^*} be the positive definite matrix satisfying Assumption 1 and let $\tilde{V}_{\theta^*} \equiv diag(V_{\theta^*})$. Suppose $\widehat{\theta}_{ML} \to \theta_0$ in $f_{0,n}$-probability. If $\theta_0 \neq \theta^*$ and $n\epsilon_n \to \varepsilon \in (0,\infty)$ then*

$$\alpha_n^* \;\to\; \frac{p}{p + \varepsilon(\theta_0 - \theta^*)^\top V_{\theta^*}(\theta_0 - \theta^*)} < 1, \tag{24}$$

$$\tilde{\alpha}_n^* \;\to\; \frac{p}{\mathrm{tr}(\tilde{V}_{\theta^*} V_{\theta^*}^{-1}) + \varepsilon(\theta_0 - \theta^*)^\top \tilde{V}_{\theta^*}(\theta_0 - \theta^*)} < 1, \tag{25}$$

where the convergence is in $f_{0,n}$-probability.

We now discuss the meaning and implications of Theorem 3. Equation (24) says that, for a properly tuned value of α, the α-posterior is—with high probability and in large samples—more robust than the regular posterior. The limit of α_n^* suggests that the properly tuned value of α decreases as the probability of misspecification increases. This makes conceptual sense: it seems reasonable to down-weight the likelihood if it is known that, very likely, it is misspecified. On the other extreme, if the likelihood is known to be correct there is no gain from down-weighting the likelihood, as this simply creates a difference in the scaling of the α-posterior relative to the true posterior.

The formula also says that if the misspecification implied by the model is large (the difference between θ_0 and θ^* is large), then the properly tuned value of α must be small.

Equation (25) refers to the properly calibrated value of α for the variational approximations of α-posteriors. The analysis is similar to what we have already discussed, but here we need to take into account that variational approximations tend to further distort the variance matrix. Indeed, Theorem 2 has shown that the asymptotic variance of the variational approximations is $\tilde{V}_{\theta^*}^{-1}/\alpha n$, as opposed to $V_{\theta^*}^{-1}/\alpha n$—where we have defined $\tilde{V}_{\theta^*} \equiv \text{diag}(V_{\theta^*})$. Since $\tilde{V}_{\theta^*} = \text{diag}(V_{\theta^*})$ for positive definite V_{θ^*}, it is not difficult to prove that $\text{tr}(\tilde{V}_{\theta^*} V_{\theta^*}^{-1}) \geq p$, which establishes the inequality in (25).

There is one additional derivation associated to the formulae in Theorem 3. Consider the optimized expected value of the KL divergence based on the asymptotic approximations to α-posteriors and their variational approximations. From the definition of r_n^* and Theorem 3:

$$2 \lim_{n\to\infty} r_n^*(\alpha_n^*) = -\, p \log(p) + p \log\left(p + \varepsilon(\theta_0 - \theta^*)^\top V_{\theta^*}(\theta_0 - \theta^*)\right). \tag{26}$$

Likewise,

$$2 \lim_{n\to\infty} \tilde{r}_n^*(\tilde{\alpha}_n^*) =$$

$$-\, p \log(p) + p \log\left(\text{tr}\left(\tilde{V}_{\theta^*} V_{\theta^*}^{-1}\right) + \varepsilon(\theta_0 - \theta^*)' \tilde{V}_{\theta^*}(\theta_0 - \theta^*)\right) + \log\left(\frac{|V_{\theta^*}|}{|\tilde{V}_{\theta^*}|}\right). \tag{27}$$

Compare these two equations with the expected KL of the usual posterior ($\alpha = 1$). In this case, the KL under correct specification equals zero, so that the only relevant piece is the KL distance under misspecification. Algebra shows that

$$2 \lim_{n\to\infty} r_n^*(1) = \varepsilon(\theta_0 - \theta^*)^\top V_{\theta^*}(\theta_0 - \theta^*).$$

Thus, the expected KL distance for the regular posterior increases linearly in the magnitude of the misspecification (the distance between θ_0 and θ^*).The the optimized KL for both the α-posteriors and their variational approximations is also monotonically increasing in this term, but its growth is logarithmic.

One final remark concerns the well-known result that the asymptotic variance of Bayesian posteriors under misspecified models does not coincide with the asymptotic 'sandwich' covariance matrix of the Maximum Likelihood Estimator under misspecification. This suggests that instead of targeting the true (perhaps infeasible) posterior, it might make more sense to target the artificial posterior suggested by Müller (2013), which is a normal centered at the Maximum Likelihood estimator but with sandwich covariance matrix. This choice of target distribution does not change our results. The reason is that under our assumptions the part of the expected Kullback Leibler under misspecification only depends on the difference between true and pseudo-true parameter and the covariance matrix of the reported posterior.

5 Illustrative Example

In order to illustrate our main results, this section presents a linear regression model with omitted variables. Our objective is twofold. First, we would like to present a simple environment where the high-level assumptions of our main theorems can be easily verified and discussed. Second, an appropriate choice of priors in this model yields closed-form solutions for the α-posteriors and their (Gaussian mean-field) variational approximations. Thus, it is possible to provide additional details about the nature of the Bernstein-von Mises theorems that we have established, as well as the optimal choice of α.

5.1 True model and misspecified model

Consider a random sample of an outcome variable Y_i and control variables $W_i \in \mathbb{R}^p$ and $Z_i \in \mathbb{R}^d$. The true data generating process is an homoskedastic Gaussian linear regression model

$$Y_i = \theta_0^\top W_i + \gamma_0^\top Z_i + \varepsilon_i,$$

where $\epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$ independently of W_i and Z_i. The joint distribution of W_i and Z_i is assumed to have a density $h(w_i, z_i)$ w.r.t to Lebesgue's measure in $\mathbb{R}^{p+d}$.

The statistician faces an omitted variables problem, in that he would like to estimate θ_0 but only observes (Y_i, W_i). The statistician's misspecified model posits that

$$Y_i = \theta^\top W_i + u_i,$$

where u_i is assumed to be univariate normal with mean zero and a pressumedly known variance σ_u^2 independently of W_i. For simplicity, we assume that the statistician has correctly specified the marginal distribution of W_i.

5.2 Pseudo-true parameter and LAN assumption

If we denote the data as $X^n \equiv \{(Y_i, W_i)\}_{i=1}^n$, the likelihood is

$$f(X^n|\theta) = \frac{1}{(2\pi\sigma_u^2)^{n/2}} \exp\left(-\frac{1}{2\sigma_u^2}\sum_{i=1}^{n}(Y_i - \theta^\top W_i)^2\right) \Pi_{i=1}^n h(W_i),$$

and the Maximum Likelihood estimator is simply the least-squares estimator of θ:

$$\widehat{\theta}_{ML} = \left(\frac{1}{n}\sum_{i=1}^{n} W_i W_i^\top\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} W_i Y_i.$$

It is straightforward to show that under mild assumptions on the joint distribution of (W_i, Z_i), the Maximum Likelihood estimator is $\sqrt{n}$-asymptotically normal around the pseudo-true parameter:

$$\theta^* \equiv \theta_0 + \left(\mathbb{E}[W_i W_i^\top]^{-1}\mathbb{E}[W_i Z_i^\top]\right)\gamma_0,$$

which equals the true parameter, θ_0, plus the usual omitted variable bias formula. Algebra shows that as long as the sample second moments of W_i converge in probability (under the true model) to the positive definite matrix $\mathbb{E}[W_i W_i^\top]$, the stochastic LAN assumption is satisfied with

$$V_{\theta^*} \equiv \frac{\mathbb{E}[W_i W_i^\top]}{\sigma_u^2}.$$

This means that in this example the curvature of the likelihood around the pseudo-true parameter does not depend on θ^*.

5.3 α-posteriors and their variational approximations

Consider now a commonly used Gaussian prior π on θ. In particular, suppose

$$\theta \sim \mathcal{N}(\mu_\pi, \sigma_u^2 \Sigma_\pi^{-1}).$$

The computation of α-posterior in this set-up is straightforward, as the α-power of the Gaussian likelihood is itself Gaussian with the new scale divided by α. Thus, algebra shows that the α-posterior for the linear regression model is also multivariate normal with mean parameter

$$\mu_{n,\alpha} \equiv \left(\frac{1}{n} \sum_{i=1}^{n} W_i W_i^\top + \frac{1}{\alpha n} \Sigma_\pi \right)^{-1} \left(\frac{1}{\alpha n} \Sigma_\pi \mu_\pi + \frac{1}{n} \sum_{i=1}^{n} W_i Y_i \right), \tag{28}$$

and covariance matrix

$$\Sigma_{n,\alpha} \equiv \frac{\sigma_u^2}{\alpha n} \left(\frac{1}{n} \sum_{i=1}^{n} W_i W_i^\top + \frac{1}{\alpha n} \Sigma_\pi \right)^{-1}. \tag{29}$$

Because we have closed-form solutions for the α-posteriors, one can readily show that they concentrate at rate $\sqrt{n}$ around θ^* for any fixed $(\alpha, \mu_\pi, \Sigma_\pi)$. This is shown in Appendix B.

Since the assumptions of Theorem 1 are met, then the total variation distance between the α-posterior and the multivariate normal

$$\mathcal{N}\left(\widehat{\theta}_{ML}, \frac{\sigma_u^2}{\alpha n} \mathbb{E}[W_i W_i^\top]^{-1} \right), \tag{30}$$

must converge in probability to zero. In fact, in this example, it is possible to establish a stronger result: the KL distance between $\pi_{n,\alpha}$ and the distribution in (30) converges in probability to zero. This example thus raises the question of whether *entropic* Bernstein-von Mises theorems are more generally available for α-posteriors in misspecified models.[16]

[16] Clarke (1999a) showed that—in a smooth parametric model with a well-behaved prior—the relative entropy between a posterior density and an appropriate normal tends to zero in probability and in mean. If an analogous result were available for standard posterior distributions in misspecified parametric models, it might be possible to extend it to cover α-posteriors.

This simple linear regression example also shows that the Bernstein-von Mises theorem is not likely to hold if α is chosen in a way that approaches zero very quickly.[17] In particular, consider a sequence α_n for which $\alpha_n n$ converges to a strictly positive constant. Then, in this simple example the total variation distance between the α_n-posterior and the distribution in (30) will be bounded away from zero.[18]

Finally, in this example the mean-field Gaussian variational approximation of the α-posterior also has a closed-form expression. Algebra shows that the variational approximation has exactly the same mean as the α-posterior, but variance

$$\tilde{\Sigma}_{n,\alpha} \equiv \frac{\sigma_u^2}{\alpha n}\left(\operatorname{diag}\left(\frac{1}{n}\sum_{i=1}^{n} W_i W_i^{\top} + \frac{1}{\alpha n}\Sigma_\pi\right)\right)^{-1}. \tag{31}$$

In this example it is possible to show that the Bernstein-von Mises theorem holds for the variational approximation to the α-posterior (not only in total variation distance, but also in KL divergence). The assumptions of Theorem 2 are verified in Appendix B.

5.4 Expected KL and Optimal α

Finally, we discuss the expected KL criterion and the choice of α in the context of our example. In Section 4, we defined $r_n(\alpha)$ as the expected KL divergence between the true posterior of θ and the α-posterior. As we therein explained, this expected KL measure is typically difficult to compute because—with the exception of some stylized examples such as our linear regression model with omitted variables—the α-posteriors and their variational approximations are not available in closed-form.

This motivated us to work with a surrogate measure $r_n^*(\alpha)$, which motivated by the Bernstein-von Mises theorem, replaces the true posterior and the α-posterior by their asymp-

[17] Although, as this example and the proof of Theorem 1 suggests, the result is likely to go through for any sequence α_n that converges to a positive constant

[18] This can be shown using the typical squared-Hellinger distance lower bound for the total variation distance. See Appendix B.3 for details.

totic approximations. In our linear regression example it is possible to formalize the relation between $r_n(\alpha)$ and $r_n^*(\alpha)$. Algebra shows that for any fixed α and any Gaussian prior

$$r_n(\alpha) - r_n^*(\alpha) \to 0 \quad \text{as} \quad n \to \infty,$$

as one would have expected. One of the challenges in generalizing this result is that the Bernstein-von Mises theorems we have established are in total variation, which we cannot use directly to analyze the behavior of the expected KL divergence.

