# Skewed Bernstein-von Mises theorem and online skew-modal approximations

Francesco Pozza

Capitanio Lecture - Skew 2026

Funded by the European Union

European Research Council
Established by the European Commission

# Introduction

This lecture is given in honor of Antonella Capitanio. Although I did not have the opportunity to know her personally, her work has had a fundamental impact on my research:

- Azzalini, A., & Capitanio, A. (1999). **Statistical applications of the multivariate skew normal distribution.** Journal of the Royal Statistical Society: Series B (Statistical Methodology), 61(3), 579–602.
- Azzalini, A., & Capitanio, A. (2003). **Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution.** Journal of the Royal Statistical Society: Series B (Statistical Methodology), 65(2), 367–389.
- Azzalini, A. & Capitanio, A. (2013). **The Skew-Normal and Related Families**. Cambridge University Press.

# Introduction

- Research on **asymmetric** families of distributions: motivated by the need to generalize the Gaussian distribution to model **non-Gaussian** data (Azzalini 1985; Azzalini and Dalla Valle 1996)

# Introduction

- Research on **asymmetric** families of distributions: motivated by the need to generalize the Gaussian distribution to model **non-Gaussian** data (Azzalini 1985; Azzalini and Dalla Valle 1996)
- Development of **flexible** families which remain **tractable** while capturing non-Gaussian features such as skewness, multimodality, heavy tails.. etc (Azzalini and Capitanio 2003; Ma and Genton 2004)

# Introduction

- Research on **asymmetric** families of distributions: motivated by the need to generalize the Gaussian distribution to model **non-Gaussian** data (Azzalini 1985; Azzalini and Dalla Valle 1996)
- Development of **flexible** families which remain **tractable** while capturing non-Gaussian features such as skewness, multimodality, heavy tails.. etc (Azzalini and Capitanio 2003; Ma and Genton 2004)
- This combination of flexibility and tractability is useful in **Bayesian computational statistics** to approximate intractable posterior distributions (Durante 2019; Onorati and Liseo 2022; Tan 2023)

# Introduction

- Research on **asymmetric** families of distributions: motivated by the need to generalize the Gaussian distribution to model **non-Gaussian** data (Azzalini 1985; Azzalini and Dalla Valle 1996)
- Development of **flexible** families which remain **tractable** while capturing non-Gaussian features such as skewness, multimodality, heavy tails.. etc (Azzalini and Capitanio 2003; Ma and Genton 2004)
- This combination of flexibility and tractability is useful in **Bayesian computational statistics** to approximate intractable posterior distributions (Durante 2019; Onorati and Liseo 2022; Tan 2023)

# Skewed Bernstein–von Mises theorem

# Framework and basic definitions

- Bayesian statistics: likelihood $p(y_1, \ldots, y_n \mid \theta)$, $\theta \in \mathbb{R}^d$, prior $\pi(\theta)$, **posterior distribution**

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

  where $m(y_1, \ldots, y_n) = \int p(y_1, \ldots, y_n \mid \theta)\pi(\theta)d\theta$

- Usually $\pi_n(\theta)$ is **intractable** and functionals of interest such as mean, variance, etc are not analytically available

# Framework and basic definitions

If $\pi_n$ is intractable two common approaches are:

- **Markov chain Monte Carlo** methods: approximate samples from the posterior are obtained by iteratively applying a Markov transition kernel
  - Arbitrarily accurate (under mild conditions)
  - Can be slow in complex models

# Framework and basic definitions

If $\pi_n$ is intractable two common approaches are:

- **Markov chain Monte Carlo** methods: approximate samples from the posterior are obtained by iteratively applying a Markov transition kernel
  - Arbitrarily accurate (under mild conditions)
  - Can be slow in complex models
- **Deterministic methods**: the posterior is approximated with a simpler distribution
  - Usually faster than MCMC
  - Not always easy to quantify their accuracy

# Framework and basic definitions

Deterministic approximations typically rely on symmetric densities, often taken to be **Gaussian**

# Framework and basic definitions

Deterministic approximations typically rely on symmetric densities, often taken to be **Gaussian**

$\Longrightarrow$ Gaussianity justified by Bernstein–Von Mises type results (Van der Vaart 2000)

