On the Robustness to Misspecification of α-Posteriors and Their Variational Approximations

Marco Avella Medina marco.avella@columbia.edu
Department of Statistics
Columbia University
New York, NY 10027, USA

José Luis Montiel Olea jm4474@columbia.edu
Department of Economics
Columbia University
New York, NY 10027, USA

Cynthia Rush cynthia.rush@columbia.edu
Department of Statistics
Columbia University
New York, NY 10027, USA

Amilcar Velez amilcare@u.northwestern.edu
Department of Economics
Northwestern University
Evanston, IL 60208, USA

Editor: Emtiyaz Khan

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, reduce the Kullback-Leibler (KL) divergence from the true, but perhaps infeasible, posterior distribution when there is potential parametric model misspecification. To make this point, we derive a Bernstein-von Mises theorem showing convergence in total variation distance of α-posteriors and their variational approximations to limiting Gaussian distributions. We use these limiting distributions to evaluate the KL divergence between true and reported posteriors. We show the KL divergence is minimized by choosing α strictly smaller than one, assuming there is a vanishingly small probability of model misspecification. The optimized value of α becomes smaller as the misspecification becomes more severe. The optimized KL divergence increases logarithmically in the magnitude of misspecification and not linearly as with the usual posterior. Moreover, the optimized variational approximations of α-posteriors can induce additional robustness to model misspecification, beyond that obtained by optimally downweighting the likelihood.

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

1. Introduction

A recent body of work in Bayesian statistics (Grünwald, 2011; Holmes and Walker, 2017; Grünwald, 2012; Bhattacharya et al., 2019; Miller and Dunson, 2019; Knoblauch et al., 2019) and probabilistic machine learning (Huang et al., 2018; Higgins et al., 2017; Burgess et al., 2018) has analyzed the statistical properties of α-posteriors and their variational approximations. The α-posteriors—also known as fractional, tempered, or power posteriors—are proportional to the product of the prior and the α-power of the likelihood (Ghosal and Van der Vaart, 2017, Chapter 8.6). Their variational approximations are defined as those distributions within some tractable subclass that minimize the Kullback-Leibler (KL) divergence to the α-posterior; see Alquier and Ridgway (2020), Definition 1.2.

We contribute to this growing literature by investigating the robustness-to-misspecification of α-posteriors and their variational approximations, with a focus on low-dimensional, parametric models. Our analysis—motivated by the seminal work of Gustafson (2001)—is based on a simple idea. Suppose two different procedures lead to *incorrect* a posteriori inference (either due to a likelihood misspecification or computational considerations). Define one procedure to be more robust than the other if it is closer—in terms of KL divergence—to the *true* posterior. Is it true that α-posteriors (or their variational approximations) are more robust than standard Bayesian inference?

We answer this question using asymptotic 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). Our result allows for both model misspecification and non i.i.d. data. The main assumptions are that the likelihood ratio of the presumed 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 $\sqrt{n}$. Our theorem generalizes 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, but under more primitive conditions.

These asymptotic distributions allows us to study our suggested measure of robustness by computing the KL divergence between multivariate Gaussians with 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, with no effect on the mean. 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 estima-

tor. Thus, the 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. Under the assumption 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) that we believe speak to the robustness of α-posteriors and their variational approximations.

The first result shows that 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^* decreases as both the probability of misspecification ϵ_n and the difference between the parameter that generated the data—which we denote θ_0—and the pseudo-true parameter—which we denote θ^*—increase, where by pseudo-true parameter we mean the point in the parameter space that provides the best approximation (in terms of KL divergence) to the distribution that generated the data. 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.

The second result demonstrates 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 $\widetilde{\alpha}_n^* \in (0, 1)$. 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 (corresponding to $\alpha = 1$). Our result also provides some theoretical support for the recent work of Higgins et al. (2017) that suggests introducing an additional hyperparameter β to temper the posterior when implementing variational inference (see Eq. 9 and the discussion that follows it).

The final result contrasts the expected KL of the optimized α-posterior (and also of the optimized α-variational approximation) against the expected KL of the regular posterior and its variational approximation. We find that the latter increases linearly in the magnitude of misspecification (in a sense we make precise), while the former do 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 (relative to standard

posteriors and their variational approximations). Our results also show that the optimized variational approximations of α-posteriors can induce additional robustness to model misspecification, beyond that obtained by the optimized α-posteriors (see Section 4.4).

1.1 Related Work

The idea that Bayesian procedures can be enhanced by adding an α parameter to decrease the influence of the likelihood in the posterior calculation has been suggested independently by many authors in the past, predating the more recent research studying its robustness properties under the name of α-posteriors. This includes work by Vovk (1990), McAllester (2003), Barron and Cover (1991), Walker and Hjort (2001), and Zhang et al. (2006).

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. 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 when the standard posterior does not. 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) also derive concentration rates for α-posteriors for high-dimensional and non-parametric models. Although comparison of the conditions required to verify concentration properties is natural in nonparametric or high-dimensional models, for the analysis of low-dimensional parametric models there are alternative suggestions in the literature. For example, Wang and Blei (2019a) compare the posterior predictive distribution based on the regular posterior and also its variational approximation, and 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 can outperform regular posteriors.

Our work complements this recent body of work by demonstrating robustness in a different sense, while still connecting to previous work on the subject. While it is true that one may be content with just 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. Yang et al. (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. Our results exploit heavily the BvM theorem which, in a sense, is based on 'second order' approximations. We believe that our results may be useful in studying other models where BvM results are available, but are more complicated than that considered in this paper, in that 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 an additional rationale for the use of variational inference methods due to their robustness to model misspecification, as opposed to solely based on computational considerations. We hope that our results will encourage the use of variational methods in fields like applied statistics and econometrics where variational inference remains somewhat underutilized despite its tremendous impact in machine learning. Recent applications of variational inference in econometrics include Bonhomme (2021), Mele and Zhu (2019), and Koop and Korobilis (2018).

We finally mention that the origins of studying the asymptotic normality of posterior distributions go back at least to the early 1900s; see Lehmann and Casella (2006); Ghosh and Ramamoorthi (2003) for historical references. To our knowledge, our Theorem 1 provides the first BvM result specialized to α-posteriors which relies on completely standard assumptions and techniques to account for model misspecification. It is worth mentioning, however, that BvM for α-posteriors can be derived from the asymptotic results for *generalized posteriors* in Miller (2021) or from the BvMs for *quasi-posteriors* in Chernozhukov and Hong (2003), Hooker and Vidyashankar (2014); Ghosh and Basu (2016); Matsubara et al. (2021). These papers make different assumptions than us, and obtain different statements for the form of convergence. For example, Miller (2021) shows almost sure convergence of total variation (while our results only guarantee convergence in probability). However, his more general results require slightly more restrictive assumptions. We also note that our BvM for variational approximations of α-posteriors in Theorem 2 and its proof (which relies on Lemma 6 in Appendix B) are novel. An alternative proof for this result could be obtained by generalizing the results of Wang and Blei (2019b), using the Γ-convergence strategy in Lu et al. (2017) for studying Gaussian approximations to target distributions found by minimizing the Kullback-Leibler divergence. We think that our results could be useful for studying variational approximations to generalized posteriors or quasi-posteriors.

1.2 Organization

The rest of this paper is organized as follows. Section 2 presents definitions and notation. Section 3 presents the Bernstein-von Mises theorem for α-posteriors and their variational

approximations. Section 4 presents our main results on the theoretical analysis of our suggested measure of robustness. Section 5 presents an example concerning a Gaussian linear regression model with omitted variables and some numerical experiments. Section 6 presents some concluding remarks and discussion. The proofs of our main results are collected in Appendix A.

2. Notation General Framework

2.1 Paper Notation and Definitions

We begin by introducing notation and definitions that we will use throughout the paper. We use $\phi(\cdot|\mu, \Sigma)$ to denote the density of a multivariate normal random vector with mean μ and covariance matrix Σ. The indicator function of event A, namely the function that takes the value 1 if event A occurs and 0 otherwise is denoted $\mathbf{1}\{A\}$. If p and q are two densities with respect to Lebesgue measure in $\mathbb{R}^p$, the total variation distance between them, denoted $d_{\text{TV}}(p,\, q)$, equals

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

See Section 2.9 in Van der Vaart (2000) for more details on the total variation distance. The KL divergence between the two distributions with densities p and q, denoted $\mathcal{K}(p \,||\, q)$, is defined as

$$\mathcal{K}(p \,||\, q) \equiv \int p(u) \log\left(\frac{p(u)}{q(u)}\right) du\,. \tag{2}$$

For a sequence of distributions P_n on random variables (or matrices) X_n, we say that $X_n = o_{P_n}(1)$ if $\lim_n P_n(\|X_n\| > \epsilon) = 0$ for every $\epsilon > 0$ where $\|X_n\| := \|X_n\|_2$ is the standard Euclidean norm (or Frobenius norm) and that the sequence X_n is 'bounded in P_n-probability' if for every $\epsilon > 0$ there exists $M_\epsilon > 0$ such that $P_n(\|X_n\| < M_\epsilon) \geq 1 - \epsilon$.

2.2 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

$$\widehat{\theta}_{\text{ML-}\mathcal{F}_n} \equiv \arg\max_{\theta \in \Theta} f_n(X^n\,|\,\theta)\,. \tag{3}$$

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. Sufficient conditions for the consistency and asymptotic normality of the ML estimator under model misspecification with i.i.d. data can be found in Huber (1967); White (1982); Kleijn and Van der Vaart (2012). Examples of other papers establishing consistency and asymptotic normality under misspecification for certain types of non i.i.d. data are given in Pouzo et al. (2016).

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* 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 the $\sqrt{n}$-stochastic local asymptotic normality (LAN) condition of Kleijn and Van der Vaart (2012), around θ^* (which, again, is assumed to belong to the interior of the parameter space).

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*

$$R_n(h) \equiv \log\left(\frac{f_n\left(X^n \,|\, \theta^* + h/\sqrt{n}\right)}{f_n\left(X^n \,|\, \theta^*\right)}\right) - h^\top V_{\theta^*}\,\Delta_{n,\theta^*} + \frac{1}{2}\,h^\top V_{\theta^*}\,h\,, \tag{4}$$

satisfies

$$\sup_{h\in K} |R_n(h)| \to 0\,, \tag{5}$$

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

Noticing that

$$h^\top V_{\theta^*}\,\Delta_{n,\theta^*} - \frac{1}{2}\,h^\top V_{\theta^*}\,h = \log\left(\frac{\phi(\theta^* + h/\sqrt{n} \,|\, \widehat{\theta}_{\text{ML-}\mathcal{F}_n},\, (nV_{\theta^*})^{-1})}{\phi(\theta^* \,|\, \widehat{\theta}_{\text{ML-}\mathcal{F}_n}, (nV_{\theta^*})^{-1})}\right)\,, \tag{6}$$

we can see that Assumption 1 implies that the corresponding likelihood ratio process for $\mathcal{F}_n$ approximates, asymptotically, that of a normal random variable.

2.3 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:

$$\pi_{n,\alpha}(\theta \,|\, X^n) \equiv \frac{[f_n(X^n \,|\, \theta)]^\alpha \, \pi(\theta)}{\int [f_n(X^n \,|\, \theta)]^\alpha \, \pi(\theta) \, d\theta} \,. \tag{7}$$

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

The projection of (7)—in Kullback-Leibler (KL) divergence—onto the space of probability distributions with independent marginals, also referred to as the mean-field family and denoted $\mathcal{Q}_{\text{MF}}$, provides the *mean-field variational approximation* to the α-posterior:

$$\widetilde{\pi}_{n,\alpha}(\cdot \,|\, X^n) \in \underset{q \,\in\, \mathcal{Q}_{\text{MF}}}{\arg\min} \, \mathcal{K}(q \,||\, \pi_{n,\alpha}(\cdot \,|\, X^n)) \,. \tag{8}$$

There is a trade-off between choosing a flexible and rich enough domain for the optimization in (8), 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. The projection onto the Gaussian mean-field family, studied in equation (8), 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. A key insight of the variational framework is that minimizing the KL divergence in (8) is equivalent to solving the program

$$\widetilde{\pi}_{n,\alpha}(\cdot \,|\, X^n) \equiv \underset{q \,\in\, \mathcal{Q}_{\text{MF}}}{\arg\min} \left\{ \int q(\theta) \log \left(f_n(X^n \,|\, \theta) \right) \, d\theta - (1/\alpha) \, \mathcal{K}(q \,||\, \pi) \right\} . \tag{9}$$

The objective function in (9) 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 scheme in (9) has been the subject of recent work in the representation learning literature in Burgess et al. (2018) and Higgins et al. (2017) under the name of the β-variational autoencoder. 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). More generally, the objective function in (9) with an arbitrary divergence function is analogous to the so-called divergence preferences studied in Maccheroni et al. (2006). This literature could be potentially useful in

understanding the role of the different divergence functions in penalizations, as well as the multiplier parameter α. We think that relating these ideas from economics and decision theory to what has been found in investigations of various forms of regularization coming from the statistical machine learning community, for example Alquier (2021); Knoblauch (2019); Knoblauch et al. (2019), is an interesting subject for future work.

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. We show that the total variation distance between the studied distribution and its limiting Gaussian distribution converges in probability to zero with growing sample size.

3.1 BvM for α-posteriors

We say that the α-posterior, $\pi_{n,\alpha}(\cdot \,|\, X^n)$ defined in (7), concentrates at rate $\sqrt{n}$ around θ^* if for every sequence of constants $r_n \to \infty$,

$$\mathbb{E}_{f_{0,n}}\left[\int \mathbf{1}\left\{\, \|\sqrt{n}\,(\theta-\theta^*)\| > r_n \,\right\}\, \pi_{n,\alpha}(\theta \,|\, X^n)\, d\theta\right] \to 0\,. \tag{10}$$

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 *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* (10)*. If Assumption 1 holds, then*

$$d_{TV}\left(\pi_{n,\alpha}(\cdot \,|\, X^n),\, \phi(\cdot \,|\, \widehat{\theta}_{ML\text{-}\mathcal{F}_n},\, V_{\theta^*}^{-1}/(\alpha n))\right) \to 0\,, \tag{11}$$

in $f_{0,n}$-probability, where $d_{TV}(\cdot,\cdot)$ denotes the total variation distance and is defined in (1) *and V_{θ^*} is the positive definite matrix satisfying Assumption 1.*

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 its inverse 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.

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 (10). 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 having 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, since $\pi_{n,\alpha}(\cdot \mid X^n)$ is not exactly multivariate normal, 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.

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, a BvM in total variation specialized to alpha posteriors 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).

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 $\widetilde{\pi}_{n,\alpha}$ will converge to the projection of such multivariate normal distribution onto the mean-field family $\mathcal{Q}_{\text{MF}}$.

To formalize this argument, let $\mathcal{Q}_{\text{GMF}-p}$ 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 $\widetilde{\mu}_n, \widetilde{\Sigma}_n$ satisfying

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

We note that the variational parameters $(\widetilde{\mu}_n, \widetilde{\Sigma}_n) := (\widetilde{\mu}_{n,\alpha}, \widetilde{\Sigma}_{n,\alpha})$ depend also on α, however, with a slight abuse of notation, in what follows we drop the explicit dependence in the subscript to simplify notation.

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

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

Note we are also abusing notation when we write (μ_n^*, Σ_n^*) instead of $(\mu_{n,\alpha}^*, \Sigma_{n,\alpha}^*)$. Algebra shows that the optimization problem (15) has a simple (and unique) closed-form solution when V_{θ^*} is positive definite; namely:

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

In words, $q(\cdot \,|\, \mu_n^*, \Sigma_n^*)$ is the distribution in the Gaussian mean-field family having the same mean as the limiting distribution of $\pi_{n,\alpha}(\cdot \,|\, X^n)$ but, as we will show, underestimates the covariance. We would like to show that the total variation distance between the distributions $\widetilde{\pi}_{n,\alpha}(\cdot \,|\, X^n)$ 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 *For any sequence (μ_n, Σ_n) such that $(\sqrt{n}(\mu_n - \theta^*), n\Sigma_n)$ is bounded in $f_{0,n}$-probability, the residual $R_n(h)$ in Assumption 1 and the prior π are such that*

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

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.

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

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

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 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 an asymptotic representation result stated formally in Lemma 6 in Appendix B showing 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

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

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, the proof of Theorem 2 shows that

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

in $f_{0,n}$-probability. The statement in (19) then 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. (See part iii) of Lemma B.1 in Ghosal and Van der Vaart (2017) for a textbook reference on Pinsker's inequality.)

4. Robustness to Misspecification

In this section we introduce our measure of robustness and our main results related to it. As we will explain below, the main idea is to measure robustness by computing the KL divergence between the α-posterior (or its variational approximations) and the *true* posterior.