Regarding the choice of α, we have shown that the limit of the optimal choice is

$$\alpha^* = \frac{p}{p + \varepsilon(\theta_0 - \theta^*)^\top V_{\theta^*}(\theta_0 - \theta^*)},$$

which we derived in Theorem 3, under the assumption that the true parameter, θ_0, is different to the pseudo-true parameter θ^*. In our example, this happens whenever $\gamma_0 \neq 0$ (there are indeed omitted variables) and the omitted variables are correlated with the observed controls (i.e., when the omitted variable bias is different from zero).

Since in the linear regression example it is possible to compute explicitly $r_n(\alpha)$, it is also possible to choose α to minimize this expression. Algebra shows that for any $\alpha' \neq \alpha^*$

$$r_n(\alpha^*) < r_n(\alpha')$$

for sufficiently large n.

6 Concluding Remarks and Discussion

We studied the robustness to model misspecification of α-posteriors and their variational approximations with a focus on parametric, low-dimensional models. To formalize the notion of robustness we built on the seminal work of Gustafson (2001) and his suggested measure of sensitivity to parametric model misspecification. To state it simply, if two different procedures both lead to incorrect a posteriori inference (either due to model misspecification or computational considerations), one procedure is more robust (or less sensitive) than the other if it is closer—in terms of KL divergence—to the *true* posterior. Thus, we analyzed the KL divergence between true posteriors and the distributions reported by either the α-posterior approach or their variational approximations.

Obtaining general results about the properties of the KL divergence between true and reported posteriors is quite challenging, as this will typically depend on the priors, the data, the statistical model, and the form of misspecification. We were able to make progress by relying on asymptotic approximations to α-posteriors and their variational approximations.

In particular, we established a Bernstein-von Mises (BvM) theorem in total variation distance for α-posteriors (Theorem 1) and for their (Gaussian mean-field) variational approximations (Theorem 2). Our results provided a generalization of the results in Wang and Blei (2019a,b), who focus on the case in which $\alpha = 1$. We also extend the results of Li et al. (2019), who establish the BvM theorem for α-posteriors under a weaker norm (weak convergence), but under more primitive conditions.

We think these asymptotic approximations have value per se. For example, we learned that relative to the BvM theorem for the standard posterior or its variational approximation, the choice of α only re-scales the limiting variance. The new scaling acts as if the observed sample size were $\alpha \cdot n$ instead of n, but the location for the Gaussian approximation continues to be the Maximum Likelihood estimator. Since choosing $\alpha < 1$ inflates the α-posterior's variance relative to the usual posterior, then the tempering parameter might help correct the variance understatement of standard variational approximations to the posterior.

The main use of the asymptotic approximations in our paper, however, is to facilitate the computation of the suggested measure of robustness. This requires elementary calculations once we have multivariate Gaussians with parameters that depend on the data, the sample size, and the 'curvature' of the likelihood.

An important caveat of our results is that we focused on analyzing *the KL divergence between the limiting distributions, as opposed to the limit of KL divergence between reported and true posteriors*. Although in some simple models the two are equivalent (for example, a linear regression model with omitted variables and Gaussian priors), the general result requires further exploration. Unfortunately, we do not yet have a good solution. It is easy to show that the function $(P, Q) \mapsto \mathcal{K}(P||Q)$ is lower semi-continuous in total variation distance, so it might be possible to get a bound on the KL divergence we want to study with the KL divergence of the Gaussian limits. However, the interesting part is not the divergences themselves, but rather the α's that minimize them. We think that perhaps the use of the Theorem of the Maximum (Berge, 1963) and its generalizations could be useful for this analysis. It is possible that the continuity results can be strengthened in our case, since we are dealing with Gaussians in the limit,

but we do not yet have answers. Also, a formal analysis might require the derivation of *entropic* BvM theorems, where the distance is measured using KL divergence, as in Clarke (1999b).

Finally, even though our paper has a theoretical prescription for choosing the tempering parameter α, further research is needed to translate this into a practical recommendation. As discussed in detail, our calculations suggest that α_n^* tends to be smaller as both the probability of misspecification ϵ_n and the difference between the true and pseudo-true parameters increase. It might be possible to hypothesize some value for the probability of misspecification. However, the true parameter is not known (and cannot be estimated consistently under misspecification). We leave the question of how to optimally choose the tempering parameter for α-posterior and their approximations for future research.

References

Alquier, P. and J. Ridgway (2020): "Concentration of tempered posteriors and of their variational approximations," *Annals of Statistics*, 48, 1475–1497.

Berge, C. (1963): *Topological Spaces: including a treatment of multi-valued functions, vector spaces, and convexity*, Dover publications, 1st english ed.

Bhattacharya, A., D. Pati, and Y. Yang (2019): "Bayesian fractional posteriors," *Annals of Statistics*, 47, 39–66.

Burgess, C. P., I. Higgins, A. Pal, L. Matthey, N. Watters, G. Desjardins, and A. Lerchner (2018): "Understanding disentangling in β-VAE," *arXiv preprint arXiv:1804.03599*.

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

Clarke, B. S. (1999a): "Asymptotic normality of the posterior in relative entropy," *IEEE Transactions on Information Theory*, 45, 165–176.

——— (1999b): "Asymptotic normality of the posterior in relative entropy," *IEEE Transactions on Information Theory*, 45, 165–176.

DasGupta, A. (2008): *Asymptotic theory of statistics and probability*, Springer Science & Business Media.

Ghosal, S. and A. Van der Vaart (2017): *Fundamentals of nonparametric Bayesian inference*, Cambridge University Press.

Grünwald, P. (2012): "The safe Bayesian," in *International Conference on Algorithmic Learning Theory*, Springer, 169–183.

Grünwald, P. and T. Van Ommen (2017): "Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it," *Bayesian Analysis*, 12, 1069–1103.

Gustafson, P. (2001): "On measuring sensitivity to parametric model misspecification," *Journal of the Royal Statistical Society: Series B*, 63, 81–94.

Hansen, L. P. and T. J. Sargent (2001): "Robust control and model uncertainty," *American Economic Review*, 91, 60–66.

Higgins, I., L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner (2017): "β-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework." in *Proceedings of the 5th International Conference on Learning Representations*, vol. 2, 6.

Horn, R. A. and C. R. Johnson (2012): *Matrix Analysis*, Cambridge University Press.

Huber, P. J. (1967): "The behavior of maximum likelihood estimates under nonstandard conditions," in *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability*, Berkeley, CA: Univ. California Press, vol. 1, 221–233.

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

Li, Y., N. Wang, and J. Yu (2019): "Improved marginal likelihood estimation via power posteriors and importance sampling," *SMU Economics and Statistics Working Paper Series, Paper No. 16-2019.*

Maccheroni, F., M. Marinacci, and A. Rustichini (2006): "Ambiguity aversion, robustness, and the variational representation of preferences," *Econometrica*, 74, 1447–1498.

Miller, J. W. and D. B. Dunson (2019): "Robust Bayesian inference via coarsening," *Journal of the American Statistical Association*, 114, 1113–1125.

Müller, U. K. (2013): "Risk of Bayesian inference in misspecified models, and the sandwich covariance matrix," *Econometrica*, 81, 1805–1849.

Pouzo, D., Z. Psaradakis, and M. Sola (2016): "Maximum likelihood estimation in possibly misspecified dynamic models with time inhomogeneous Markov regimes," *Available at SSRN 2887771*.

Strzalecki, T. (2011): "Axiomatic foundations of multiplier preferences," *Econometrica*, 79, 47–73.

Van der Vaart, A. W. (2000): *Asymptotic statistics*, vol. 3, Cambridge university press.

Wang, Y. and D. Blei (2019a): "Variational Bayes under model misspecification," in *Advances in Neural Information Processing Systems*, 13357–13367.

Wang, Y. and D. M. Blei (2019b): "Frequentist Consistency of Variational Bayes," *Journal of the American Statistical Association*, 114, 1147–1161.

White, H. (1982): "Maximum likelihood estimation of misspecified models," *Econometrica*, 50, 1–25.

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

A Proofs of Main Results

A.1 Proof of Theorem 1

This proof follows Theorem 2.1 in Kleijn and Van der Vaart (2012), which shows that the posterior under misspecification is asymptotically normal, but it is adapted and simplified for the α-posterior framework. In what follows, we let $\Delta_{n,\theta^*} \equiv \sqrt{n}\,(\widehat{\theta}_{\text{ML-}\mathcal{F}_n} - \theta^*)$ as in Assumption 1.

By the assumption that θ^* is in the interior of $\Theta \subset \mathbb{R}^p$, there exists a sufficiently small $\delta > 0$ such that the open ball $B_{\theta^*}(\delta) \equiv \{\theta : ||\theta - \theta^*|| < \delta\}$ is a neighborhood of the (pseudo-)true parameter θ^* in Θ. In particular, we choose δ such that $B_{\theta^*}(\delta)$ belongs to the neighborhood of θ^* in which it is assumed π is continuous and positive. Note that for any compact set $K_0 \subset \mathbb{R}^p$ including the origin, we can find an integer $N_0 \equiv N_0(K_0, B_{\theta^*}(\delta))$ sufficiently large such that for any vector $h \in K_0$, we have that the perturbation of θ^* in the direction $h/\sqrt{n}$, meaning $\theta^* + h/\sqrt{n}$, belongs to $B_{\theta^*}(\delta)$ whenever $n \geq N_0$.

The goal is to show that the total variation distance between the α-posterior of θ, denoted $\pi_{n,\alpha}(\cdot \mid X^n)$, and a multivariate Normal distribution with mean $\widehat{\theta}_{\text{ML}-\mathcal{F}_n}$ and variance $V_{\theta^*}^{-1}/(n\alpha)$ goes to zero in $f_{0,n}$-probability. Because the total variation distance is invariant to re-centering and scaling of both measures being compared (Van der Vaart, 2000), it is more convenient to work with the α-posterior of the transformation $\sqrt{n}\,(\theta - \theta^*)$ and compare it to a suitably re-centered and scaled multivariate Normal distribution, i.e. one with mean Δ_{n,θ^*} and variance $V_{\theta^*}^{-1}/\alpha$.

For vectors $g, h \in K_0$, the following random variable will be used to bound the average total variation distance between the α-posterior and the alleged multivariate normal limit:

$$f_n(g,h) \equiv \left\{1 - \frac{\phi_n(h)}{\pi_{n,\alpha}^{LAN}(h \mid X^n)}\,\frac{\pi_{n,\alpha}^{LAN}(g \mid X^n)}{\phi_n(g)}\right\}^+, \tag{32}$$

where $\phi_n(h) \equiv n^{-1/2}\phi(h \mid \Delta_{n,\theta^*},\, V_{\theta^*}^{-1}/\alpha)$ and $\pi_{n,\alpha}^{LAN}(h \mid X^n) \equiv n^{-1/2}\pi_{n,\alpha}(\theta^* + h/\sqrt{n} \mid X^n)$, are scaled versions of the densities that we want to compare using the total variation distance, and $\{x\}^+ = \max\{0, x\}$ denotes the positive part of x. Define also $\pi_n(h) \equiv n^{-1/2}\,\pi(\theta^* + h/\sqrt{n})$ to be the density of the prior distribution of the transformation $\sqrt{n}\,(\theta - \theta^*)$. It follows that f_n is well-defined on $K_0 \times K_0$ for all $n > N_0$, as in this regime $\pi_{n,\alpha}^{LAN}(h \mid X^n)$ is guaranteed to be positive since $\theta^* + h/\sqrt{n}$ belongs to $B_{\theta^*}(\delta)$ as discussed above.

