Skew-symmetric approximations of posterior distributions

Francesco Pozza

Bocconi Institute for Data Science and Analytics

Joint work with: Daniele Durante and Botond Szabo

AHIDI2024-Verona
November 8, 2024

Introduction

- **Framework**: Bayesian parametric model

$$\pi_n(\theta) = \frac{p(y_1, \dots, y_n \mid \theta)\pi(\theta)}{m(y_1, \dots, y_n)}$$

  where $\theta \in \mathbb{R}^d$ and $\pi_n(\theta)$ is **intractable**

- Common to use of Gaussian (or symmetric) **deterministic** approximations of $\pi_n(\theta)$

  $\implies$ Gaussianity justified in asymptotic regimes by Bernstein–Von Mises type results (e.g., Van der Vaart, 2000)

- In non-asymptotic settings the posterior distribution often displays substantial asymmetries

Introduction

- Recent research proposes more flexible classes of asymmetric approximating densities

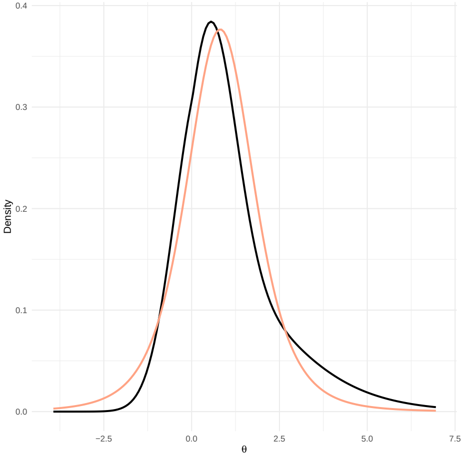
  $\Longrightarrow$ model specific solutions, higher computational complexity, fewer theoretical guarantees

- **Aim**: To derive class of asymmetric approximations that is:

  1. broadly applicable
  2. computationally efficient
  3. theoretically supported

Skew-symmetric approximations: derivation

Starting point: approximate the posterior distribution $\pi_n(\theta)$ with a generic density $f^*_{\tilde{\theta}}(\theta)$, symmetric about $\tilde{\theta}$.

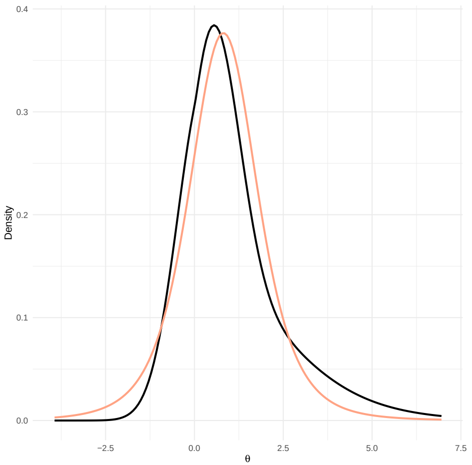


Skew-symmetric approximations: derivation

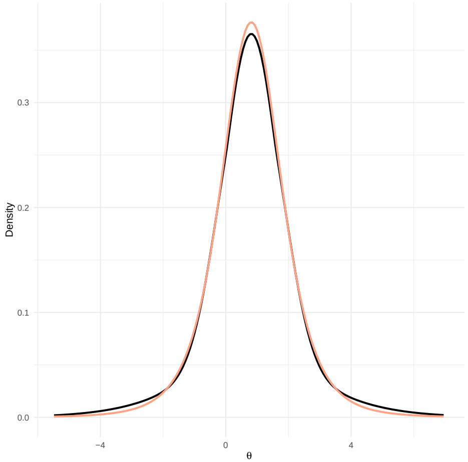
- If the posterior is asymmetric with respect to $\tilde{\theta}$ the quality of the approximation will be always sub-optimal
- Would $f^*_{\tilde{\theta}}(\theta)$ provide a better approximation of a **symmetrized** version of $\pi_n(\theta)$?
- Many different options but, probably, the most natural (see, e.g., Schuster, 1975) is

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

Skew-symmetric approximations: derivation



(a) **Target density** and **Approximating density**



(b) **Target density** and **Skew-symmetric approximating density**

Skew-symmetric approximations: derivation

- Clearly, we aim to approximate $\pi_n(\theta)$ not $\bar{\pi}_{n,\tilde{\theta}}(\theta)$
- **Key point**: $\pi_n(\theta)$ and $\bar{\pi}_{n,\tilde{\theta}}(\theta)$ are related by