## Theorem

*Let $\tilde{\theta}$ be the maximum at posteriori (MAP) and $N_d(\tilde{\theta}; J_{\tilde{\theta}}^{-1})$ the Gaussian Laplace approximation where $J_{\tilde{\theta}}$ is the observed information matrix at $\tilde{\theta}$. Then, in regular parametric models,*

$$\mathcal{D}_{\mathrm{TV}}(\pi_n, N_d(\tilde{\theta}; J_{\tilde{\theta}}^{-1})) = O_P(n^{-1/2})$$

*with $\mathcal{D}_{\mathrm{TV}}(\cdot, \cdot)$ denoting the total variation distance*

# Framework and basic definitions

- $\tilde{\theta}$ can be replaced with any efficient estimator (e.g. the MLE)
- Results can be generalized to different settings, e.g. high-dimensional models
- Asymptotically, the posterior is Gaussian **but**:

# Framework and basic definitions

- $\tilde{\theta}$ can be replaced with any efficient estimator (e.g. the MLE)
- Results can be generalized to different settings, e.g. high-dimensional models
- Asymptotically, the posterior is Gaussian **but**: with finite $n$ the posterior distribution can **present substantial departures from Gaussianity** such as skewness and heavy tails

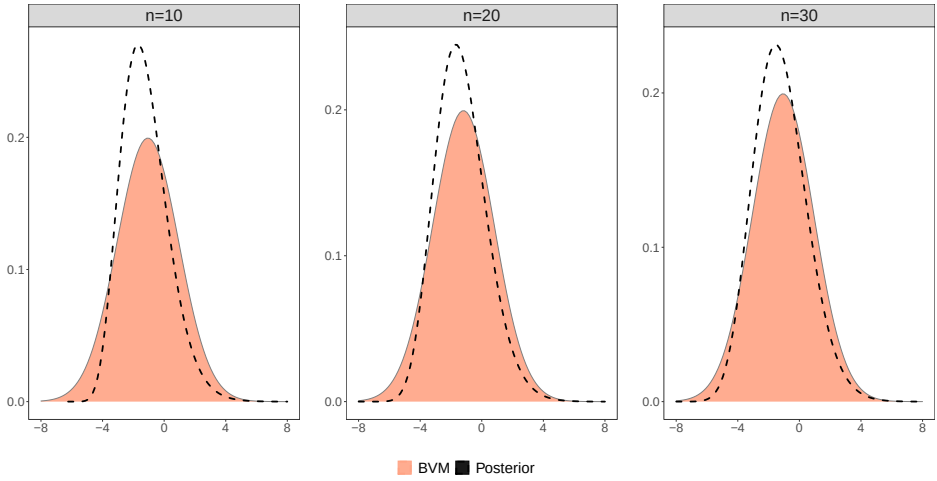



Figure: Posterior vs Gaussian BvM approximation for a 1-dimensional exponential model with different sample sizes

# Framework and basic definitions

- Recent research has proposed more flexible classes of approximating densities that incorporate skewness

  $\Longrightarrow$ these solutions are often model-specific and not always supported by rigorous theoretical guarantees

# Framework and basic definitions

- Recent research has proposed more flexible classes of approximating densities that incorporate skewness

  $\Longrightarrow$ these solutions are often model-specific and not always supported by rigorous theoretical guarantees

- **Aim**: develop a skewed version of the Bernstein–von Mises theorem that:

  1. theoretically justifies the need for asymmetric approximations
  2. clarifies which forms of asymmetry are relevant in Bayesian models

- Reference: Durante, D., Pozza, F., & Szabo, B. (2024). Skewed Bernstein–von Mises theorem and skew-modal approximations. *The Annals of Statistics*, 52(6), 2714-2737.

# Framework and basic definitions

**Definition (skew-symmetric distributions - Azzalini and Capitanio (2003))**

A random variable $\theta$ is skew-symmetric if its density takes the form

$$2p_{\tilde{\theta}}(\theta)F(\alpha(\theta-\tilde{\theta})),$$

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

**Sampling**: $\theta' \sim p_{\tilde{\theta}}(\theta)$ then $\theta = \theta'$ with probability $F(\alpha(\theta' - \tilde{\theta}))$ otherwise $\theta = 2\tilde{\theta} - \theta'$

# Skewed Bernstein–von Mises theorem

Reasons for the **inaccuracies** of the Laplace approximation:

- Let $d = 1$. The Laplace approximation is obtained from a second-order Taylor expansion around $\tilde{\theta}$:

$$\log \pi_n(\theta) = -\frac{1}{2} j_{\tilde{\theta}} (\theta - \tilde{\theta})^2 + r_n(\theta),$$

  where $r_n(\theta) = O_P(n^{-1/2})$.