Let $\overline{B}_{\mathbf{0}}(r_n)$ denote a closed ball of radius r_n around $\mathbf{0}$. Since $d_{TV} \leq 1$ and the expectation is linear, we have that for any sequence r_n and for any $\eta > 0$:

$$
\begin{aligned}
&\mathbb{E}_{f_{0,n}}\left[d_{\mathrm{TV}}\left(\pi_{n,\alpha}^{LAN}(\cdot \mid X^n),\, \phi_n(\cdot)\right)\right] \\
&\quad \leq \mathbb{E}_{f_{0,n}}\left[d_{\mathrm{TV}}\left(\pi_{n,\alpha}^{LAN}(\cdot \mid X^n),\, \phi_n(\cdot)\right) 1\left\{\sup_{g,h \,\in\, \overline{B}_{\mathbf{0}}(r_n)} f_n(g,h) \leq \eta\right\}\right] \\
&\qquad + \mathbb{P}_{f_{0,n}}\left(\sup_{g,h \,\in\, \overline{B}_{\mathbf{0}}(r_n)} f_n(g,h) > \eta\right).
\end{aligned}
\tag{33}
$$

Lemma 1 in Appendix C implies

$$
\begin{aligned}
d_{\mathrm{TV}}\left(\pi_{n,\alpha}^{LAN}(\cdot \mid X^n),\, \phi_n\right) \leq &\sup_{g,h \,\in\, \overline{B}_{\mathbf{0}}(r_n)} f_n(g,h) \\
&+ \max\left\{\int_{||h||>r_n} \pi_{n,\alpha}^{LAN}(\cdot \mid X^n)\, dh,\, \int_{||h||>r_n} \phi_n(h)\, dh\right\},
\end{aligned}
\tag{34}
$$

and since $\max\{A, B\} \leq A + B$ when $A, B > 0$,

$$
\begin{aligned}
&\mathbb{E}_{f_{0,n}}\left[d_{\mathrm{TV}}\left(\pi_{n,\alpha}^{LAN}(\cdot \mid X^n),\, \phi_n(\cdot)\right) \mathbf{1}\left\{\sup_{g,h \,\in\, \overline{B}_{\mathbf{0}}(r_n)} f_n(g,h) \leq \eta\right\}\right] \\
&\leq \eta + \mathbb{E}_{f_{0,n}}\left[\int_{||h||>r_n} \pi_{n,\alpha}^{LAN}(h \mid X^n)\, dh\right] + \mathbb{E}_{f_{0,n}}\left[\int_{||h||>r_n} \phi_n(h)\, dh\right],
\end{aligned}
\tag{35}
$$

is an upper bound for the expectation on the right side of equation (33).[19]

[19]The bound in (35) uses that

$$
\mathbb{E}_{f_{0,n}}\left[\int_{||h||>r_n} \phi_n(h)\, dh\, \mathbf{1}\left\{\sup_{g,h \,\in\, \overline{B}_{\mathbf{0}}(r_n)} f_n(g,h) \leq \eta\right\}\right] \leq \mathbb{E}_{f_{0,n}}\left[\int_{||h||>r_n} \phi_n(h)\, dh\right],
$$

by the non-negativity of $\phi_n(\cdot)$ and a similar upper bound for the second term on the right side of (35).

Lemma 2 in Appendix C implies that for a given $\eta, \epsilon > 0$, there exists a sequence $r_n \to +\infty$ and $N(\eta, \epsilon)$ such that the second term on the right side of equation (33) is small for $n > N(\eta, \epsilon)$; that is, for all $n > N(\eta, \epsilon)$,

$$\mathbb{P}_{f_{0,n}}\left(\sup_{g,h \in \overline{B}_{\mathbf{0}}(r_n)} f_n(g,h) > \eta\right) \leq \epsilon\,, \tag{36}$$

which uses the stochastic LAN condition in Assumption 1. In addition—by the concentration assumption in equation (9) given in Theorem 1—there also exists an integer $N_1(\eta, \epsilon) > N(\eta, \epsilon)$ such that, for all $n > N_1(\eta, \epsilon)$,

$$\mathbb{E}_{f_{0,n}}\left[\int_{||h||>r_n} \pi^{LAN}_{n,\alpha}(h \mid X^n)\, dh\right] < \epsilon. \tag{37}$$

Also, by Lemma 5.2 in (Kleijn and Van der Vaart, 2012) which exploits properties of the multivariate normal distribution, we have that for all $n > N_1(\eta, \epsilon)$,

$$\mathbb{E}_{f_{0,n}}\left[\int_{||h||>r_n} \phi_n(h)\, dh\right] < \epsilon\,. \tag{38}$$

Now plugging (37) and (38) into (35), we find

$$\mathbb{E}_{f_{0,n}}\left[d_{\mathrm{TV}}\left(\pi^{LAN}_{n,\alpha}(\cdot \mid X^n),\, \phi_n(\cdot)\right) 1\left\{\sup_{g,h\,\in\,\overline{B}_{\mathbf{0}}(r_n)} f_n(g,h) \leq \eta\right\}\right] \leq \eta + 2\epsilon. \tag{39}$$

Finally we conclude from (33), using the bound in (39) and the bound in (36) for the second term of (33), that for all $n > N_1(\eta, \epsilon)$,

$$\mathbb{E}_{f_{0,n}}\left[d_{\mathrm{TV}}\left(\pi^{LAN}_{n,\alpha}(\cdot \mid X^n),\, \phi_n(\cdot)\right)\right] \leq \eta + 2\epsilon + \epsilon = \eta + 3\epsilon\,.$$

A standard application of Markov's inequality gives the desired result.

A.2 Proof of Theorem 2

Let $\tilde{\pi}_{n,\alpha}(\cdot \mid X^n) = q(\cdot \mid \tilde{\mu}_n, \tilde{\Sigma}_n)$ be the *Gaussian mean-field approximation to the α-posterior* defined in (14). We will prove that

$$\mathcal{K}(\tilde{\pi}_{n,\alpha}(\cdot \mid X^n) \,||\, \phi(\cdot \mid \widehat{\theta}_{\text{ML-}\mathcal{F}_n}, \operatorname{diag}(V_{\theta^*})^{-1}/(\alpha n))) \to 0 \tag{40}$$

in $f_{0,n}$-probability. The statement in (19) follows by Pinsker's inequality as it ensures that convergence in Kullback-Leibler divergence implies convergence in total variation distance.

Note that the KL divergence between two p-dimensional Gaussian distributions can be computed explicitly as

$$\begin{aligned}
&\mathcal{K}(\phi(\cdot \mid \mu_1, \Sigma_1) \,||\, \phi(\cdot \mid \mu_2, \Sigma_2)) \\
&= \frac{1}{2}\left(\log\left(\frac{|\Sigma_2|}{|\Sigma_1|}\right) + \operatorname{tr}\left(\Sigma_2^{-1}\Sigma_1\right) + (\mu_2 - \mu_1)^\top \Sigma_2^{-1}(\mu_2 - \mu_1) - p\right).
\end{aligned} \tag{41}$$

Applying the identity (41) we see that

$$\mathcal{K}(\tilde{\pi}_{n,\alpha}(\cdot \mid X^n) \,||\, \phi(\cdot \mid \widehat{\theta}_{\text{ML-}\mathcal{F}_n}, \operatorname{diag}(\alpha n V_{\theta^*})^{-1})) = T_1 + T_2$$

where

$$T_1 = \frac{1}{2}(\widehat{\theta}_{\text{ML-}\mathcal{F}_n} - \tilde{\mu}_n)^\top \alpha\, n \operatorname{diag}(V_{\theta^*})(\widehat{\theta}_{\text{ML-}\mathcal{F}_n} - \tilde{\mu}_n) \tag{42}$$

and

$$T_2 = \frac{1}{2}\operatorname{tr}(\alpha\, n \operatorname{diag}(V_{\theta^*})\tilde{\Sigma}_n) - \frac{p}{2} + \frac{1}{2}\log\left(\frac{|\tilde{\Sigma}_n|^{-1}}{|\alpha\, n \operatorname{diag}(V_{\theta^*})|}\right). \tag{43}$$

We claim that both $T_1 = o_{f_{0,n}(1)}$ and $T_2 = o_{f_{0,n}(1)}$. The proof of these claims below completes the proof of the theorem as they entail (40). The key step for establishing these results is the asymptotic representation result of Lemma 3. It shows that, under Assumptions 1 and 2, for any sequence (μ_n, Σ_n)—possibly dependent on the data—that is bounded in $f_{0,n}$-probability we have that:

$$\mathcal{K}(q(\cdot \mid \mu_n,\, \Sigma_n) \,||\, \pi_{n,\alpha}(\cdot \mid X^n)) = \mathcal{K}\left(q(\cdot \mid \mu_n,\, \Sigma_n) \,||\, \phi(\cdot \mid \widehat{\theta}_{\text{ML-}\mathcal{F}_n},\, V_{\theta^*}^{-1}/(\alpha n))\right) + o_{f_{0,n}}(1).$$

This means that the KL divergence between any normal density $q(\cdot|\mu_n, \Sigma_n)$ and the α-posterior $\pi_{n,\alpha}(\cdot|X^n)$ is eventually close to the KL divergence between such density and the α-posterior's total variation limit (which we have characterized in Theorem 1).

We use two intermediate steps to relate $(\tilde{\mu}_n, \tilde{\Sigma}_n)$ to $(\widehat{\theta}_{\text{ML}-\mathcal{F}_n}, \text{diag}(V_{\theta^*})^{-1}/(\alpha n))$.

Claim 1. We start by showing that $T_1 = o_{f_{0,n}}(1)$. Since $\tilde{\mu}_n$ and $\tilde{\Sigma}_n$ are defined as the parameters that solve the variational approximation problem in (6), then for every n

$$\mathcal{K}(\tilde{\pi}_{n,\alpha}(\cdot \mid X^n) \,||\, \pi_{n,\alpha}(\cdot \mid X^n)) \leq \mathcal{K}(q(\cdot \mid \widehat{\theta}_{\text{ML-}\mathcal{F}_n}, \tilde{\Sigma}_n) \,||\, \pi_{n,\alpha}(\cdot \mid X^n)).$$

This simply reflects the fact that $\tilde{\mu}_n$ and $\tilde{\Sigma}_n$ correspond to the parameters that minimize the KL divergence between the Gaussian mean-field family and the α-posterior.

Using Lemma 3, we can evaluate each of the KL divergences above up to an $o_{f_{0,n}}(1)$ term. In fact, we have that

$$\begin{aligned}
&\mathcal{K}\left(q(\cdot \mid \tilde{\mu}_n,\, \tilde{\Sigma}_n) \,||\, \phi(\cdot \mid \widehat{\theta}_{\text{ML-}\mathcal{F}_n},\, V_{\theta^*}^{-1}/(\alpha n))\right) \\
&\quad \leq \mathcal{K}\left(q(\cdot \mid \widehat{\theta}_{\text{ML-}\mathcal{F}_n},\, \tilde{\Sigma}_n) \,||\, \phi(\cdot \mid \widehat{\theta}_{\text{ML-}\mathcal{F}_n},\, V_{\theta^*}^{-1}/(\alpha n))\right) + o_{f_{0,n}}(1).
\end{aligned} \tag{44}$$

Therefore, it follows from (41) and (44) that

$$\begin{aligned}
&\mathcal{K}\left(q(\cdot \mid \tilde{\mu}_n,\, \tilde{\Sigma}_n) \,||\, \phi(\cdot \mid \widehat{\theta}_{\text{ML-}\mathcal{F}_n},\, (\alpha n V_{\theta^*})^{-1})\right) - \mathcal{K}\left(q(\cdot \mid \widehat{\theta}_{\text{ML-}\mathcal{F}_n},\, \tilde{\Sigma}_n) \,||\, \phi(\cdot \mid \widehat{\theta}_{\text{ML-}\mathcal{F}_n},\, (\alpha n V_{\theta^*})^{-1})\right) \\
&= \frac{1}{2}(\widehat{\theta}_{\text{ML-}\mathcal{F}_n} - \tilde{\mu}_n)^\top (\alpha n V_{\theta^*})(\widehat{\theta}_{\text{ML-}\mathcal{F}_n} - \tilde{\mu}_n) \\
&\leq o_{f_{0,n}}(1).
\end{aligned} \tag{45}$$

This shows that $\tilde{\mu}_n$ concentrates around $\widehat{\theta}_{\text{ML}-\mathcal{F}_n}$ From the last inequality and (42) we conclude that $T_1 = o_{f_{0,n}}(1)$.