4.1 Data Generating Process under Model Misspecification

In Section 2.2 we introduced the parametric family of densities $\mathcal{F}_n \equiv \{f_n(\cdot \,|\, \theta) : \theta \in \Theta \subseteq \mathbb{R}^p\}$ to denote the model used by the statistician. While it is standard to assume that the statistician—as a modeler or decision maker—is either misspecified or well-specified, we depart from the literature by instead considering a decision maker that does not know ex-ante whether its model $\mathcal{F}_n$ is correct or not. To do this, we introduce an alternative family of densities $\mathcal{G}_n \equiv \{g_n(\cdot \,|\, \theta, \gamma) : \theta \in \Theta, \gamma \in \Gamma\}$ and assume that the data observed by the statistician is generated as follows.[1]

A non-adversarial nature picks parameters (θ_0, γ_0) and then draws a Bernoulli random variable, T_n, with success probability ϵ_n. If $T_n = 0$, nature draws data $X^n(0)$ according to the statistician's model—that is, $X^n(0) \sim f_n(\cdot|\theta_0)$. However, if $T_n = 1$, nature draws $X^n(1)$ independently according to the alternative model $\mathcal{G}_n$; in particular, $X^n(1) \sim g_n(\cdot|\theta_0, \gamma_0)$. If $g_n(\cdot|\theta_0, \gamma_0) \notin \mathcal{F}_n$, then we can interpret ϵ_n as the probability that the statistician's model is misspecified.

The data observed by the statistician can be represented as $X^n = (1-T_n)X^n(0)+T_nX^n(1)$. Thus, our model posits that the true data generating process can be viewed as a mixture with density:

$$(1-\epsilon_n)f_n(\cdot|\theta_0) + \epsilon_n g_n(\cdot|\theta_0, \gamma_0),$$

1. The parameter γ could be of finite or infinite dimension (thus, $\mathcal{G}_n$ could be a nonparametric model). We do not require $\mathcal{F}_n$ to be nested in $\mathcal{G}_n$. That is, we do not posit the existence of a parameter $\gamma(\theta)$ for which $f_n(\cdot|\theta) = g_n(\cdot|\theta, \gamma(\theta))$. The main restriction of our framework is that θ is well defined in both $\mathcal{F}_n$ and $\mathcal{G}_n$.

which is conceptually analogous to the seminal ϵ-contamination model of Huber (1964) ($f_n(\cdot|\theta_0)$ is a baseline model known to the statistician, and $g_n(\cdot|\theta_0, \gamma_0)$ is an unknown contamination distribution). As in Huber's original formulation, the departure from $f_n(\cdot|\theta_0)$ is not a consequence of the modeler's ignorance: it arises from the very nature of the experiment.

There are two justifications for our choice. First, it seems reasonable to think that even a modeler that has been extremely careful in choosing a statistical model $\mathcal{F}_n$ may still doubt it, and may believe the model is incorrect with some probability. Second, our framework nests the standard set-up where the model is either misspecified or well-specified. In fact, we will argue later that $\epsilon_n = 0$ or $\epsilon_n = 1$ leads to obvious conclusions (asymptotically) regarding the use of α-posteriors: if the model is well specified, $\alpha = 1$ is the best choice; if the model is incorrectly specified, then $\alpha = 0$ is optimal.

4.2 Robustness Measures

To assess the robustness of α-posteriors we would like to compare the KL divergence between *reported* and *true* posteriors. To do this, it is convenient to explain what we mean by true posterior, and how this meaning changes depending on whether the statistician is misspecified or not.

Suppose first that $T_n = 1$. In this case $X^n = X^n(1)$ is drawn according to $g_n(\cdot|\theta_0, \gamma_0)$. We take π^* to be a prior over $\Theta \times \Gamma$ and, in a slight abuse of notation, use $\pi_n^*(\theta \mid X^n)$ to denote the posterior for θ based on the model $\mathcal{G}_n$, the prior π^*, and the data $X^n(1)$. Consequently, when $T_n = 1$, we refer to $\pi_n^*(\theta|X^n)$ as the true posterior. If the statistician reports $\pi_{n,\alpha}(\theta|X^n)$, then the quality of this report can be measured using the KL divergence $\mathcal{K}(\pi_n^*(\theta|X^n)||\pi_{n,\alpha}(\theta|X^n))$.

Suppose now that $T_n = 0$. In this case $X^n = X^n(0)$ is drawn according to $f(\cdot|\theta_0)$. In this case, the true posterior is simply $\pi_{n,1}$, which corresponds to the posterior for θ based on the model $\mathcal{F}_n$, the prior π, and the data $X^n(0)$. If the statistician reports $\pi_{n,\alpha}(\theta|X^n)$, the quality of this report can be measured using the KL divergence $\mathcal{K}(\pi_{n,1}(\theta|X^n) \,||\, \pi_{n,\alpha}(\theta|X^n))$.

Thus, the expected value of KL divergence—as we average over the realizations of T_n, but keep fixed the values of $X^n(1)$ and $X^n(0)$—becomes:

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

where, for the sake of notation, we have replaced $X^n(1)$ in the left KL divergence and $X^n(0)$ in the right KL divergence by X^n, since the dataset should be clear depending on whether we are comparing the α-posterior to the correctly-specified posterior or not. The α-posteriors corresponding to values of α that lead to smaller values of $r_n(\alpha)$ are said to be more robust to

parametric specification.[2] To the best of our knowledge, the idea of using the KL divergence between reported posteriors and true posteriors was first introduced by Gustafson (2001).

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

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

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 $\widetilde{\pi}_{n,\alpha}(\theta|X^n)$ in (22). Both $r_n(\alpha)$ and $\widetilde{r}_n(\alpha)$ are random variables, as they depend on the sampled data $X^n(1)$ and $X^n(0)$ that are used to construct the posterior distributions. The magnitudes of $r_n(\alpha)$ and $\widetilde{r}_n(\alpha)$ depend on the likelihoods 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 (10)), the true, infeasible posterior $\pi_n^*(\theta|X^n)$ is close in total variation distance to the p.d.f. of a $\mathcal{N}(\widehat{\theta}_{\text{ML}}, \Omega_0^{-1}/n)$ random variable, where $\widehat{\theta}_{\text{ML}}$ denotes the ML estimator of θ based on $\mathcal{G}_n$, and Ω_0 is a matrix that depends on (θ_0, γ_0) and the model $\mathcal{G}_n$. The same theorem also implies that when $X^n \sim f_n(\cdot|\theta_0)$, then $\pi_{n,1}(\theta|X^n)$ is close in total variation distance to the p.d.f. of a $\mathcal{N}(\widehat{\theta}_{\text{ML}-\mathcal{F}_n}, V_{\theta_0}^{-1}/n)$. These results, combined with our Theorems 1 and 2, naturally suggest surrogates for (21) and (22) where we replace the densities by their asymptotically normal approximations. When the data is generated by $g_n(\theta_0, \gamma_0)$, but analyzed using $\mathcal{F}_n$, our Bernstein-von Mises theorem will depend on the pseudo-true parameter (White, 1982) as

$$\theta^* \equiv \text{argmin}_{\theta\in\Theta} \mathcal{K}\left(g_n(\cdot|\theta_0,\gamma_0) \,||\, f_n(\cdot|\theta)\right), \tag{23}$$

which we will assume to be unique and to belong to the interior of Θ.

Replacing true and reported posteriors by their asymptotically normal approximations, we can define the surrogate measures $r_n^*(\alpha)$ and $\widetilde{r}_n^*(\alpha)$ where

$$\begin{aligned} r_n^*(\alpha) \equiv{}& \epsilon_n \mathcal{K}\left(\phi(\cdot\,|\,\widehat{\theta}_{\text{ML}}, \Omega_0^{-1}/n) \,||\, \phi(\cdot\,|\,\widehat{\theta}_{\text{ML-}\mathcal{F}_n}, V_{\theta^*}^{-1}/(\alpha n))\right) \\ &+ (1-\epsilon_n)\mathcal{K}\left(\phi(\cdot\,|\,\widehat{\theta}_{\text{ML-}\mathcal{F}_n}, V_{\theta_0}^{-1}/n) \,||\, \phi(\cdot\,|\,\widehat{\theta}_{\text{ML-}\mathcal{F}_n}, V_{\theta_0}^{-1}/(\alpha n))\right), \end{aligned} \tag{24}$$

2. Because $r_n(\cdot)$ is a measure of robustness, we decided to use the letter r to remind a reader that we are using this function to measure robustness.

and

$$\begin{aligned}\widetilde{r}_n^*(\alpha) \equiv \epsilon_n \mathcal{K}\left(\phi(\cdot \,|\, \widehat{\theta}_{\text{ML}}, \Omega_0^{-1}/n) \,||\, \phi(\cdot \,|\, \widehat{\theta}_{\text{ML-}\mathcal{F}_n}, \text{diag}(V_{\theta^*})^{-1}/(\alpha n))\right) \\ + (1-\epsilon_n)\mathcal{K}\left(\phi(\cdot \,|\, \widehat{\theta}_{\text{ML-}\mathcal{F}_n}, V_{\theta_0}^{-1}/n) \,||\, \phi(\cdot \,|\, \widehat{\theta}_{\text{ML-}\mathcal{F}_n}, \text{diag}(V_{\theta_0})^{-1}/(\alpha n))\right).\end{aligned} \tag{25}$$

We provide a brief discussion of the terms that appear in Equations (24)-(25).

Consider first the surrogate measure $r_n^*(\alpha)$ in Equation (24). The first term in the formula for the expected KL captures the behavior of reported α-posteriors under misspecification by comparing two normal distributions with different locations and different variances. One of these normal distributions is centered at the Maximum Likelihood estimator based on the true (but infeasible) model $\mathcal{G}_n$; while the other is centered at the Maximum Likelihood estimator based on the postulated model $\mathcal{F}_n$. Both variances converge to zero at rate $1/n$, but they are different because as they are based on different models. Note that the normal approximation for the reported α-posterior depends on the pseudo-true parameter θ^*. The second term in the formula for the expected KL captures the behavior of reported α-posteriors under correct specification, which now involves the comparison of two normal distributions with only a different scaling (α-posteriors adjust the variance by a factor $1/\alpha$). Because the data was generated by $f(\cdot|\theta_0)$, the parameter θ_0 determines the covariance matrix in these approximations.

The interpretation of the surrogate measure $\widetilde{r}_n^*(\alpha)$ is analogous. One important difference is that in the well-specified regime the difference between the variances is not only an issue of scaling. The covariance matrix for the variational approximation of α-posteriors is diagonal, while that of the true posterior need not be. The goal is also to find the value of α that makes the expected KL in Equation (25) as small as possible.

4.3 Optimal tuning of α-posteriors and their variational approximations

Our goal is to find the value of α that makes the expected KL in Equations (24) and (25) as small as possible. Denote these values as

$$\alpha_n^* \equiv \arg\min_{\alpha \geq 0} r_n^*(\alpha)\,, \qquad \widetilde{\alpha}_n^* \equiv \arg\min_{\alpha \geq 0} \widetilde{r}_n^*(\alpha)\,. \tag{26}$$

The following result provides exact, non-asymptotic expressions for α_n^* and $\tilde{\alpha}_n^*$:

Lemma 1 *Let $p \equiv dim(\theta)$. Let $\widehat{\theta}_{ML\text{-}\mathcal{F}_n}$ and $\widehat{\theta}_{ML}$ denote the Maximum Likelihood estimators based on $\mathcal{F}_n$ and $\mathcal{G}_n$, respectively. Let $\Omega_0, V_{\theta^*}, V_{\theta_0}$ be the positive definite matrices in Equations*

(24) *and* (25)*. For any* $n \in \mathbb{N}$ *and* $\epsilon_n \in [0,1]$*:*

$$\begin{aligned}
\alpha_n^* &= \frac{p}{\epsilon_n tr(V_{\theta^*}\Omega_0^{-1}) + (1-\epsilon_n)p + n\epsilon_n(\widehat{\theta}_{ML\text{-}\mathcal{F}_n} - \widehat{\theta}_{ML})^\top V_{\theta^*}(\widehat{\theta}_{ML\text{-}\mathcal{F}_n} - \widehat{\theta}_{ML})}\,, \\
\widetilde{\alpha}_n^* &= \frac{p}{\epsilon_n tr(\widetilde{V}_{\theta^*}\Omega_0^{-1}) + (1-\epsilon_n) tr(\widetilde{V}_{\theta_0}V_{\theta_0}^{-1}) + n\epsilon_n(\widehat{\theta}_{ML\text{-}\mathcal{F}_n} - \widehat{\theta}_{ML})^\top \widetilde{V}_{\theta^*}(\widehat{\theta}_{ML\text{-}\mathcal{F}_n} - \widehat{\theta}_{ML})}\,,
\end{aligned}$$

where

$$\widetilde{V}_{\theta_0} = diag\,(V_{\theta_0})\,, \quad \widetilde{V}_{\theta^*} \equiv diag\,(V_{\theta^*})\,. \tag{27}$$

The lemma shows that the optimal tuning for α-posteriors and their variational approximations will vary depending on how large the probability of misspecification is relative to the sample size. For example, when $n\epsilon_n$ is large (which will be the case when n is large and the misspecified regime has probability one) both α_n^* and $\tilde{\alpha}_n^*$ will be close to zero, provided $\widehat{\theta}_{\mathrm{ML}}$ and $\widehat{\theta}_{\mathrm{ML}-\mathcal{F}_n}$ are sufficiently different. Likewise, when $n\epsilon_n$ is small (which will be the case when the well-specified regime has probability close to one) α_n^* will be close to one, and $\widetilde{\alpha}_n^*$ will be close to $p/\mathrm{tr}(\widetilde{V}_{\theta_0}V_{\theta_0}^{-1})$ which can be shown to be less than or equal to one. We think these cases are of limited interest, because they lead to obvious recommendations for the tuning of α-posteriors (either ignore the likelihood completely, or take it at face value).

Consequently, we choose to focus on the case in which $n\epsilon_n$ converges to a strictly positive constant. Our main result is stated in Theorem 3 below, where we show that choices of α_n^* and $\widetilde{\alpha}_n^*$ strictly in $(0,1)$ when we consider sequences ϵ_n such that $n\epsilon_n \to \varepsilon$. This asymptotic regime aims to approximate situations in which the probability of misspecification is not too small relative to sample size. This regime is interesting because whenever the probability of misspecification is too small or too large, α_n^* and $\widetilde{\alpha}_n^*$ have trivial limits that are unlikely to be useful in finite samples. As we have discussed in the previous paragraph, if the probability of misspecification is too large (i.e., if $n\epsilon_n \to \infty$), then both α_n^* and $\widetilde{\alpha}_n^*$ converge to zero. We now present our main result.

Theorem 3 *Let* $p \equiv dim(\theta)$*. Let* $V_{\theta^*}, V_{\theta_0}$ *be the positive definite matrices in Equations* (24) *and* (25)*, where* θ_0 *and* θ^* *are the true and pseudo-true parameters respectively. If* $n\epsilon_n \to \varepsilon \in (0,\infty)$ *then*

$$\alpha_n^* \to \alpha^* \equiv \frac{p}{p + \varepsilon(\theta_0 - \theta^*)^\top V_{\theta^*}(\theta_0 - \theta^*)} \leq 1\,, \tag{28}$$

$$\widetilde{\alpha}_n^* \to \widetilde{\alpha}^* \equiv \frac{p}{\mathrm{tr}(\widetilde{V}_{\theta_0}V_{\theta_0}^{-1}) + \varepsilon(\theta_0 - \theta^*)^\top \widetilde{V}_{\theta^*}(\theta_0 - \theta^*)} \leq 1\,, \tag{29}$$

where the convergence is in probability, relative to the data generating process described in Section 4.1. Moreover, if $\theta_0 \neq \theta^*$ *then the inequalities above are strict.*

We now discuss the meaning and implications of Theorem 3. Equation (28) 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 (29) 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 $\widetilde{V}_\theta^{-1}/\alpha n$, as opposed to $V_\theta^{-1}/\alpha n$, where we have defined $\widetilde{V}_\theta \equiv \text{diag}(V_\theta)$, for positive definite V_θ and $\theta \in \{\theta^*, \theta_0\}$ depending on whether the Bernstein-von Mises theorem is derived for a correctly specified model or not. It is well known that the variational approximation understates the variances of each coordinate of θ (Blei et al. (2017)), hence $[(\widetilde{V}_{\theta_0})^{-1}]_{jj} \leq [V_{\theta_0}^{-1}]_{jj}$ for $j = 1, \ldots, p$. Therefore, $\text{tr}(\widetilde{V}_{\theta_0} V_{\theta_0}^{-1}) \geq p$, which establishes the inequality in (29).

4.4 Additional Robustness of Variational Inference

In this section, we present additional derivations showing that the optimized variational approximation to α-posteriors can be more robust than the optimized α-posteriors.