# Skewed Bernstein–von Mises theorem

Reasons for the **inaccuracies** of the Laplace approximation:

- Let $d = 1$. The Laplace approximation is obtained from a second-order Taylor expansion around $\tilde{\theta}$:

$$\log \pi_n(\theta) = -\frac{1}{2} j_{\tilde{\theta}} (\theta - \tilde{\theta})^2 + r_n(\theta),$$

  where $r_n(\theta) = O_P(n^{-1/2})$.

- Using $\exp(x) = 1 + O(x)$,

$$\pi_n(\theta) \propto \exp\left( -\frac{1}{2} j_{\tilde{\theta}} (\theta - \tilde{\theta})^2 \right) \left\{ 1 + O_P(n^{-1/2}) \right\}.$$

# Skewed Bernstein–von Mises theorem

Reasons for the **inaccuracies** of the Laplace approximation:

- Let $d = 1$. The Laplace approximation is obtained from a second-order Taylor expansion around $\tilde{\theta}$:

$$\log \pi_n(\theta) = -\frac{1}{2} j_{\tilde{\theta}} (\theta - \tilde{\theta})^2 + r_n(\theta),$$

  where $r_n(\theta) = O_P(n^{-1/2})$.

- Using $\exp(x) = 1 + O(x)$,

$$\pi_n(\theta) \propto \exp\left( -\frac{1}{2} j_{\tilde{\theta}} (\theta - \tilde{\theta})^2 \right) \left\{ 1 + O_P(n^{-1/2}) \right\}.$$

- Hence,

$$\pi_n(\theta) \simeq \varphi\left( \theta; \tilde{\theta}, j_{\tilde{\theta}}^{-1} \right) + O_P(n^{-1/2}).$$

- The **leading error** term is

$$r_n(\theta) = \frac{1}{6}\ell^{(3)}_{\tilde{\theta}}(\theta - \tilde{\theta})^3 + O_P(n^{-1})$$

where $\ell^{(3)}_{\tilde{\theta}} = (\partial^3/\partial^3\theta)\log \pi_n(\theta) \mid_{\theta=\tilde{\theta}}$

- The **leading error** term is

$$r_n(\theta) = \frac{1}{6}\ell^{(3)}_{\tilde{\theta}}(\theta - \tilde{\theta})^3 + O_P(n^{-1})$$

  where $\ell^{(3)}_{\tilde{\theta}} = (\partial^3/\partial^3\theta) \log \pi_n(\theta) \mid_{\theta=\tilde{\theta}}$

- **Classical** solutions to improve accuracy (Johnson 1970)

$$\begin{aligned}\pi_n(\theta) \simeq & \exp\Big( -\frac{1}{2} j_{\tilde{\theta}}(\theta - \tilde{\theta})^2 + (1/6)\ell^{(3)}_{\tilde{\theta}}(\theta - \tilde{\theta})^3 + O_P(n^{-1})\Big) \\ \simeq & \varphi(\theta; \tilde{\theta}, j^{-1}_{\tilde{\theta}})\Big\{1 + \frac{1}{6}\ell^{(3)}_{\tilde{\theta}}(\theta - \tilde{\theta})^3\Big\} + O_P(n^{-1})\end{aligned}$$

- The **leading error** term is

$$r_n(\theta) = \frac{1}{6}\ell^{(3)}_{\tilde{\theta}}(\theta - \tilde{\theta})^3 + O_P(n^{-1})$$

  where $\ell^{(3)}_{\tilde{\theta}} = (\partial^3/\partial^3\theta) \log \pi_n(\theta) \mid_{\theta=\tilde{\theta}}$

- **Classical** solutions to improve accuracy (Johnson 1970)

$$\begin{aligned}\pi_n(\theta) \simeq & \exp\Big( -\frac{1}{2} j_{\tilde{\theta}}(\theta - \tilde{\theta})^2 + (1/6)\ell^{(3)}_{\tilde{\theta}}(\theta - \tilde{\theta})^3 + O_P(n^{-1})\Big) \\ \simeq & \varphi(\theta; \tilde{\theta}, j^{-1}_{\tilde{\theta}})\Big\{1 + \frac{1}{6}\ell^{(3)}_{\tilde{\theta}}(\theta - \tilde{\theta})^3\Big\} + O_P(n^{-1})\end{aligned}$$