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

  where

$$w^*_{\tilde{\theta}}(\theta - \tilde{\theta}) = \frac{\pi_n(\theta)}{\pi_n(\theta) + \pi_n(2\tilde{\theta} - \theta)}$$

  does **not depend** on the normalizing constant

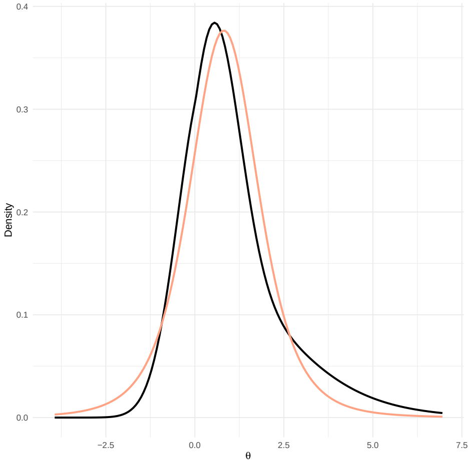
Skew-symmetric approximations

- In many Bayesian problems, $w_{\tilde{\theta}^*}(\cdot)$ is available in closed form
- This suggests the approximation

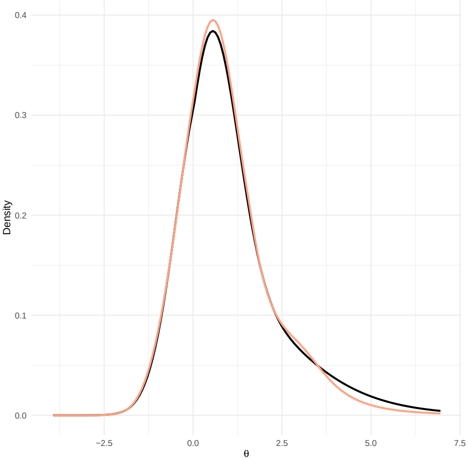
$$q^*_{\tilde{\theta}}(\theta) = 2 f^*_{\tilde{\theta}}(\theta) w^*_{\tilde{\theta}}(\theta - \tilde{\theta})$$

  which can be shown to be a proper **skew-symmetric** density (Azzalini and Capitanio, 2003)
- If simulating from $f^*_{\tilde{\theta}}(\theta)$ is easy then drawing a sample from $q^*_{\tilde{\theta}}(\theta)$ can be done at the same cost of evaluating $w^*_{\tilde{\theta}}(\theta - \tilde{\theta})$

Skew-symmetric approximations: derivation



(a) **Target density** and **Approximating density**



(b) **Target density** and **Skew-symmetric approximating density**

Skew-symmetric approximations: theory

Theorem (Finite–sample accuracy)

Let $\pi_n(\theta)$ be the posterior, $f^*_{\tilde{\theta}}(\theta)$ be an approximation symmetric about $\tilde{\theta} \in \Theta$ and $q^*_{\tilde{\theta}}(\theta) = 2f^*_{\tilde{\theta}}(\theta)w^*_{\tilde{\theta}}(\theta - \tilde{\theta})$. Then

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

for any $\tilde{\theta} \in \Theta$ and n, where $\bar{\pi}_{n,\tilde{\theta}}(\theta)$ is the symmetrized posterior and $\mathcal{D}$ is either the total variation distance or any α-divergence.

Asymptotic properties: when $f^*_{\tilde{\theta}}(\theta) = \phi_d(\theta; \tilde{\theta}, J^{-1}_{\tilde{\theta}})$ with $J_{\tilde{\theta}} = -(\partial^2/\partial\theta\partial\theta^\top)\log \pi_n(\theta)$,

$$\mathcal{D}_{\mathrm{TV}}[\pi_n(\theta) \mid\mid q^*_{\tilde{\theta}}(\theta)] = O_P(d^3/n)$$

up to a logarithmic term

Application: logistic regression

We compare Gaussian and skew-symmetric approximations on 3 Logistic regression models with Gaussian prior, i.e, $\pi_n(\theta) = \phi_d(\theta; 0, \sigma^2 \mathbb{I}_d) \prod_{i=1}^n p_i^{y_i}(1 - p_i)^{1-y_i}$, $y_i \in \{0, 1\}$, $p_i = 1/(1 + \exp(-x_i^\top \theta))$

1. **Glioma** $n = 839$ $d = 24$
2. **Musk** $n = 476$ $d = 167$
3. **Sonar** $n = 208$ $d = 1831$

Symmetric approximations: Gaussian Laplace (2nd order Taylor around posterior mode), Gaussian variational Bayes and Gaussian expectation propagation

Summary statistics: ratio between mean absolute error in estimating the posterior mean and median made by the Gaussian approximations and their skew-symmetric counterparts

Application: logistic regression

	MEDIAN.BIAS	BIAS
Glioma $n = 839$ $d = 24$		
LA/SKE-LA	2.60	2.40
GVB/SKE-GVB	1.59	1.57
EP/SKE-EP	5.81	1.68
Musk $n = 476$ $d = 167$		
LA/SKE-LA	1.20	1.20
GVB/SKE-GVB	1.09	1.09
EP/SKE-EP	1.17	1.01
Sonar $n = 208$ $d = 1831$		
LA/SKE-LA	1.13	1.13
EP/SKE-EP	1.06	1.10

Conclusions

- A general methodological framework for obtaining asymmetric approximations of the the posterior distribution is introduced
- The proposed methods provably perform better than standard symmetric approximations not only asymptotically but also in finite samples regimes
- **Ongoing work/ future directions**: Develop symmetric approximations directly targeting $\bar{\pi}_{n,\tilde{\theta}}(\theta)$

References

Azzalini, A. and A. Capitanio (2003). "Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 65.2, pp. 367–389.