In order to formalize our argument, 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:

$$2 \lim_{n\to\infty} r_n^*(\alpha_n^*) = -\, p \log(p) + p \log\left(p + \varepsilon(\theta_0 - \theta^*)^\top V_{\theta^*}(\theta_0 - \theta^*)\right) . \tag{30}$$

Likewise,

$$2 \lim_{n\to\infty} \widetilde{r}_n^*(\widetilde{\alpha}_n^*) = -p \log(p) + p \log\left(\text{tr}\left(\widetilde{V}_{\theta_0} V_{\theta_0}^{-1}\right) + \varepsilon(\theta_0 - \theta^*)^\top \widetilde{V}_{\theta^*}(\theta_0 - \theta^*)\right) + \log\left(\frac{|V_{\theta_0}|}{|\widetilde{V}_{\theta_0}|}\right) . \tag{31}$$

We first compare these two equations with the expected KL of the usual posterior ($\alpha = 1$) and its standard variational approximation. Algebra shows that

$$2 \lim_{n\to\infty} r_n^*(1) = \varepsilon(\theta_0 - \theta^*)^\top V_{\theta^*}(\theta_0 - \theta^*)\,, \tag{32}$$

and

$$2 \lim_{n\to\infty} \widetilde{r}_n^*(1) = \mathrm{tr}\left(\widetilde{V}_{\theta_0} V_{\theta_0}^{-1}\right) + \varepsilon(\theta_0 - \theta^*)^\top \widetilde{V}_{\theta^*}(\theta_0 - \theta^*) + \log\left(\frac{|V_{\theta^*}|}{|\widetilde{V}_{\theta^*}|}\right)\,. \tag{33}$$

These expressions immediately show that the expected KL distance for the regular posterior and its variational approximation increases linearly in $\varepsilon \cdot \|\theta_0 - \theta^*\|^2$, which we refer to as the *magnitude of misspecification*. The optimized KL for both the α-posteriors and their variational approximations is also monotonically increasing in this term, but its growth is logarithmic.

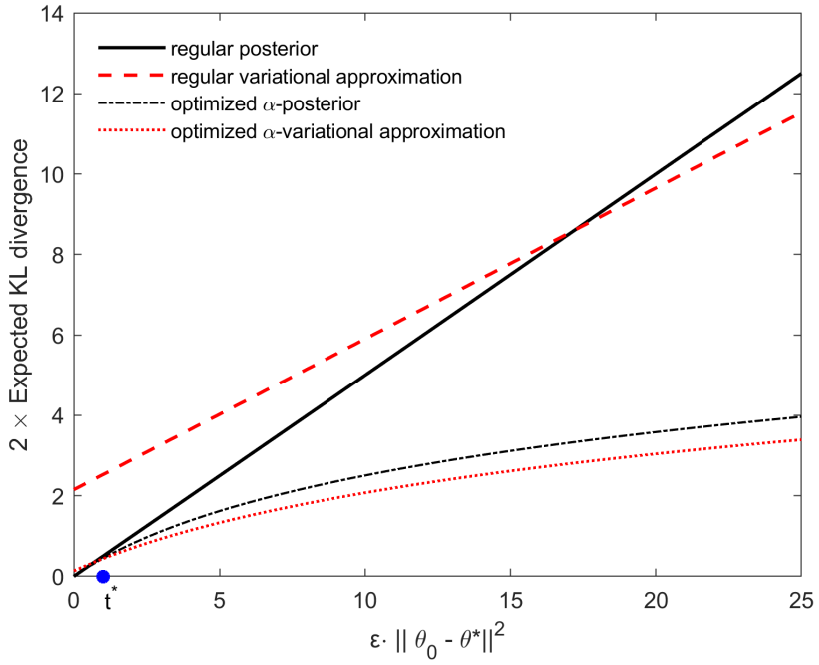
Equations (30)-(33) thus allow us to conclude that when α-posteriors and their corresponding variational approximations are optimally tuned they can be considerably more robust to model misspecification than standard variational inference, provided the magnitued of misspecification is large. Note, however, that Equations (30) and (31) are only equal when $p = 1$. Consequently, optimally tuned variational approximations to α-posteriors can be more or less robust to model misspecification than optimally tuned α-posteriors depending on the values of $\varepsilon, \theta_0, \theta^*, V_{\theta^*}$, and V_{θ_0}.

In order to illustrate this point, Figure 1 plots Equations (30)-(33) as a function of $\varepsilon \cdot \|\theta_0 - \theta^*\|^2$. Equation (32), which represents the expected KL of the standard posterior, is the solid straight line starting at the origin (since the expected KL is zero, when $\varepsilon = 0$). Equation (33), which represents the expected KL of the variational approximation of the standard posterior, is a dashed straight line, with a strictly positive ordinate at the origin (when $\varepsilon = 0$, the variational approximation distorts the variance of the posterior and thus incurs in a positive KL). The plot considers the case in which the slope of the solid line is larger than the slope of the dashed line; that is:

$$\Delta \equiv (\theta_0 - \theta^*)^\top V_{\theta^*}(\theta_0 - \theta^*)/\|\theta_0 - \theta^*\|^2 > \widetilde{\Delta} \equiv (\theta_0 - \theta^*)^\top \widetilde{V}_{\theta^*}(\theta_0 - \theta^*)/\|\theta_0 - \theta^*\|^2. \tag{34}$$

Interestingly, Lemma 7 in Appendix C shows that whenever Equation (34) holds there exists a threshold t^* such that Equation (30) is strictly larger than Equation (31) if and only if $\varepsilon\|\theta_0 - \theta^*\|^2 > t^*$. In the figure above, t^* can be found at the intersection of the two lines at the bottom of the figure, which represent the optimally tuned α-posterior and the optimally tuned α-variational approximation.

Figure 1: Robustness of α-posteriors and their variational approximations



Notes: This plot use the following parameters: $p = 2$, $\theta_0 = [-0.626\,, 0.184]$, $\theta^* = \theta_0 + (5/6) \cdot [1\,, 1]'$, $V_{\theta^*} = V_{\theta_0} = 0.25 \cdot [1\,, 1/2; 1/2\,, 1]$. These are the same as in Section 5.5.

Lemma 7 thus shows that the optimized variational approximation to α-posteriors can be more robust than the optimized α-posteriors, provided the magnitude of misspecification is large enough. It is helpful to provide some further graphical intuition for this result. Consider Figure 2 below where we plot the difference between Equations (30) and (31) as a function of $\varepsilon \cdot \|\theta_0 - \theta^*\|^2$.

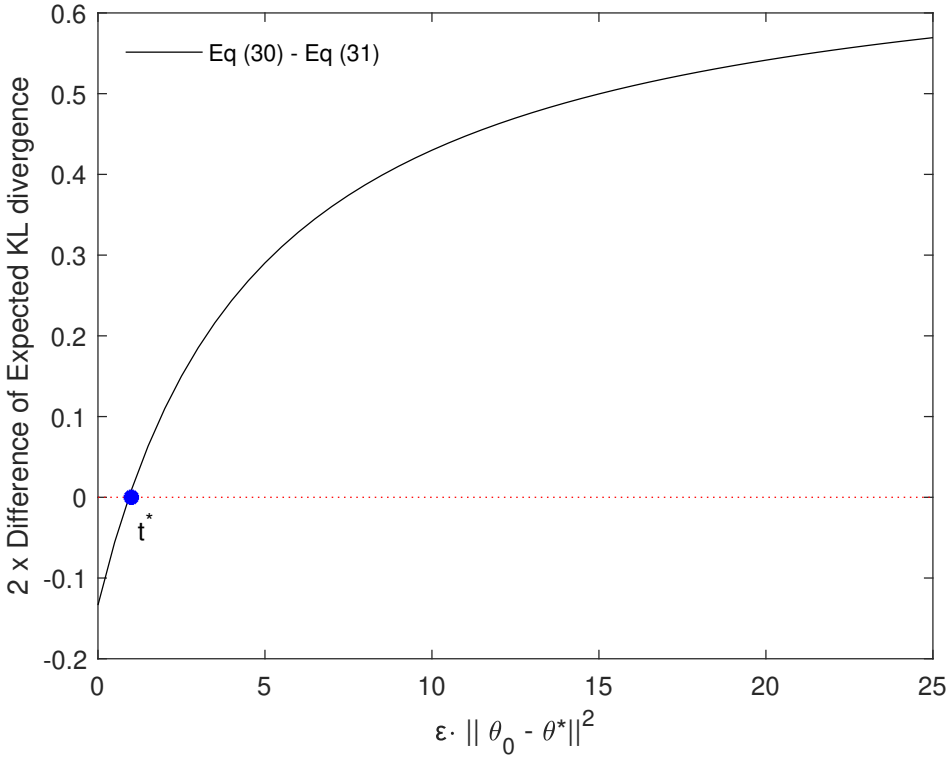
When $\varepsilon = 0$, the optimized α-posterior achieves an expected KL of zero by setting $\alpha^* = 1$. Optimizing the variational approximation in this scenario yields a positive expected KL. Thus, because the difference between these two equations is continuous near zero, we can conclude that the optimized α-posterior will be more robust than the optimized variational approximation when the magnitude of misspecification is small.

When $\varepsilon \cdot \|\theta_0 - \theta^*\|^2$ is large, the difference between Equations (30) and (31) asymptotes to

$$p \log(\Delta) - p \log(\widetilde{\Delta}) - \log\left(\frac{|V_{\theta_0}|}{|\widetilde{V}_{\theta_0}|}\right) \,,$$

which can be shown to be positive whenever Equation (34) holds. Thus, since the difference between Equations (30) and (31) is continuous in $\varepsilon \cdot \|\theta_0 - \theta^*\|^2$, there must be a magnitude

Figure 2: Difference between the expected KL between optimized α-posteriors and optimized variational approximations as a function of the magnitude of misspecification



Notes: This plot use the following parameters: $p = 2$, $\theta_0 = [-0.626\,, 0.184]$, $\theta^* = \theta_0 + (5/6) \cdot [1\,, 1]'$, $V_{\theta^*} = V_{\theta_0} = 0.25 \cdot [1\,, 1/2; 1/2\,, 1]$. These are the same as in Section 5.5.

of misspecification large enough for which the optimized variational approximation is more robust.

We still need to show there is a threshold t^* that identifies the regions in which Equation (30) is strictly larger than Equation (31). This follows because under Equation (34), the difference between the two equations is strictly increasing in $\varepsilon \cdot \|\theta_0 - \theta^*\|^2$. As we have discussed before, Equation (34) is also the condition under which standard variational inference becomes more robust than the usual posterior inference.

4.5 Additional Remarks on our results

L^q-Likelihood: Ferrari and Yang (2010) consider a generalization of the MLE that down-weights the likelihood, but their proposal is different from α-posteriors. Indeed, the ML$_q$E estimator of Ferrari and Yang (2010) can be interpreted as a weighted likelihood estimator where each individual contribution to the likelihood in the estimating equations, i.e.

$\nabla_\theta \log f(x_i, \theta)$, is multiplied by $f(x_i, \theta)^{1-q}$ for $q < 1$. Hence observations that are associated with low probabilities of the model are downweighted. One can therefore expect this proposal to give some robustness towards outliers. This is very different from the downweighting of α-posteriors since in this case the whole likelihood is downweighted and there is no observation specific downweighting. We also note that the Bayesian estimators of (Wang et al., 2017b; Wang and Blei, 2018) can be viewed as a generalization of α-posteriors where the ith contribution of the likelihood gets an observation specific weight α_i and is in that sense more similar to the ML_qE. In fact, one of the proposals in Wang et al. (2017b) leads to an individual specific downweighting scheme where observations with small values of $f(x_i, \theta)$ get downweighted; see section 2.2 and eq. (8).

'Sandwich' Covariance Matrix: It is a 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

To illustrate our main results, this section studies an example of model misspecification in the form of a linear regression model with omitted variables. Our objective is twofold. First, we 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 $\mathcal{F}_n$ and $\mathcal{G}_n$

Consider a random sample of an outcome variable Y_i with control variables $W_i \in \mathbb{R}^p$ and $Z_i \in \mathbb{R}^d$. Suppose that $\mathcal{G}_n$ is a homoskedastic Gaussian linear regression model

$$Y_i = \theta^\top W_i + \gamma^\top Z_i + \varepsilon_i \,,$$

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 $h(w_i, z_i)$ with respect to the Lebesgue measure on $\mathbb{R}^{p+d}$.

The statistician faces an omitted variables problem, in that she would like to estimate θ but only observes (Y_i, W_i). The statistician's misspecified model, $\mathcal{F}_n$, posits that

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

where u_i is assumed to be univariate normal with mean zero and a presumably known variance σ_u^2 independently of W_i. For simplicity, we assume that the statistician's specification for marginal distribution of W_i is given by the marginal of $h(w_i, z_i)$.

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

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

and the maximum likelihood estimator is simply the least-squares estimator of θ:

$$\widehat{\theta}_{\mathrm{ML}-\mathcal{F}_n} = \left(\frac{1}{n} \sum_{i=1}^n W_i W_i^\top \right)^{-1} \frac{1}{n} \sum_{i=1}^n W_i Y_i \,. \tag{35}$$

It is straightforward to show that under mild assumptions on the joint distribution of (W_i, Z_i), and provided the data is generated by (θ_0, γ_0), the Maximum Likelihood estimator in (35) is $\sqrt{n}$-asymptotically normal around the pseudo-true parameter:

$$\theta^* \;\equiv\; \theta_0 + \left(\mathbb{E}[W_i \, W_i^\top]^{-1} \, \mathbb{E}[W_i \, Z_i^\top] \right) \gamma_0 \,, \tag{36}$$

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

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

This means that, in this example, the curvature of the likelihood around the pseudo-true parameter does not depend on θ^*. In particular, the positive definite matrix V_{θ_0} defined in Section 4 is equal to V_{θ^*}.

5.3 α-posteriors and their variational approximations

Consider now the setting where we assume a commonly used Gaussian prior, denoted π, for θ. In particular, suppose

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

The computation of the 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

$$\mu_{n,\alpha} \equiv \left(\frac{1}{n} \sum_{i=1}^{n} W_i W_i^\top + \frac{1}{\alpha n}\, \Sigma_\pi \right)^{-1} \left(\frac{1}{\alpha n}\, \Sigma_\pi \mu_\pi + \frac{1}{n} \sum_{i=1}^{n} W_i Y_i \right) , \tag{37}$$

and covariance matrix

$$\Sigma_{n,\alpha} \equiv \frac{\sigma_u^2}{\alpha n} \left(\frac{1}{n} \sum_{i=1}^{n} W_i W_i^\top + \frac{1}{\alpha n}\, \Sigma_\pi \right)^{-1} . \tag{38}$$

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 C.1.

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

$$\mathcal{N}\left(\widehat{\theta}_{\mathrm{ML}-\mathcal{F}_n},\, \frac{\sigma_u^2}{\alpha n}\, \mathbb{E}[W_i\, W_i^\top]^{-1} \right) , \tag{39}$$

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 (39) converges in probability to zero (see Appendix C.2). This example thus raises the question of whether *entropic*

Bernstein-von Mises theorems are more generally available for α-posteriors in misspecified models.[3]

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. 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 (39) will be bounded away from zero. This is shown in Appendix C.3.

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 equal to

$$\widetilde{\Sigma}_{n,\alpha} \equiv \frac{\sigma_u^2}{\alpha n}\left(\text{diag}\left(\frac{1}{n}\sum_{i=1}^{n} W_i W_i^\top + \frac{1}{\alpha n}\Sigma_\pi\right)\right)^{-1}. \tag{40}$$

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 C.4.

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.

For this reason, we chose to work instead 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 asymptotic approximations. In our linear regression example, it is possible to formalize the relationship between $r_n(\alpha)$ and $r_n^*(\alpha)$. Algebra shows that for any fixed α and any Gaussian prior, $r_n(\alpha) - r_n^*(\alpha) \to 0$ as $n \to \infty$, as one would have expected (see Appendix C.6). One of the challenges in generalizing this result is that the Bernstein-von Mises theorems

3. Clarke (1999) 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.

we have established are in total variation, thus we cannot use them directly to analyze the behavior of the expected KL divergence.

Regarding the choice of α, we have shown in Theorem 3 that the limit of the optimal choice is

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

If θ_0 is different from the pseudo-true parameter θ^* then $\alpha^* < 1$. In our example, from (36), we see that $\theta_0 - \theta^* = \left(\mathbb{E}[W_i\, W_i^\top]^{-1}\, \mathbb{E}[W_i\, Z_i^\top]\right) \gamma_0$, and hence equals zero whenever $\gamma_0 = 0$ (i.e., there are indeed no omitted variables) or when the omitted variables are uncorrelated with the observed controls (i.e., when the omitted variable bias is exactly zero).

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

5.5 Numerical experiments

In our simulations, we let $p = 2$ and $d = 1$, thus $g_n(\cdot|\theta_0, \gamma_0)$ is given by the model

$$Y_i = \theta_{0,1} W_{i,1} + \theta_{0,2} W_{i,2} + \gamma_0 Z_i + \varepsilon_i\,.$$

We select $\gamma_0 = 5$ and set $\theta_0 = (-0.626, 0.184)$.[4] We use $\sigma_\epsilon = 2$ and the vector $(W_{i,1}, W_{i,2}, Z_i) \sim \mathcal{N}(0, \Sigma)$ where

$$\Sigma = \begin{bmatrix} 1 & \rho & \rho/2 \\ \rho & 1 & \rho/2 \\ \rho/2 & \rho/2 & 1 \end{bmatrix}, \qquad \text{with} \qquad \rho = 1/2\,. \tag{41}$$

The model $f_n(\cdot|\theta_0)$ omits Z_i:

$$Y_i = \theta_{0,1} W_{i,1} + \theta_{0,2} W_{i,2} + u_i\,,$$

where $u_i \sim \mathcal{N}(0, \sigma_\epsilon)$.

