Generalised Bayes with intractable distances

David T. Frazier (Monash), Jeremias Knoblauch (UCL), and Christopher Drovandi (QUT)

05/02/24

Outline

1. Generalised Bayesian beliefs
2. Loss-based generalised Bayes
3. Computational challenges & Examples
4. Novel asymptotic theory

Today’s take-away in a nutshell

1. Generalised Bayes with distances (between model & data-generating process) is attractive
2. Challenge: often, such distances are intractable & require simulation (KDE, MMD, DPD)
    1. Turner and Sederberg (2014) - KDE
    2. Cherief-Abdellatif and Alquier (2020) - MMD
    3. Jewson et al. (2018) - Beta Divergence
    4. Pacchiardi et al. (2022) - Kernel Scores
    5. Legramanti et al. (2023) - IPMs
3. Show that this requires analysing novel posterior object
4. Need to make sure we are (asymptotically) targeting the correct object

Standard Bayesian beliefs

Standard Bayes posterior

Empirical measure, $P_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$

$$\pi(\theta \mid \widehat{\mathrm{KL}}_n) = \frac{\prod_{i=1}^{n} p_\theta(x_i)\pi(\theta)}{\int \prod_{i=1}^{n} p_\theta(x_i)\pi(\theta)d\theta} = \mathrm{argmin}_{q\in\mathscr{P}(\Theta)} \left\{ \mathbb{E}_{\theta\sim q}\left[n \cdot \widehat{\mathrm{KL}}_n\left(P_n, p_\theta\right)\right] + \mathrm{KL}(q,\pi) \right\}$$

$$\widehat{\mathrm{KL}}_n\left(P_n, p_\theta\right) = -\frac{1}{n}\sum_{i=1}^{n} \log p_\theta(x_i) + C$$

A general framework for updating belief distributions, Bissiri, P.G., Holmes, C., & Walker, S.G. (2016)
An Optimization-centric View on Bayes' Rule: Reviewing and Generalizing Variational Inference, Knoblauch, J., Jewson, J., & Damoulas, T. (2022).

Loss-based generalised Bayesian beliefs

Loss-based generalised posterior

$$\pi(\theta \mid \widehat{\mathrm{D}}_n) = \frac{\exp\{-n \cdot \mathrm{D}(P_n, p_\theta)\}\pi(\theta)}{\int \exp\{-n \cdot \mathrm{D}(P_n, p_\theta)\}\pi(\theta)d\theta} = \operatorname{argmin}_{q \in \mathscr{P}(\Theta)} \left\{ \mathbb{E}_{\theta \sim q} \left[n \cdot \mathrm{D}\left(P_n, p_\theta\right) \right] + \mathrm{KL}(q, \pi) \right\}$$

Estimator for some loss function D

An Optimization-centric View on Bayes' Rule: Reviewing and Generalizing Variational Inference, Knoblauch, J., Jewson, J., & Damoulas, T. (2022).
Principles of Bayesian Inference using General Divergence Criteria, Jewson, J., Smith, J., & Holmes, C., (2016)

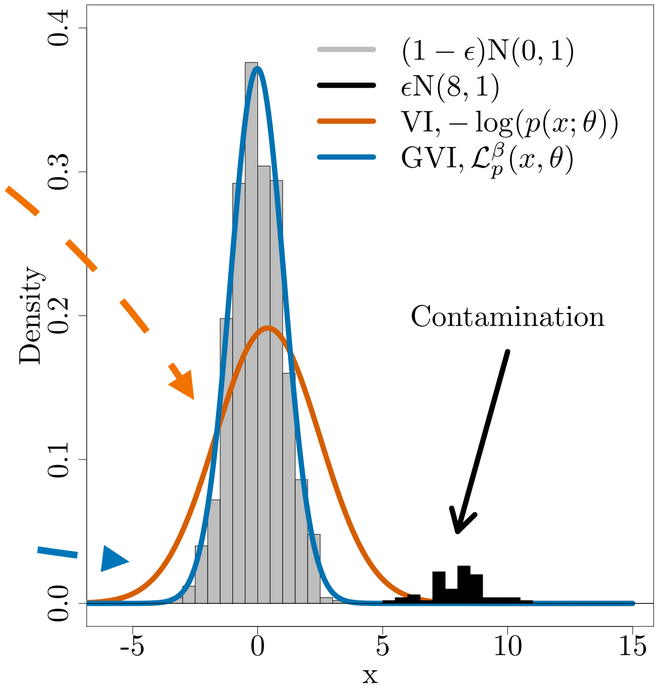
Why distance-based generalised Bayesian beliefs? (Robustness)

$$\pi(\theta \,|\, \mathrm{D}_n) = \frac{\exp\{-n \cdot D(P_n, p_\theta)\}\pi(\theta)}{\int \exp\{-n \cdot D(P_n, p_\theta)\}\pi(\theta) d\theta}$$

1. **Standard Bayes**: brittle/unreliable/non-robust under model misspecification (because KL is not a robust measure of distance)

2. But we want to keep Bayesian advantages (UQ, expert input via priors)

Solution: belief distributions on the parameter through **robust distances**



Generalised Bayesian inference for discrete intractable likelihood, Matsubara, T., Knoblauch, J., Briol, F.X., and Oates, C. (2022)
Robust generalised Bayesian inference for intractable likelihoods, Matsubara, T., Knoblauch, J., Briol, F.X., and Oates, C. (2022)
Generalised Bayesian likelihood-free inference using scoring rule estimators, Pacchiardi, L. & Dutta, R. (2021)
MMD-Bayes: Robust Bayesian estimation via maximum mean discrepancy, Cherrief-Abdellatif, B.E. & Alquier, P. (2020)
Principles of Bayesian Inference using General Divergence Criteria, Jewson, J., Smith, J., & Holmes, C., (2016)
Robust Bayes estimation using the density power divergence, Ghosh, A. & Basu, A. (2016)
Bayesian model robustness via disparities, Hooker, G. & Vidyashankar, A. N. (2014)
...

When are distance-based generalised Bayesian beliefs difficult?

$$\pi(\theta \mid \widehat{\mathrm{KL}}_n) = \frac{\prod_{i=1}^n p_\theta(x_i)\pi(\theta)}{\int \prod_{i=1}^n p_\theta(x_i)\pi(\theta)d\theta}$$

Need to evaluate $p_\theta(x)$ for $x \in \{x_i : 1 \leq i \leq n\}$

Why loss-based generalised Bayesian beliefs?

$$\pi(\theta \mid \widehat{\mathrm{KL}}_n) = \frac{\prod_{i=1}^n p_\theta(x_i)\pi(\theta)}{\int \prod_{i=1}^n p_\theta(x_i)\pi(\theta)d\theta}$$

Need to evaluate $p_\theta(x)$ for $x \in \{x_i : 1 \leq i \leq n\}$

$$\pi(\theta \mid \mathrm{D}_n) = \frac{\exp\{-n \cdot D(P_n, p_\theta)\}\pi(\theta)}{\int \exp\{-n \cdot D(P_n, p_\theta)\}\pi(\theta)d\theta}$$

