Theoretical Foundations of Post Bayes Belief Updates: Asymptotic Analysis

Post-Bayes Workshop

David T. Frazier (Monash)

25/02/25

A quick promotion

Main topics to cover

1. Gibbs Measures (aka Generalised Bayesian belief updates)
2. Concentration
    1. “Old School” - Limit theory
    2. “New School” - PAC-Bayes (V.cool, read Pierre’s papers on it)
3. Coverage results and posterior asymptotic normality results
4. Examples/Applications of theory.

Key Papers

Concentration

1. Syring and Martin (2023): arxiv:2012.04505
2. Alquier and Ridgeway (2020): arXiv:1706.09293.

Asymptotic Normality

1. Miller (2021): https://www.jmlr.org/papers/volume22/20-469/20-469.pdf
2. Martin and Syring (2022): arxiv:2203.09381

“Applications” of theory

1. Prediction: McLatchie et al (2024) (arxiv:2408.08806)
2. Estimated loss functions: Frazier et al. (2024) (arXiv:2404.15649)

Standard Bayesian beliefs

Standard Bayes posterior

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

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

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

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 Bayesian beliefs

Gibbs measure

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

Examples: $\mathrm{D}_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} D(y_i, \theta), \quad \mathrm{D}_n(\theta) = \mathrm{D}(P_n, P_\theta)$

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)

Is $\pi(\theta \mid \mathrm{D}_n)$ a reasonable set of beliefs?

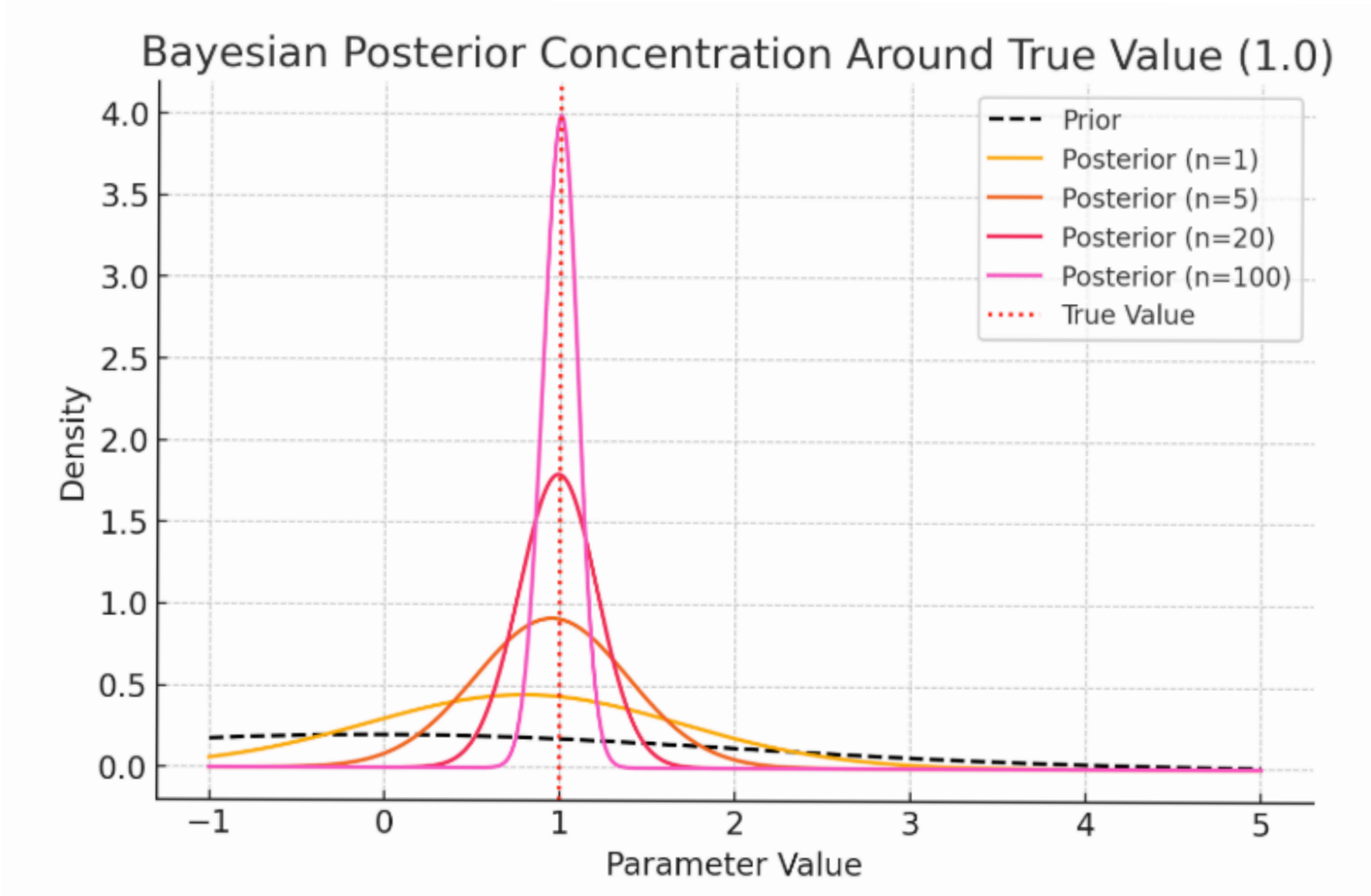
- Are inferences based on $\pi(\theta \mid \mathrm{D}_n)$ are “reliable/reasonable/useful”?
- One way we measured “reasonable-ness” is to consider “large sample”/ “average”/“high-probability” behaviour.
  - Why?
  - Std. Bayes has nice regular behaviour. Want something similar for $\pi(\theta \mid \mathrm{D}_n)$.
  - **Posterior Concentration**: Want $\pi(\theta \mid \mathrm{D}_n)$ to assign mass to regions where loss is small.
  - **Posterior Normality (asymp)**: $\pi(\theta \mid \mathrm{D}_n)$ should be roughly Gaussian in large samples (Bernstein-von Mises phenomenon).