We mention that under this model, α-optimized variational inference provides *additional* robustness beyond that given by optimized α-posteriors alone when ϵ_n is large, as described in Section 4. Indeed, in Appendix C.7 we prove that the condition in (34) is satisfied. This is also demonstrated empirically in Figure 3.

4. These are a single draw from independent, standard Gaussians using R, with set.seed(1).

The Maximum Likelihood estimator with respect to $\mathcal{F}_n$ is $\sqrt{n}$-asymptotically normal around the pseudo-true parameter when the data is generated by $g(\cdot|\theta_0, \gamma_0)$. In this example, the pseudo-true parameter is given by

$$\theta^* \equiv \theta_0 + \left(\mathbb{E}[W_i\, W_i^\top]^{-1}\, \mathbb{E}[W_i\, Z_i]\right) \gamma_0 = \theta_0 + \left(\begin{bmatrix} 1 & 1/2 \\ 1/2 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 1/4 \\ 1/4 \end{bmatrix} \right) 5 = \theta_0 + \frac{5}{6} \begin{bmatrix} 1 \\ 1 \end{bmatrix},$$

and the stochastic LAN assumption is satisfied with the covariance matrix

$$V_{\theta^*} \equiv \mathbb{E}[W_i W_i^\top]/\sigma_u^2 = \frac{1}{4} \begin{bmatrix} 1 & 1/2 \\ 1/2 & 1 \end{bmatrix}.$$

Now we assume a Gaussian prior for (θ, γ), denoted π^*. In particular, suppose $(\theta, \gamma) \sim \mathcal{N}(\mathbf{0},\, \sigma_u^2\, \Sigma^{-1})$, where covariance structure is chosen randomly from the Wishart ensemble. We assume the prior π over θ used by the statistician is the marginal distribution of the joint prior π^* specified above. Using π, we can calculate the α-posterior using (37) and (38) and we prove in Appendix C.1 that the α-posterior concentrates at rate $\sqrt{n}$ around θ^* for any fixed α.

Recall from Section 4 that $r_n(\alpha)$ denotes the expected KL divergence between the true posterior of θ and the α-posterior and $r_n^*(\alpha)$ denotes the expected KL divergence when one replaces the true posterior and the α-posterior by their asymptotic approximations. As discussed in Section 5.4, for any fixed α and any Gaussian prior, $r_n(\alpha) - r_n^*(\alpha) \to 0$ as $n \to \infty$. Similarly, $\widetilde{r}_n(\alpha)$ denotes the expected KL divergence between the true posterior of θ and the variational approximation to the α-posterior with $\widetilde{r}_n^*(\alpha)$ similarly replacing finite n quantities with their Gaussian asymptotic approximations.

Finally, in Figure 3, for $n \in \{1000, 100000, \infty\}$ and $\alpha \in (0, 1)$, we plot $r_n(\alpha)$ in blue and $r_n^*(\alpha)$ in black on the left and $\widetilde{r}_n(\alpha)$ in blue and $\widetilde{r}_n^*(\alpha)$ in black on the right, where denoting $\widetilde{V}_{\theta^*} := \mathrm{diag}(V_{\theta^*})$, we have that

$$\begin{aligned} r_n^*(\alpha) &= \frac{1}{2}\left(\alpha A_n(V_{\theta^*}) - p\log(\alpha) + B_n(V_{\theta^*})\right), \\ \widetilde{r}_n^*(\alpha) &= \frac{1}{2}\left(\alpha A_n(\widetilde{V}_{\theta^*}) - p\log(\alpha) + B_n(\widetilde{V}_{\theta^*})\right), \end{aligned} \tag{42}$$

with

$$\begin{aligned} A_n(V) &\equiv \epsilon_n \mathrm{tr}(V\Omega^{-1}) + (1-\epsilon_n)\mathrm{tr}(VV_{\theta^*}^{-1}) + n\epsilon_n(\widehat{\theta}_{\mathrm{ML}\text{-}\mathcal{F}_n} - \widehat{\theta}_{\mathrm{ML}})^\top V(\widehat{\theta}_{\mathrm{ML}\text{-}\mathcal{F}_n} - \widehat{\theta}_{\mathrm{ML}}), \\ B_n(V) &\equiv -p + \epsilon_n \log(|\Omega||V|^{-1}) + (1-\epsilon_n)\log(|V_{\theta^*}||V|^{-1}). \end{aligned} \tag{43}$$

In (43), Ω^{-1} is the $p \times p$ upper submatrix of $\sigma_u^2 \Sigma^{-1}$ for Σ in (41), $\widehat{\theta}_{\mathrm{ML}-\mathcal{F}_n}$ is given in (35), and $\widehat{\theta}_{\mathrm{ML}}$ is the first p elements of

$$\left(\frac{1}{n} \sum_{i=1}^{n} (W_i, Z_i)^\top (W_i, Z_i) \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} (W_i, Z_i)^\top Y_i \, .$$

We mention that the curves in (42) are minimized with α^* values described by Lemma 1 in Section 4.

The red curves in both plots of Figure 3 are the limiting curves, $r_\infty^*(\alpha)$ and $\widetilde{r}_\infty^*(\alpha)$ given by

$$\begin{aligned} r_\infty^*(\alpha) &= \frac{1}{2} \left(\alpha p + \alpha \epsilon (\theta^* - \theta_0)^\top V_{\theta^*} (\theta^* - \theta_0) - p \log(\alpha) - p \right), \\ \widetilde{r}_\infty^*(\alpha) &= \frac{1}{2} \left(\alpha \mathrm{tr}(\widetilde{V}_{\theta^*} V_{\theta^*}^{-1}) + \alpha \epsilon (\theta^* - \theta_0)^\top \widetilde{V}_{\theta^*} (\theta^* - \theta_0) - p \log(\alpha) - p + \log(|V_{\theta^*}||\widetilde{V}_{\theta^*}|^{-1}) \right). \end{aligned} \tag{44}$$

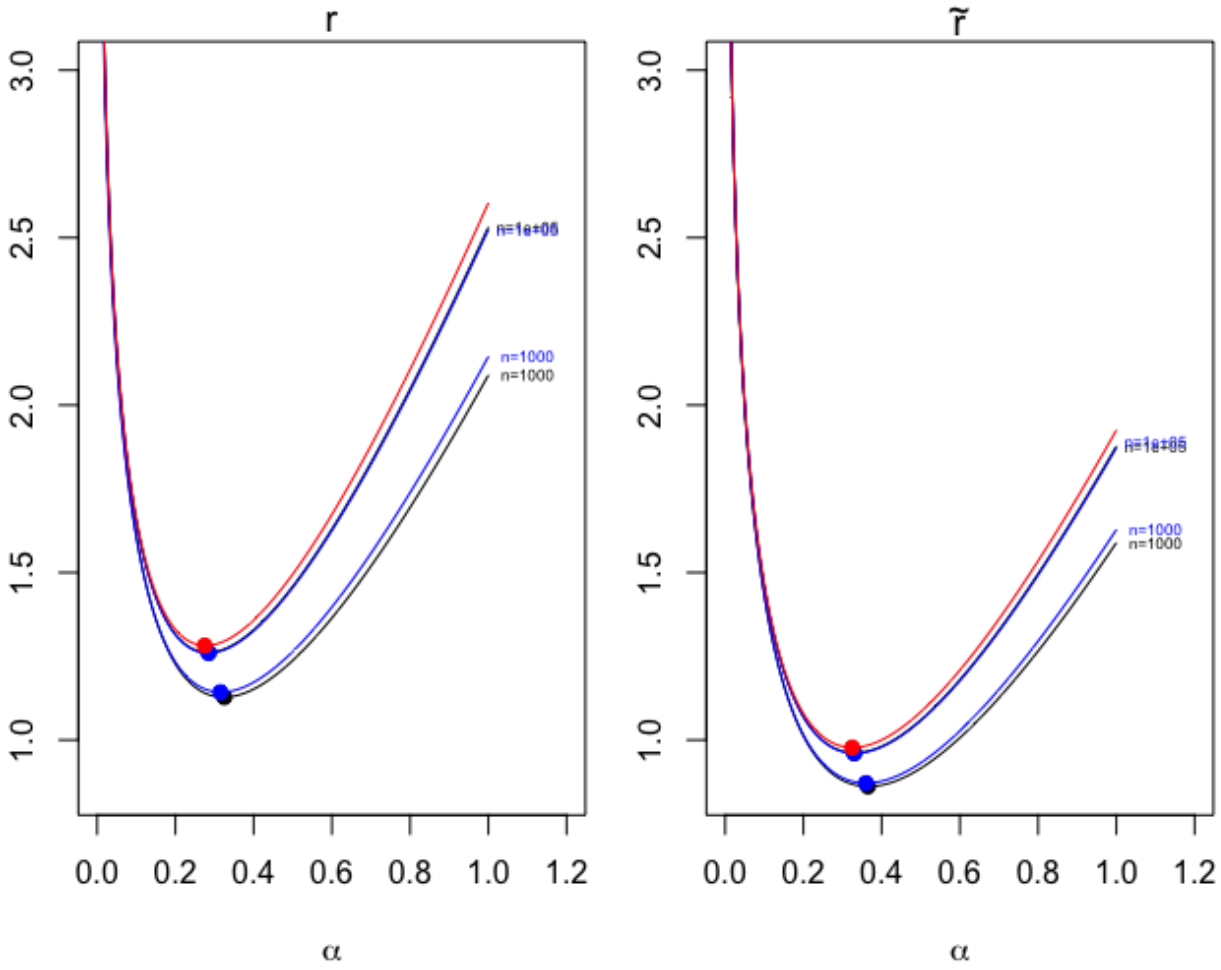
In these plots we have again selected $\epsilon_n = 10/n$ so that $n\epsilon_n \to \epsilon = 10$. The points emphasized on the curves are their minimum values.

We see from Figure 3 that the $\widetilde{r}_n(\alpha)$ and $\widetilde{r}_n^*(\alpha)$ curves are generally below the $r_n(\alpha)$ and $r_n^*(\alpha)$ curves, and indeed $\widetilde{r}_\infty^*(\alpha^*)$ is smaller than $r_\infty^*(\alpha^*)$, meaning this is an example where the variational approximation to the posterior is *more robust* than the α-posterior itself. We also see from this plot that $r_n(\alpha)$ and $r_n^*(\alpha)$ as well as $\widetilde{r}_n(\alpha)$ and $\widetilde{r}_n^*(\alpha)$ are close to each other for all value of α already at $n = 1000$. Moreover, the finite n curves are quite close to the $n = \infty$ curves when $n = 100000$. However, even at $n = 1000$, we have that the minimizing α values indicated by the points at the bottom of the curve are close to their limiting values (the points in red), suggesting that the limiting α's may be used to indicate the appropriate levels of tempering even in finite samples.

6. Concluding Remarks and Discussion

In this work, we have 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,

Figure 3: Plots of $r_n(\alpha)$ (blue, left), $r_n^*(\alpha)$ (black, left) and $\widetilde{r}_n(\alpha)$ (blue, right), $\widetilde{r}_n^*(\alpha)$ (black, right) for $n \in \{1000, 100000, \infty\}$ with $\alpha \in (0,1)$ on the x-axis and points indicating minimum values on the curves. Red curves are $r_\infty^*(\alpha)$ (left) and $\widetilde{r}_\infty^*(\alpha)$ (right) given in (44). Plotted with $\epsilon_n = 10/n$ so that $n\epsilon_n \to \epsilon = 10$.



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 corrects some of the variance understatement of standard variational approximations to the posterior. Also, there is some recent work considering variational approximations using the α-Rényi divergence instead of the KL divergence (Jaiswal et al., 2020). It might be interesting to explore whether it is possible to derive a result analogous to Theorem 2, where we approximate the limiting distribution of α-Rényi approximate posteriors by projecting the limiting distribution obtained in Theorem 1.

The main use of the asymptotic approximations in our paper, however, was simply to facilitate the computation of the suggested measure of robustness. This required 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 (1999).

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. One promising line of work is to do a full Bayesian treatment of the problem as in Wang et al. (2017a).

Acknowledgments

We would like to thank Pierre Alquier, David Blei, Stéphane Bonhomme, Matias Cattaneo, Yun Ju, Emitiyaz Khan, Jason Klusowski, Jeffrey Miller, Debdeep Pati, Yixin Wang, Mark Watson, and three anonymous referees for extremely helpful comments and suggestions. All errors remain our own. Cynthia Rush would like to acknowledge support for this project from the National Science Foundation (NSF CCF #1849883), the Simons Institute for the Theory of Computing, and NTT Research.

Appendix 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 our proof 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, where θ^* is assumed to be an element of the interior of $\Theta \subset \mathbb{R}^p$.

Since 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 $\pi_{n,\alpha}(\cdot \,|\, X^n)$, 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

the simultaneous 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 the similarly 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:

$$f_n(g,h) \equiv \left\{1 - \frac{\phi_n(h)}{\pi_{n,\alpha}^{LAN}(h \mid X^n)}\,\frac{\pi_{n,\alpha}^{LAN}(g \mid X^n)}{\phi_n(g)}\right\}^+, \tag{45}$$

where $\phi_n(h) \equiv n^{-1/2}\phi(h \,|\, \Delta_{n,\theta^*},\, V_{\theta^*}^{-1}/\alpha)$ and $\pi_{n,\alpha}^{LAN}(h \mid X^n) \equiv n^{-1/2}\pi_{n,\alpha}(\theta^* + h/\sqrt{n} \,|\, X^n)$, 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 in (45) is well-defined on $K_0 \times K_0$ for all $n > N_0$, as in this regime $\pi_{n,\alpha}^{LAN}(h \,|\, X^n)$ 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 $\eta > 0$:

$$\begin{aligned}
&\mathbb{E}_{f_{0,n}}\left[d_{\mathrm{TV}}\left(\pi_{n,\alpha}^{LAN}(\cdot \mid X^n),\, \phi_n(\cdot)\right)\right] \\
&\leq \mathbb{E}_{f_{0,n}}\left[d_{\mathrm{TV}}(\pi_{n,\alpha}^{LAN}(\cdot \mid X^n),\, \phi_n(\cdot))\mathbf{1}\left\{\sup_{g,h \in \overline{B}_{\mathbf{0}}(r_n)} f_n(g,h) \leq \eta\right\}\right] + \mathbb{P}_{f_{0,n}}\left(\sup_{g,h \in \overline{B}_{\mathbf{0}}(r_n)} f_n(g,h) > \eta\right).
\end{aligned} \tag{46}$$

The proof is completed by bounding the two terms on the right side of (46).

First we bound the expectation on the right side of (46). Lemma 4 in Appendix B implies

$$d_{\mathrm{TV}}(\pi_{n,\alpha}^{LAN}(\cdot \mid X^n),\, \phi_n) \leq \sup_{g,h \in \overline{B}_{\mathbf{0}}(r_n)} f_n(g,h) + \int_{||h||>r_n} \pi_{n,\alpha}^{LAN}(\cdot \mid X^n)\, dh + \int_{||h||>r_n} \phi_n(h)\, dh\,, \tag{47}$$

therefore

$$\begin{aligned}
&\mathbb{E}_{f_{0,n}}\left[d_{\mathrm{TV}}(\pi_{n,\alpha}^{LAN}(\cdot \mid X^n),\, \phi_n(\cdot))\mathbf{1}\left\{\sup_{g,h \in \overline{B}_{\mathbf{0}}(r_n)} f_n(g,h) \leq \eta\right\}\right] \\
&\qquad \leq \eta + \mathbb{E}_{f_{0,n}}\left[\int_{||h||>r_n} \pi_{n,\alpha}^{LAN}(h \mid X^n)\, dh\right] + \mathbb{E}_{f_{0,n}}\left[\int_{||h||>r_n} \phi_n(h)\, dh\right].
\end{aligned} \tag{48}$$

The bound in (48) uses that by the non-negativity of $\phi_n(\cdot)$, which implies

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

and a similar upper bound for the third term on the right side of (48). In addition, by the concentration assumption of the theorem (defined in equation (10)), there exists an integer $N_1(\eta,\epsilon)$ such that, for all $n > N_1(\eta,\epsilon)$,

$$\mathbb{E}_{f_{0,n}}\left[\int_{||h||>r_n}\pi^{LAN}_{n,\alpha}(h\,|\,X^n)\,dh\right]<\epsilon. \tag{49}$$

Also, by Lemma 5.2 in Kleijn and Van der Vaart (2012) which exploits properties of the multivariate normal distribution, there exists an integer $N_2(\eta,\epsilon)$, such that for all $n > N_2(\eta,\epsilon)$,

$$\mathbb{E}_{f_{0,n}}\left[\int_{||h||>r_n}\phi_n(h)\,dh\right]<\epsilon\,. \tag{50}$$

Now plugging (49) and (50) into (48), defining $\widetilde{N}(\eta,\epsilon)=\max\{N_1(\eta,\epsilon),N_2(\eta,\epsilon)\}$, we find for all $n > \widetilde{N}(\eta,\epsilon)$,

$$\mathbb{E}_{f_{0,n}}\left[d_{\mathrm{TV}}\left(\pi^{LAN}_{n,\alpha}(\cdot\,|\,X^n),\,\phi_n(\cdot)\right)1\left\{\sup_{g,h\in\overline{B}_{\mathbf{0}}(r_n)}f_n(g,h)\leq\eta\right\}\right]\leq\eta+2\epsilon. \tag{51}$$

Lemma 5 in Appendix B shows 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 (46) is small for $n > N(\eta,\epsilon)$; that is, for all $n > N(\eta,\epsilon)$,

$$\mathbb{P}_{f_{0,n}}\left(\sup_{g,h\in\overline{B}_{\mathbf{0}}(r_n)}f_n(g,h)>\eta\right)\leq\epsilon\,. \tag{52}$$

We mention that it is in the proof of Lemma 5 that we make use the stochastic LAN condition in Assumption 1.