Need to evaluate/approximate $D(P_n, p_\theta)$

E.g., MMD^2 for simulators,
KSD^2 for unnormalised models

Generalised Bayesian inference for discrete intractable likelihood, Matsubara, T., Knoblauch, J., Briol, F.X., and Oates, C. (2022)
Robust generalised Bayesian inference for intractable likelihoods, Matsubara, T., Knoblauch, J., Briol, F.X., and Oates, C. (2022)
Generalised Bayesian likelihood-free inference using scoring rule estimators, Pacchiardi, L. & Dutta, R. (2021)

Computational challenges (MMD)

$$\pi(\theta \,|\, \mathrm{D}_n) \propto \exp\{-n \cdot D(P_n, p_\theta)\}\pi(\theta)$$

Actual target: $\mathrm{MMD}^2(P_n, p_\theta) = \mathbb{E}_{X\sim P_n, X'\sim P_n}[k(X, X')] - 2\mathbb{E}_{X\sim P_n, Y\sim p_\theta}[k(X, Y)] + \mathbb{E}_{Y\sim p_\theta, Y'\sim p_\theta}[k(Y, Y')]$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} k(x_i, x_j) \quad -\frac{2}{n}\sum_{i=1}^{n} \mathbb{E}_{Y\sim p_\theta}[k(x_i, Y)] \quad +\mathbb{E}_{Y\sim p_\theta, Y'\sim p_\theta}[k(Y, Y')]$$

Intractable expectations

Generalised Bayesian likelihood-free inference using scoring rule estimators, Pacchiardi, L. & Dutta, R. (2021)
MMD-Bayes: Robust Bayesian estimation via maximum mean discrepancy, Cherrief-Abdellatif, B.E. & Alquier, P. (2020)

Computation: challenges (MMD/kernel score example)

$$\pi(\theta \,|\, \mathrm{D}_n) \propto \exp\{-n \cdot D(P_n, p_\theta)\}\pi(\theta)$$

Actual target: $\mathrm{MMD}^2(P_n, p_\theta) = \mathbb{E}_{X\sim P_n, X'\sim P_n}[k(X, X')] - 2\mathbb{E}_{X\sim P_n, Y\sim p_\theta}[k(X, Y)] + \mathbb{E}_{Y\sim p_\theta, Y'\sim p_\theta}[k(Y, Y')]$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} k(x_i, x_j) \quad -\frac{2}{n}\sum_{i=1}^{n} \mathbb{E}_{Y\sim p_\theta}[k(x_i, Y)] \quad +\mathbb{E}_{Y\sim p_\theta, Y'\sim p_\theta}[k(Y, Y')]$$

Intractable expectations

What we compute: $\mathrm{MMD}^2(P_n, P_{m,\theta}) = \mathbb{E}_{X\sim P_n, X'\sim P_n}[k(X, X')] - 2\mathbb{E}_{X\sim P_n, Y\sim P_{m,\theta}}[k(X, Y)] + \mathbb{E}_{Y\sim P_{m,\theta}, Y'\sim P_{m,\theta}}[k(Y, Y')]$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} k(x_i, x_j) \quad -2\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} k(x_i, y_j) \quad +\frac{1}{m^2}\sum_{j=1}^{m}\sum_{j=1}^{m} k(y_i, y_j)$$

Generalised Bayesian likelihood-free inference using scoring rule estimators, Pacchiardi, L. & Dutta, R. (2021)
MMD-Bayes: Robust Bayesian estimation via maximum mean discrepancy, Cherrief-Abdellatif, B.E. & Alquier, P. (2020)

Computation: challenges

Some other examples for D with this problem:

F_n^{-1} = empirical quantile function
$F_{p_\theta}^{-1}$ = model quantile function

$$\mathrm{W}_p^p(P_n, p_\theta) = \left[\int_{x \in \mathcal{X}} \left| F_{p_\theta}^{-1}(x) - F_n^{-1}(x) \right| dx \right]^{1/p} \qquad \text{if } \mathcal{X} \subset \mathbb{R}$$

Computation: challenges

Some other examples for D with this problem:

F_n^{-1} = empirical quantile function
$F_{p_\theta}^{-1}$ = model quantile function

$$\mathrm{W}_p^p(P_n, p_\theta) = \left[\int_{x \in \mathcal{X}} \left| F_{p_\theta}^{-1}(x) - F_n^{-1}(x) \right| dx \right]^{1/p} \quad \text{if } \mathcal{X} \subset \mathbb{R}$$

F_n = empirical cdf
F_{p_θ} = model cdf

$$\mathrm{KS}(P_n, p_\theta) = \sup_{x \in \mathcal{X}} \left| F_{p_\theta}(x) - F_n(x) \right|$$

$$\mathrm{CvM}(P_n, p_\theta) = \int_{x \in \mathcal{X}} \left| F_{p_\theta}(x) - F_n(x) \right|^2 dF_n(x)$$

Computation: challenges

Some other examples for D with this problem:

F_n^{-1} = empirical quantile function
$F_{p_\theta}^{-1}$ = model quantile function

$$\mathrm{W}_p^p(P_n, p_\theta) = \left[\int_{x \in \mathcal{X}} \left| F_{p_\theta}^{-1}(x) - F_n^{-1}(x) \right| dx \right]^{1/p} \quad \text{if } \mathcal{X} \subset \mathbb{R}$$

F_n = empirical cdf
F_{p_θ} = model cdf

$$\mathrm{KS}(P_n, p_\theta) = \sup_{x \in \mathcal{X}} \left| F_{p_\theta}(x) - F_n(x) \right|$$

$$\mathrm{CvM}(P_n, p_\theta) = \int_{x \in \mathcal{X}} \left| F_{p_\theta}(x) - F_n(x) \right|^2 dF_n(x)$$

Also: general IPMs, $\beta/\alpha/\gamma$-divergences, …, any intractable models as well...

Computation: challenges

What we actually want: $\pi(\theta \mid \mathrm{D}_n) \propto \exp\{-n \cdot D(P_n, p_\theta)\}\pi(\theta)$

What we can compute: $\pi(\theta \mid \widehat{\mathrm{D}}_{m,n}, y_{1:m}) \propto \exp\{-n \cdot \widehat{D}(P_n, P_{m,\theta})\}\pi(\theta)$

Estimator $\widehat{D} \approx D$

Two pieces: n-sample data points, m-simulated data points

Sample-based approximation of p_θ,

$$P_{m,\theta} = \frac{1}{m}\sum_{j=1}^{m} \delta_{y_j} \quad y_{1:m} \overset{i.i.d.}{\sim} p_\theta$$

Computation: challenges

What we actually want: $\pi(\theta \mid \mathrm{D}_n) \propto \exp\{-n \cdot D(P_n, p_\theta)\}\pi(\theta)$

What we (can) compute: $\pi(\theta \mid \widehat{\mathrm{D}}_{m,n}, y_{1:m}) \propto \exp\{-n \cdot \widehat{D}(P_n, P_{m,\theta})\}\pi(\theta)$