Claim 2. We now show that $T_2 = o_{f_{0,n}}(1)$ by relating $\tilde{\Sigma}_n$ to $\text{diag}(V_{\theta^*})^{-1}/(\alpha n)$. The optimality of $\tilde{\mu}_n$ and $\tilde{\Sigma}_n$ once again implies that for every n:

$$\mathcal{K}(q(\cdot \mid \tilde{\mu}_n, \tilde{\Sigma}_n) \,||\, \pi_{n,\alpha}(\cdot \mid X^n)) \leq \mathcal{K}(q(\cdot \mid \widehat{\theta}_{\text{ML-}\mathcal{F}_n}, \text{diag}(V_{\theta^*})^{-1}/(\alpha n)) \,||\, \pi_{n,\alpha}(\cdot \mid X^n)). \tag{46}$$

This inequality and Lemma 3 imply that

$$\begin{aligned}
&\mathcal{K}(q(\cdot \mid \tilde{\mu}_n, \tilde{\Sigma}_n) \,||\, q(\cdot \mid \widehat{\theta}_{\text{ML-}\mathcal{F}_n}, (\alpha n V_{\theta^*})^{-1})) \\
&\leq \mathcal{K}(q(\cdot \mid \widehat{\theta}_{\text{ML-}\mathcal{F}_n}, \text{diag}(V_{\theta^*})^{-1}/(\alpha n)) \,||\, q(\cdot \mid \widehat{\theta}_{\text{ML-}\mathcal{F}_n}, (\alpha n V_{\theta^*})^{-1})) + o_{f_{0,n}}(1).
\end{aligned}$$

Using this last inequality, the formula (41) and equation (45) we obtain

$$\begin{aligned}
&\frac{1}{2}\left[\text{tr}(\alpha n V_{\theta^*}\tilde{\Sigma}_n) - p + \log\left(\frac{|V_{\theta^*}^{-1}/(\alpha n)|}{|\tilde{\Sigma}_n|}\right)\right] \\
&\leq \frac{1}{2}\left[\text{tr}(\alpha n V_{\theta^*}\text{diag}(V_{\theta^*})^{-1}/(\alpha n)) - p + \log\left(\frac{|V_{\theta^*}^{-1}/(\alpha n)|}{|\text{diag}(V_{\theta^*})^{-1}/(\alpha n)|}\right)\right] + o_{f_{0,n}}(1).
\end{aligned}$$

Further noting that $\text{tr}(V_{\theta^*}\text{diag}(V_{\theta^*})^{-1}) = p$, we note that the last inequality is equivalent to

$$\frac{1}{2}\left[\text{tr}(\alpha n V_{\theta^*}\tilde{\Sigma}_n) - p + \log\left(\frac{|\tilde{\Sigma}_n|^{-1}}{|\alpha\, n\, \text{diag}(V_{\theta^*})|}\right)\right] \leq o_{f_{0,n}}(1). \tag{47}$$

Since $\tilde{\Sigma}_n$ is diagonal

$$\text{tr}(\alpha n V_{\theta^*}\tilde{\Sigma}_n) = \text{tr}(\alpha\, n\, \text{diag}(V_{\theta^*})\tilde{\Sigma}_n).$$

Consequently, the left-hand side of (47) equals T_2, which has been defined in (43). The term T_2 is nonnegative, as it equals the KL divergence between two normals with the same mean, but variances $\tilde{\Sigma}_n$ and $\text{diag}(V_{\theta^*})^{-1}/\alpha n$. We conclude that $T_2 = o_{f_{0,n}}(1)$.

A.3 Proof of Theorem 3

We note that (41) allows us to compute $r_n^*(\alpha)$ and $\tilde{r}^*(\alpha)$ explicitly as

$$
\begin{aligned}
r_n^*(\alpha) =& \frac{1}{2}\Big(\alpha\, A_n(V_{\theta^*}) - p\log(\alpha) + B_n(V_\theta^*)\Big),\\
\tilde{r}^*(\alpha) =& \frac{1}{2}\Big(\alpha\, A_n(\tilde{V}_{\theta^*}) - p\log(\alpha) + B_n(\tilde{V}_\theta^*)\Big),
\end{aligned}
$$

where

$$
\begin{aligned}
A_n(\Sigma) \equiv& \epsilon_n \mathrm{tr}(\Sigma\Omega) + (1-\epsilon_n)\mathrm{tr}(\Sigma V_{\theta^*}^{-1}) + n\epsilon_n(\widehat{\theta}_{\text{ML-}\mathcal{F}_n} - \widehat{\theta}_{\text{ML}})^\top \Sigma (\widehat{\theta}_{\text{ML-}\mathcal{F}_n} - \widehat{\theta}_{\text{ML}})\\
B_n(\Sigma) \equiv& -p + \epsilon_n \log(|\Omega^{-1}||\Sigma|^{-1}) + (1-\epsilon_n)\log(|V_{\theta^*}||\Sigma|^{-1})
\end{aligned}
$$

Using this notation, we see that r_n^* and $\tilde{r}^*$ are convex functions on α. This implies that first order conditions pin-down the optimal α_n^* and $\tilde{\alpha}_n^*$ defined in (23). These are equal to

$$
\alpha_n^* = \frac{p}{A_n(V_{\theta^*})} \quad \text{and} \quad \tilde{\alpha}_n^* = \frac{p}{A_n(\tilde{V}_{\theta^*})} \tag{48}
$$

By assumption we have $\widehat{\theta}_{\text{ML}} \xrightarrow{p} \theta_0$ and $n\epsilon_n \to \varepsilon \in (0,\infty)$. In addition, we known that $\widehat{\theta}_{\text{ML-}\mathcal{F}_n} \xrightarrow{p} \theta^*$. It follows that

$$
A_n(\Sigma) \xrightarrow{p} \mathrm{tr}(\Sigma V_{\theta^*}^{-1}) + \varepsilon(\theta^* - \theta_0)^\top \Sigma (\theta^* - \theta_0).
$$

Denote α^* and $\tilde{\alpha}^*$ the limits in $f_{0,n}$-probability of α_n^* and $\tilde{\alpha}_n^*$. These limits can be computed replacing the expression above for $\Sigma = V_{\theta^*}, \tilde{V}_{\theta^*}$ and taking limit on (48), it implies

$$
\begin{aligned}
\alpha_n^* \;&\xrightarrow{p}\; \alpha^* \equiv \frac{p}{p + \varepsilon(\theta_0 - \theta^*)^\top V_{\theta^*}(\theta_0 - \theta^*)},\\
\tilde{\alpha}_n^* \;&\xrightarrow{p}\; \tilde{\alpha}^* \equiv \frac{p}{\mathrm{tr}\left(\tilde{V}_{\theta^*} V_{\theta^*}^{-1}\right) + \varepsilon(\theta_0 - \theta^*)^\top \tilde{V}_{\theta^*}(\theta_0 - \theta^*)}.
\end{aligned}
$$

To conclude $\tilde{\alpha}^* < 1$ is sufficient to prove that $\mathrm{tr}(\tilde{V}_{\theta^*} V_{\theta^*}^{-1}) \geq p$

One way of verifying this, remember that trace of a matrix is the sum of the eigenvalues and the determinant is the product. Applying the Arithmetic Mean-Geometric Mean inequality, we have

$$\frac{1}{p}\mathrm{tr}(\tilde{V}_{\theta^*} V_{\theta^*}^{-1}) \geq \left(|\tilde{V}_{\theta^*} V_{\theta^*}^{-1}| \right)^{1/p}.$$

Then since $|\tilde{V}_{\theta^*} V_{\theta^*}^{-1}| = |\tilde{V}_{\theta^*}||V_{\theta^*}^{-1}| = |\tilde{V}_{\theta^*}||V_{\theta^*}|^{-1}$, it will be sufficient to prove that $|\tilde{V}_{\theta^*}| \geq |V_{\theta^*}|$, where $\tilde{V}_{\theta^*} = \mathrm{diag}(V_{\theta^*})$. Because V_{θ^*} is a semi-definite matrix, this is exactly Hadamard's inequality (see Theorem 7.8.1 in Horn and Johnson (2012)).

B Verification of the assumptions for the illustrative example

B.1 Verification of α-posterior concentration

We first want to prove that the α-posterior concentrates at the rate $\sqrt{n}$ around θ^* as defined in (9). In other words need to show that for every sequence $r_n \to \infty$, we have

$$\mathbb{E}_{f_{0,n}} \left[\mathbb{P}_{\pi_{n,\alpha}(\cdot \,|\, X^n)} \left(\|\sqrt{n}\,(\theta - \theta^*)\| > r_n \right) \right] \to 0. \tag{49}$$

Step 1: Compute an upper bound for the probability inside the brackets using Markov's inequality.

Note that in our illustrative example, we have $\pi_{n,\alpha}(\theta \,|\, X^n) \sim \mathcal{N}(\mu_{n,\alpha},\, \Sigma_{n,\alpha})$, where $\mu_{n,\alpha}$ and $\Sigma_{n,\alpha}$ were defined in (28) and (29). Therefore, by Markov's inequality and the normal distribution of the α-posterior we have

$$\begin{aligned}\mathbb{P}_{\pi_{n,\alpha}(\cdot\,|\,X^n)}\left(\|\theta-\theta^*\|^2 > r_n^2/n\right) &\geq \frac{n}{r_n^2}\mathbb{E}_{\pi_{n,\alpha}(\cdot\,|\,X^n)}\left[\|\theta-\theta^*\|^2\right] \\ &= \frac{n}{r_n^2}\left[\|\mu_{n,\alpha}-\theta^*\|^2+\operatorname{tr}(\Sigma_{n,\alpha})\right], \end{aligned} \tag{50}$$

which defines the upper bound for the probability inside the brackets.

Step 2: Conclude that the expected value of the probability goes to zero. Lemma 4 in Appendix C implies that the sequence $(\sqrt{n}(\mu_{n,\alpha}-\theta^*),\, n\Sigma_{n,\alpha})$ is bounded in $f_{0,n}$-probability. It follows that both $\|\sqrt{n}(\mu_{n,\alpha}-\theta^*)\|^2$ and $\operatorname{tr}(n\Sigma_{n,\alpha})$ are bounded in $f_{0,n}$-probability. This means that for every $\epsilon>0$, there exists an $M_\epsilon>0$ such that

$$\mathbb{P}_{\pi_{n,\alpha}(\cdot\,|\,X^n)}\left(A_\epsilon\right) \leq \epsilon, \tag{51}$$

where $A_\epsilon = \{\|\sqrt{n}(\mu_{n,\alpha}-\theta^*)\|^2+\operatorname{tr}(n\Sigma_{n,\alpha}) > M_\epsilon\}$. By the law of total probabiliy and linearity of the expectation, we have that for any sequence r_n and for any $\epsilon>0$:

$$\begin{aligned}&\mathbb{E}_{f_{0,n}}[\mathbb{P}_{\pi_{n,\alpha}(\cdot\,|\,X^n)}\left(\|\sqrt{n}\,(\theta-\theta^*)\|>r_n\right)] \\ &\qquad\leq \mathbb{E}_{f_{0,n}}\left[\mathbb{P}_{\pi_{n,\alpha}(\cdot\,|\,X^n)}\left(\|\sqrt{n}\,(\theta-\theta^*)\|>r_n\right)1\{A_\epsilon^c\}\right]+\mathbb{E}_{f_{0,n}}\left[1\{A_\epsilon\}\right], \end{aligned} \tag{52}$$

where $1\{A_\epsilon\}$ is the indicator function of the event A_ϵ. The first term in (52) can be bounded using (50) and the definition of A_ϵ, leading to

$$\begin{aligned}\mathbb{E}_{f_{0,n}}\left[\mathbb{P}_{\pi_{n,\alpha}(\cdot\,|\,X^n)}\left(\|\sqrt{n}\,(\theta-\theta^*)\|>r_n\right)1\{A_\epsilon^c\}\right] &\leq \mathbb{E}_{f_{0,n}}\left[\frac{1}{r_n^2}\left[\|\sqrt{n}(\mu_{n,\alpha}-\theta^*)\|^2+\operatorname{tr}(n\Sigma_{n,\alpha})\right]1\{A_\epsilon^c\}\right] \\ &\leq \frac{M_\epsilon}{r_n^2}\end{aligned}$$

Using (51), we see that the second term in (52) is smaller than ϵ. Hence, we conclude that

$$\mathbb{E}_{f_{0,n}}[\mathbb{P}_{\pi_{n,\alpha}(\cdot\,|\,X^n)}\left(\|\sqrt{n}\,(\theta-\theta^*)\| > r_n\right)] \leq \frac{M_\epsilon}{r_n^2} + \epsilon,$$

which is sufficiently small since $\epsilon > 0$ was arbitrary and $r_n \to \infty$. This verifies (49).