Posterior Concentration

First Basic Requirement

Posterior Concentration

- Most basic property one could really want.
- More data = More precise inferences
- Posteriors become more peaked



How to formalise?

Clearly, nothing like ``truth" for Gibbs measures. Population loss minimiser:

$$\theta^\star := \arg\min_{\theta \in \Theta} \mathbb{E}\left[D_n(\theta)\right]$$

Hope is that $\pi(\theta \mid \mathrm{D}_n)$ concentrates onto $\theta^\star$.

Definition. The Gibbs posterior $\pi(\theta \mid \mathrm{D}_n)$ concentrates around $\theta_\star$ at rate (at least) ε_n, with respect to a metric $d(\theta, \theta')$, if

$$\mathbb{E}\,\Pi\left(\theta : d\left(\theta; \theta^\star\right) > M_n \varepsilon_n \mid \mathrm{D}_n\right) \to 0, \quad \text{as } n \to \infty, \quad M_n > 0$$

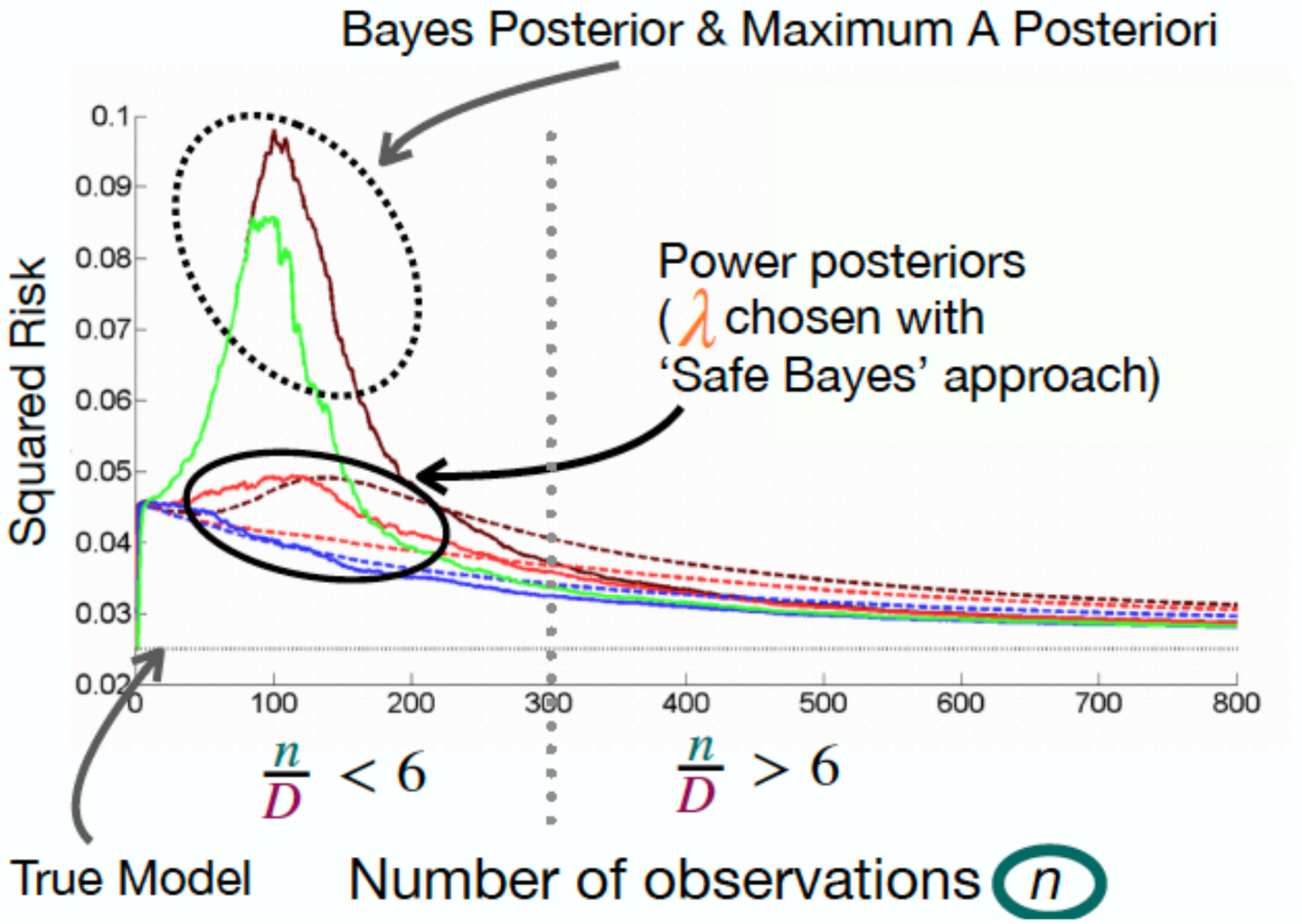
where $M_n \to \infty$ arbitrarily slowly or is a sufficiently large constant.

Why should I care?