Estimator $\widehat{D} \approx D$

Sample-based approximation of p_θ,

$$P_{m,\theta} = \frac{1}{m}\sum_{j=1}^{m} \delta_{y_j} \quad y_{1:m} \overset{i.i.d.}{\sim} p_\theta$$

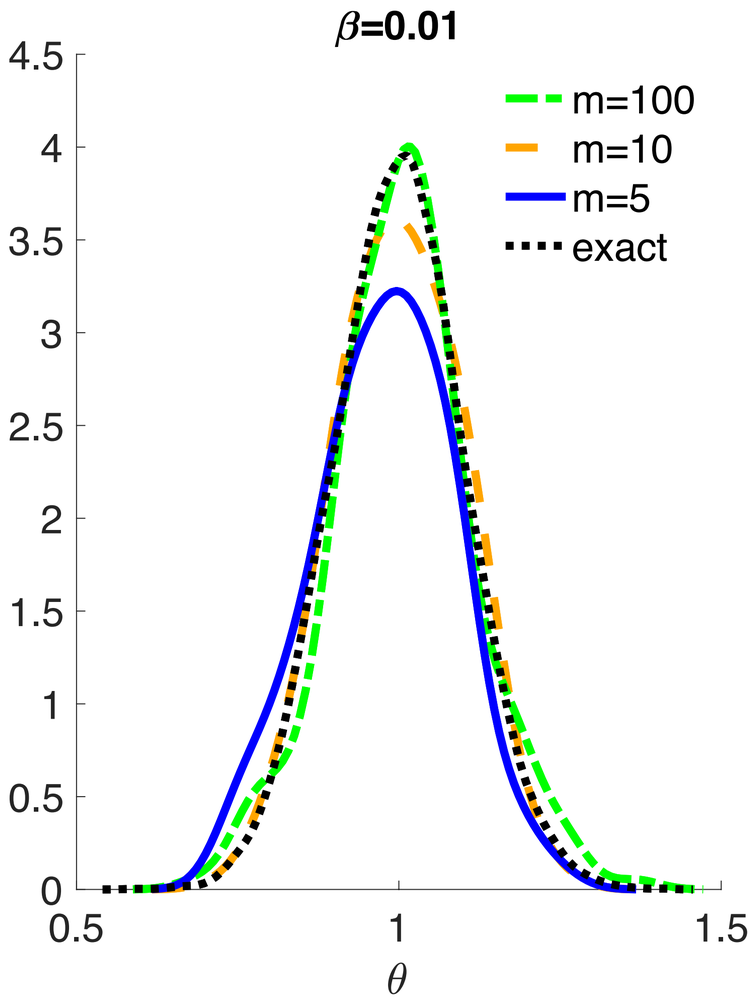
Research questions we tackle:

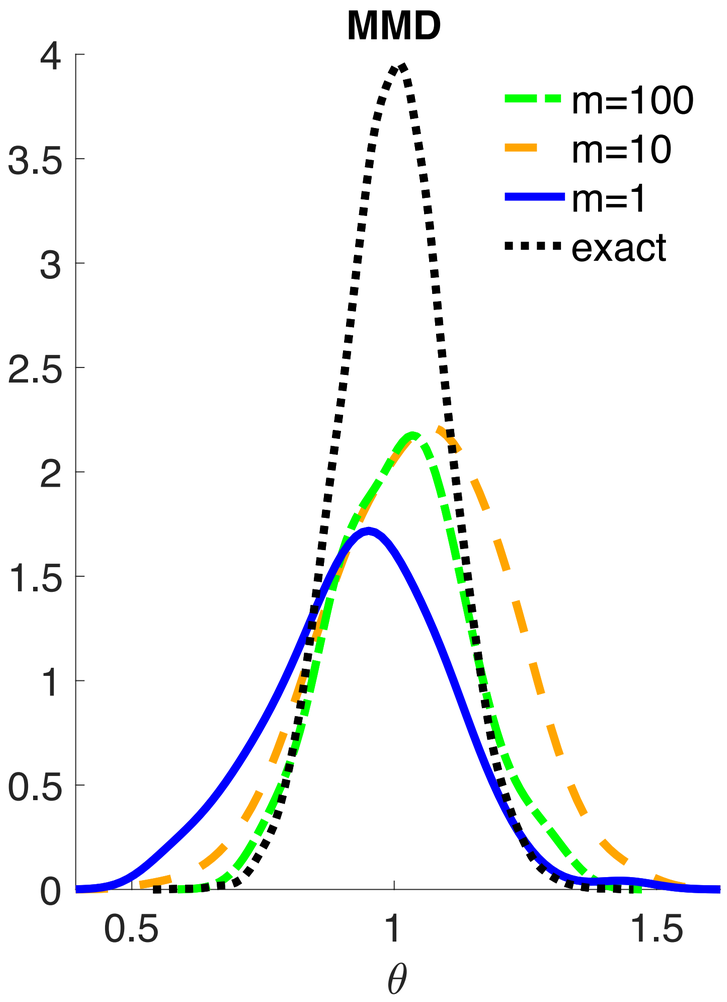
(1) How different is $\pi(\theta \mid \mathrm{D}_{m,n}, y_{1:m})$ from $\pi(\theta \mid \widehat{\mathrm{D}}_{m,n}, y_{1:m})$?

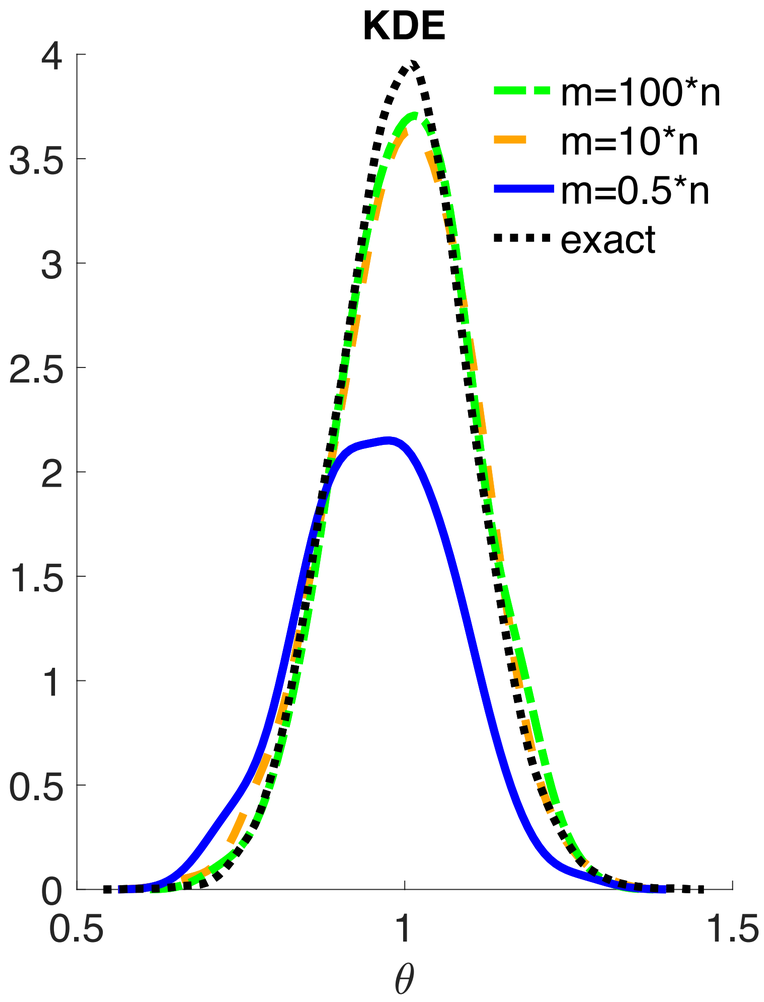
(2) How does $\widehat{\mathrm{D}}_{m,n}$ dictate behaviour of and $\pi(\theta \mid \widehat{\mathrm{D}}_{m,n}, y_{1:m})$?