B.2 KL distance for Theorem 1

In our illustrative example, we can compute the KL distance between the α-posterior distribution $\pi_{n,\alpha}$ and the distribution defined in (30). This is possible since both distribution are multivariate normal. Using (41), we obtain

$$\frac{1}{2}\left(-p + \log\left(\frac{|V_{\theta^*}\alpha n|^{-1}}{|\Sigma_{n,\alpha}|}\right) + \mathrm{tr}(V_{\theta^*}\alpha n\Sigma_{n,\alpha}) + (\mu_{n,\alpha} - \widehat{\theta}_{ML})^\top V_{\theta^*}\alpha n(\mu_{n,\alpha} - \widehat{\theta}_{ML})\right).$$

Lemma 4 implies that the previous expression converge to 0 in $f_{0,n}$-probability. This means that the KL distance between $\pi_{n,\alpha}$ and the distribution in (30) goes to zero in $f_{0,n}$-probability.

B.3 What if α goes to zero very quickly?

Suppose that the sequence of α_n verifies that $n \cdot \alpha_n \to \alpha_0 > 0$. This implies that

$$\begin{aligned}
\mu_{n,\alpha_n} \to& \left(\mathbb{E}[W_i W_i^\top] + \frac{1}{\alpha_0}\Sigma_\pi\right)^{-1}\left(\frac{1}{\alpha_0}\Sigma_\pi\mu_\pi + \mathbb{E}[W_i W_i^\top]\theta^*\right),\\
\Sigma_{n,\alpha_n} \to& \frac{\sigma_u^2}{\alpha_0}\left(\mathbb{E}[W_i W_i^\top] + \frac{1}{\alpha_0}\Sigma_\pi\right)^{-1},
\end{aligned}$$

where the convergence is in $f_{0,n}$-probability. Using these different limits, we show that the total variation distance between the α_n-posterior distribution π_{n,α_n} and the distribution in (30) is bounded away from zero. We show this using that the square of the Hellinger distance is a lower bound for the total variation distance.

In our illustrative example, the α_n-posterior distribution π_{n,α_n} and the distribution in (30) are both multivariate normal distributions. Then, we can compute the square of the

Hellinger distance between these distributions. Using Lemma B.1 part ii) of Ghosal and Van der Vaart (2017), we obtain

$$1-\frac{|\Sigma_{n,\alpha_n}|^{1/4}|(V_{\theta^*}\alpha_n n)^{-1}|^{1/4}}{|(\Sigma_{n,\alpha_n}+(V_{\theta^*}\alpha_n n)^{-1})/2|^{1/2}}\times$$
$$\exp\left(-\frac{1}{8}(\mu_{n,\alpha_n}-\widehat{\theta}_{ML})^\top\frac{\Sigma_{n,\alpha_n}+(V_{\theta^*}\alpha_n n)^{-1}}{2}(\mu_{n,\alpha_n}-\widehat{\theta}_{ML})\right)$$

which converge in $f_{0,n}$-probability to a positive number. To verify this, notice that Σ_{n,α_n} and $(V_{\theta^*}\alpha_n n)^{-1}$ converge to different limits, which guarantees

$$\lim_{n\to\infty}\frac{|\Sigma_{n,\alpha_n}|^{1/4}|(V_{\theta^*}\alpha_n n)^{-1}|^{1/4}}{|(\Sigma_{n,\alpha_n}+(V_{\theta^*}\alpha_n n)^{-1})/2|^{1/2}}<1,$$

where the limit is taken in $f_{0,n}$-probability and the inequality follows by applying the Arithmetic Mean-Geometric Mean inequality.

B.4 Verification of Assumption 2

In our illustrative example:

$$\pi(\theta)\sim\mathcal{N}(\mu_\pi,\sigma_u^2\Sigma_\pi^{-1}),$$
$$R_n(h)=h^\top Q_n\Delta_{n,\theta^*}-\frac{1}{2}h^\top Q_n h,$$

where $\Delta_{n,\theta^*}=\sqrt{n}(\widehat{\theta}_{ML}-\theta^*)$ and

$$Q_n\equiv\frac{\sum_{i=1}^n W_iW_i^\top}{n\sigma_u^2}-V_{\theta^*}.$$

Q_n converges to zero in $f_{0,n}$-probability in our illustrative example since $V_{\theta^*}=\mathbb{E}[W_iW_i^\top]/\sigma_u^2$.

Let us take a sequence (μ_n,Σ_n) such that $(\sqrt{n}(\mu_n-\theta^*),n\Sigma_n)$ is bounded in $f_{0,n}$-probability.

Then, equation (17) in Assumption 2 becomes

$$\begin{aligned}&\int \phi(h \mid \sqrt{n}(\mu_n - \theta^*), n\Sigma_n) \left(-\frac{1}{2} \frac{h^\top}{\sqrt{n}} \frac{\Sigma_\pi}{\sigma_u^2} \frac{h}{\sqrt{n}} + \frac{h^\top}{\sqrt{n}} \frac{\Sigma_\pi}{\sigma_u^2} (\mu_\pi - \theta^*) \right) dh \\ &= -\frac{1}{2n} \overline{\mu}_n^\top \frac{\Sigma_\pi}{\sigma_u^2} \overline{\mu}_n - \frac{1}{2n} \mathrm{tr} \left(n\Sigma_n \frac{\Sigma_\pi}{\sigma_u^2} \right) + \frac{1}{\sqrt{n}} \overline{\mu}_n^\top \frac{\Sigma_\pi}{\sigma_u^2} (\mu_\pi - \theta^*),\end{aligned} \tag{53}$$

where $\overline{\mu}_n = \sqrt{n}(\mu_n - \theta^*)$. By assumption, the sequence $(\overline{\mu}_n, n\Sigma_n)$ is bounded in $f_{0,n}$-probability. This implies that (53) goes to zero in $f_{0,n}$-probability.

Equation (18) in Assumption 2 can be computed explicitly as

$$\int \phi(h|\sqrt{n}(\mu_n - \theta^*), n\Sigma_n) \left(h^\top Q_n \Delta_{n,\theta^*} - \frac{1}{2} h^\top Q_n h \right) dh = \overline{\mu}_n^\top Q_n \Delta_{n,\theta^*} - \frac{1}{2} \mu_n^\top Q_n \overline{\mu}_n - \frac{1}{2} \mathrm{tr}(Q_n n\Sigma_n).$$

By assumption, the sequence $(\overline{\mu}_n, n\Sigma_n)$ is bounded in $f_{0,n}$-probability. Since Q_n converge to zero in $f_{0,n}$-probability, the expression above goes to zero in $f_{0,n}$-probability.

B.5 KL distance for Theorem 2

Lemma 4 shows that in our illustrative example, the sequence $(\mu_{n,\alpha}, \Sigma_{n,\alpha})$ verifies that $(\sqrt{n}(\mu_{n,\alpha} - \theta^*),\ n\Sigma_{n,\alpha})$ is bounded in $f_{0,n}$-probability. Then, we can apply Theorem 2. The proof presented in Section A.2 shows that

$$\mathcal{K}(\tilde{\pi}_{n,\alpha}(\cdot \mid X^n) \mid\mid \phi(\cdot \mid \widehat{\theta}_{\text{ML-}\mathcal{F}_n},\ \mathrm{diag}(V_{\theta^*})^{-1}/(\alpha n))) \to 0,$$

in $f_{0,n}$-probability. This means that Bernstein-von Mises Theorem holds for the variational approximation to the α-posterior in KL divergence.

B.6 About the optimal α for robustness

Let us recall the definition of $r_n(\alpha)$ presented in Section 4 in equation (21):

$$r_n(\alpha) = \epsilon_n \mathcal{K}\left(\phi(\cdot \mid \nu_n^*,\, \Omega_n^*) \,||\, \phi(\cdot \mid \mu_{n,\alpha},\, \Sigma_{n,\alpha})\right) \\ + (1-\epsilon_n)\mathcal{K}\left(\phi(\cdot \mid \mu_{n,1},\, \Sigma_{n,1}) \,||\, \phi(\cdot \mid \mu_{n,\alpha},\, \Sigma_{n,\alpha})\right),$$

where ν_n^* is the true posterior mean and Ω_n^* is the true posterior covariance matrix. The expression above is equal to

$$\begin{aligned} r_n(\alpha) =& \frac{1}{2}\Big\{\epsilon_n \mathrm{tr}(\Sigma_{n,\alpha}^{-1}\Omega_n^*) + \alpha(1-\epsilon_n)\mathrm{tr}((\alpha n \Sigma_{n,\alpha})^{-1} n \Sigma_{n,1}) \\ &+ \alpha n \epsilon_n (\mu_{n,\alpha} - \nu_n^*)^\top (\alpha n \Sigma_{n,\alpha})^{-1}(\mu_{n,\alpha} - \nu_n^*) \\ &+ \alpha n (1-\epsilon_n)(\mu_{n,1} - \mu_{n,\alpha})^\top (\alpha n \Sigma_{n,\alpha})^{-1}(\mu_{n,1} - \mu_{n,\alpha}) \\ &- p\log(\alpha) - p + \epsilon_n \log\left(\frac{|\Sigma_{n,\alpha}|}{|\Omega_n^*|}\right) + (1-\epsilon_n)\log\left(\frac{|\Sigma_{n,\alpha}|}{|\Sigma_{n,1}|}\right)\Big\}. \end{aligned}$$

Notice that $\Sigma_{n,\alpha}^{-1}\Omega_n^* \to \alpha V_{\theta^*}\Omega$ in $f_{0,n}$-probability, since it can be proved that $n\Omega_n^* \to \Omega$ in the well-specified model, for some definite positive matrix Ω. Lemma 4 implies that $(\alpha n \Sigma_{n,\alpha})^{-1} n \Sigma_{n,1} \to \mathbb{I}_p$ in $f_{0,n}$-probability. Moreover, we have that $n(\mu_{n,1} - \mu_{n,\alpha})$ is bounded in $f_{0,n}$-probability, which implies that

$$(\mu_{n,1} - \mu_{n,\alpha})^\top \Sigma_{n,\alpha}(\mu_{n,1} - \mu_{n,\alpha}) \to 0$$

in $f_{0,n}$-probability. Since $n\epsilon_n \to \varepsilon$, we conclude that

$$r_n(\alpha) \to r_\infty(\alpha) \equiv \frac{1}{2}\left(\alpha p + \alpha\varepsilon(\theta^* - \theta_0)^\top V_{\theta^*}(\theta^* - \theta_0) - p\log(\alpha) - p\right)$$

in $f_{0,n}$-probability.

Using the notation introduced in the proof of Theorem 3 in Section A.3, we have

$$r_n^*(\alpha) = \frac{1}{2}\Big(\alpha\, A_n(V_{\theta^*}) - p\log(\alpha) + B_n(V_\theta^*)\Big)$$

where

$$\begin{aligned}
A_n(\Sigma) \equiv& \epsilon_n \mathrm{tr}(\Sigma\Omega) + (1-\epsilon_n)\mathrm{tr}(\Sigma V_{\theta^*}^{-1}) + n\epsilon_n(\widehat{\theta}_{\text{ML-}\mathcal{F}_n} - \widehat{\theta}_{\text{ML}})^\top \Sigma(\widehat{\theta}_{\text{ML-}\mathcal{F}_n} - \widehat{\theta}_{\text{ML}}),\\
B_n(\Sigma) \equiv& -p + \epsilon_n \log(|\Omega^{-1}||\Sigma|^{-1}) + (1-\epsilon_n)\log(|V_{\theta^*}||\Sigma|^{-1}).
\end{aligned}$$

Notice that $A_n(V_{\theta^*}) \to p + \varepsilon(\theta^* - \theta_0)^\top V_{\theta^*}(\theta^* - \theta_0)$ and $B_n(V_{\theta^*}) \to -p$ in $f_{0,n}$-probability. This implies that

$$r_n^*(\alpha) \to r_\infty(\alpha) = \frac{1}{2}\left(\alpha p + \alpha\varepsilon(\theta^* - \theta_0)^\top V_{\theta^*}(\theta^* - \theta_0) - p\log(\alpha) - p\right).$$

Then, we can conclude that

$$r_n(\alpha) - r_n^*(\alpha) \to 0$$

in $f_{0,n}$-probability. In particular, for $\alpha = \alpha^*$ defined in Theorem 3 and any $\alpha' \neq \alpha^*$, we have that $r_n(\alpha^*)$ is close to $r_\infty(\alpha^*)$ and $r_n(\alpha')$ is close to $r_\infty(\alpha')$. Since $r_\infty(\alpha^*) < r_\infty(\alpha')$ by definition of α^*, it follows that for large n, $r_n(\alpha^*) < r_n(\alpha')$.