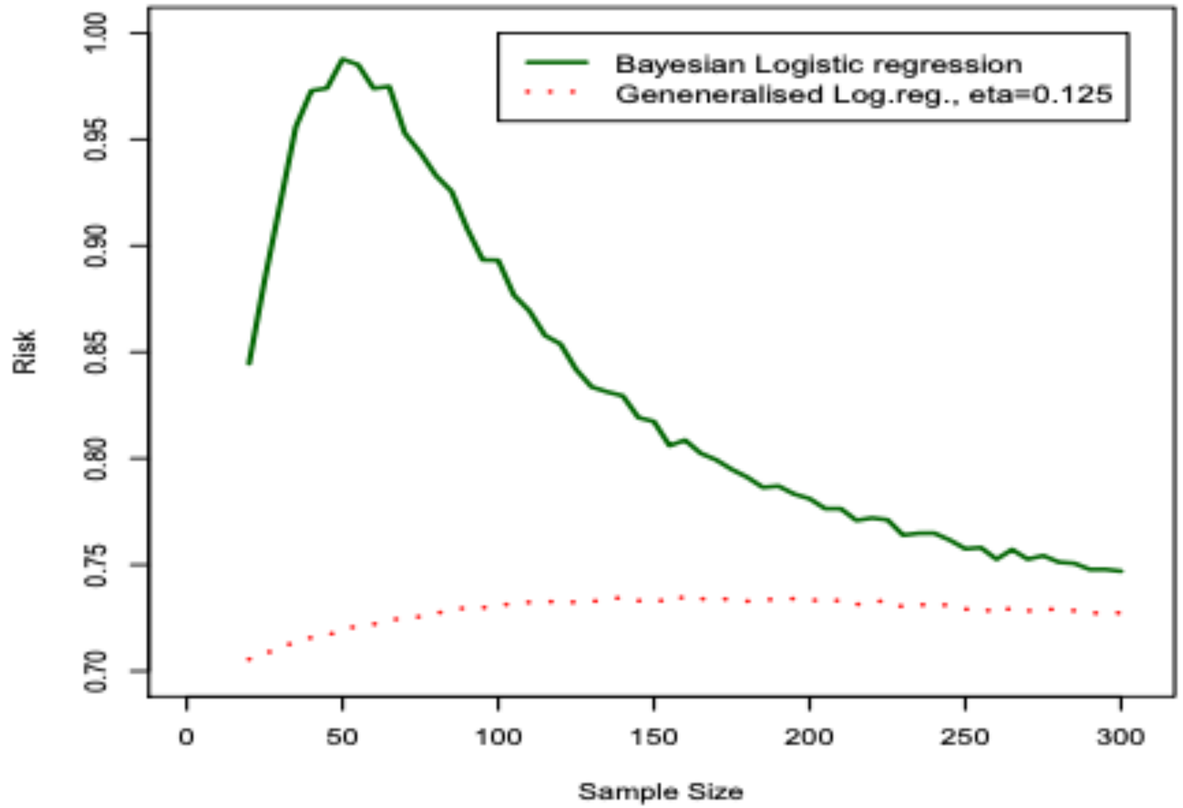
- ϵ_n tells us how fast we can expect our posterior inferences to concentrate around $\theta^\star$ - for a fixed learning rate.
- Makes clear the link between prior and its impact on the posterior: bad priors lead to slower concentration.
- Imbeds link between model size and sample size: larger model, slower ϵ_n
  - For example: if $\theta \in \Theta \subseteq \mathbb{R}^d$, then $\epsilon_n \asymp 1/\sqrt{n}$
  - If $d \to \infty$ as $n \to \infty,$ then ϵ_n is slower.

Concentration Examples

The **'Safe Bayes' effect** (see Grünwald, 2012)
(picture from Grünwald & van Ommen, 2017)



Concentration Examples



When is this likely to be satisfied?

Two Key Conditions: Prior mass and well-behaved loss

Prior mass condition: for certain sets $\mathscr{B}_n$ with radius with radius ϵ_n^r, $r > 0$,

$$\Pi[\theta \in \mathscr{B}_n] \geq e^{-n\epsilon_n^r}$$

- The sets $\mathscr{B}_n$ depend on the expected loss $\mathrm{D}(\theta) = \mathbb{E}[D(Y, \theta)]$ and depend on the type of argument used in the proof.
  - Common sets: $\mathscr{B}_n := \left\{\theta : m\left(\theta, \theta^\star\right) \vee v\left(\theta, \theta^\star\right) \leq \varepsilon_n^r\right\}$,
  - $m(\theta, \theta^\star) = D(\theta) - D(\theta^\star)$
  - $\nu(\theta, \theta^\star)$ is like variance of loss diff.
  - If $\nu(\theta, \theta^\star)$ is bounded, then $\mathscr{B}_n := \left\{\theta : m\left(\theta, \theta^\star\right) \leq \varepsilon_n^r\right\}$