Preliminary answers to (1) and (2)

Data generated iid from $N(\theta,1)$, $\theta = 1$







Computation: accounting for model samples

Question (1): How different is $\pi(\theta \mid \mathrm{D}_n)$ from $\pi(\theta \mid \widehat{\mathrm{D}}_{m,n}, y_{1:m})$?

$$\pi(\theta \mid \widehat{\mathrm{D}}_{m,n}, , y_{1:m}) \propto \exp\{-n \cdot \widehat{D}(P_n, P_{m,\theta}, , y_{1:m})\}\pi(\theta)$$

Sample-based approximation of p_θ,

$$P_{m,\theta} = \frac{1}{m}\sum_{j=1}^{m} \delta_{y_j} \quad y_{1:m} \overset{i.i.d.}{\sim} p_\theta$$

But hold on! $y_{1:m}$ itself is random! (not fixed/conditioned on!)

Computation: accounting for model samples

Question (1): How different is $\pi(\cdot \mid D_n)$ from $\pi(\cdot \mid \widehat{D}_{m,n})$?

$$\pi(\theta \mid \widehat{\mathrm{D}}_{m,n}, y_{1:m}) \propto \exp\{-n \cdot \widehat{D}(P_n, P_{m,\theta})\}\pi(\theta)$$

$\pi(\theta \mid \widehat{\mathrm{D}}_{m,n}, y_{1:m})$ constructed from draws $y_{1:m} \overset{i.i.d.}{\sim} p_\theta$

is itself random, and an (unbiased) estimate of $\bar{\pi}(\theta \mid \widehat{\mathrm{D}}_{m,n})$

$$\bar{\pi}(\theta \mid \widehat{\mathrm{D}}_{m,n}) \propto \pi(\theta) \cdot \mathbb{E}_{y_{1:m} \sim p_\theta}\left[\exp\{-n \cdot \widehat{D}(P_n, P_{m,\theta})\}\right] \quad \left(\neq \pi(\theta \mid \mathrm{D}_n)\right)$$

The pseudo-marginal approach for efficient Monte Carlo computations, Andrieu, C. & Roberts, G. O. (2009)

Computation: accounting for model samples

Question (1): How different is $\pi(\cdot \mid D_n)$ from $\pi(\cdot \mid \widehat{D}_{m,n})$?

$$\pi(\theta \mid \widehat{\mathrm{D}}_{m,n}) \propto \exp\{-n \cdot \widehat{D}(P_n, P_{m,\theta})\}\pi(\theta)$$

$\pi(\theta \mid \widehat{\mathrm{D}}_{m,n})$ constructed from draws $y_{1:m} \overset{i.i.d.}{\sim} p_\theta$

is itself random, and an (unbiased) estimate of $\bar{\pi}(\theta \mid \widehat{\mathrm{D}}_{m,n})$

$$\bar{\pi}(\theta \mid \widehat{\mathrm{D}}_{m,n}) \propto \pi(\theta) \cdot \underbrace{\mathbb{E}_{y_{1:m} \sim P_\theta}\left[\exp\{-n \cdot \widehat{D}(P_n, P_{m,\theta})\}\right]}_{= \text{like an intractable likelihood!}} \qquad \left(\neq \pi(\theta \mid \mathrm{D}_n) \right)$$

The pseudo-marginal approach for efficient Monte Carlo computations, Andrieu, C. & Roberts, G. O. (2009)

Computation: accounting for model samples

Question (1): How different is $\pi(\cdot \mid D_n)$ from $\pi(\cdot \mid \widehat{D}_{m,n})$?

$$\bar{\pi}(\theta \mid \widehat{\mathrm{D}}_{m,n}) \propto \pi(\theta) \cdot \mathbb{E}_{y_{1:m} \sim p_\theta} \left[\exp\{-n \cdot \widehat{D}(P_n, P_{m,\theta})\} \right]$$

$$= \pi(\theta) \cdot \int_{\mathcal{X}^m} \left(\prod_{j=1}^{m} p_\theta(y_j) \right) \cdot \underbrace{\exp\{-n \cdot \widehat{D}(P_n, P_{m,\theta})\}}_{= \text{like a kernel in ABC/BSL}} \, dy_{1:m}$$