C Technical Lemmas

Lemma 1 *Consider sequences of densities ϕ_n and ψ_n. For a given compact set $K \subset \mathbb{R}^p$, suppose that the densities ψ_n and ϕ_n are positive on K. Then,*

$$d_{TV}(\psi_n, \phi_n) \leq \sup_{g,h \in K} f_n(g,h) + \max\left\{\int_{\mathbb{R}^p \setminus K} \psi_n(h)\, dh, \int_{\mathbb{R}^p \setminus K} \phi_n(h)\, dh\right\},$$

where we have defined the random variable

$$f_n(g,h) = \left\{1 - \frac{\phi_n(h)}{\psi_n(h)} \frac{\psi_n(g)}{\phi_n(g)}\right\}^+. \tag{54}$$

Proof. First, denote $a_n = \left\{\int_K \psi_n(g)\, dg\right\}^{-1}$ and $b_n = \left\{\int_K \phi_n(g)\, dg\right\}^{-1}$. Notice that both are well defined since ϕ and ψ are assumed positive on K. We will assume throughout that $a_n \geq b_n$, without loss of generality.

First, by definition, $d_{\mathrm{TV}}(\psi_n, \phi_n) = \frac{1}{2}\int |\psi_n(h) - \phi_n(h)|\, dh$. Then since $|x| = 2\{x\}^+ - x$, it follows that the total variation is equal to

$$\begin{aligned}
d_{\mathrm{TV}}(\psi_n, \phi_n) &= \frac{1}{2}\int |\psi_n(h) - \phi_n(h)|\, dh \\
&= \int_{\mathbb{R}^p} \left\{\psi_n(h) - \phi_n(h)\right\}^+ dh + \frac{1}{2}\int_{\mathbb{R}^p} (\psi_n(h) - \phi_n(h))\, dh \\
&= \int_{\mathbb{R}^p} \left\{\psi_n(h) - \phi_n(h)\right\}^+ dh\,.
\end{aligned}$$

Next notice that

$$\begin{aligned}
&\int_{\mathbb{R}^p} \left\{\psi_n(h) - \phi_n(h)\right\}^+ dh \\
&\quad = \int_K \left\{\psi_n(h) - \phi_n(h)\right\}^+ dh + \int_{\mathbb{R}^p \setminus K} \left\{\psi_n(h) - \phi_n(h)\right\}^+ dh.
\end{aligned} \tag{55}$$

Now, since ψ_n, ϕ_n are non-negative on all of $\mathbb{R}^p$, it follows that $\{\psi_n(h) - \phi_n(h)\}^+ \leq \psi_n(h)$.

This provides a bound for the second term on the right side of (55):

$$\begin{aligned}\int_{\mathbb{R}^p\setminus K}\Big\{\psi_n(h)-\phi_n(h)\Big\}^+ dh &\le \int_{\mathbb{R}^p\setminus K}\psi_n(h)\,dh\\ &= \max\left\{\int_{\mathbb{R}^p\setminus K}\psi_n(h)\,dh, \int_{\mathbb{R}^p\setminus K}\phi_n(h)\,dh\right\}.\end{aligned} \tag{56}$$

In the above, the final step uses the fact that $a_n \ge b_n$ which implies $\int_K \psi_n(g)\,dg = a_n^{-1} \le b_n^{-1} = \int_K \phi_n(g)\,dg$ and therefore,

$$\int_{\mathbb{R}^p\setminus K}\psi_n(h)\,dh \ge \int_{\mathbb{R}^p\setminus K}\phi_n(h)\,dh\,.$$

Finally, for all $h \in K$, we have the following identity,

$$\frac{\phi_n(h)}{\psi_n(h)} = \frac{a_n}{b_n}\int_K \frac{\phi_n(h)}{\psi_n(h)}\,\frac{\psi_n(g)}{\phi_n(g)}\,b_n\phi_n(g)\,dg\,.$$

Thus, we can rewrite the first term on the right side of (55) as follows,

$$\begin{aligned}&\int_K\Big\{\psi_n(h)-\phi_n(h)\Big\}^+ dh = \int_K\left\{1-\frac{\phi_n(h)}{\psi_n(h)}\right\}^+ \psi_n(h)\,dh\\ &= \int_K\left\{1-\frac{a_n}{b_n}\int_K\frac{\phi_n(h)}{\psi_n(h)}\,\frac{\psi_n(g)}{\phi_n(g)}\,b_n\phi_n(g)\,dg\right\}^+\psi_n(h)\,dh\,.\end{aligned} \tag{57}$$

Now, by applying Jensen's inequality on the convex function $f(x)=\{1-x\}^+$, we have that $\{1-\mathbb{E}[X]\}^+ \le \mathbb{E}\left[\{1-X\}^+\right]$. Applying this to the final expression above, and recalling the definition of $f_n(g,h)$ in (54), we find

$$\begin{aligned}&\int_K\left(\left\{1-\frac{a_n}{b_n}\int_K\frac{\phi_n(h)}{\psi_n(h)}\,\frac{\psi_n(g)}{\phi_n(g)}\,b_n\phi_n(g)\,dg\right\}^+\right)\psi_n(h)\,dh\\ &\le \int_K\left(\int_K\left\{1-\frac{a_n}{b_n}\,\frac{\phi_n(h)}{\psi_n(h)}\,\frac{\psi_n(g)}{\phi_n(g)}\right\}^+ b_n\phi_n(g)\,dg\right)\psi_n(h)\,dh\\ &\le \int_K\int_K f_n(g,h)\,b_n\phi_n(g)\,\psi_n(h)\,dg\,dh\,,\end{aligned} \tag{58}$$

where the final step uses that when $a_n/b_n \ge 1$, we have $(1-(a_n/b_n)x)^+ \le (1-x)^+$ for $x \ge 0$.

We finally note that,

$$\begin{aligned}
&\int_K \int_K f_n(g,h)\, b_n\phi_n(g)\, \psi_n(h)\, dg\, dh \\
&\le \sup_{g,h\,\in K} f_n(g,h) \int_K \int_K b_n\phi_n(g)\, \psi_n(h)\, dg\, dh \\
&= \frac{1}{a_n} \sup_{g,h\,\in K} f_n(g,h) \le \sup_{g,h\,\in K} f_n(g,h)\,,
\end{aligned}$$

The final equality uses that $a_n^{-1} = \int_K \psi_n(g)\, dg \le 1$.

Lemma 2 *Assume that there is a $\delta > 0$ such that the prior density π is continuous and positive on $B_{\theta^*}(\delta)$, the closed ball of radius δ around θ^*. For any $\eta, \epsilon > 0$, there exists a sequence $r_n \to +\infty$ and an integer $N(\eta,\epsilon) > 0$, such that for all $n > N(\eta,\epsilon)$ and $f_n(g,h)$ defined in* (32)*,*

$$\mathbb{P}_{f_{0,n}}\Big(\sup_{g,h\,\in \overline{B}_{\mathbf{0}}(r_n)} f_n(g,h) > \eta \Big) \le \epsilon\,,$$

where $\overline{B}_{\mathbf{0}}(r_n)$ denotes a closed ball of radius r_n around $\mathbf{0}$.

Proof: The proof has two steps. In Step 1, we prove the claim for any fixed $r > 0$, instead of a sequence r_n. In Step 2, we construct a sequence of r_n using equation (63).

Step 1: First notice that for any $r > 0$, there exists an integer $N_0(r) > 0$ such that $\theta^* + h/\sqrt{n} \in B_{\theta^*}(\delta)$ whenever $h \in \overline{B}_{\mathbf{0}}(r)$ and $n \ge N_0(r)$.

Recalling the definition of the α-posterior $\pi_{n,\alpha}$ in (5), as well as the scaled densities $\pi_{n,\alpha}^{LAN}(g_n\,|X^n) = n^{-1/2}\pi_{n,\alpha}(\theta^* + h/\sqrt{n}\,|X^n)$ and $\pi_n(h) = n^{-1/2}\pi(\theta^* + h/\sqrt{n})$ introduced in the proof of Theorem 1, we see that for any two sequences $(h_n), (g_n)$ in $\overline{B}_{\mathbf{0}}(r)$ such that

$n > N_0(r)$,

$$\frac{\pi^{LAN}_{n,\alpha}(g_n\,|X^n)}{\pi^{LAN}_{n,\alpha}(h_n\mid X^n)}=\frac{\pi_{n,\alpha}\left(\theta^*+\frac{g_n}{\sqrt{n}}\,\middle|\,X^n\right)}{\pi_{n,\alpha}\left(\theta^*+\frac{h_n}{\sqrt{n}}\,\middle|\,X^n\right)}=\frac{\left[f_n\left(X^n\,\middle|\,\theta^*+\frac{g_n}{\sqrt{n}}\right)\right]^\alpha\pi\left(\theta^*+\frac{g_n}{\sqrt{n}}\right)}{\left[f_n\left(X^n\,\middle|\,\theta^*+\frac{h_n}{\sqrt{n}}\right)\right]^\alpha\pi\left(\theta^*+\frac{h_n}{\sqrt{n}}\right)}$$
$$=\frac{\left[f_n\left(X^n\,\middle|\,\theta^*+\frac{g_n}{\sqrt{n}}\right)\right]^\alpha\pi_n\left(g_n\right)}{\left[f_n\left(X^n\,\middle|\,\theta^*+\frac{h_n}{\sqrt{n}}\right)\right]^\alpha\pi_n\left(h_n\right)}.$$

Thus, with the notation $s_n(h_n)=[f_n(X^n|\theta^*+h_n/\sqrt{n})/f_n(X^n|\theta^*)]^\alpha$ and $\phi_n(h_n)=\phi(h_n|\Delta_{n,\theta^*},V_{\theta^*}^{-1}/\alpha)$, we have

$$f_n(g_n,h_n)=\Big\{1-\frac{\phi_n(h_n)}{\pi^{LAN}_{n,\alpha}(h_n\mid X^n)}\,\frac{\pi^{LAN}_{n,\alpha}(g_n\mid X^n)}{\phi_n(g_n)}\Big\}^+=\Big\{1-\frac{\phi_n(h_n)s_n(g_n)\pi_n(g_n)}{\phi_n(g_n)s_n(h_n)\pi_n(h_n)}\Big\}^+.$$

Next, recall that $X_n=o_{f_{0,n}}(1)$ if $\lim_n P_{f_{0,n}}(|X_n|>\epsilon)=0$ for every $\epsilon>0$. Since the sequence $h_n\in\overline{B}_{\mathbf{0}}(r)$, Assumption 1 implies

$$\log(s_n(h_n))=\alpha\log\left(\frac{f_n(X^n\mid\theta^*+\frac{h_n}{\sqrt{n}})}{f_n(X^n\mid\theta^*)}\right)=h_n^\top\alpha V_{\theta^*}\Delta_{n,\theta^*}-\frac{1}{2}h_n^\top\alpha V_{\theta^*}h_n+o_{f_{0,n}}(1),\tag{59}$$

and algebra shows that the log-likelihood of the normal density ϕ_n can be written as

$$\log\phi_n(h_n)=-\frac{p}{2}\log(2\pi)+\frac{1}{2}\log(\det(\alpha V_{\theta^*}))-\frac{1}{2}(h_n-\Delta_{n,\theta^*})^\top\alpha V_{\theta^*}(h_n-\Delta_{n,\theta^*}).$$