- **Limitations of this approach**: $1 + \frac{1}{6}\ell^{(3)}_{\tilde{\theta}}(\theta - \tilde{\theta})^3$ can assume negative values
  $\Longrightarrow$ approximation **not** a proper density function

- $\ell^{(3)}_{\tilde{\theta}}(\theta - \tilde{\theta})^3$ **odd** in $(\theta - \tilde{\theta}) \implies$ associated to **skewness**
- **Key point**: assume that the term

$$\varphi(\theta; \tilde{\theta}, j^{-1}_{\tilde{\theta}}) \Big\{ 1 + \frac{1}{6} \ell^{(3)}_{\tilde{\theta}} (\theta - \tilde{\theta})^3 \Big\}$$

as a Taylor approximation of a **proper skew-symmetric** distribution

Recall that a skew-symmetric density takes the form

$$2p(\theta - \tilde{\theta})F(\alpha(\theta - \tilde{\theta}))$$

Let $F(\cdot) = \Phi(\cdot)$. Then

Recall that a skew-symmetric density takes the form

$$2p(\theta - \tilde{\theta})F(\alpha(\theta - \tilde{\theta}))$$

Let $F(\cdot) = \Phi(\cdot)$. Then

- $p(\theta - \tilde{\theta}) = \varphi(\theta; \tilde{\theta}, j_{\tilde{\theta}}^{-1})$
- $2\Phi(\alpha(\theta - \tilde{\theta})) = \left\{1 + \frac{1}{6}\ell_{\tilde{\theta}}^{(3)}(\theta - \tilde{\theta})^3\right\} + O_P(n^{-1})$

gives the **skew-modal approximation**:

$$\pi_n(\theta) \simeq 2\varphi(\theta; \tilde{\theta}, j_{\tilde{\theta}}^{-1})\Phi\Big(\frac{\sqrt{2\pi}}{12}\ell_{\tilde{\theta}}^{(3)}(\theta - \tilde{\theta})^3\Big) + O_P(n^{-1})$$

# Skew-modal approximation

- The extension to the **multivariate** case is straightforward

$$2\varphi_d(\theta;\tilde{\theta},J_{\tilde{\theta}}^{-1})\Phi\Big(\frac{\sqrt{2\pi}}{12}\sum_{s,t,l}^{d}\ell^{(3)}_{\tilde{\theta},stl}(\theta-\tilde{\theta})_s(\theta-\tilde{\theta})_t(\theta-\tilde{\theta})_l\Big)$$

  where $\ell^{(3)}_{\tilde{\theta},stl} = (\partial^3/\partial\theta_s\partial\theta_t\partial\theta_l)\log\pi_n(\theta)\mid_{\theta=\tilde{\theta}}$

- Since the symmetric component is Gaussian it is easy to obtain i.i.d samples

# Skew-modal approximation: theory

## Theorem (Skewed Bernstein–von Mises)

*Let*

$$p_{SM}(\theta) = 2\varphi_d(\theta; \tilde{\theta}, J_{\tilde{\theta}}^{-1})\Phi\Big(\frac{\sqrt{2\pi}}{12}\sum_{s,t,l}^{d}\ell^{(3)}_{\tilde{\theta},stl}(\theta-\tilde{\theta})_s(\theta-\tilde{\theta})_t(\theta-\tilde{\theta})_l\Big)$$

*Then, in regular parametric models,*

$$\mathcal{D}_{\mathrm{TV}}(\pi_n, p_{SM}) = O_P(n^{-1})$$

*with $\mathcal{D}_{\mathrm{TV}}(\cdot,\cdot)$ denoting the total variation distance.*

# Skew-modal approximation: theory

## Theorem (Skewed Bernstein–von Mises)

*Let*

$$p_{SM}(\theta) = 2\varphi_d(\theta; \tilde{\theta}, J_{\tilde{\theta}}^{-1})\Phi\Big(\frac{\sqrt{2\pi}}{12}\sum_{s,t,l}^{d} \ell^{(3)}_{\tilde{\theta},stl}(\theta-\tilde{\theta})_s(\theta-\tilde{\theta})_t(\theta-\tilde{\theta})_l\Big)$$

*Then, in regular parametric models,*

$$\mathcal{D}_{\mathrm{TV}}(\pi_n, p_{SM}) = O_P(n^{-1})$$

*with* $\mathcal{D}_{\mathrm{TV}}(\cdot,\cdot)$ *denoting the total variation distance.*

- improvement of order $O_P(n^{-1/2})$ over classical Gaussian BvM $\implies$ better accuracy