Finally we conclude from (46), using the bounds in (51) and (52), that for all $n > \max\{N(\eta,\epsilon),\widetilde{N}(\eta,\epsilon)\}$,

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

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

A.2 Proof of Theorem 2

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

$$\mathcal{K}\left(\widetilde{\pi}_{n,\alpha}\left(\cdot \,|\, X^n\right) \,||\, q(\cdot \,|\, \mu_n^*, \Sigma_n^*)\right) = \mathcal{K}\left(\phi(\cdot \,|\, \widetilde{\mu}_n, \widetilde{\Sigma}_n) \,||\, \phi(\cdot \,|\, \widehat{\theta}_{\text{ML-}\mathcal{F}_n}, \operatorname{diag}(\alpha n V_{\theta^*})^{-1})\right) \to 0\,, \quad (53)$$

in $f_{0,n}$-probability where the equality follow since $q(\cdot \,|\, \mu_n^*, \Sigma_n^*)$ is Gaussian with mean and covariance defined in (16). 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

$$\mathcal{K}\left(\phi(\cdot \,|\, \mu_1,\, \Sigma_1) \,||\, \phi(\cdot \,|\, \mu_2,\, \Sigma_2)\right) = \frac{1}{2}\left[\log\left(\frac{|\Sigma_2|}{|\Sigma_1|}\right) + \operatorname{tr}\left(\Sigma_2^{-1}\Sigma_1\right) + (\mu_2 - \mu_1)^\top \Sigma_2^{-1}(\mu_2 - \mu_1) - p\right]. \quad (54)$$

Applying the identity (54) we see that

$$\mathcal{K}\left(\phi(\cdot \,|\, \widetilde{\mu}_n, \widetilde{\Sigma}_n) \,||\, \phi(\cdot \,|\, \widehat{\theta}_{\text{ML-}\mathcal{F}_n}, \operatorname{diag}(\alpha n V_{\theta^*})^{-1})\right) = T_1 + T_2\,,$$

where

$$T_1 = \frac{1}{2}(\widehat{\theta}_{\text{ML-}\mathcal{F}_n} - \widetilde{\mu}_n)^\top \operatorname{diag}(\alpha n V_{\theta^*})(\widehat{\theta}_{\text{ML-}\mathcal{F}_n} - \widetilde{\mu}_n)\,, \quad (55)$$

$$T_2 = \frac{1}{2}\operatorname{tr}(\operatorname{diag}(\alpha n V_{\theta^*})\widetilde{\Sigma}_n) - \frac{p}{2} + \frac{1}{2}\log\left(\frac{|\widetilde{\Sigma}_n|^{-1}}{|\operatorname{diag}(\alpha n V_{\theta^*})|}\right). \quad (56)$$

To prove (53), we will prove that both $T_1 = o_{f_{0,n}(1)}$ and $T_2 = o_{f_{0,n}(1)}$. The key step to establishing these results is the asymptotic representation result of Lemma 6. 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:

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

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 the same density and the α-posterior's total variation limit (which we have characterized in Theorem 1).

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

Claim 1. We start by showing that $T_1 = o_{f_{0,n}}(1)$. Since $\widetilde{\mu}_n$ and $\widetilde{\Sigma}_n$ are the parameters that solve the variational approximation in (14), they are thus the parameters that minimize the KL divergence between the Gaussian mean-field family and the α-posterior. It follows that for every n,

$$\mathcal{K}(q(\cdot \,|\, \widetilde{\mu}_n,\, \widetilde{\Sigma}_n) \,||\, \pi_{n,\alpha}(\cdot \,|\, X^n)) \leq \mathcal{K}(q(\cdot \,|\, \widehat{\theta}_{\mathrm{ML}\text{-}\mathcal{F}_n}, \widetilde{\Sigma}_n) \,||\, \pi_{n,\alpha}(\cdot \,|\, X^n))\,. \tag{57}$$

Using Lemma 6, we can evaluate each of the KL divergences above up to an $o_{f_{0,n}}(1)$ term using the asymptotic Gaussian limit of $\pi_{n,\alpha}(\cdot \,|\, X^n)$ given in Thm 1. Indeed, since both $(\sqrt{n}(\widetilde{\mu}_n - \theta^*),\, n\widetilde{\Sigma}_n)$ and $(\sqrt{n}(\widehat{\theta}_{\mathrm{ML}\text{-}\mathcal{F}_n} - \theta^*),\, n\widetilde{\Sigma}_n)$ are bounded in $f_{0,n}$-probability, we can apply Lemma 6 to find

$$\begin{aligned}
\mathcal{K}(q(\cdot \,|\, \widetilde{\mu}_n,\, \widetilde{\Sigma}_n) \,||\, \pi_{n,\alpha}(\cdot \,|\, X^n)) &= \mathcal{K}(q(\cdot \,|\, \widetilde{\mu}_n,\, \widetilde{\Sigma}_n) \,||\, \phi(\cdot \,|\, \widehat{\theta}_{\mathrm{ML}\text{-}\mathcal{F}_n},\, V_{\theta^*}^{-1}/(\alpha n))) + o_{f_{0,n}}(1)\,, \\
\mathcal{K}(q(\cdot \,|\, \widehat{\theta}_{\mathrm{ML}\text{-}\mathcal{F}_n}, \widetilde{\Sigma}_n) \,||\, \pi_{n,\alpha}(\cdot \,|\, X^n)) &= \mathcal{K}(q(\cdot \,|\, \widehat{\theta}_{\mathrm{ML}\text{-}\mathcal{F}_n}, \widetilde{\Sigma}_n) \,||\, \phi(\cdot \,|\, \widehat{\theta}_{\mathrm{ML}\text{-}\mathcal{F}_n},\, V_{\theta^*}^{-1}/(\alpha n))) + o_{f_{0,n}}(1)\,,
\end{aligned}$$

and plugging the above into (57) gives

$$\begin{aligned}
&\mathcal{K}\left(q(\cdot \,|\, \widetilde{\mu}_n,\, \widetilde{\Sigma}_n) \,||\, \phi(\cdot \,|\, \widehat{\theta}_{\mathrm{ML}\text{-}\mathcal{F}_n},\, V_{\theta^*}^{-1}/(\alpha n))\right) \\
&\quad \leq \mathcal{K}\left(q(\cdot \,|\, \widehat{\theta}_{\mathrm{ML}\text{-}\mathcal{F}_n},\, \widetilde{\Sigma}_n) \,||\, \phi(\cdot \,|\, \widehat{\theta}_{\mathrm{ML}\text{-}\mathcal{F}_n},\, V_{\theta^*}^{-1}/(\alpha n))\right) + o_{f_{0,n}}(1).
\end{aligned} \tag{58}$$

Therefore, the result follows from (54) and (58) that

$$\begin{aligned}
T_1 &= \frac{1}{2}(\widehat{\theta}_{\mathrm{ML}\text{-}\mathcal{F}_n} - \widetilde{\mu}_n)^\top (\alpha n V_{\theta^*})(\widehat{\theta}_{\mathrm{ML}\text{-}\mathcal{F}_n} - \widetilde{\mu}_n) \\
&= \mathcal{K}\left(q(\cdot \,|\, \widetilde{\mu}_n,\, \widetilde{\Sigma}_n) \,||\, \phi(\cdot \,|\, \widehat{\theta}_{\mathrm{ML}\text{-}\mathcal{F}_n},\, V_{\theta^*}^{-1}/(\alpha n))\right) - \mathcal{K}\left(q(\cdot \,|\, \widehat{\theta}_{\mathrm{ML}\text{-}\mathcal{F}_n},\, \widetilde{\Sigma}_n) \,||\, \phi(\cdot \,|\, \widehat{\theta}_{\mathrm{ML}\text{-}\mathcal{F}_n},\, V_{\theta^*}^{-1}/(\alpha n))\right) \\
&\leq o_{f_{0,n}}(1)\,.
\end{aligned} \tag{59}$$

Claim 2. We now show that $T_2 = o_{f_{0,n}}(1)$ by relating $\widetilde{\Sigma}_n$ to $\mathrm{diag}(V_{\theta^*})^{-1}/(\alpha n)$. The optimality of $\widetilde{\mu}_n$ and $\widetilde{\Sigma}_n$ defined by (14) once again implies that for every n:

$$\mathcal{K}(q(\cdot \,|\, \widetilde{\mu}_n, \widetilde{\Sigma}_n) \,||\, \pi_{n,\alpha}(\cdot \,|\, X^n)) \leq \mathcal{K}(q(\cdot \,|\, \widehat{\theta}_{\mathrm{ML}\text{-}\mathcal{F}_n}, \mathrm{diag}(V_{\theta^*})^{-1}/(\alpha n)) \,||\, \pi_{n,\alpha}(\cdot \,|\, X^n)). \tag{60}$$

As in the work of Claim 1, this inequality and Lemma 6 imply that

$$\begin{aligned}&\mathcal{K}(q(\cdot \,|\, \widetilde{\mu}_n, \widetilde{\Sigma}_n) \,||\, \phi(\cdot \,|\, \widehat{\theta}_{\text{ML-}\mathcal{F}_n},\, V_{\theta^*}^{-1}/(\alpha n)))\\ &\leq \mathcal{K}(q(\cdot \,|\, \widehat{\theta}_{\text{ML-}\mathcal{F}_n}, \text{diag}(V_{\theta^*})^{-1}/(\alpha n)) \,||\, \phi(\cdot \,|\, \widehat{\theta}_{\text{ML-}\mathcal{F}_n},\, V_{\theta^*}^{-1}/(\alpha n))) + o_{f_{0,n}}(1).\end{aligned} \tag{61}$$

Next, using the inequality in (61), applying (54) to each term, and noting that (59) implies $\frac{1}{2}(\widehat{\theta}_{\text{ML-}\mathcal{F}_n} - \widetilde{\mu}_n)^\top(\alpha n V_{\theta^*})(\widehat{\theta}_{\text{ML-}\mathcal{F}_n} - \widetilde{\mu}_n) \leq o_{f_{0,n}}(1)$, we obtain

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

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

$$\frac{1}{2}\left[\text{tr}(\alpha n V_{\theta^*}\widetilde{\Sigma}_n) - p + \log\left(\frac{|V_{\theta^*}^{-1}/(\alpha n)|}{|\widetilde{\Sigma}_n|}\right) - \log\left(\frac{|V_{\theta^*}^{-1}/(\alpha n)|}{|\text{diag}(V_{\theta^*})^{-1}/(\alpha n)|}\right)\right] \leq o_{f_{0,n}}(1)\,. \tag{62}$$

Since $\widetilde{\Sigma}_n$ is diagonal,

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

and consequently, the left-hand side of (62) equals T_2, which has been defined in (56). Moreover, since the term T_2 is nonnegative, as it equals the KL divergence between two normals with the same mean, but variances $\widetilde{\Sigma}_n$ and $\text{diag}(V_{\theta^*})^{-1}/\alpha n$, we conclude that $T_2 = o_{f_{0,n}}(1)$.

A.3 Proof of Lemma 1

We note that r_n^* and $\widetilde{r}_n^*$ defined in (24) and (25) are both expectations involving KL divergences of two multivariate Gaussian distributions, hence (51) allows us to compute $r_n^*(\alpha)$ and $\widetilde{r}^*(\alpha)$ explicitly as

$$\begin{aligned}r_n^*(\alpha) =& \frac{1}{2}\Big(\alpha A_n(V_{\theta^*}, V_{\theta_0}) - p\log(\alpha) + B_n(V_{\theta^*}, V_{\theta_0})\Big),\\ \widetilde{r}_n^*(\alpha) =& \frac{1}{2}\Big(\alpha A_n(\widetilde{V}_{\theta^*}, \widetilde{V}_{\theta_0}) - p\log(\alpha) + B_n(\widetilde{V}_{\theta^*}, \widetilde{V}_{\theta_0})\Big)\,,\end{aligned}$$

where

$$\begin{aligned}A_n(\Sigma, \Psi) &\equiv \epsilon_n \text{tr}(\Sigma\Omega_0^{-1}) + (1-\epsilon_n)\text{tr}(\Psi V_{\theta_0}^{-1}) + n\epsilon_n(\widehat{\theta}_{\text{ML-}\mathcal{F}_n} - \widehat{\theta}_{\text{ML}})^\top\Sigma(\widehat{\theta}_{\text{ML-}\mathcal{F}_n} - \widehat{\theta}_{\text{ML}})\,,\\ B_n(\Sigma, \Psi) &\equiv -p + \epsilon_n\log(|\Omega_0||\Sigma^{-1}|) + (1-\epsilon_n)\log(|\Psi^{-1}||V_{\theta_0}|)\,.\end{aligned}$$

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

$$\alpha_n^* = \frac{p}{A_n(V_{\theta^*}, V_{\theta_0})} \quad \text{and} \quad \widetilde{\alpha}_n^* = \frac{p}{A_n(\widetilde{V}_{\theta^*}, \widetilde{V}_{\theta_0})}\,. \tag{63}$$

A.4 Proof of Theorem 3

Denote by $f_{0,n}$ the probability density relative to the data generating process described in Section 4.1. By assumption we have $\widehat{\theta}_{\mathrm{ML}} \to \theta_0$ in $f_{0,n}$-probability and $n\epsilon_n \to \varepsilon \in (0, \infty)$. In addition, we know that $\widehat{\theta}_{\mathrm{ML}\text{-}\mathcal{F}_n} \to \theta^*$, and therefore it follows that $A_n(\Sigma, \Psi) \to \mathrm{tr}(\Psi V_{\theta_0}^{-1}) + \varepsilon(\theta^* - \theta_0)^\top \Sigma(\theta^* - \theta_0)$, both in $f_{0,n}$-probability (where A_n is defined as in the proof of Lemma 1, Section A.3). Denote α^* and $\widetilde{\alpha}^*$ the limits in $f_{0,n}$-probability of α_n^* and $\widetilde{\alpha}_n^*$. These limits can be computed replacing (Σ, Ψ) in the expressions above with $(V_{\theta^*}, V_{\theta_0})$ or $(\widetilde{V}_{\theta^*}, \widetilde{V}_{\theta_0})$ and taking limit of (63). This implies, in $f_{0,n}$-probability, that

$$\alpha_n^* \to \alpha^* \equiv \frac{p}{p + \varepsilon(\theta_0 - \theta^*)^\top V_{\theta^*}(\theta_0 - \theta^*)}\,, \qquad \widetilde{\alpha}_n^* \to \widetilde{\alpha}^* \equiv \frac{p}{\mathrm{tr}(\widetilde{V}_{\theta_0} V_{\theta_0}^{-1}) + \varepsilon(\theta_0 - \theta^*)^\top \widetilde{V}_{\theta^*}(\theta_0 - \theta^*)}\,.$$

To conclude $\widetilde{\alpha}^* \leq 1$ is sufficient to prove that $\mathrm{tr}(\widetilde{V}_{\theta_0} V_{\theta_0}^{-1}) \geq p$.

Recall that the 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

$$\frac{1}{p}\mathrm{tr}\left(\widetilde{V}_{\theta_0} V_{\theta_0}^{-1}\right) \geq \left(|\widetilde{V}_{\theta_0} V_{\theta_0}^{-1}|\right)^{1/p}. \tag{64}$$

Then since $|\widetilde{V}_{\theta_0} V_{\theta_0}^{-1}| = |\widetilde{V}_{\theta_0}||V_{\theta_0}^{-1}| = |\widetilde{V}_{\theta_0}||V_{\theta_0}|^{-1}$, it will be sufficient to prove that $|\widetilde{V}_{\theta_0}| \geq |V_{\theta_0}|$, where $\widetilde{V}_{\theta_0} = \mathrm{diag}(V_{\theta_0})$. Because V_{θ_0} is a positive definite matrix, this is exactly Hadamard's inequality (see Theorem 7.8.1 in Horn and Johnson (2012)).