Using these expressions for the sequence $g_n\in\overline{B}_{\mathbf{0}}(r)$, and letting

$$b_n(g_n,h_n)\equiv\frac{\phi_n(h_n)s_n(g_n)\pi_n(g_n)}{\phi_n(g_n)s_n(h_n)\pi_n(h_n)}\,,\tag{60}$$

we conclude

$$\log(b_n(g_n,h_n))=\log\Big(\frac{\phi_n(h_n)s_n(g_n)\pi_n(g_n)}{\phi_n(g_n)s_n(h_n)\pi_n(h_n)}\Big)=o_{f_{0,n}}(1)\,,\tag{61}$$

where we have used that $\pi_n(g_n),\pi_n(h_n)\to\pi(\theta^*)$ as $n\to\infty$. Since h_n,g_n are any sequences, the result in (61) is equivalent to saying that for any fixed r, there exists an integer $\tilde{N}_0(r,\epsilon,\eta) \geq N_0(r)$ such that for any $n > \tilde{N}_0(r,\epsilon,\eta)$:

$$P_{f_{0,n}}\left(\sup_{g_n,h_n\,\in\,\overline{B}_{\mathbf{0}}(r)} |\log(b_n(g_n,h_n))| > \eta\right) \leq \epsilon\,. \tag{62}$$

Next, notice that

$$\begin{aligned}
|\log(b_n(g_n,h_n))| &\geq |\log(\min\{1, b_n(g_n,h_n)\})| \\
&= |\log(1 - f_n(g_n,h_n))| \geq f_n(g_n,h_n)
\end{aligned}$$

where the equality follows since $f_n(g,h) = 1 - \min\{b_n(g,h),1\}$ by definition, and the final inequality follows by noting that the function $f(x) = |\log(1-x)| - x$ is increasing for $x \in [0,1]$ and $f(0) = 0$. Thus, by (62), denoting $\tilde{N}(r,\delta,\eta,\epsilon) = \max\{\widetilde{N}_0(r,\eta,\epsilon), 4r^2/\delta^2\}$, for all $n > \tilde{N}(r,\delta,\eta,\epsilon)$,

$$\begin{aligned}
&P_{f_{0,n}}\left(\sup_{g,h\,\in\,\overline{B}_{\mathbf{0}}(r)} f_n(g,h) > \eta\right) \\
&\leq P_{f_{0,n}}\left(\sup_{g_n,h_n\,\in\,\overline{B}_{\mathbf{0}}(r)} |\log(b_n(g_n,h_n))| > \eta\right) \leq \epsilon\,.
\end{aligned} \tag{63}$$

Step 2: In the previous step, we have shown how to control the function $f_n(g,h)$ over a ball of fixed radius r around zero. In this step we show that it is possible to construct a sequence $r_n \to \infty$ and an integer-valued function $N^*(\delta,\eta,\epsilon)$ for which $n > N^*(\delta,\eta,\epsilon)$ implies

$$P_{f_{0,n}}\left(\sup_{g,h\,\in\,\overline{B}_{\mathbf{0}}(r_n)} f_n(g,h) > \eta\right) < \epsilon. \tag{64}$$

For this purpose, define the integer-valued function:

$$N^*_{\delta,\eta,\epsilon}(r) \equiv$$
$$\min\left\{ n_0 \in \mathbb{N} \;\middle|\; n_0 \geq 4r^2/\delta^2 \text{ and } \mathbb{P}_{f_{0,n}}\left(\sup_{g,h\in \overline{B}_0(r)} f_n(g,h) > \eta\right) \leq \epsilon,\ \forall n > n_0 \right\}.$$

This function is well-defined by Step 1, and in fact, $N^*_{\delta,\eta,\epsilon}(r) \leq \tilde{N}(r,\delta,\eta,\epsilon)$ for any $r > 0$; see the definition of $\tilde{N}(r,\delta,\eta,\epsilon)$ at the end of Step 1. This function is non-decreasing.[20] Moreover, it has countably many discontinuity points that we can enumerate as $\tilde{r}_1 < \tilde{r}_2 < \ldots$ and $\tilde{r}_n \to \infty$ as $n \to \infty$.

We use the sequence $\tilde{r}_1 < \tilde{r}_2 < \ldots$ to find a threshold $N^*(\delta,\eta,\epsilon)$ and a diverging sequence $\{r_n\}_{n\in\mathbb{N}}$ for which (64) holds. We start by setting

$$N^*(\delta,\eta,\epsilon) \equiv N^*_{\delta,\eta,\epsilon}(\tilde{r}_1), \tag{65}$$

and for any $n > N^*(\delta,\eta,\epsilon)$ we define:

$$\tau(n) \equiv \min\left\{ j \in \mathbb{N} \;\middle|\; n \leq N^*_{\delta,\eta,\epsilon}(\tilde{r}_{j+1}) \right\}.$$

This means that—by construction—$N^*_{\delta,\eta,\epsilon}(\tilde{r}_{\tau(n)}) < n \leq N^*_{\delta,\eta,\epsilon}(\tilde{r}_{\tau(n)+1})$ and this implies $\tau(n) \to \infty$ as $n \to \infty$. We build a sequence $\{r_n\}_{n\in\mathbb{N}}$, where for $n > N^*(\delta,\eta,\epsilon)$ we set

$$r_n = \tilde{r}_{\tau(n)}. \tag{66}$$

Thus $r_n \to \infty$ as $n \to \infty$, since $\tilde{r}_n$ and $\tau(n)$ both go to infinity. To verify that our suggested

[20] To verify this, observe that $\sup_{g,h\in\overline{B}_0(r)} f_n(g,h) \leq \sup_{g,h\in\overline{B}_0(s)} f_n(g,h)$ when $r < s$. This implies

$$P_{f_{0,n}}\left(\sup_{g,h\,\in\,\overline{B}_{\mathbf{0}}(r)} f_n(g,h) > \eta\right) \leq P_{f_{0,n}}\left(\sup_{g_n,h_n\,\in\,\overline{B}_{\mathbf{0}}(s)} f_n(g,h) > \eta\right),$$

and taking $n = N^*_{\delta,\eta,\epsilon}(s)$, we conclude that right side in the inequality above is lower than ϵ. Then, $N^*_{\delta,\eta,\epsilon}(r) \leq N^*_{\delta,\eta,\epsilon}(s)$.

sequence $\{r_n\}_{n\in\mathbb{N}}$ and the threshold $N^*(\delta,\eta,\epsilon)$ satisfy (64) simply take any $n > N^*(\delta,\eta,\epsilon)$ and evaluate

$$P_{f_{0,n}}\left(\sup_{g,h\,\in\,\overline{B}_{\mathbf{0}}(r_n)} f_n(g,h) > \eta\right) = P_{f_{0,n}}\left(\sup_{g,h\,\in\,\overline{B}_{\mathbf{0}}(\tilde{r}_{\tau(n)})} f_n(g,h) > \eta\right),$$

where the equality follows since $r_n = \tilde{r}_{\tau(n)}$. We finally notice that, by construction $N^*_{\delta,\eta,\epsilon}(\tilde{r}_{\tau(n)}) < n$, hence the right side of the above is upper bounded by ϵ, since for any r, the function $N^*_{\delta,\eta,\epsilon}(r)$ gives the smallest integer n_0 such that

$$n_0 \geq 4r^2/\delta^2 \quad \text{ and } \quad \mathbb{P}_{f_{0,n}}\left(\sup_{g,h\in\overline{B}_0(r)} f_n(g,h) > \eta\right) \leq \epsilon,\ \forall n > n_0.$$

This implies that for any $n > N^*(\delta,\eta,\epsilon)$,

$$P_{f_{0,n}}\left(\sup_{g,h\,\in\,\overline{B}_{\mathbf{0}}(r_n)} f_n(g,h) > \eta\right) < \epsilon.$$

Lemma 3 *Suppose Assumption 1 and 2 holds. Let (μ_n, Σ_n) be a sequence such that $(\sqrt{n}(\mu_n - \theta^*), n\Sigma_n)$ is bounded in $f_{0,n}$-probability. Then,*

$$\mathcal{K}(\phi(\cdot\,|\,\mu_n,\,\Sigma_n)\,||\,\pi_{n,\alpha}(\cdot\,|\,X^n)) = \mathcal{K}\left(\phi(\cdot\,|\,\mu_n,\,\Sigma_n)\,||\,\phi(\cdot\,|\,\widehat{\theta}_{ML\text{-}\mathcal{F}_n},\,V_{\theta^*}^{-1}/(\alpha n))\right) + o_{f_{0,n}}(1)$$

Proof: Using the change of variables $h \equiv \sqrt{n}(\theta - \theta^*)$ and reparametrizing $\overline{\mu}_n \equiv \sqrt{n}(\mu_n - \theta^*)$, we have that

$$\begin{aligned}\mathcal{K}\left(\phi(\cdot\,|\,\mu_n,\,\Sigma_n)\,||\,\pi_{n,\alpha}(\cdot\,|\,X^n)\right) &= \int \phi(h\,|\,\overline{\mu}_n,\,n\Sigma_n)\log\left(\frac{n^{p/2}\phi(h\,|\,\overline{\mu}_n,\,n\Sigma_n)}{\pi_{n,\alpha}(\theta^* + h/\sqrt{n}\,|\,X^n)}\right)\mathrm{d}h \\ &= I_1 + I_2 + I_3, \end{aligned} \tag{67}$$

where

$$
\begin{aligned}
I_1 &\equiv \int \phi(h \,|\, \overline{\mu}_n,\, n\Sigma_n)\ \log \phi(h \,|\, \overline{\mu}_n,\, n\Sigma_n) dh \\
I_2 &\equiv -\int \phi(h \,|\, \overline{\mu}_n,\, n\Sigma_n)\ \log\left(\frac{\pi_{n,\alpha}(\theta^* + h/\sqrt{n} \mid X^n)}{\pi_{n,\alpha}(\theta^* \mid X^n)}\right) dh \\
I_3 &\equiv -\left[\log\left(\frac{\pi_{n,\alpha}(\theta^* \mid X^n)}{\phi(\theta^* \,|\, \widehat{\theta}_{\text{ML-}\mathcal{F}_n},\, V_{\theta^*}^{-1}/(\alpha n))}\right) + \log\left(\frac{\phi(\theta^* \,|\, \widehat{\theta}_{\text{ML-}\mathcal{F}_n},\, V_{\theta^*}^{-1}/(\alpha n))}{n^{p/2}}\right)\right]
\end{aligned}
$$

We will compute these three terms separately and show that their sum gives indeed the desired result. The first term I_1 is the negative of the entropy of a Gaussian distribution with mean μ_n and covariance-matrix Σ_n. A direct computation of this quantity gives

$$
I_1 = -\frac{p}{2} - \frac{p}{2}\log(2\pi) - \frac{1}{2}\log|n\Sigma_n|. \tag{68}
$$

To compute the second term I_2, notice that

$$
\frac{\pi_{n,\alpha}(\theta^* + h/\sqrt{n} \mid X^n)}{\pi_{n,\alpha}(\theta^* \mid X^n)} = \frac{f_n(X^n \mid \theta^* + h/\sqrt{n})^\alpha\, \pi(\theta^* + h/\sqrt{n})}{f_n(X^n \mid \theta^*)^\alpha\, \pi(\theta^*)}.
$$