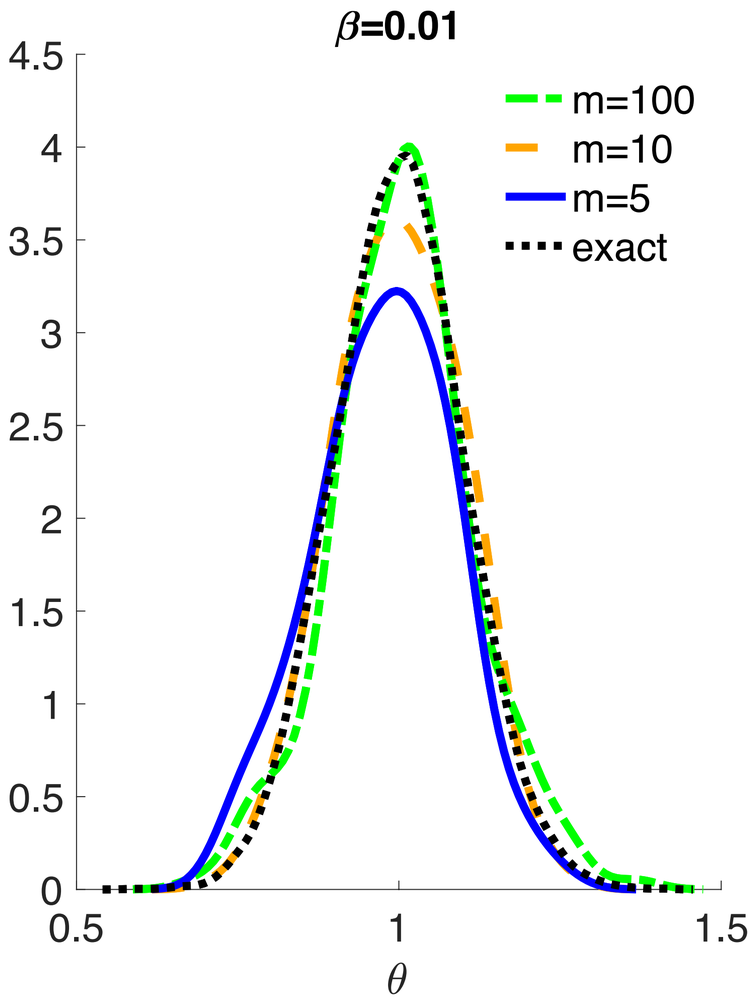
$$\neq \pi(\theta \mid D_n)$$

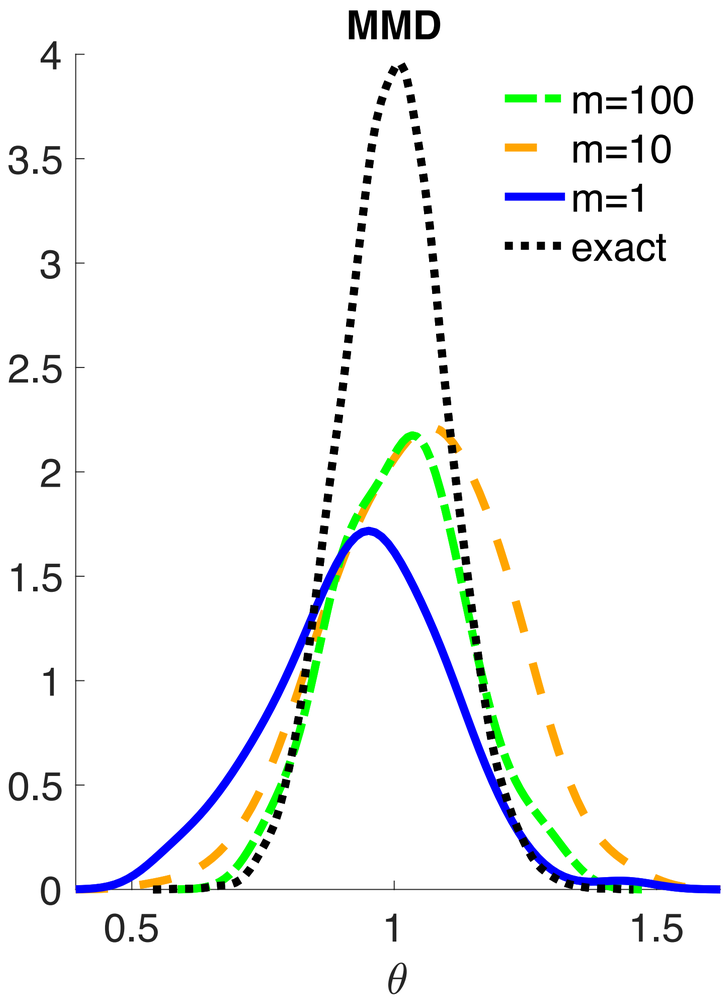
However: clearly $\bar{\pi}_n^D \neq \pi_n^{D,\text{ideal}}$; so can we quantify/bound the difference?

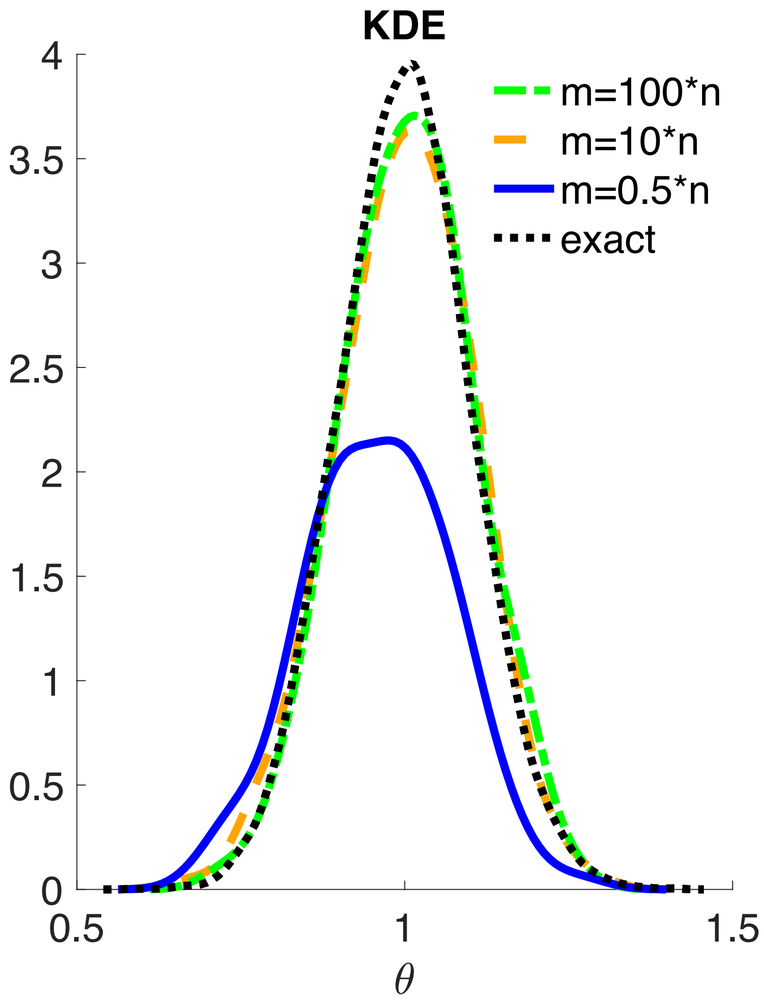
The pseudo-marginal approach for efficient Monte Carlo computations, Andrieu, C. & Roberts, G. O. (2009)

Preliminary answers to (1) and (2)

Data generated iid from $N(\theta,1)$, $\theta = 1$







Computation: accounting for model samples

Theorem 1:

Under mild regularity conditions on the moments of $\widehat{D}(P_n, P_{m,\theta})$ (in $y_{1:m}$) and $\pi_n^{D,\text{ideal}}$, then for all $\gamma \in [0,2]$ and as $n, m \to \infty$ with $m \gg n$,

$$\int_\Theta \|\theta\|^\gamma \left| \pi(\theta \mid D_n) - \bar{\pi}(\theta \mid \widehat{\mathrm{D}}_{m,n}) \right| d\theta = O_p(\max\{m^{-\kappa_1}, m^{-\kappa_2}\}),$$

which immediately implies that

$$\mathrm{TVD}\left(\pi_n^{D,\text{ideal}}, \bar{\pi}_n^D\right) = O_p(\max\{m^{-\kappa_1}, m^{-\kappa_2}\}), \qquad \left\| \mathbb{E}_{\theta\sim\pi_n^{D,\text{ideal}}}[\theta] - \mathbb{E}_{\theta\sim\bar{\pi}_n^D}[\theta] \right\| = O_p(\max\{m^{-\kappa_1}, m^{-\kappa_2}\}).$$

Computation: accounting for model samples

Theorem 1:

Under mild regularity conditions on the moments of $\widehat{D}(P_n, P_{m,\theta})$ (in $y_{1:m}$) and $\pi_n^{D,\text{ideal}}$, then for all $\gamma \in [0,2]$ and as $n, m \to \infty$ with $m \gg n$,

Bias Term

Variance Term

$$\int_\Theta \|\theta\|^\gamma \left| \pi(\theta \mid D_n) - \bar{\pi}(\theta \mid \widehat{\mathrm{D}}_{m,n}) \right| d\theta = O_p(\max\{m^{-\kappa_1}, m^{-\kappa_2}\}),$$

which immediately implies that

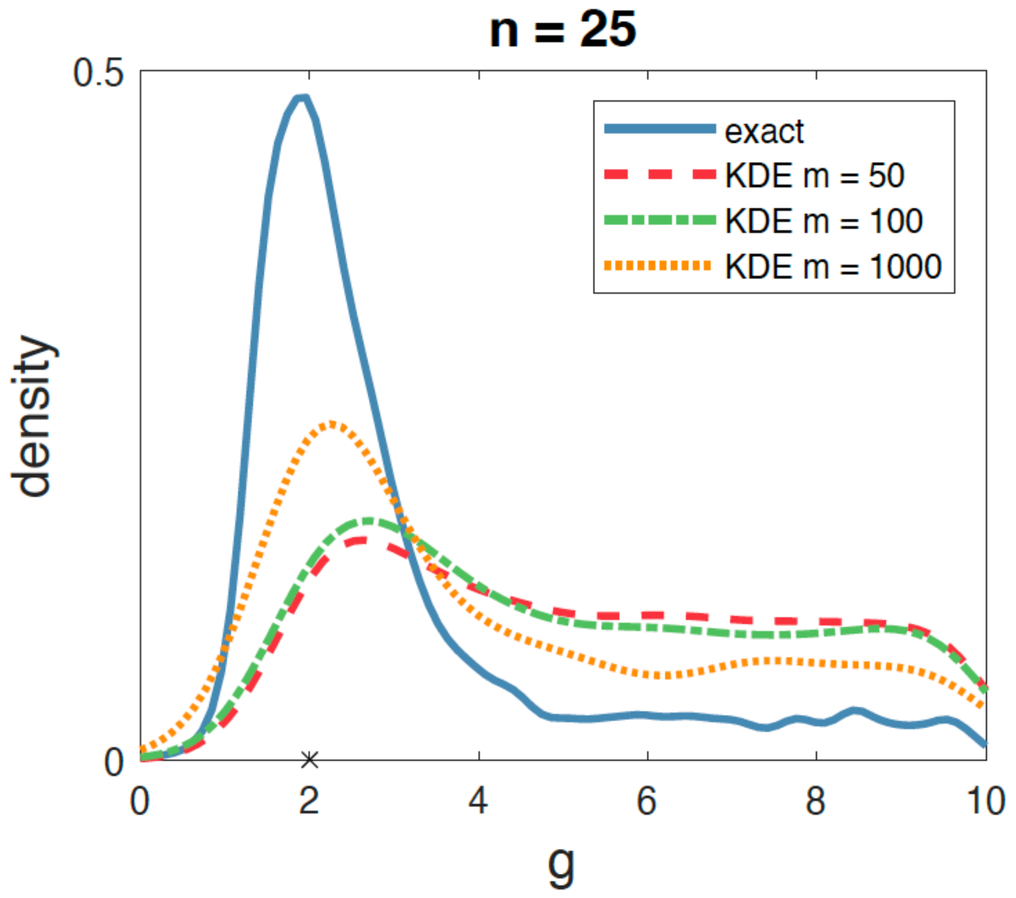
$$\mathrm{TVD}\left(\pi_n^{D,\text{ideal}}, \bar{\pi}_n^D\right) = O_p(\max\{m^{-\kappa_1}, m^{-\kappa_2}\}), \qquad \left\| \mathbb{E}_{\theta\sim\pi_n^{D,\text{ideal}}}[\theta] - \mathbb{E}_{\theta\sim\bar{\pi}_n^D}[\theta] \right\| = O_p(\max\{m^{-\kappa_1}, m^{-\kappa_2}\}).$$

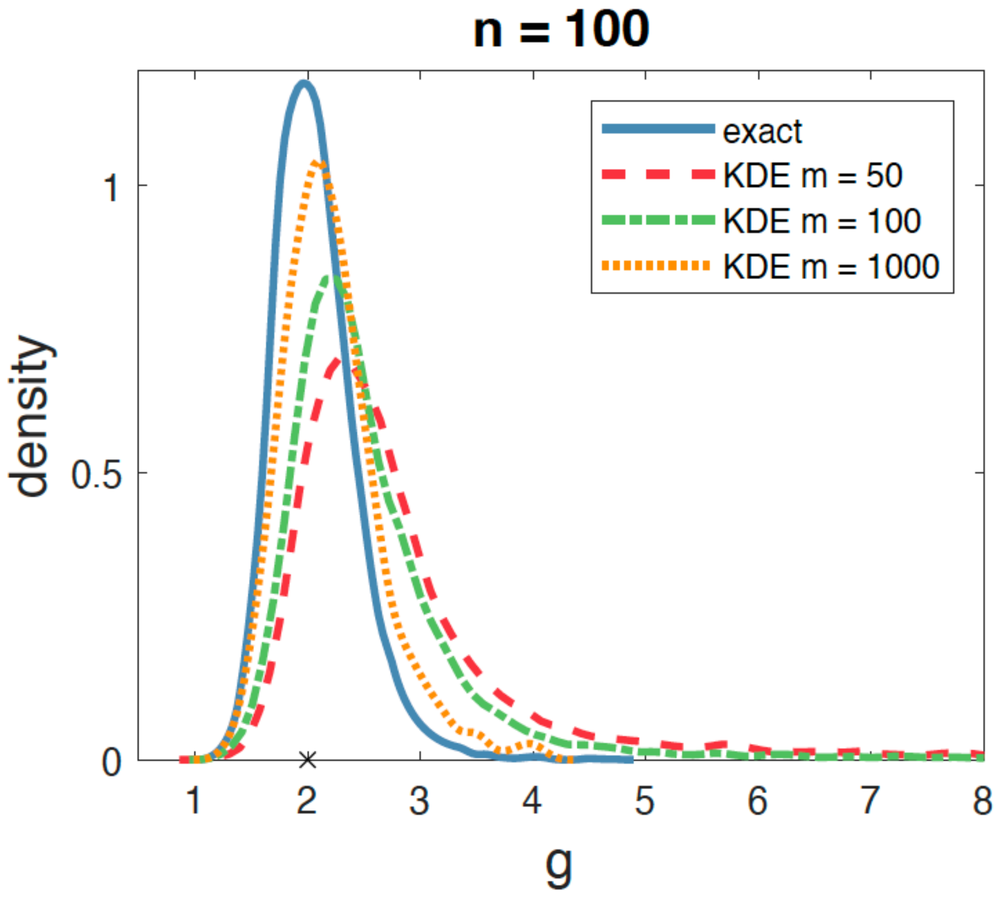
Tells us that implicitly targeting $\bar{\pi}_n^D$ instead of $\pi_n^{D,\text{ideal}}$ is (generally) fine
(bias vanishes at linear rate for larger model-samples)

Asymptotic & choice of distance

Question (2): How does D dictate behaviour of $\bar{\pi}_n^D$ and $\pi_n^{D,\text{ideal}}$?