Finally, if $\theta_0 \neq \theta^*$, it follows that $\varepsilon(\theta_0 - \theta^*)^\top V_{\theta^*}(\theta_0 - \theta^*) > 0$, which implies that $\alpha^* < 1$. To conclude $\widetilde{\alpha}^* < 1$, note that all the diagonal terms of $\widetilde{V}_{\theta^*}$ are strictly positive since $\widetilde{V}_{\theta^*} = \mathrm{diag}(V_{\theta^*})$ and V_{θ^*} is a positive definite matrix. Otherwise, by Hadamard's inequality we will have $|V_{\theta^*}| = 0$, which is not possible. This implies $\varepsilon(\theta_0 - \theta^*)^\top \widetilde{V}_{\theta^*}(\theta_0 - \theta^*) > 0$.

Appendix B. Technical Lemmas

Lemma 4 *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,*

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

where we have defined the function

$$f_n(g,h) = \Big\{1 - \frac{\phi_n(h)}{\psi_n(h)}\,\frac{\psi_n(g)}{\phi_n(g)}\Big\}^+. \tag{65}$$

Proof. First, denote $a_n = \big\{\int_K \psi_n(g)\, dg\big\}^{-1}$ and $b_n = \big\{\int_K \phi_n(g)\, dg\big\}^{-1}$. 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

$$\begin{aligned} d_{\mathrm{TV}}\,(\psi_n,\, \phi_n) &= \frac{1}{2}\int |\psi_n(h) - \phi_n(h)|\, dh \\ &= \int_{\mathbb{R}^p} \Big\{\psi_n(h) - \phi_n(h)\Big\}^+ dh + \frac{1}{2}\int_{\mathbb{R}^p} (\psi_n(h) - \phi_n(h))\, dh \\ &= \int_{\mathbb{R}^p} \Big\{\psi_n(h) - \phi_n(h)\Big\}^+ dh \\ &= \int_K \Big\{\psi_n(h) - \phi_n(h)\Big\}^+ dh + \int_{\mathbb{R}^p\setminus K} \Big\{\psi_n(h) - \phi_n(h)\Big\}^+ dh\,. \end{aligned} \tag{66}$$

Now, since ψ_n, ϕ_n are non-negative on all of $\mathbb{R}^p$, it follows that $\{\psi_n(h) - \phi_n(h)\}^+ \leq \psi_n(h) + \phi_n(h)$. This provides a bound for the second term on the right side of (66):

$$\int_{\mathbb{R}^p\setminus K} \Big\{\psi_n(h) - \phi_n(h)\Big\}^+ dh \leq \int_{\mathbb{R}^p\setminus K} \psi_n(h)\, dh \,+\, \int_{\mathbb{R}^p\setminus K} \phi_n(h)\, dh\,. \tag{67}$$

To complete the proof we show that the first term on the right side of (66) is upper bounded by $\sup_{g,h\,\in K} f_n(g,h)$. 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 (66) as follows,

$$\begin{aligned}\int_K \Big\{\psi_n(h) - \phi_n(h)\Big\}^+ dh &= \int_K \left\{1 - \frac{\phi_n(h)}{\psi_n(h)}\right\}^+ \psi_n(h)\, dh \\ &= \int_K \left\{1 - \frac{a_n}{b_n}\int_K \frac{\phi_n(h)}{\psi_n(h)}\,\frac{\psi_n(g)}{\phi_n(g)}\, b_n\phi_n(g)\, dg\right\}^+ \psi_n(h)\, dh\,.\end{aligned} \tag{68}$$

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 (65), we find

$$\begin{aligned}&\int_K \left(\left\{1 - \frac{a_n}{b_n}\int_K \frac{\phi_n(h)}{\psi_n(h)}\,\frac{\psi_n(g)}{\phi_n(g)}\, b_n\phi_n(g)\, dg\right\}^+\right)\psi_n(h)\, dh \\ &\le \int_K \left(\int_K \left\{1 - \frac{a_n}{b_n}\,\frac{\phi_n(h)}{\psi_n(h)}\,\frac{\psi_n(g)}{\phi_n(g)}\right\}^+ b_n\phi_n(g)\, dg\right)\psi_n(h)\, dh \le \int_K\int_K f_n(g,h)\, b_n\phi_n(g)\,\psi_n(h)\, dg\, dh\,,\end{aligned} \tag{69}$$

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,

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

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

Lemma 5 *Assume there exists 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^* and that Assumption 1 holds. For any $\eta, \epsilon > 0$, there exists a sequence $r_n \to +\infty$ and an integer $N(\eta,\epsilon) > 0$, such that for all $n > N(\eta,\epsilon)$, with $f_n(g,h)$ defined in (45),*

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

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 (75).

Step 1: First notice that for any $r > 0$, there exists an integer $N_0(r) := \lceil 4r^2/\delta^2\rceil > 0$ such that $\theta^* + h/\sqrt{n} \in B_{\theta^*}(\delta)$ whenever $h \in \overline{B}_{\mathbf{0}}(r)$ and $n \ge N_0(r)$. To see that this is true, notice

that if $||h|| \leq r$ and $n \geq 4r^2/\delta^2$, then $||h||/\sqrt{n} \leq \delta/2 < \delta$. This will ensure that the function $f_n(g,h)$ is well-defined whenever $g, h \in \overline{B}_{\mathbf{0}}(r)$.

Recall the definition of the α-posterior $\pi_{n,\alpha}$ in (7), as well as that of the scaled densities $\pi^{LAN}_{n,\alpha}(h \,|\, X^n) = n^{-1/2}\pi_{n,\alpha}(\theta^* + h/\sqrt{n} \,|\, X^n)$ and $\pi_n(h) = n^{-1/2}\pi(\theta^* + h/\sqrt{n})$ introduced in the proof of Theorem 1. Then we see that for any two sequences $\{h_n\}, \{g_n\}$ in $\overline{B}_{\mathbf{0}}(r)$ and $n > N_0(r)$,

$$\begin{aligned}\frac{\pi^{LAN}_{n,\alpha}(g_n \,|\, X^n)}{\pi^{LAN}_{n,\alpha}(h_n \,|\, X^n)} = \frac{\pi_{n,\alpha}\left(\theta^* + \frac{g_n}{\sqrt{n}} \,\middle|\, X^n\right)}{\pi_{n,\alpha}\left(\theta^* + \frac{h_n}{\sqrt{n}} \,\middle|\, X^n\right)} &= \frac{\left[f_n\left(X^n \,\middle|\, \theta^* + \frac{g_n}{\sqrt{n}}\right)\right]^\alpha \pi\left(\theta^* + \frac{g_n}{\sqrt{n}}\right)}{\left[f_n\left(X^n \,\middle|\, \theta^* + \frac{h_n}{\sqrt{n}}\right)\right]^\alpha \pi\left(\theta^* + \frac{h_n}{\sqrt{n}}\right)} \\ &= \frac{\left[f_n\left(X^n \,\middle|\, \theta^* + \frac{g_n}{\sqrt{n}}\right)\right]^\alpha \pi_n\left(g_n\right)}{\left[f_n\left(X^n \,\middle|\, \theta^* + \frac{h_n}{\sqrt{n}}\right)\right]^\alpha \pi_n\left(h_n\right)}. \end{aligned} \tag{70}$$

Thus by the definition in (45), with the notation $s_n(h_n) = [f_n(X^n \,|\, \theta^* + h_n/\sqrt{n})/f_n(X^n \,|\, \theta^*)]^\alpha$,

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

Recall that $\phi_n(h_n) = \phi(h_n \,|\, \Delta_{n,\theta^*}, V_{\theta^*}^{-1}/\alpha)$ and notice that since $||h_n||/\sqrt{n} < \delta$ as discussed at the beginning of the step 1 proof, the above is well-defined (i.e. $\pi^{LAN}_{n,\alpha}(h_n \,|\, X^n)$ is positive).

Next, for any sequence $h_n \in \overline{B}_{\mathbf{0}}(r)$, Assumption 1 implies

$$\log(s_n(h_n)) = \alpha \log\left(\frac{f_n(X^n \,|\, \theta^* + \frac{h_n}{\sqrt{n}})}{f_n(X^n \,|\, \theta^*)}\right) = h_n^\top \alpha V_{\theta^*} \Delta_{n,\theta^*} - \frac{1}{2} h_n^\top \alpha V_{\theta^*} h_n + o_{f_{0,n}}(1), \tag{71}$$

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

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

hence

$$\log\left(\frac{s_n(h_n)}{\phi_n(h_n)}\right) = o_{f_{0,n}}(1) + \frac{p}{2}\log(2\pi) - \frac{1}{2}\log(\det(\alpha V_{\theta^*})) + \frac{1}{2}\Delta_{n,\theta^*}^\top \alpha V_{\theta^*} \Delta_{n,\theta^*}$$

Now, for any sequence $g_n \in \overline{B}_{\mathbf{0}}(r)$, define

$$b_n(g_n, h_n) \equiv \frac{\phi_n(h_n) s_n(g_n) \pi_n(g_n)}{\phi_n(g_n) s_n(h_n) \pi_n(h_n)}. \tag{72}$$

Then we conclude

$$\log(b_n(g_n, h_n)) = \log\left(\frac{\phi_n(h_n)s_n(g_n)\pi_n(g_n)}{\phi_n(g_n)s_n(h_n)\pi_n(h_n)}\right) = o_{f_{0,n}}(1)\,, \tag{73}$$

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 arbitrary sequences in $B_{\mathbf{0}}(r)$, the result in (73) is equivalent to saying that for any fixed r, there exists an integer $\widetilde{N}_0(r, \epsilon, \eta)$, such that for $n > \max\{\widetilde{N}_0(r, \epsilon, \eta), N_0(r)\}$:

$$P_{f_{0,n}}\left(\sup_{g_n, h_n \in \overline{B}_{\mathbf{0}}(r)} |\log\left(b_n\left(g_n, h_n\right)\right)| > \eta\right) \leq \epsilon\,. \tag{74}$$

Next, notice that

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

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 (74), for all $n > \max\{\widetilde{N}_0(r, \eta, \epsilon), N_0(r)\}$,

$$\begin{aligned} P_{f_{0,n}}\left(\sup_{g,h \in \overline{B}_{\mathbf{0}}(r)} f_n(g, h) > \eta\right) &\leq P_{f_{0,n}}\left(\sup_{g_n, h_n \in \overline{B}_{\mathbf{0}}(r)} f_n(g, h) > \eta\right) \\ &\leq P_{f_{0,n}}\left(\sup_{g_n, h_n \in \overline{B}_{\mathbf{0}}(r)} |\log(b_n(g_n, h_n))| > \eta\right) \leq \epsilon\,. \end{aligned} \tag{75}$$

Step 2: Let $N^*(\epsilon, \eta) = \max\{\widetilde{N}_0(1, \epsilon, \eta), 4/\delta^2\}$ for $\widetilde{N}_0(r, \epsilon, \eta)$ defined just above (74) and for all $n > N^*(\epsilon, \eta)$ let

$$r_n = \max\{\, r \in \mathbb{R} \,|\, r \leq \delta\sqrt{n}/2 \text{ and } n > \widetilde{N}_0(r, \epsilon, \eta)\,\}\,.$$

Notice this is well defined since for any $n > N^*(\epsilon, \eta)$ the choice $r = 1$ is valid as $\frac{\delta}{2}(\sqrt{n}) > 1$ and $n > \widetilde{N}_0(1, \epsilon, \eta)$ by the definition of $N^*(\epsilon, \eta)$. For $n \leq N^*(\epsilon, \eta)$ we can define r_n arbitrarily, e.g. $r_n = 1$. First notice $r_n \to \infty$. Indeed,

$$r_n = \max\{\, r \in \mathbb{R} \,|\, r \leq \delta\sqrt{n}/2 \text{ and } n > \widetilde{N}_0(r, \epsilon, \eta)\,\} \leq \min\{\delta\sqrt{n}/2,\, \max\{r \in \mathbb{R} \,|\, n > \widetilde{N}_0(r, \epsilon, \eta)\}\}\,.$$

Clearly $\delta\sqrt{n}/2 \to \infty$, so $r_n \to \infty$ if $\max\{r \in \mathbb{R} \mid n > \widetilde{N}_0(r, \epsilon, \eta)\} \to \infty$. This is true, since if it were not, there must be some r_{max} such that $\widetilde{N}_0(r, \epsilon, \eta) = \infty$ for $r > r_{max}$. However

Assumption 1 holds for any compact set $K \in \mathbb{R}^p$, meaning for any ball $\overline{B}_{\mathbf{0}}(r)$ with $r < \infty$, hence $\widetilde{N}_0(r, \epsilon, \eta) < \infty$ for any $r < \infty$.

Finally, following the proof in step 1, we show that for all $n > N^*(\epsilon, \eta)$,

$$\mathbb{P}_{f_{0,n}}\Big(\sup_{g,h \in \overline{B}_{\mathbf{0}}(r_n)} f_n(g,h) > \eta \Big) \leq \epsilon \,. \tag{76}$$

First, $\theta^* + h_n/\sqrt{n} \in B_{\theta^*}(\delta)$ whenever $h_n \in \overline{B}_{\mathbf{0}}(r_n)$ since $||h_n||/\sqrt{n} \leq r_n/\sqrt{n} \leq \delta/2$. This guarantees $f_n(g_n, h_n)$ is well-defined, as discussed in step 1. Then for all $n > N^*(\epsilon, \eta)$, by the r_n definition, $n > \widetilde{N}_0(r_n, \epsilon, \eta)$, hence by the work in (74)-(75), the bound in (76) holds.

Lemma 6 *Suppose Assumptions 1 and 2 hold. 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

$$\begin{aligned}
\mathcal{K}\left(\phi(\cdot \,|\, \mu_n,\, \Sigma_n) \,||\, \pi_{n,\alpha}(\cdot \,|\, X^n)\right) &= \int \phi(\theta \,|\, \mu_n,\, \Sigma_n) \log\left(\frac{\phi(\theta \,|\, \mu_n,\, \Sigma_n)}{\pi_{n,\alpha}(\theta \,|\, X^n)}\right) d\theta \\
&= \int \phi(h \,|\, \overline{\mu}_n,\, n\Sigma_n) \log\left(\frac{n^{p/2}\phi(h \,|\, \overline{\mu}_n,\, n\Sigma_n)}{\pi_{n,\alpha}(\theta^* + h/\sqrt{n} \,|\, X^n)}\right) dh \\
&= I_1 + I_2 + I_3 \,,
\end{aligned} \tag{77}$$

where

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

We will compute the three terms separately and show that their sum gives the desired result. The first term I_1 is the negative of the entropy of a Gaussian distribution with mean $\overline{\mu}_n$ and covariance-matrix $n\Sigma_n$. A direct computation of this quantity gives

$$I_1 = -\frac{p}{2} - \frac{p}{2}\log(2\pi) - \frac{1}{2}\log|n\Sigma_n| \,. \tag{78}$$

To compute the second term I_2, notice that

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

hence

$$\begin{aligned} I_2 = &-\alpha \int \phi(h \,|\, \overline{\mu}_n, \, n\Sigma_n) \log \left(\frac{f_n(X^n \,|\, \theta^* + h/\sqrt{n})}{f_n(X^n \,|\, \theta^*)} \right) dh \\ &- \int \phi(h \,|\, \overline{\mu}_n, \, n\Sigma_n) \log \left(\frac{\pi(\theta^* + h/\sqrt{n})}{\pi(\theta^*)} \right) dh \,, \end{aligned} \tag{79}$$

and we consider the two terms on the right side of (79) separately. First, notice that by Assumption 2 the second term is $o_{f_{0,n}}(1)$. For the first term, consider Assumption 1, and we find

$$\begin{aligned} &- \alpha \int \phi(h \,|\, \overline{\mu}_n, \, n\Sigma_n) \log \left(\frac{f_n(X^n \,|\, \theta^* + h/\sqrt{n})}{f_n(X^n \,|\, \theta^*)} \right) dh \\ &= -\alpha \int \phi(h \,|\, \overline{\mu}_n, \, n\Sigma_n) \, \left(h^\top V_{\theta^*} \, \Delta_{n,\theta^*} - \frac{1}{2} \, h^\top V_{\theta^*} \, h + R_n(h) \right) dh \\ &= -\alpha \int \phi(h \,|\, \overline{\mu}_n, \, n\Sigma_n) \, \left(h^\top V_{\theta^*} \, \Delta_{n,\theta^*} - \frac{1}{2} \, h^\top V_{\theta^*} \, h \right) dh + o_{f_{0,n}}(1) \,. \end{aligned} \tag{80}$$

Plugging this into (79) and solving the integral, we find

$$\begin{aligned} I_2 &= -\alpha \int \phi(h \,|\, \overline{\mu}_n, \, n\Sigma_n) \, \left(h^\top V_{\theta^*} \, \Delta_{n,\theta^*} - \frac{1}{2} \, h^\top V_{\theta^*} \, h \right) dh + o_{f_{0,n}}(1) \\ &= -\alpha \, \overline{\mu}_n^\top \, V_{\theta^*} \, \Delta_{n,\theta^*} + \frac{\alpha}{2} \left(\overline{\mu}_n^\top \, V_{\theta^*} \, \overline{\mu}_n + \mathrm{tr}(n\Sigma_n V_{\theta^*}) \right) + o_{f_{0,n}}(1) \\ &= \frac{\alpha}{2} \left((\Delta_{n,\theta^*} - \overline{\mu}_n)^\top \, V_{\theta^*} \, (\Delta_{n,\theta^*} - \overline{\mu}_n) - \Delta_{n,\theta^*}^\top \, V_{\theta^*} \, \Delta_{n,\theta^*} + \mathrm{tr}(n\Sigma_n V_{\theta^*}) \right) + o_{f_{0,n}}(1) \,. \end{aligned} \tag{81}$$