# Skew-modal approximation: theory

**Different** skew-symmetric approximations with the same accuracy can be obtained by considering Taylor expansions around **any efficient estimator** $\hat{\theta}$

$$p_{SS,\hat{\theta}}(\theta) = 2\,\varphi_d\left(\theta; \mu_{\hat{\theta}}, \Omega_{\hat{\theta}}^{-1}\right) \Phi\left(\frac{\sqrt{2\pi}}{12} \sum_{s,t,l=1}^{d} \ell^{(3)}_{\hat{\theta},stl} \left\{(\theta-\mu_{\hat{\theta}})_s(\theta-\mu_{\hat{\theta}})_t(\theta-\mu_{\hat{\theta}})_l + 3(\theta-\mu_{\hat{\theta}})_s \xi_{\hat{\theta},t}\xi_{\hat{\theta},l}\right\}\right)$$

# Skew-modal approximation: theory

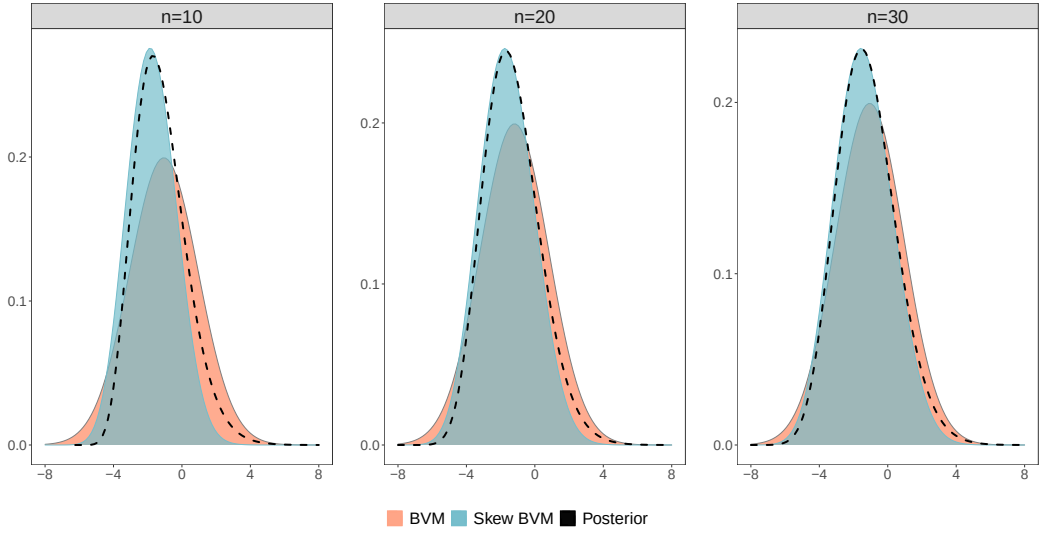
**Different** skew-symmetric approximations with the same accuracy can be obtained by considering Taylor expansions around **any efficient estimator** $\hat{\theta}$

$$p_{SS,\hat{\theta}}(\theta) = 2\,\varphi_d\left(\theta; \mu_{\hat{\theta}}, \Omega_{\hat{\theta}}^{-1}\right) \Phi\left(\frac{\sqrt{2\pi}}{12} \sum_{s,t,l=1}^{d} \ell^{(3)}_{\hat{\theta},stl} \left\{(\theta-\mu_{\hat{\theta}})_s(\theta-\mu_{\hat{\theta}})_t(\theta-\mu_{\hat{\theta}})_l + 3(\theta-\mu_{\hat{\theta}})_s \xi_{\hat{\theta},t}\xi_{\hat{\theta},l}\right\}\right)$$