Data generated iid from G-and-K dist

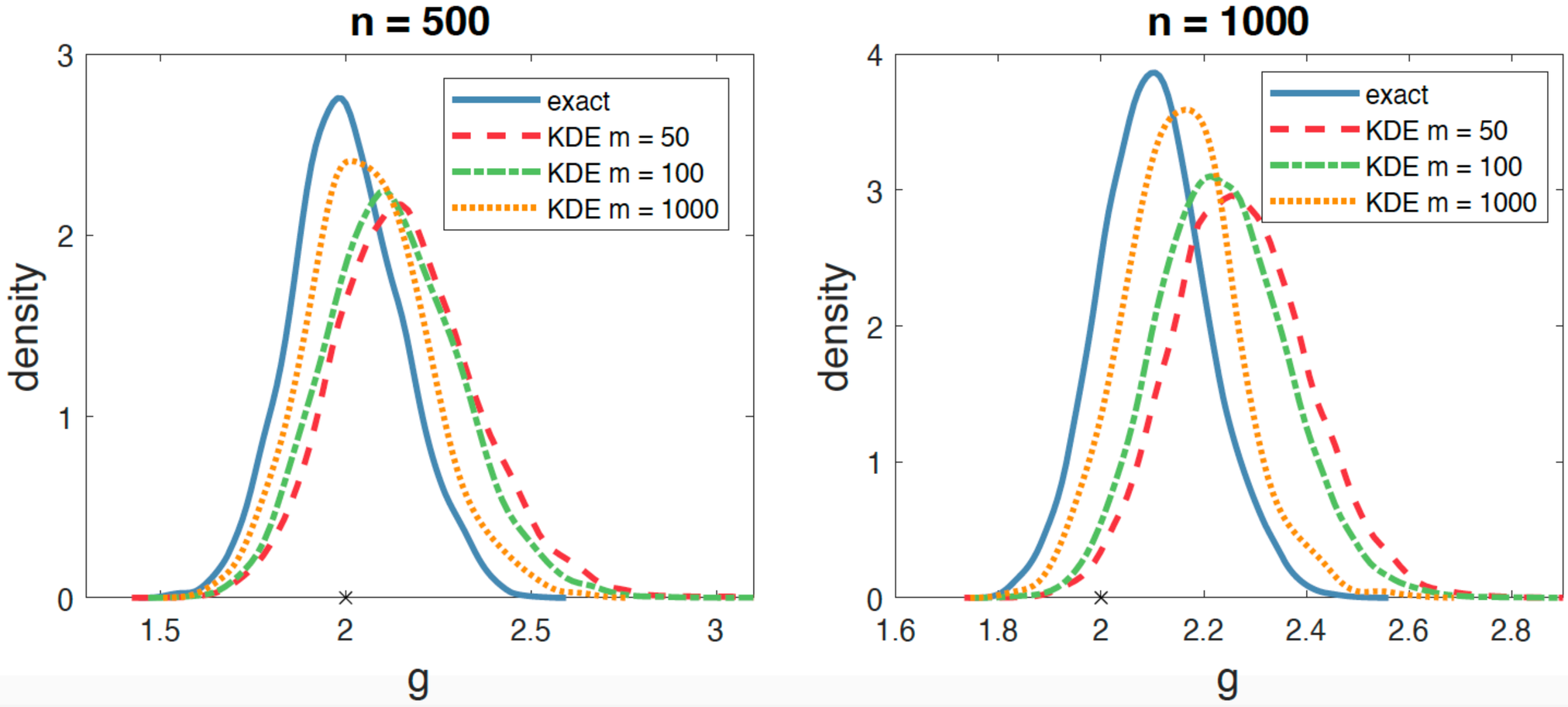




Asymptotic & choice of distance

Question (2): How does D dictate behaviour of $\bar{\pi}_n^D$ and $\pi_n^{D,\text{ideal}}$?