Next, replacing $\Delta_{n,\theta^*} = \sqrt{n}(\widehat{\theta}_{\mathrm{ML}\text{-}\mathcal{F}_n} - \theta^*)$ and $\overline{\mu}_n = \sqrt{n}(\mu_n - \theta^*)$ in (81) yields

$$\begin{aligned} I_2 = &\frac{1}{2} (\widehat{\theta}_{\mathrm{ML}\text{-}\mathcal{F}_n} - \mu_n)^\top (\alpha n V_{\theta^*}) (\widehat{\theta}_{\mathrm{ML}\text{-}\mathcal{F}_n} - \mu_n) + \frac{1}{2} \mathrm{tr}(\Sigma_n \alpha n V_{\theta^*}) \\ &- \frac{1}{2} (\widehat{\theta}_{\mathrm{ML}\text{-}\mathcal{F}_n} - \theta^*)^\top (\alpha n V_{\theta^*}) (\widehat{\theta}_{\mathrm{ML}\text{-}\mathcal{F}_n} - \theta^*) + o_{f_{0,n}}(1) \,. \end{aligned} \tag{82}$$

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

$$\log \left(\frac{\pi_{n,\alpha}(\theta^* \,|\, X^n)}{\phi(\theta^* \,|\, \widehat{\theta}_{\mathrm{ML}\text{-}\mathcal{F}_n}, \, V_{\theta^*}^{-1}/(\alpha n))} \right) = o_{f_{0,n}}(1) \,, \tag{83}$$

and we will prove this in what follows. Result (83) implies that

$$\begin{aligned} I_3 &= -\log\Big(n^{-p/2}\phi(\theta^* \,|\, \widehat{\theta}_{\text{ML-}\mathcal{F}_n},\, V_{\theta^*}^{-1}/(\alpha n))\Big) + o_{f_{0,n}}(1) \\ &= \frac{p}{2}\log(2\pi) - \frac{1}{2}\log|\alpha V| + \frac{1}{2}(\widehat{\theta}_{\text{ML-}\mathcal{F}_n} - \theta^*)^\top(\alpha n V_{\theta^*})(\widehat{\theta}_{\text{ML-}\mathcal{F}_n} - \theta^*) + o_{f_{0,n}}(1)\,. \end{aligned} \tag{84}$$

Finally, combining (77), (82) and (84) we conclude that

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

where the last equality follows from (54).

It therefore remains to verify (83) in order to complete the proof. We first notice that the change of variables $h \equiv \sqrt{n}(\theta - \theta^*)$ leads to

$$\frac{\pi_{n,\alpha}(\theta^* \,|\, X^n)}{\phi(\theta^* \,|\, \widehat{\theta}_{\text{ML-}\mathcal{F}_n},\, V_{\theta^*}^{-1}/(\alpha n))} = \frac{\pi_{n,\alpha}^{LAN}(0 \,|\, X^n)}{\phi_n(0)}\,, \tag{85}$$

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

$$b_n(g,h) \equiv \left(\frac{\pi_{n,\alpha}^{LAN}(h \,|\, X^n)}{\phi_n(h)}\right)\left(\frac{\phi_n(g)}{\pi_{n,\alpha}^{LAN}(g \,|\, X^n)}\right)\,,$$

and note that it is the same as (72). Therefore, from using (74), we can guarantee the existence of r_n such that for any $\eta > 0$,

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

This implies that $\sup_{g,h\in\overline{B}_0(r_n)} b_n(g,h) = 1 + o_{f_{0,n}}(1)$. Define $c_n = (\int_{\overline{B}_0(r_n)} \pi_{n,\alpha}^{LAN}(h \,|\, X^n)dh)^{-1}$ and $d_n = (\int_{\overline{B}_0(r_n)} \phi_n(h)dh)^{-1}$. Using this notation, we have that (85) is equal to

$$\frac{d_n}{c_n}\int_{\overline{B}_0(r_n)}\left(\frac{\pi_{n,\alpha}^{LAN}(0 \,|\, X^n)}{\phi_n(0)}\right)\left(\frac{\phi_n(g)}{\pi_{n,\alpha}^{LAN}(g \,|\, X^n)}\right)\, c_n\, \pi_{n,\alpha}^{LAN}(g \,|\, X^n)\, dg\,,$$

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 (10) 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

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

In a similar way, we can conclude

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

Using (85), and the two inequalities above, we can conclude (83).

Lemma 7 *Suppose the following condition holds*

$$\Delta \equiv (\theta_0 - \theta^*)^\top V_{\theta^*}(\theta_0 - \theta^*)/\|\theta_0 - \theta^*\|^2 > \widetilde{\Delta} \equiv (\theta_0 - \theta^*)^\top \widetilde{V}_{\theta^*}(\theta_0 - \theta^*)/\|\theta_0 - \theta^*\|^2 . \quad (86)$$

Then, there exists a threshold $t^* = t^*(\theta_0, \theta^*, V_{\theta^*}, \widetilde{V}_{\theta^*}, V_{\theta_0}, \widetilde{V}_{\theta_0})$ *such that Equation* (30) *is strictly larger than Equation* (31) *if and only if* $\varepsilon\|\theta_0 - \theta^*\|^2 > t^*$.

Proof For each $t = \varepsilon\|\theta_0 - \theta^*\|^2 \geq 0$, denote the difference of equations (30) and (31) by

$$\Psi(t) \equiv\ p \log(p + t\Delta) - p \log\left(\mathrm{tr}\left(\widetilde{V}_{\theta_0} V^{-1}_{\theta_0}\right) + t\widetilde{\Delta}\right) - \log\left(\frac{|V_{\theta_0}|}{|\widetilde{V}_{\theta_0}|}\right) ,$$

where Δ and $\widetilde{\Delta}$ are defined in (86). Since $|\widetilde{V}_{\theta_0} V^{-1}_{\theta_0}| = |\widetilde{V}_{\theta_0}||V^{-1}_{\theta_0}| = |\widetilde{V}_{\theta_0}||V_{\theta_0}|^{-1}$, the Arithmetic Mean-Geometric Mean inequality (64) implies

$$\frac{1}{p}\mathrm{tr}\left(\widetilde{V}_{\theta_0} V^{-1}_{\theta_0}\right) \geq \left(|\widetilde{V}_{\theta_0} V^{-1}_{\theta_0}|\right)^{1/p} ,$$

Therefore

$$\Psi(0) = p \log(p) - p \log\left(\mathrm{tr}\left(\widetilde{V}_{\theta_0} V^{-1}_{\theta_0}\right)\right) - \log\left(\frac{|V_{\theta_0}|}{|\widetilde{V}_{\theta_0}|}\right) \leq 0 .$$

Further note that

$$\frac{\partial}{\partial t}\Psi(t) = \frac{p\Delta}{p + t\Delta} - \frac{p\widetilde{\Delta}}{\text{tr}\left(\widetilde{V}_{\theta_0} V_{\theta_0}^{-1}\right) + t\widetilde{\Delta}} \ .$$

Since $\text{tr}\left(\widetilde{V}_{\theta_0} V_{\theta_0}^{-1}\right) \geq p$, as shown at the end of the proof of Theorem 3 in Section A.4, and $\Delta > \widetilde{\Delta}$, by assumption (86), we can conclude that $\Psi(\cdot)$ is an increasing function on t since $\frac{\partial}{\partial t}\Psi(t) > 0$.

Finally, using again the assumption $\Delta > \widetilde{\Delta}$, we have for large t

$$p \log\left(p + t\Delta\right) - p \log\left(\text{tr}\left(\widetilde{V}_{\theta_0} V_{\theta_0}^{-1}\right) + t\widetilde{\Delta}\right) > 0 \ .$$

Since V_{θ_0} is a semi-definite matrix, Hadamard's inequality (see Theorem 7.8.1 in Horn and Johnson (2012)) implies $\log\left(|V_{\theta_0}||\widetilde{V}_{\theta_0}|^{-1}\right) \leq 0$. This implies that for large t, $\Psi(t) > 0$. Because $\Psi(\cdot)$ is an increasing function, it must exists a threshold $t^* = t^*(\theta_0, \theta^*, V_{\theta^*}, \widetilde{V}_{\theta^*}, V_{\theta_0}, \widetilde{V}_{\theta_0})$ such that $\Psi(t) > 0$ for any $t > t^*$. ■

References

Pierre Alquier. Non-exponentially weighted aggregation: regret bounds for unbounded loss functions. In *International Conference on Machine Learning*, pages 207–218. PMLR, 2021.

Pierre Alquier and James Ridgway. Concentration of tempered posteriors and of their variational approximations. *Annals of Statistics*, 48(3):1475–1497, 2020.

Andrew R Barron and Thomas M Cover. Minimum complexity density estimation. *IEEE transactions on information theory*, 37(4):1034–1054, 1991.

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

Anirban Bhattacharya, Debdeep Pati, and Yun Yang. Bayesian fractional posteriors. *Annals of Statistics*, 47(1):39–66, 2019.

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

Stéphane Bonhomme. Teams: Heterogeneity, sorting, and complementarity. *University of Chicago, Becker Friedman Institute for Economics Working Paper*, (2021-15), 2021.

Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-VAE. *arXiv preprint arXiv:1804.03599*, 2018.

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

Victor Chernozhukov and Han Hong. An mcmc approach to classical estimation. *Journal of Econometrics*, 115(2):293–346, 2003.

Bertrand S. Clarke. Asymptotic normality of the posterior in relative entropy. *IEEE Transactions on Information Theory*, 45(1):165–176, 1999.

Anirban DasGupta. *Asymptotic theory of statistics and probability*. Springer Science & Business Media, 2008.

Davide Ferrari and Yuhong Yang. Maximum L_q-likelihood method. *Annals of Statistics*, 38 (2):753–783, 2010.

Subhashis Ghosal and Aad Van der Vaart. *Fundamentals of nonparametric Bayesian inference*. Cambridge University Press, 2017.

Abhik Ghosh and Ayanendranath Basu. Robust bayes estimation using the density power divergence. *Annals of the Institute of Statistical Mathematics*, 68(2):413–437, 2016.

JK Ghosh and RV Ramamoorthi. Introduction: Why bayesian nonparametrics—an overview and summary. *Bayesian Nonparametrics*, pages 1–8, 2003.

Peter Grünwald. Safe learning: bridging the gap between bayes, mdl and statistical learning theory via empirical convexity. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 397–420. JMLR Workshop and Conference Proceedings, 2011.

Peter Grünwald. The safe Bayesian. In *International Conference on Algorithmic Learning Theory*, pages 169–183. Springer, 2012.

Peter Grünwald and Thijs Van Ommen. Inconsistency of bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Analysis*, 12(4):1069–1103, 2017.

Paul Gustafson. On measuring sensitivity to parametric model misspecification. *Journal of the Royal Statistical Society: Series B*, 63(1):81–94, 2001.

Lars Peter Hansen and Thomas J Sargent. Robust control and model uncertainty. *American Economic Review*, 91(2):60–66, 2001.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. β-VAE: Learning basic visual concepts with a constrained variational framework. In *Proceedings of the 5th International Conference on Learning Representations*, volume 2, page 6, 2017.

CC Holmes and SG Walker. Assigning a value to a power likelihood in a general bayesian model. *Biometrika*, 104(2):497–503, 2017.

Giles Hooker and Anand N Vidyashankar. Bayesian model robustness via disparities. *Test*, 23(3):556–584, 2014.

Roger A Horn and Charles R Johnson. *Matrix Analysis*. Cambridge University Press, 2012.

Chin-Wei Huang, Shawn Tan, Alexandre Lacoste, and Aaron Courville. Improving explorability in variational inference with annealed variational objectives. pages 9724—-9734, 2018.

Peter J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964. ISSN 00034851. URL `http://www.jstor.org/stable/2238020`.

Peter J Huber. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability*, volume 1, pages 221–233. Berkeley, CA: Univ. California Press, 1967.

Prateek Jaiswal, Vinayak Rao, and Harsha Honnappa. Asymptotic consistency of α-rényi-approximate posteriors. *Journal of Machine Learning Research*, 21(156):1–42, 2020.

Bas J.K Kleijn and Aad .W. Van der Vaart. The Bernstein-von-Mises theorem under misspecification. *Electronic Journal of Statistics*, 6:354–381, 2012.

J. Knoblauch, J. Jewson, and T. Damoulas. Generalized variational inference: Three arguments for deriving new posteriors. *arXiv:1904.02063*, 2019.

Jeremias Knoblauch. Frequentist consistency of generalized variational inference. *arXiv preprint arXiv:1912.04946*, 2019.

Gary Koop and Dimitris Korobilis. Variational bayes inference in high-dimensional time-varying parameter models. 2018.

Erich L Lehmann and George Casella. *Theory of point estimation*. Springer Science & Business Media, 2006.

Yong Li, Nianling Wang, and Jun Yu. Improved marginal likelihood estimation via power posteriors and importance sampling. *SMU Economics and Statistics Working Paper Series, Paper No. 16-2019*, 2019.

Yulong Lu, Andrew Stuart, and Hendrik Weber. Gaussian approximations for probability measures on r^d. *SIAM/ASA Journal on Uncertainty Quantification*, 5(1):1136–1165, 2017.

Fabio Maccheroni, Massimo Marinacci, and Aldo Rustichini. Ambiguity aversion, robustness, and the variational representation of preferences. *Econometrica*, 74(6):1447–1498, 2006.

Takuo Matsubara, Jeremias Knoblauch, François-Xavier Briol, Chris Oates, et al. Robust generalised bayesian inference for intractable likelihoods. *arXiv preprint arXiv:2104.07359*, 2021.

David A McAllester. Pac-bayesian stochastic model selection. *Machine Learning*, 51(1):5–21, 2003.

Angelo Mele and Lingjiong Zhu. Approximate variational estimation for a model of network formation. *Available at SSRN 2909829*, 2019.

Jeffrey W Miller. Asymptotic normality, concentration, and coverage of generalized posteriors. *Journal of Machine Learning Research*, 22(168):1–53, 2021.

Jeffrey W Miller and David B Dunson. Robust Bayesian inference via coarsening. *Journal of the American Statistical Association*, 114(527):1113–1125, 2019.

Ulrich K Müller. Risk of Bayesian inference in misspecified models, and the sandwich covariance matrix. *Econometrica*, 81(5):1805–1849, 2013.

Demian Pouzo, Zacharias Psaradakis, and Martin Sola. Maximum likelihood estimation in possibly misspecified dynamic models with time inhomogeneous markov regimes. *Available at SSRN 2887771*, 2016.

Tomasz Strzalecki. Axiomatic foundations of multiplier preferences. *Econometrica*, 79(1): 47–73, 2011.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Volodimir G Vovk. Aggregating strategies. *Proc. of Computational Learning Theory, 1990*, 1990.

Stephen Walker and Nils Lid Hjort. On bayesian consistency. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(4):811–821, 2001.

Chong Wang and David M Blei. A general method for robust bayesian modeling. *Bayesian Analysis*, 13(4):1163–1191, 2018.

Yixin Wang and David Blei. Variational bayes under model misspecification. In *Advances in Neural Information Processing Systems*, pages 13357–13367, 2019a.

Yixin Wang and David M. Blei. Frequentist consistency of variational Bayes. *Journal of the American Statistical Association*, 114(527):1147–1161, 2019b. doi: 10.1080/01621459.2018.1473776. URL `https://doi.org/10.1080/01621459.2018.1473776`.

Yixin Wang, Alp Kucukelbir, and David M. Blei. Robust probabilistic modeling with Bayesian data reweighting. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3646–3655. PMLR, 06–11 Aug 2017a.

Yixin Wang, Alp Kucukelbir, and David M Blei. Robust probabilistic modeling with bayesian data reweighting. In *International Conference on Machine Learning*, pages 3646–3655. PMLR, 2017b.

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

Yun Yang, Debdeep Pati, and Anirban Bhattacharya. α-variational inference with statistical guarantees. *Annals of Statistics*, 48(2):886–905, 2020.

Tong Zhang et al. From ε-entropy to kl-entropy: Analysis of minimum information complexity density estimation. *The Annals of Statistics*, 34(5):2180–2210, 2006.