where

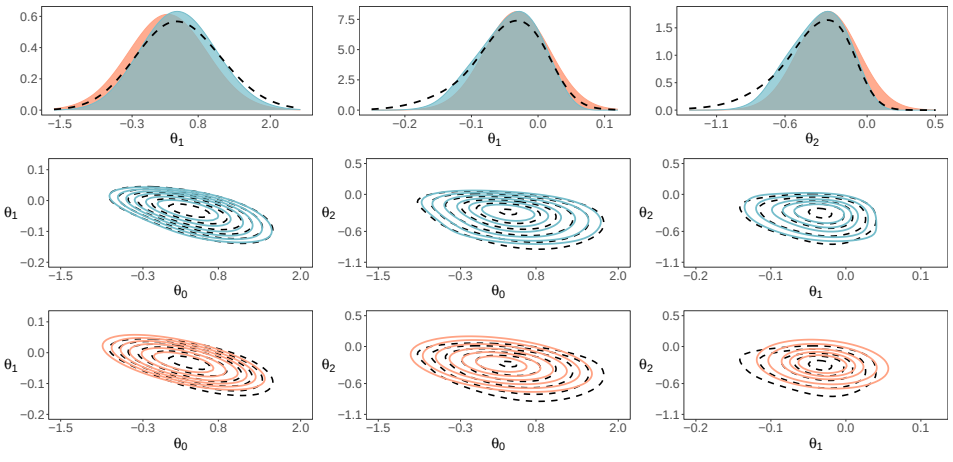
- $\xi_{\hat{\theta}} = \left(-\ell^{(2)}_{\hat{\theta}}\right)^{-1} \ell^{(1)}_{\hat{\theta}}$
- $\mu_{\hat{\theta}} = \hat{\theta} + \xi_{\hat{\theta}}$
- $\Omega^{-1}_{\hat{\theta},st} = -\ell^{(2)}_{\hat{\theta},st} - \sum_{l=1}^{d} \xi_{\hat{\theta},l} \ell^{(3)}_{\hat{\theta},stl}$

# Skew-modal approximation: examples



# Skew-modal approximation: examples

Bivariate (+ intercept) Probit regression with $n = 27$



# Skew-modal approximation: examples

Logistic regression with $n = 333$ and $d = 135$

**Summary statistics**:

- $\text{BIAS} = (1/d) \sum_{i=1}^{n} |\mu_{\pi_n,i} - \mu_{\text{APP},i}|$ with $\mu_p = \int \theta p(\theta) d\theta$
- $\text{TV} = (1/d) \sum_{i=1}^{n} \mathcal{D}_{\text{TV}}(\pi_{n,i}, p_{\text{APP},i})$, where $i$ denotes the $i$-th marginal

| | BIAS | TV |
|---|---|---|
| SKEW−M | **0.139** | **0.104** |
| LA | 0.425 | 0.145 |

# Summary and further developments

- A general theoretical and methodological framework justifying the use of asymmetric approximations of the posterior distribution is introduced
- Convergence to the posterior at a rate one order of magnitude faster for skew-modal approximations than for the standard Laplace approximation

# Summary and further developments

- A general theoretical and methodological framework justifying the use of asymmetric approximations of the posterior distribution is introduced
- Convergence to the posterior at a rate one order of magnitude faster for skew-modal approximations than for the standard Laplace approximation
- **Further extensions**:
  1. Development of similar corrections for other symmetric approximations (Daniele's talk tomorrow)
  2. Skew-modal approximations for **online learning** problems (Dolmeta et al., 2026+)

# Online skew-symmetric approximations

# Online learning

- Data $y_1, \ldots, y_t$ are collected sequentially.
- When a new observation $y_t$ arrives, the posterior distribution must be updated:

$$\pi_{t-1}(\theta) \xrightarrow{y_t} \pi_t(\theta)$$

  Doing it from scratch at each iteration is computationally expensive

# Online learning

- Data $y_1, \ldots, y_t$ are collected sequentially.
- When a new observation $y_t$ arrives, the posterior distribution must be updated:

$$\pi_{t-1}(\theta) \xrightarrow{y_t} \pi_t(\theta)$$

  Doing it from scratch at each iteration is computationally expensive

- The same issue arises for the skew-modal approximation:
  - $\tilde{\theta}^t$, $J^t_{\tilde{\theta}}$, $\ell^{(3),t}_{\tilde{\theta}}$: computational costs linear in $n$ and cubic in $d$

# Online skew-symmetric approximation

**Aim**: develop efficient computational strategies for skew-symmetric approximations in the online setting

# Online skew-symmetric approximation

**Aim**: develop efficient computational strategies for skew-symmetric approximations in the online setting

**Assumption**: the posterior mode $\tilde{\theta}^t$ is replaced with an online estimator $\hat{\theta}^t$

# Online skew-symmetric approximation

**Aim**: develop efficient computational strategies for skew-symmetric approximations in the online setting

**Assumption**: the posterior mode $\tilde{\theta}^t$ is replaced with an online estimator $\hat{\theta}^t$

If $\hat{\theta}^t$ is not the posterior mode, the skew-symmetric approximation takes the form