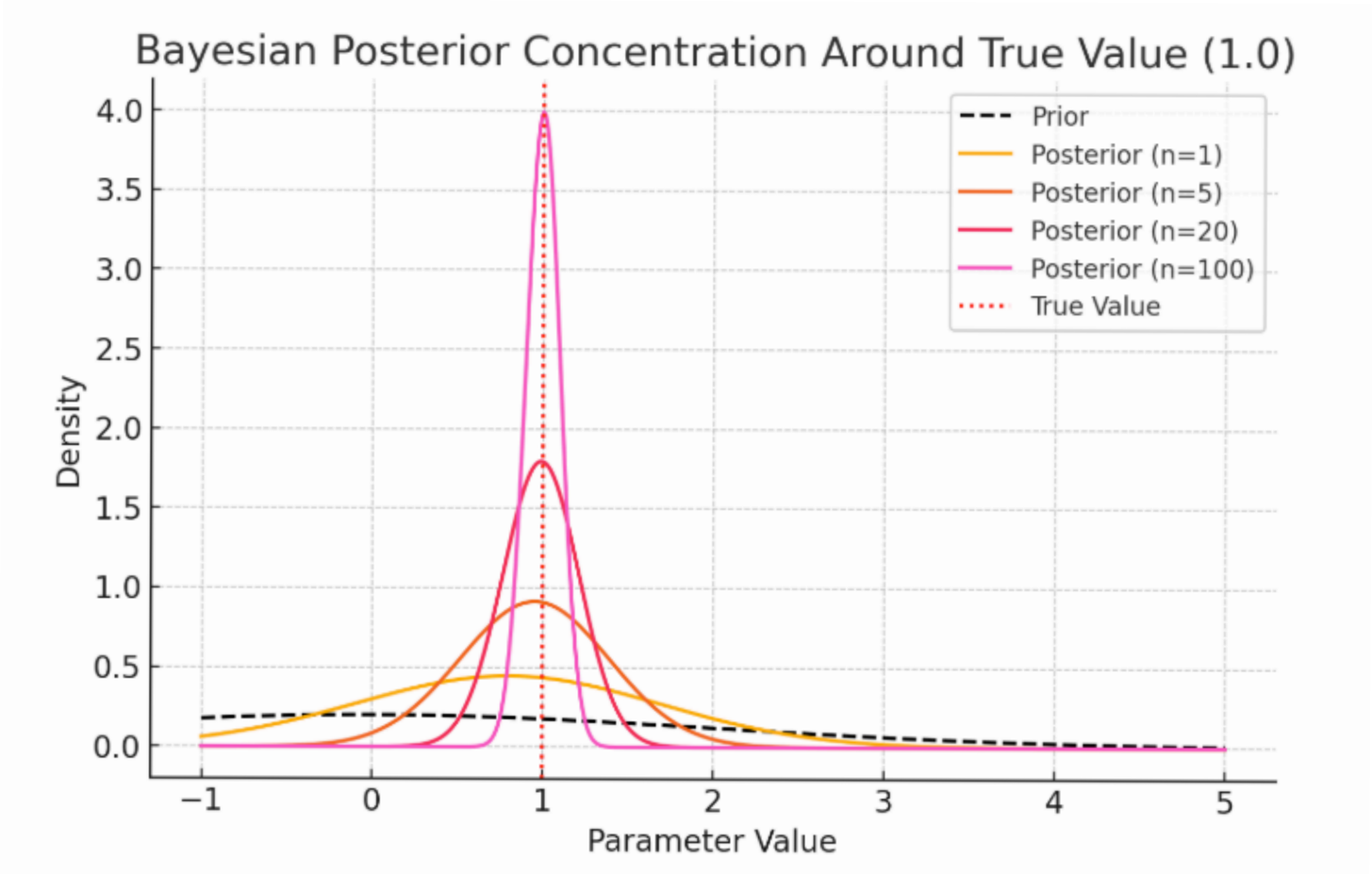
Well-Behaved loss condition

- Gibbs measures look like $e^{\{-n\omega D_n(\theta)\}} \times \pi(\theta)$
- Hence, we really need $e^{\{-n\omega[D_n(\theta)-D_n(\theta^\star)]\}}$ to exist, and be well-behaved.
- Many different ways to accomplish this:
  - Loss has an exponential moment: $\mathbb{E}e^{\{-\omega[D(Y,\theta)-D(Y,\theta^\star)]\}} < 1$
  - Hoeffding or Bernstein-type condition: $\int_\Theta \mathbb{E}e^{\{-n\omega[D_n(\theta)-D_n(\theta^\star)]}\pi(\theta)d\theta < e^{\omega^2 C/n}$
  - Uniform CLT + Well-separated points: for all $s_n > 0$, s.th. $\omega_n n s_n^2 \to \infty$,

$$\mathbb{P}\left[\sup_{\theta:d(\theta,\theta^\star)\geq s_n} n\omega_n\left\{D_n(\theta) - D_n\left(\theta^\star\right)\right\} > -c_1 n s_n^2 \omega_n\right] = o(1)$$

  - Similar concentration rate under each of these…

In Summary: Well-behaved loss + Prior Mass= Posterior concentration



What about intractable loss functions?

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

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

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

Intractable expectations

Can 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: accounting for model samples