Appendix C. Online supplementary material for the illustrative example

C.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 (10). In other words need to show that for every sequence $r_n \to \infty$, we have

$$\mathbb{E}_{f_{0,n}}\left[\mathbb{P}_{\pi_{n,\alpha}(\cdot\,|\,X^n)}\left(\|\sqrt{n}\,(\theta-\theta^*)\| > r_n\right)\right] \to 0\,. \tag{87}$$

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 (37) and (38). Therefore, by Markov's inequality and the normal distribution of the α-posterior we have

$$\mathbb{P}_{\pi_{n,\alpha}(\cdot\,|\,X^n)}\left(\|\theta-\theta^*\|^2 > r_n^2/n\right) \le \frac{n}{r_n^2}\mathbb{E}_{\pi_{n,\alpha}(\cdot\,|\,X^n)}\left[\|\theta-\theta^*\|^2\right] = \frac{n}{r_n^2}\left[\|\mu_{n,\alpha}-\theta^*\|^2 + \mathrm{tr}(\Sigma_{n,\alpha})\right]\,, \tag{88}$$

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 8 below 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 $\mathrm{tr}(n\Sigma_{n,\alpha})$ are bounded in $f_{0,n}$-probability. This means that for every $\epsilon > 0$, there exists an $M_\epsilon > 0$ such that

$$\mathbb{P}_{f_{0,n}}(A_\epsilon) \le \epsilon\,, \tag{89}$$

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

$$\begin{aligned}&\mathbb{E}_{f_{0,n}}\left[\mathbb{P}_{\pi_{n,\alpha}(\cdot\,|\,X^n)}\left(\|\sqrt{n}\,(\theta-\theta^*)\| > r_n\right)\right]\\ &\quad\le \mathbb{E}_{f_{0,n}}\left[\mathbb{P}_{\pi_{n,\alpha}(\cdot\,|\,X^n)}\left(\|\sqrt{n}\,(\theta-\theta^*)\| > r_n\right)1\{A_\epsilon^c\}\right] + \mathbb{E}_{f_{0,n}}\left[1\{A_\epsilon\}\right]\,,\end{aligned} \tag{90}$$

where $1\{A_\epsilon\}$ is the indicator function of the event A_ϵ. The first term in (90) can be bounded using (88) and the definition of A_ϵ, leading to

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

Using (89), we see that the second term in (90) 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, M_ϵ is constant, and $r_n\to\infty$. This verifies (87).

Lemma 8 *Let $(\mu_{n,\alpha},\Sigma_{n,\alpha})$ be the sequence defined in (37) and (38). 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 (37), 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

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

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

Using (38), we have that $n\Sigma_{n,\alpha}$ is equal to

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

and this converges in $f_{0,n}$-probability to $\frac{\sigma_u^2}{\alpha}(\mathbb{E}[W_i W_i^\top])^{-1} = \frac{1}{\alpha}V_{\theta^*}^{-1}$. 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 $(\mathbb{E}[W_i W_i^\top])^{-1}(\frac{1}{\alpha}\Sigma_\pi(\mu_\pi - \widehat{\theta}_{ML}))$. This implies that $n(\mu_{n,\alpha} - \widehat{\theta}_{ML})$ is bounded in $f_{0,n}$-probability.

C.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 (39) since both distribution are multivariate normal. Using (54), we find the KL divergence ot equal

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

Lemma 8 implies that the previous expression converges to 0 in $f_{0,n}$-probability. This means that the KL distance between $\pi_{n,\alpha}$ and the distribution in (39) goes to zero in $f_{0,n}$-probability.

C.3 BvM theorem does not hold if α goes to zero very quickly

Suppose that the sequence of α_n satisfies $n\alpha_n \to \alpha_0 > 0$. This implies that, in $f_{0,n}$-probability,

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

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 (39) 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 (39) 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

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

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.

C.4 Verification of Assumption 2

In our illustrative example:

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

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

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

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

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

$$\begin{aligned}&\int \phi(h \mid \sqrt{n}(\mu_n-\theta^*), n\Sigma_n)\left(-\frac{1}{2}\frac{h^\top}{\sqrt{n}}\frac{\Sigma_\pi}{\sigma_u^2}\frac{h}{\sqrt{n}} + \frac{h^\top}{\sqrt{n}}\frac{\Sigma_\pi}{\sigma_u^2}(\mu_\pi-\theta^*)\right)dh \\ &= -\frac{1}{2n}\overline{\mu}_n^\top\frac{\Sigma_\pi}{\sigma_u^2}\overline{\mu}_n - \frac{1}{2n}\mathrm{tr}\left(n\Sigma_n\frac{\Sigma_\pi}{\sigma_u^2}\right) + \frac{1}{\sqrt{n}}\overline{\mu}_n^\top\frac{\Sigma_\pi}{\sigma_u^2}(\mu_\pi-\theta^*)\,,\end{aligned} \tag{91}$$

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 (91) goes to zero in $f_{0,n}$-probability.

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

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

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.

C.5 KL distance for Theorem 2

Lemma 8 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, in $f_{0,n}$-probability,

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

This means that the Bernstein-von Mises Theorem holds for the variational approximation to the α-posterior in KL divergence.

C.6 Optimal α in the illustrative example

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

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

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

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

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 8 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 $(\mu_{n,1} - \mu_{n,\alpha})^\top \Sigma_{n,\alpha}^{-1}(\mu_{n,1} - \mu_{n,\alpha}) \to 0$ in $f_{0,n}$-probability. Since $n\epsilon_n \to \varepsilon$, we conclude that, in $f_{0,n}$-probability, $r_n(\alpha) \to r_\infty(\alpha)$ where $r_\infty(\alpha) \equiv \frac{1}{2}(\alpha p + \alpha\varepsilon(\theta^* - \theta_0)^\top V_{\theta^*}(\theta^* - \theta_0) - p\log(\alpha) - p)$.

Using the notation introduced in the proof of Theorem 3 in Section A.4, we have

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

where

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

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.7 Additional robustness of VI in the illustrative example

We prove that the misspecification model described in Section 5.5 satisfies the condition in (34), hence in this scenario α-optimized variational inference provides *additional* robustness beyond that given by optimized α-posteriors alone when ϵ_n is large.

In what follows, we drop the i subscript to save space. In particular, we denote $\rho := \mathbb{E}[W_{1i}W_{2i}] = \mathbb{E}[W_1W_2]$ and assume that $\gamma_0 > 0$, so there is indeed misspecification. Our proof uses general ρ and γ_0 and provides conditions on these terms under which (34) is satisfies. Then we will see that our model satisfies those conditions. Then, from Section 5.2, we have that

$$\theta^* - \theta_0 = \left(\mathbb{E}[WW^T]^{-1}\mathbb{E}[WZ]\right)\gamma_0, \qquad \text{and} \qquad V_{\theta^*} = \frac{1}{\sigma_u^2}\mathbb{E}[WW^T].$$

Therefore, to demonstrate that the condition from Equation (34) is true in some settings, we aim to characterize when the following is strictly positive.

$$\begin{aligned}
&(\theta^* - \theta_0)^T \left(V_{\theta^*} - \operatorname{diag}(V_{\theta^*})\right)(\theta^* - \theta_0) \\
&= \frac{\gamma_0^2}{\sigma_u^2}\mathbb{E}[ZW^T]\mathbb{E}[WW^T]^{-1}\left(\mathbb{E}[WW^T] - \operatorname{diag}(\mathbb{E}[WW^T])\right)\mathbb{E}[WW^T]^{-1}\mathbb{E}[WZ] \\
&\overset{(a)}{=} \frac{\gamma_0^2\rho}{\sigma_u^2(\mathbb{E}[W_1^2]\mathbb{E}[W_2^2] - \rho^2)^2} \times \mathbb{E}[ZW^T]\begin{bmatrix} -2\rho\mathbb{E}[W_2^2] & \rho^2 + \mathbb{E}[W_1^2]\mathbb{E}[W_2^2] \\ \rho^2 + \mathbb{E}[W_1^2]\mathbb{E}[W_2^2] & -2\rho\mathbb{E}[W_1^2] \end{bmatrix}\mathbb{E}[WZ] \\
&= \frac{2\gamma_0^2\rho\mathbb{E}[ZW_1]\mathbb{E}[ZW_2]}{\sigma_u^2(\mathbb{E}[W_1^2]\mathbb{E}[W_2^2] - \rho^2)^2} \times \left(\rho^2 + \mathbb{E}[W_1^2]\mathbb{E}[W_2^2] - \rho\frac{\mathbb{E}[ZW_1]}{\mathbb{E}[ZW_2]}\mathbb{E}[W_2^2] - \rho\frac{\mathbb{E}[ZW_2]}{\mathbb{E}[ZW_1]}\mathbb{E}[W_1^2]\right).
\end{aligned} \tag{92}$$

In step (a) we have used that

$$\begin{aligned}
&\mathbb{E}[WW^T]^{-1}\left(\mathbb{E}[WW^T] - \operatorname{diag}(\mathbb{E}[WW^T])\right)\mathbb{E}[WW^T]^{-1} \\
&\qquad = \frac{\rho}{(\mathbb{E}[W_1^2]\mathbb{E}[W_2^2] - \rho^2)^2}\begin{bmatrix} -2\rho\mathbb{E}[W_2^2] & \rho^2 + \mathbb{E}[W_1^2]\mathbb{E}[W_2^2] \\ \rho^2 + \mathbb{E}[W_1^2]\mathbb{E}[W_2^2] & -2\rho\mathbb{E}[W_1^2], \end{bmatrix},
\end{aligned}$$

which can be seen by noting that for a generic, symmetric 2×2 matrix we have that

$$\begin{aligned}
\begin{bmatrix} a & b \\ b & c \end{bmatrix}^{-1}\left(\begin{bmatrix} a & b \\ b & c \end{bmatrix} - \begin{bmatrix} a & 0 \\ 0 & c \end{bmatrix}\right)\begin{bmatrix} a & b \\ b & c \end{bmatrix}^{-1} &= \frac{b}{(ac - b^2)^2}\begin{bmatrix} c & -b \\ -b & a \end{bmatrix}\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}\begin{bmatrix} c & -b \\ -b & a \end{bmatrix} \\
&= \frac{b}{(ac - b^2)^2}\begin{bmatrix} -2bc & b^2 + ac \\ b^2 + ac & -2ab \end{bmatrix}.
\end{aligned} \tag{93}$$

Finally notice that, assuming $\rho \neq 0$, then (92) is strictly positive whenever the following is true:

$$\rho \mathbb{E}[ZW_1]\mathbb{E}[ZW_2] \times \left(\rho^2 + \mathbb{E}[W_1^2]\mathbb{E}[W_2^2] - \rho \frac{\mathbb{E}[ZW_1]}{\mathbb{E}[ZW_2]}\mathbb{E}[W_2^2] - \rho \frac{\mathbb{E}[ZW_2]}{\mathbb{E}[ZW_1]}\mathbb{E}[W_1^2] \right) > 0. \quad (94)$$

One sufficient condition for the above is true, for example, is when $\rho = 1/2$ and both $\mathbb{E}[ZW_1]$ and $\mathbb{E}[ZW_2]$ take the same sign, as long as the variance of the observed features is large enough:

$$\mathbb{E}[W_1^2] > \frac{\mathbb{E}[ZW_1]}{\mathbb{E}[ZW_2]} \qquad \text{and} \qquad \mathbb{E}[W_2^2] > \frac{\mathbb{E}[ZW_2]}{\mathbb{E}[ZW_1]}.$$

In our model, we have that $\mathbb{E}[W_1^2] = \mathbb{E}[W_2^2] = 1$ and $\mathbb{E}[ZW_1] = \mathbb{E}[ZW_2] = \rho/2$. Hence, if we plug these values into (94) using $\rho = 1/2$, we find

$$\begin{aligned} \rho \mathbb{E}[ZW_1]\mathbb{E}[ZW_2] \left[\rho^2 + \mathbb{E}[W_1^2]\mathbb{E}[W_2^2] - \rho \frac{\mathbb{E}[ZW_1]}{\mathbb{E}[ZW_2]}\mathbb{E}[W_2^2] - \rho \frac{\mathbb{E}[ZW_2]}{\mathbb{E}[ZW_1]}\mathbb{E}[W_1^2] \right] &= \frac{\rho^3}{4}(\rho^2 + 1 - 2\rho) \\ &= \frac{1}{2^7} > 0. \end{aligned}$$

C.8 Additional empirical results for the illustrative example

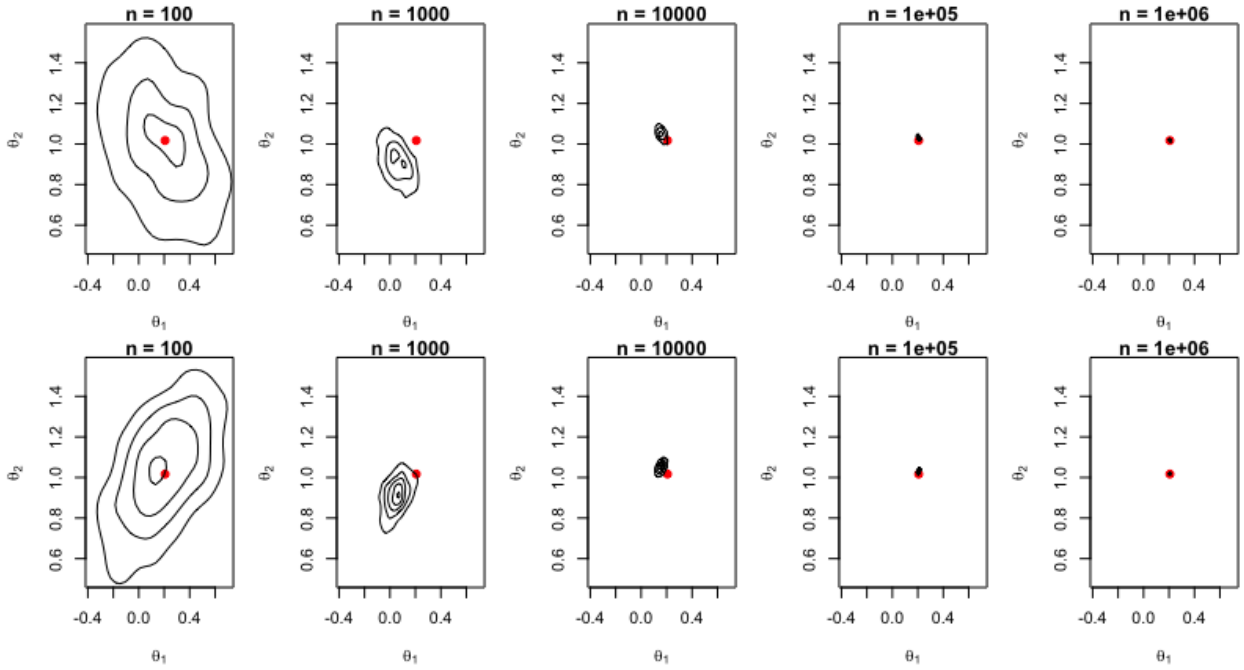
As discussed Section 5.5, we assume a Gaussian prior for (θ, γ), denoted π, i.e. $(\theta, \gamma) \sim \mathcal{N}(\mu_\pi, \sigma_u^2 \Sigma_\pi^{-1})$, where μ_π is the all-zeros vector and the covariance structure is chosen randomly from the Wishart ensemble. Using this prior, we can calculate the α-posterior using (37) and (38) and we prove in Appendix C.1 that the α-posterior concentrates at rate $\sqrt{n}$ around θ^* for any fixed α. This result is empirically verified in Figure 4. The top row of Figure 4 shows contour plots based one thousand samples from the α-posterior when $\alpha = 0.5$ and the sample size n grows. The red point is θ^*. As discussed in Section 5.3, the total variation distance between the α-posterior and the multivariate normal

$$\mathcal{N}\left(\widehat{\theta}_{ML}, \frac{\sigma_u^2}{\alpha n} \mathbb{E}[W_i W_i^\top]^{-1} \right), \quad (95)$$

converges in probability to zero and the limiting distribution in (95) also concentrates around θ^*. This can be seen from Figure 4 as well, where the top row plots sample contours of the α-posterior and the bottom row plots sample contours of the multivariate normal in (39). Both converge to θ^*, the red dot in the plots, as n grows.

In Figure 5, we empirically verify that $r_n(\alpha) - r_n^*(\alpha) \to 0$ (top row) and $\widetilde{r}_n(\alpha) - \widetilde{r}_n^*(\alpha) \to 0$ (bottom row) as $n \to \infty$ for $\alpha \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. In the plots, the x-axis is n while the y-axis is the difference and the α grows from $\alpha = 0.1$ in the leftmost column to $\alpha = 0.9$ in

Figure 4: Sample contour plots based one thousand draws from the α-posterior (top row) and its limiting Gaussian distribution given in (95) (bottom row) when $\alpha = 0.5$ for growing sample size n. The red point is θ^*.



the rightmost column. The grey line plots the average difference of fifty problem realizations and the red lines show one standard deviation above and below the average. We see that as α grows, the difference approaches 0 less quickly. Here we have chosen $\epsilon_n = 10/n$ so that $n\epsilon_n \to \epsilon = 10$.

Figure 5: Differences $r_n(\alpha) - r_n^*(\alpha)$ (top row) and $\widetilde{r}_n(\alpha) - \widetilde{r}_n^*(\alpha)$ plotted against n on the x-axis. From left to right α grows with $\alpha \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. The grey line plots the average difference of fifty problem realizations and the red lines show one standard deviation above and below the average with $\epsilon_n = 10/n$ so that $n\epsilon_n \to \epsilon = 10$.