$$2\,\varphi_d\left(\theta; \mu_{\hat{\theta}^t}, \Omega^{-1}_{\hat{\theta}^t}\right) \Phi\left(\frac{\sqrt{2\pi}}{12} \sum_{s,m,l=1}^{d} \ell^{(3)}_{\hat{\theta}^t,sml} \left\{(\theta-\mu_{\hat{\theta}^t})_s(\theta-\mu_{\hat{\theta}^t})_m(\theta-\mu_{\hat{\theta}^t})_l + 3(\theta-\mu_{\hat{\theta}^t})_s \xi_{\hat{\theta}^t,m}\xi_{\hat{\theta}^t,l}\right\}\right)$$

and depends on $\hat{\theta}^t$ and on

$$\ell^{(k)}_{\hat{\theta}^t} = \sum_{i=0}^{t} \ell^{(k)}(\hat{\theta}^t, y_i), \qquad k = 1, 2, 3$$

# Online skew-symmetric approximation

- **Full offline** approach: evaluate

$$\ell^{(k)}_{\hat{\theta}^t} = \sum_{i=0}^{t} \ell^{(k)}(\hat{\theta}^t, y_i), \qquad k = 1, 2, 3$$

for every $t$

# Online skew-symmetric approximation

- **Full offline** approach: evaluate

$$\ell^{(k)}_{\hat{\theta}^t} = \sum_{i=0}^{t} \ell^{(k)}(\hat{\theta}^t, y_i), \qquad k = 1, 2, 3$$

  for every $t$

  $\Longrightarrow$ **very expensive**

- **Full online** approach:

$$\ell^{(k)}_{\hat{\theta}^t} = \sum_{i=0}^{t} \ell^{(k)}(\hat{\theta}^i, y_i), \qquad k = 1, 2, 3$$

  for every $t$

# Online skew-symmetric approximation

- **Full offline** approach: evaluate

$$\ell^{(k)}_{\hat{\theta}^t} = \sum_{i=0}^{t} \ell^{(k)}(\hat{\theta}^t, y_i), \qquad k = 1, 2, 3$$

  for every $t$
  $\Longrightarrow$ **very expensive**

- **Full online** approach:

$$\ell^{(k)}_{\hat{\theta}^t} = \sum_{i=0}^{t} \ell^{(k)}(\hat{\theta}^i, y_i), \qquad k = 1, 2, 3$$

  for every $t$
  $\Longrightarrow$ **no theoretical guarantees**, the approximation can become **arbitrarily bad**

# Online skew-symmetric approximation

- **Window-type estimators**: each contribution is updated every $M$ lags.
  1. For $t < M$,
$$\hat{\ell}^{(k)}_{\hat{\theta}^t} = \ell^{(k)}(\hat{\theta}^1, y_1) + \ell^{(k)}(\hat{\theta}^2, y_2) + \cdots + \ell^{(k)}(\hat{\theta}^t, y_t).$$

# Online skew-symmetric approximation

- **Window-type estimators**: each contribution is updated every $M$ lags.
  1. For $t < M$,
  $$\hat{\ell}^{(k)}_{\hat{\theta}^t} = \ell^{(k)}(\hat{\theta}^1, y_1) + \ell^{(k)}(\hat{\theta}^2, y_2) + \cdots + \ell^{(k)}(\hat{\theta}^t, y_t).$$
  2. For $M + 1 \leq t \leq 2M$,
  $$\begin{aligned}\hat{\ell}^{(k)}_{\hat{\theta}^t} = \ell^{(k)}(\hat{\theta}^{M+1}, y_1) + \ell^{(k)}(\hat{\theta}^{M+2}, y_2) + \cdots + \ell^{(k)}(\hat{\theta}^t, y_t) \\ + \ell^{(k)}(\hat{\theta}^{M+1}, y_{M+1}) + \ell^{(k)}(\hat{\theta}^{M+2}, y_{M+2}) + \ldots.\end{aligned}$$

# Online skew-symmetric approximation

- **Window-type estimators**: each contribution is updated every $M$ lags.