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

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

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

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

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

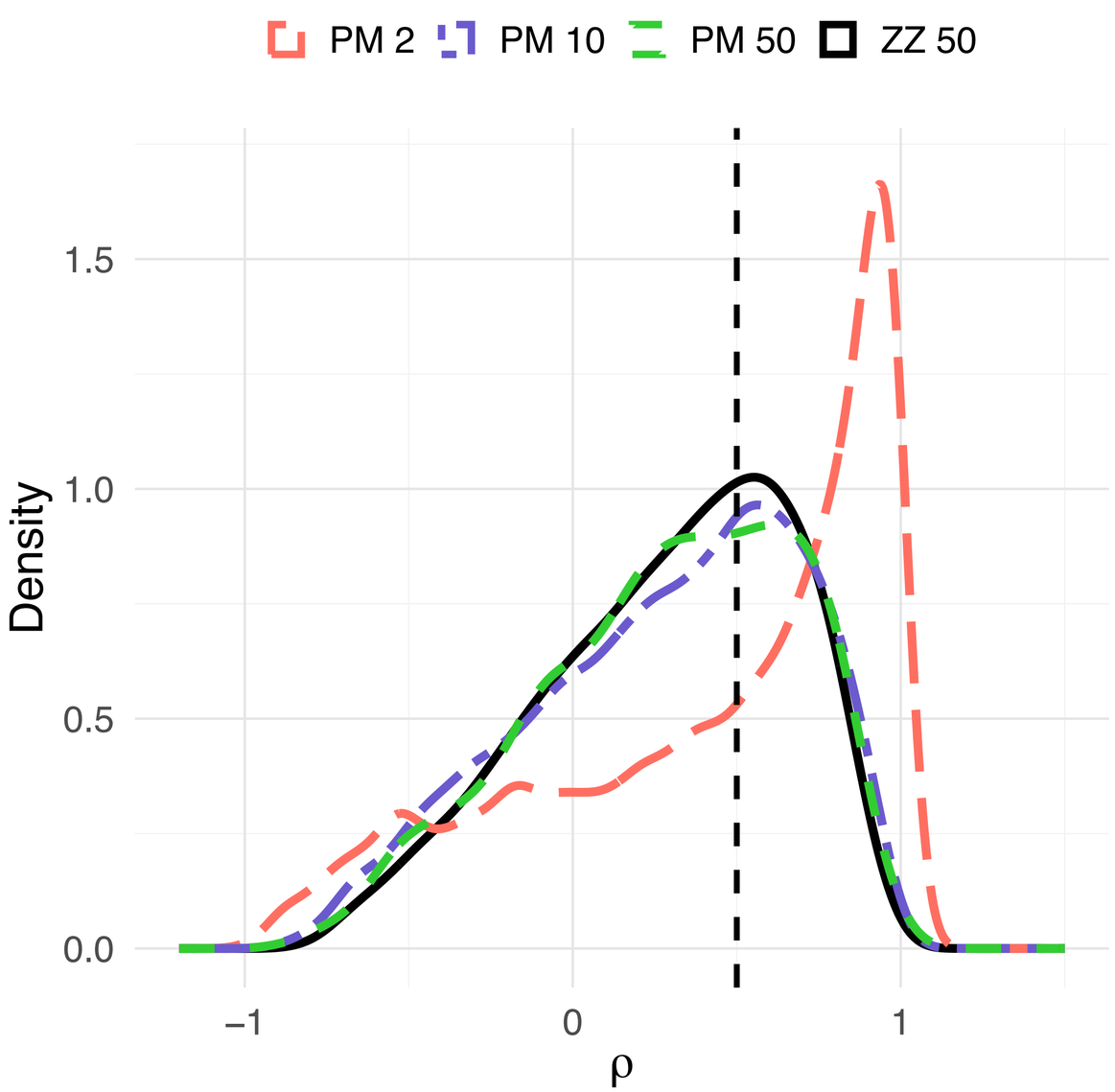
Research questions:

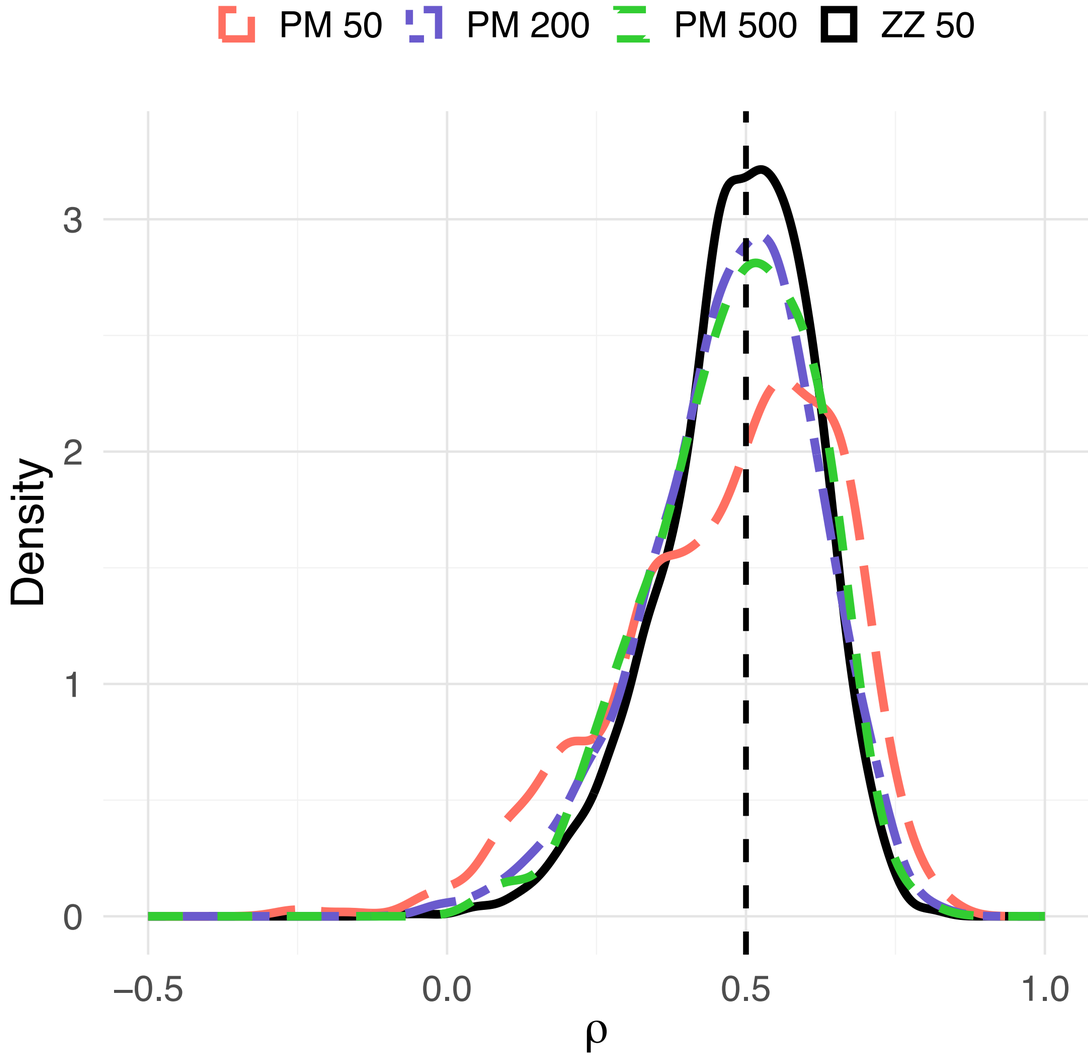
(1) Quantify how differnce b/t $\pi(\theta \mid \mathrm{D}_n)$ and $\bar{\pi}(\theta \mid \widehat{\mathrm{D}}_{m,n})$?