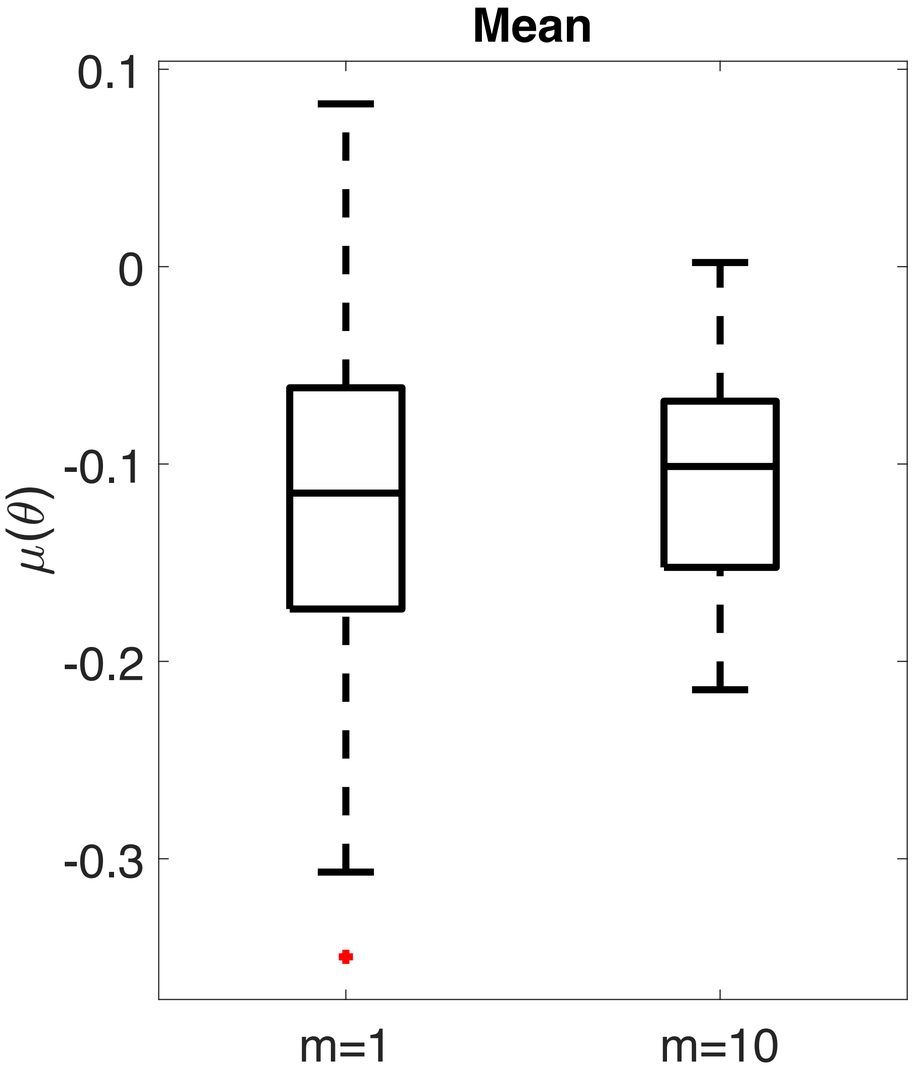
Data generated iid from G-and-K dist

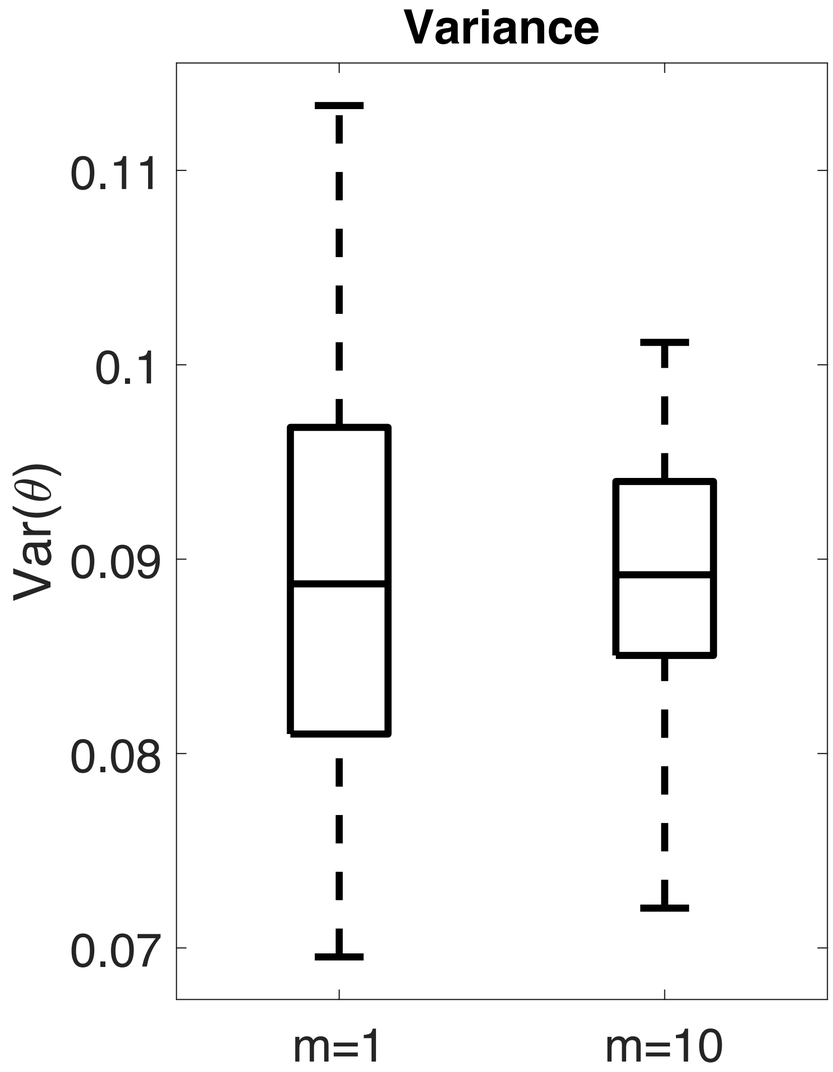


Asymptotics & choice of distance: MMD

Question (2): How does D dictate behaviour of $\bar{\pi}_n^D$?

Data generated iid from t-copula: $\theta = -0.1$, $\nu = 10$ – fixed





Summary & implications

Research questions we tackled:

(1) How different is $\bar{\pi}_n^D$ from $\pi_n^{D,\text{ideal}}$?

$\Longrightarrow$ not very! (at least for m large)

$\Longrightarrow$ Standard sampling algorithms will approximate a sensible target!

Summary & implications

Research questions we tackled:

(1) How different is $\bar{\pi}_n^D$ from $\pi_n^{D,\text{ideal}}$?

$\Longrightarrow$ not very! (see Theorem 1)

$\Longrightarrow$ Standard sampling algorithms will approximate a sensible target!

(2) What behaviour should we expect from from $\bar{\pi}_n^D$, ?

$\Longrightarrow$ Concentration depends on choice of D and in particular $\mathbb{E}_{y_{1:m}}[\widehat{\mathrm{D}_n}]$, (κ_1) and $\mathbb{V}_{y_{1:m}}[\widehat{\mathrm{D}_n}]$, (κ_2)

$\Longrightarrow$ $\bar{\pi}(\theta \in \cdot \mid \widehat{\mathrm{D}}_{m,n})$ concentration depends on bias & variance!

$\Longrightarrow$ choice of $\widehat{\mathrm{D}}_{m,n}$ impacts concentration! $\rightarrow$ Concentration occurs at slower of

$$\sqrt{n} \quad - \quad \text{normal parametric rate}$$

$$m^{-\kappa_1} \quad - \quad \text{Bias of estimator for } \widehat{\mathrm{D}}_{m,n}$$

$$m^{-\kappa_2} \quad - \quad \text{Variance of estimator for } \widehat{\mathrm{D}}_{m,n}$$

Summary & implications

Research questions we tackled:

(3) Can this inform/justify methodology?

$\Longrightarrow$ Pick D and $\widehat{\mathrm{D}}$ with small bias and low variance: MMD^2✓, KDE^2 X

$\Longrightarrow$ Given D and $\widehat{\mathrm{D}}_{m,n}$ when no bias $m \asymp \sqrt{n},$ else must choose $m \gg \sqrt{n}$

References

Generalized Bayesian Likelihood-Free Inference Using Scoring Rules Estimators, Pacchiardi, L., & Dutta, R. arXiv:2104.03889 (2022)
A general framework for updating belief distributions, Bissiri, P.G., Holmes, C., & Walker, S.G. (2016)
Principles of Bayesian Inference using General Divergence Criteria, Jewson, J., Smith, J., & Holmes, C., (2016)
The pseudo-marginal approach for efficient Monte Carlo computations, Andrieu, C. & Roberts, G. O. (2009)
Robust Generalised Bayesian Inference for Intractable Likelihoods, Matsubara, T., Knoblauch, J., Briol, F.-X., & Oates, C. JRSSB (2022)
Generalised Bayesian Inference for Discrete Intractable Likelihood, Matsubara, T., Knoblauch, J., Briol, F.-X., & Oates, C. arXiv:2206.08420 (2022)
An Optimization-centric View on Bayes' Rule: Reviewing and Generalizing Variational Inference, Knoblauch, J., Jewson, J., & Damoulas, T., JMLR (2022).
Robust Bayesian Inference for Simulator-based Models via the MMD Posterior Bootstrap, Dellaporta, C., Knoblauch, J., Damoulas, T., & Briol, F.-X. AISTATS (2022)
Generalised Posteriors in Approximate Bayesian Computation, Schmon, S.M., Cannon, P., & Knoblauch, J., AABI (2020)