  1. For $t < M$,
  $$\hat{\ell}^{(k)}_{\hat{\theta}^t} = \ell^{(k)}(\hat{\theta}^1, y_1) + \ell^{(k)}(\hat{\theta}^2, y_2) + \cdots + \ell^{(k)}(\hat{\theta}^t, y_t).$$

  2. For $M+1 \leq t \leq 2M$,
  $$\begin{aligned}\hat{\ell}^{(k)}_{\hat{\theta}^t} = \ell^{(k)}(\hat{\theta}^{M+1}, y_1) + \ell^{(k)}(\hat{\theta}^{M+2}, y_2) + \cdots + \ell^{(k)}(\hat{\theta}^t, y_t) \\ + \ell^{(k)}(\hat{\theta}^{M+1}, y_{M+1}) + \ell^{(k)}(\hat{\theta}^{M+2}, y_{M+2}) + \ldots.\end{aligned}$$

  3. In general,
  $$\hat{\ell}^{(k)}_{\hat{\theta}^t} = \sum_{i=1}^{t} \ell^{(k)}\left(\hat{\theta}^{\,t-(t-i)\%M}, y_i\right), \qquad k = 1, 2, 3,$$
  where % denotes the modulo operator

# Online skew-symmetric approximation: theory

## Theorem

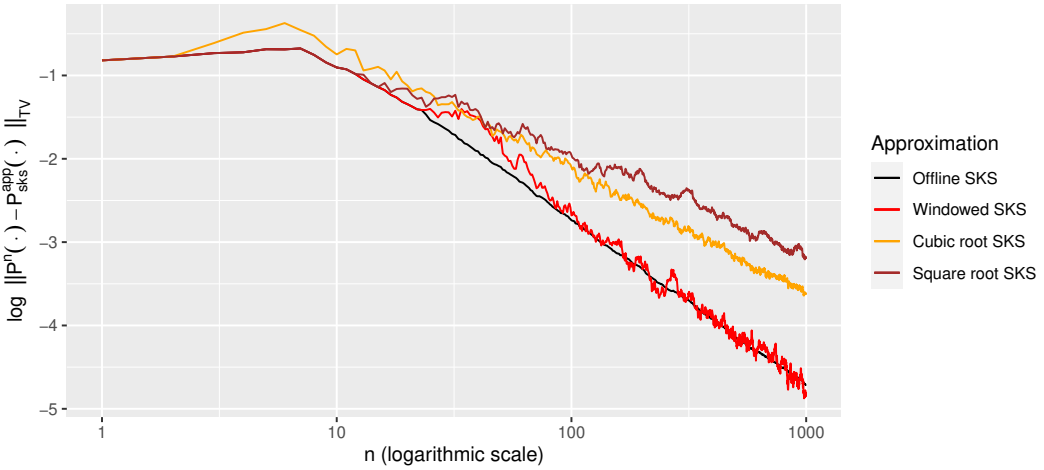
*Let $\hat{\theta}^t$ be an efficient online estimator and let $p_{SS,\hat{\theta}^t}(\theta)$ be a skew-symmetric approximation in which $\ell^{(k)}_{\hat{\theta}^t}$, $k = 1, 2, 3$, are estimated using a window-type estimator with fixed window size $M$. Then, under regular parametric models,*

$$\mathcal{D}_{\mathrm{TV}}\left(\pi_n, p_{SS,\hat{\theta}^t}(\theta)\right) = O_P\left(n^{-1}\right),$$

*where $\mathcal{D}_{\mathrm{TV}}(\cdot,\cdot)$ denotes the total variation distance*

- **Same rate** as offline skew-symmetric approximations
- The window size $M$ may increase with $n$, at the cost of a slower convergence rate

# Online skew-symmetric approximation: example




Total variation distance the posterior and online skew-symmetric approximation with three different window sizes

# References


- Azzalini, A. (1985). "A class of distributions which includes the normal ones". In: *Scandinavian journal of statistics*, pp. 171–178.
- 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.
- Azzalini, A. and A. Dalla Valle (1996). "The multivariate skew-normal distribution". In: *Biometrika* 83.4, pp. 715–726.
- Durante, D. (2019). "Conjugate Bayes for probit regression via unified skew-normal distributions". In: *Biometrika* 106.4, pp. 765–779.
- Johnson, R. A. (1970). "Asymptotic expansions associated with posterior distributions". In: *The Annals of Mathematical Statistics*, pp. 851–864.
- Ma, Y. and M. G. Genton (2004). "Flexible class of skew-symmetric distributions". In: *Scandinavian Journal of Statistics* 31.3, pp. 459–468.
- Onorati, P. and B. Liseo (2022). "An extension of the Unified Skew-Normal family of distributions and application to Bayesian binary regression". In: *arXiv preprint arXiv:2209.03474*.
- Tan, L. S. (2023). "Variational inference based on a subclass of closed skew normals". In: *arXiv preprint arXiv:2306.02813*.