Durante, D., F. Pozza, and B. Szabo (2024). "Skewed Bernstein-von Mises theorem and skew-modal approximations". In: *Annals of Statistics (forthcoming), arXiv preprint arXiv:2301.03038.*

Pozza, F., D. Durante, and B. Szabo (2024+). "Skew-symmetric approximations of posterior distributions". In: *arXiv preprint arXiv:2409.14167.*

Schuster, E. F. (1975). "Estimating the distribution function of a symmetric distribution". In: *Biometrika* 62.3, pp. 631–635.

Van der Vaart, A. W. (2000). *Asymptotic Statistics*. Vol. 3. Cambridge University Press.

Skew-symmetric distributions

Definition (skew-symmetric distribution (Azzalini and Capitanio, 2003))

A random variable $\theta \in \mathbb{R}^d$ is skew-symmetric if it has density

$$2p(\theta - \xi)w(\theta - \xi),$$

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

I.i.d samples from $2p(\theta - \xi)w(\theta - \xi)$:

1. $\theta_0 \sim p(\theta - \xi)$
2. $\theta = \theta_0$ with probability $w(\theta_0 - \xi)$ otherwise $\theta = 2\xi - \theta_0$

Skew-symmetric approximations: theory

Theorem (Optimality of the skewness–inducing factor)

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

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

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

Skew-symmetric approximations: theory

Lemma

Let $\pi_n(\theta)$ be the posterior distribution and denote with $f^*_{\tilde{\theta}}(\theta)$ an already–available approximation of $\pi_n(\theta)$ which is symmetric about the point $\tilde{\theta} \in \Theta$. Define the symmetrized posterior density about $\tilde{\theta}$ as $\bar{\pi}_{n,\tilde{\theta}}(\theta) = [\pi_n(\theta) + \pi_n(2\tilde{\theta} - \theta)]/2$ and let $q^*_{\tilde{\theta}}(\theta) = 2f^*_{\tilde{\theta}}(\theta)w^*_{\tilde{\theta}}(\theta - \tilde{\theta})$. Then

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

and

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

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

Skew-symmetric approximations: theory (Pozza et al., 2024+)

- The method improves any symmetric approximation. Natural to perturb routinely implemented approximations such as Laplace, Gaussian Expectation Propagation and Gaussian Variational Bayes.
- Laplace:
  - Mean $=$ posterior mode $\hat{\theta}$
  - Covariance matrix $= \hat{\Omega} = -(\tilde{\ell}^{(2)}_{\hat{\theta}})^{-1}$

gives the skew-symmetric approximation: $2\phi_d(\theta; \hat{\theta}, \hat{\Omega}) w^*_{\tilde{\theta}}(\theta - \tilde{\theta})$

$\implies$ Closely related to the asymptotic version given in Durante et al. (2024)

$$2\phi_d(\theta; \hat{\theta}, \hat{\Omega}) \Phi\left(\frac{\sqrt{2\pi}}{12} \tilde{\ell}_{\hat{\theta},stl} (\theta - \hat{\theta})_s (\theta - \hat{\theta})_t (\theta - \hat{\theta})_l \right)$$

(same asymptotic accuracy)

Efficient evaluation skewness-inducing factor

- $w_{\hat{\theta}}(\theta)$ requires two un-normalized posterior evaluations
- In many models, the log-likelihood has the form $\ell(\theta) = \sum_{i=1}^{n} g(x_i^\top \theta)$ where g is $O(1)$ and $x_i^\top \theta$ is $O(d)$

Algorithm: Efficient evaluation $w_{\hat{\theta}}(\theta)$

Require: $\eta_{i,\hat{\theta}} = x_i^\top \hat{\theta}$

For: $i = 1, \ldots, n$ **do**:

$\eta_i = x_i^\top (\theta - \hat{\theta})$

Return: $\sum_{i=1}^{n} g(\eta_{i,\hat{\theta}} + \eta_{i,\theta})$ and $\sum_{i=1}^{n} g(\eta_{i,\hat{\theta}} - \eta_{i,\theta})$

Application: logistic regression

	MEDIAN.BIAS	BIAS	MEDIAN.$\mu(\theta)$	MEAN.$\mu(\theta)$
Glioma $n = 839$ $d = 24$				
LA/SKE-LA	2.60	2.40	2.44	2.64
GVB/SKE-GVB	1.59	1.57	2.56	2.76
EP/SKE-EP	5.81	1.68	3.02	1.96
Musk $n = 476$ $d = 167$				
LA/SKE-LA	1.20	1.20	1.24	1.30
GVB/SKE-GVB	1.09	1.09	1.13	1.13
EP/SKE-EP	1.17	1.01	1.41	1.81
Sonar $n = 208$ $d = 1831$				
LA/SKE-LA	1.13	1.13	1.20	1.27
EP/SKE-EP	1.06	1.10	1.26	1.87