Using this expression, Assumption 1 and Assumption 2 we see that

$$
\begin{aligned}
I_2 = -\alpha \int \phi(h \,|\, \overline{\mu}_n,\, n\Sigma_n) \log\left(\frac{f_n(X^n \mid \theta^* + h/\sqrt{n})}{f_n(X^n \mid \theta^*)}\right) \mathrm{d}h \\
- \int \phi(h \,|\, \overline{\mu}_n,\, n\Sigma_n)\; \log\left(\frac{\pi(\theta^* + h/\sqrt{n})}{\pi(\theta^*)}\right) \mathrm{d}h \\
= -\alpha \int \phi(h \,|\, \overline{\mu}_n,\, n\Sigma_n)\; \left(h^\top V_{\theta^*}\, \Delta_{n,\theta^*} - \frac{1}{2}\, h^\top V_{\theta^*}\, h + R_n(h)\right) \mathrm{d}h \\
+ o_{f_{0.n}}(1) - \int \phi(h \,|\, \overline{\mu}_n,\, n\Sigma_n)\; \log\left(\frac{\pi(\theta^* + h/\sqrt{n})}{\pi(\theta^*)}\right) \mathrm{d}h \\
= -\alpha \int \phi(h \,|\, \overline{\mu}_n,\, n\Sigma_n)\; \left(h^\top V_{\theta^*}\, \Delta_{n,\theta^*} - \frac{1}{2}\, h^\top V_{\theta^*}\, h\right) \mathrm{d}h + o_{f_{0.n}}(1) \\
= -\alpha \overline{\mu}_n^\top V_{\theta^*}\, \Delta_{n,\theta^*} + \frac{\alpha}{2} \overline{\mu}_n^\top V_{\theta^*}\, \overline{\mu}_n + \frac{\alpha}{2} \mathrm{tr}(n\Sigma_n V_{\theta^*}) + o_{f_{0,n}}(1) \\
= -\frac{\alpha}{2} \Delta_{n,\theta^*}^\top V_{\theta^*}\, \Delta_{n,\theta^*} + \frac{\alpha}{2} (\Delta_{n,\theta^*} - \overline{\mu}_n)^\top V_{\theta^*}\, (\Delta_{n,\theta^*} - \overline{\mu}_n) \\
+ \frac{\alpha}{2} \mathrm{tr}(n\Sigma_n V_{\theta^*}) + o_{f_{0,n}}(1).
\end{aligned}
$$

Replacing $\Delta_{n,\theta^*} = \sqrt{n}(\widehat{\theta}_{\text{ML-}\mathcal{F}_n} - \theta^*)$ and $\overline{\mu}_n = \sqrt{n}(\mu_n - \theta^*)$ yields

$$
\begin{aligned}
I_2 = \frac{1}{2} (\widehat{\theta}_{\text{ML-}\mathcal{F}_n} - \mu_n)'(\alpha n V_{\theta^*})(\widehat{\theta}_{\text{ML-}\mathcal{F}_n} - \mu_n) + \frac{1}{2} \mathrm{tr}(\Sigma_n \alpha n V_{\theta^*}) \\
- \frac{1}{2} (\widehat{\theta}_{\text{ML-}\mathcal{F}_n} - \theta^*)'(\alpha n V_{\theta^*})(\widehat{\theta}_{\text{ML-}\mathcal{F}_n} - \theta^*) + o_{f_{0,n}}(1)
\end{aligned}
\tag{69}
$$

Let's now turn to the term I_3. We first claim that

$$
\log\left(\frac{\pi_{n,\alpha}(\theta^* \mid X^n)}{\phi(\theta^* \,|\, \widehat{\theta}_{\text{ML-}\mathcal{F}_n},\, V_{\theta^*}^{-1}/(\alpha n))}\right) = o_{f_{0,n}}(1). \tag{70}
$$

which in turn implies implies that

$$
\begin{aligned}
I_3 = -\log\left(\frac{\phi(\theta^* \,|\, \widehat{\theta}_{\text{ML-}\mathcal{F}_n},\, V_{\theta^*}^{-1}/(\alpha n))}{n^{p/2}}\right) + o_{f_{0,n}}(1) \\
= \frac{p}{2} \log(2\pi) - \frac{1}{2} \log|\alpha V| + \frac{1}{2} (\widehat{\theta}_{\text{ML-}\mathcal{F}_n} - \theta^*)^\top (\alpha n V_{\theta^*})(\widehat{\theta}_{\text{ML-}\mathcal{F}_n} - \theta^*) + o_{f_{0,n}}(1).
\end{aligned}
\tag{71}
$$

Finally, from (67), (69) and (71) we conclude that

$$
\begin{aligned}
&\mathcal{K}(\phi(\cdot \mid \mu_n,\, \Sigma_n) \,||\, \pi_{n,\alpha}(\cdot \mid X^n)) \\
&\qquad = \frac{1}{2}\Big[-\log|\alpha n V_{\theta^*}| - \log|\Sigma_n| + \mathrm{tr}(\Sigma_n \alpha n V_{\theta^*}) \\
&\qquad\quad + (\widehat{\theta}_{\text{ML-}\mathcal{F}_n} - \mu_n)^\top (\alpha n V_{\theta^*})(\widehat{\theta}_{\text{ML-}\mathcal{F}_n} - \mu_n) - p\Big] + o_{f_{0,n}}(1) \\
&\qquad = \mathcal{K}\left(\phi(\cdot \mid \mu_n,\, \Sigma_n) \,||\, \phi(\cdot \mid \widehat{\theta}_{\text{ML-}\mathcal{F}_n},\, V_{\theta^*}^{-1}/(\alpha n))\right) + o_{f_{0,n}}(1),
\end{aligned}
$$

where the last equality follows from (41). It therefore remains to verify (70) in order to complete the proof. We first notice that the change of variables $h \equiv \sqrt{n}(\theta - \theta^*)$ leads to

$$
\frac{\pi_{n,\alpha}(\theta^* \mid X^n)}{\phi(\theta^* \mid \widehat{\theta}_{\text{ML-}\mathcal{F}_n},\, V_{\theta^*}^{-1}/(\alpha n))} = \frac{\pi^{LAN}_{n,\alpha}(0 \mid X^n)}{\phi_n(0)}, \tag{72}
$$

where $\phi_n(h) \equiv n^{-p/2}\phi(h \mid \Delta_{n,\theta^*},\, V_{\theta^*}^{-1}/\alpha)$ and $\pi^{LAN}_{n,\alpha}(h \mid X^n) \equiv n^{-p/2}\pi_{n,\alpha}(\theta^* + h/\sqrt{n} \mid X^n)$, are scaled versions of the densities $\pi_{n,\alpha}$ and ϕ. Consider the function $b_n(g,h)$ defined as

$$
b_n(g,h) \equiv \frac{\pi^{LAN}_{n,\alpha}(h \mid X^n)}{\phi_n(h)} \frac{\phi_n(g)}{\pi^{LAN}_{n,\alpha}(g \mid X^n)},
$$

and note that it is the same as (60). Therefore, from using (62) we can guarantee the existence of r_n such that for any $\eta > 0$

$$
\lim P_{f_{0,n}}\left(\sup_{g,h \,\in\, \overline{B}_{\mathbf{0}}(r_n)} |\log(b_n(g,h))| > \eta\right) = 0.
$$

This implies that

$$
\sup_{g,h \in \overline{B}_0(r_n)} b_n(g,h) = 1 + o_{f_{0,n}}(1).
$$

Define $c_n = (\int_{\overline{B}_0(r_n)} \pi^{LAN}_{n,\alpha}(h \mid X^n)\mathrm{d}h)^{-1}$ and $d_n = (\int_{\overline{B}_0(r_n)} \phi_n(h)\mathrm{d}h)^{-1}$. Using this notation, we have that (72) is equal to

$$\frac{d_n}{c_n}\int_{\overline{B}_0(r_n)}\frac{\pi^{LAN}_{n,\alpha}(0\mid X^n)}{\phi_n(0)}\frac{\phi_n(g)}{\pi^{LAN}_{n,\alpha}(g\mid X^n)}c_n\pi^{LAN}_{n,\alpha}(g\mid X^n)\mathrm{d}g,$$

which is lower—by definition of b_n—than

$$\frac{d_n}{c_n}\sup_{g,h\in\overline{B}_0(r_n)} b_n(g,h)=\frac{d_n}{c_n}\left(1+o_{f_{0,n}}(1)\right).$$

By the concentration assumption of the α-posterior (9) we have that $c_n \to 1$ in $f_{0,n}$-probability. Furthermore, from Lemma 5.2 in Kleijn and Van der Vaart (2012), it follows that $d_n \to 1$ in $f_{0,n}$-probability. This implies that

$$\frac{\pi^{LAN}_{n,\alpha}(0\mid X^n)}{\phi_n(0\,|\,\Delta_{n,\theta^*},\,V_{\theta^*}^{-1}/\alpha)}\le 1+o_{f_{0,n}}(1).$$

In a similar way, we can conclude

$$\frac{\phi_n(0\,|\,\Delta_{n,\theta^*},\,V_{\theta^*}^{-1}/\alpha)}{\pi^{LAN}_{n,\alpha}(0\mid X^n)}\le 1+o_{f_{0,n}}(1).$$

Using (72), and the two inequalities above, we can conclude (70).

Lemma 4 *Let* $(\mu_{n,\alpha},\Sigma_{n,\alpha})$ *be the sequence defined in* (28) *and* (29). *Denote by* θ^* *the (pseudo-) true parameter of the illustrative example in Section 5. Then, the sequence* $(\sqrt{n}(\mu_{n,\alpha}-\theta^*),\, n\Sigma_{n,\alpha})$ *is bounded in* $f_{0,n}$*-probability. Moreover, if* $\widehat{\theta}_{ML}$ *is the maximum likelihood estimator of* θ^**, we have that* $n(\mu_{n,\alpha}-\widehat{\theta}_{ML})$ *is bounded in* $f_{0,n}$*-probability.*

Proof: Using (28), we have that $\sqrt{n}(\mu_{n,\alpha}-\theta^*)$ is equal to

$$\left(\frac{1}{n}\sum_{i=1}^{n}W_iW_i^{\top}+\frac{1}{\alpha n}\Sigma_\pi\right)^{-1}\left(\frac{1}{\alpha\sqrt{n}}\Sigma_\pi(\mu_\pi-\theta^*)+\frac{1}{n}\sum_{i=1}^{n}W_iW_i^{\top}\,\sqrt{n}(\widehat{\theta}_{ML}-\theta^*)\right),$$

where

$$\widehat{\theta}_{ML} = \left(\frac{1}{n}\sum_{i=1}^{n} W_i W_i^\top\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} W_i Y_i.$$

Since θ^* is the (pseudo-) true parameter and $\widehat{\theta}_{ML}$ is the maximum likelihood estimator, we have that $\sqrt{n}(\widehat{\theta}_{ML} - \theta^*)$ converge in distribution to a multivariate normal distribution, and $n^{-1}\sum_{i=1}^{n} W_i W_i^\top$ converge to $\mathbb{E}[W_i W_i^\top]$ in $f_{0,n}$-probability. By Slutsky's theorem, we conclude $\sqrt{n}(\mu_{n,\alpha} - \theta^*)$ converge in distribution, which implies that $\sqrt{n}(\mu_{n,\alpha} - \theta^*)$ is bounded in $f_{0,n}$-probability.

Using (29), we have that $n\Sigma_{n,\alpha}$ is equal to

$$n\Sigma_{n,\alpha} \equiv \frac{\sigma_u^2}{\alpha}\left(\frac{1}{n}\sum_{i=1}^{n} W_i W_i^\top + \frac{1}{\alpha n}\Sigma_\pi\right)^{-1},$$

and this converge in $f_{0,n}$-probability to

$$\frac{\sigma_u^2}{\alpha}\left(\mathbb{E}[W_i W_i^\top]\right)^{-1} = \frac{1}{\alpha} V_{\theta^*}^{-1}.$$

Then, it follows that $n\Sigma_{n,\alpha}$ is bounded in $f_{0,n}$-probability.

Finally, algebra shows that $n(\mu_{n,\alpha} - \widehat{\theta}_{ML})$ is equal to

$$\left(\frac{1}{n}\sum_{i=1}^{n} W_i W_i^\top + \frac{1}{\alpha n}\Sigma_\pi\right)^{-1}\left(\frac{1}{\alpha}\Sigma_\pi(\mu_\pi - \widehat{\theta}_{ML})\right),$$

which converge in $f_{0,n}$-probability to

$$\left(\mathbb{E}[W_i W_i^\top]\right)^{-1}\left(\frac{1}{\alpha}\Sigma_\pi(\mu_\pi - \widehat{\theta}_{ML})\right).$$

This implies that $n(\mu_{n,\alpha} - \widehat{\theta}_{ML})$ is bounded in $f_{0,n}$-probability.