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

Does loss estimation matter?

Data generated iid from Gaussian copula $\rho = 0.5$





Concentration: accounting for loss estimation

Theorem 2:

Under a Bernstein condition on the loss, and other regularity conditions, if $m, n \to \infty$, for $M_n > 0$, possibly $M_n \to \infty$ slowly and $m = m(n) \to \infty$

$$\mathbb{E}\left(\int_{\Theta} |D(\theta) - D(\theta^\star)| \; \pi(\theta \mid \widehat{D}_{m,n}) d\theta \;\; > \;\; \frac{\log(n) M_n}{\min\{n^{1/2}, m^{1/2}\}} \right) \longrightarrow 0.$$

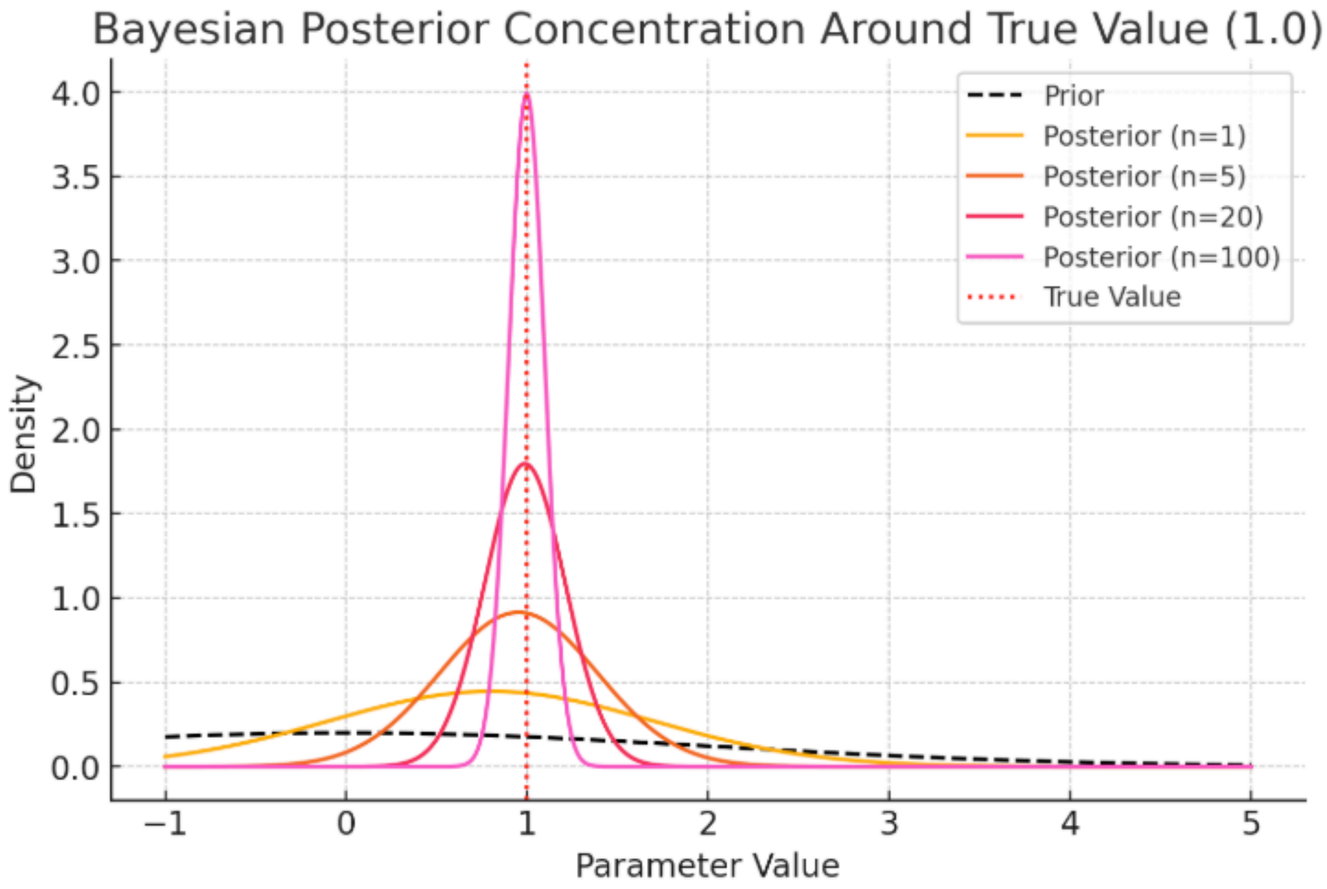
Asymptotic Posterior Normality (BvM)

A Next Step

- Concentration is helpful, but if $\theta \in \Theta \subseteq \mathbb{R}^d$, then $\epsilon_n \approx 1/\sqrt{n}$, doesn't really tell us much…
- Need something more informative.
- What behaviour should we expect?

A Next Step

- Concentration is helpful, but if $\theta \in \Theta \subseteq \mathbb{R}^d$, then $\epsilon_n \approx 1/\sqrt{n}$, doesn't really tell us much…
- Need something more informative.
- What behaviour should we expect?
- What do you see?



Why BvM? Brief History Lesson

- Need some external notion of reliability with which to compare posteriors.
- Enter BvM: “older” view…

Why BvM? Second order behavior…

- In large samples, and parametric models, $\pi(\theta \mid D_n)$, behaves like
- $\pi(\theta \mid \mathrm{D}_n) \propto |\det\{\mathrm{H}_n(\theta_n)\}|^{-1/2} \exp\left\{-\frac{\omega n}{2}(\theta-\theta_n)^\top \mathrm{H}_n(\theta_n)(\theta-\theta_n) + R_n(\theta,\theta_n)\right\}\pi(\theta),$
- For some remainder term $R_n(\theta)$ that can be suitably controlled, and $\mathrm{H}_n(\theta) = \nabla^2_\theta D_n(\theta)$. (Note: $\mathrm{H}_n(\theta_n)$ must be psd, fine as a min!)
- This follows from a 2nd-order TSE of $D_n(\theta)$ around θ_n or $\theta^\star$
- $D_n(\theta) - D_n(\theta_n) = (\theta-\theta_n)^\top \nabla_\theta D_n(\theta_n) + \frac{n}{2}(\theta-\theta_n)^\top \mathrm{H}_n(\theta_n)(\theta-\theta_n) + R_n(\theta,\theta_n)$
- Therefore, variability of $\pi(\theta \mid D_n)$ is determined by $\mathrm{H}_n(\theta) = \nabla^2_\theta D_n(\theta)$

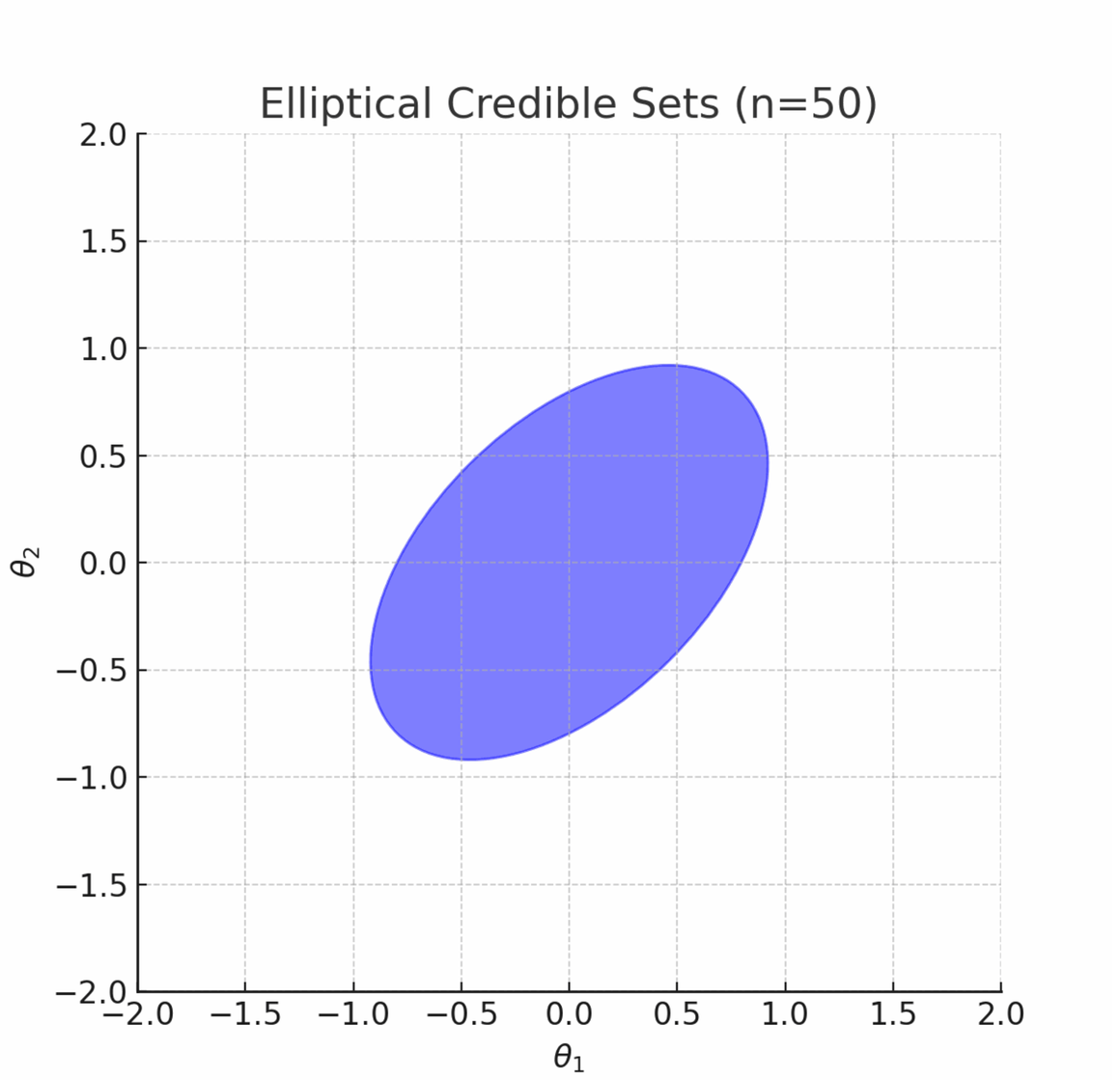
Why BvM? Second order behavior

Requirements

- Essentially, any loss that admits a well-behaved quadratic expansion in $\sqrt{n}(\theta - \theta^\star)$ will ensure the Gibbs measure is asymptotically normal.
  - This is not strictly necessary as this can be weakened using equi-continuity
  - Lots of examples satisfy this condition. From Miller (2021) (arxiv:1907.09611)
    - Losses like: composite likelihoods, quasi-likelihoods,
    - Specific examples: GLMs, Gaussian Markov random fields, Boltzmann machines, Cox models
  - And many more losses/models will satisfy the conditions necessary for a BvM.

Visually?

- For reasonable sample sizes, the prior washes out.
- Posterior concentrates onto a single point in model space.
- Credible sets shrink like $\mathrm{H}_n^{-1}/\sqrt{n\omega^2}$

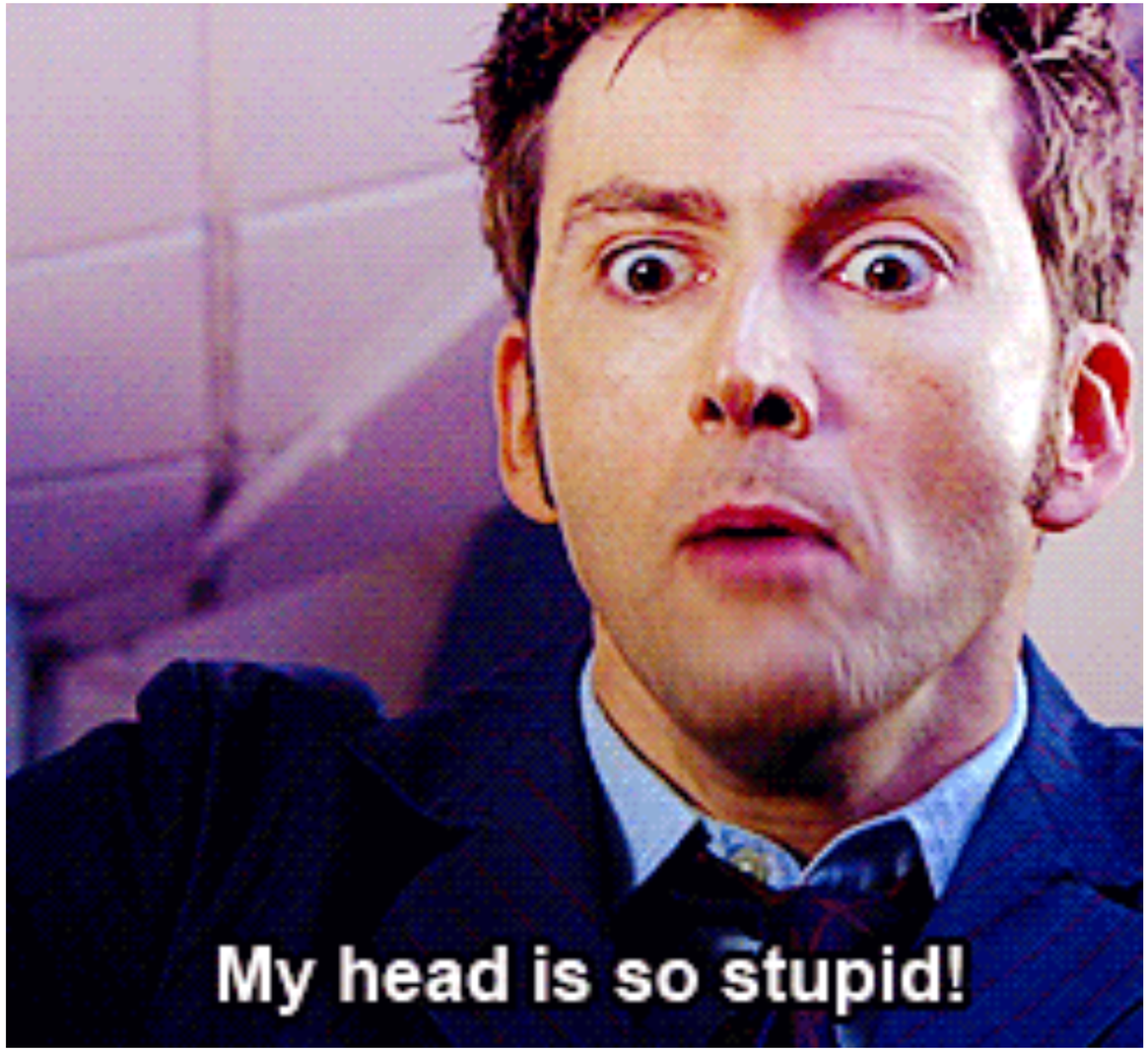


Implications?

- A Gibbs-based credible set for $\theta^\star$, say $C^{\mathrm{H}_n}_{1-\alpha}$, has width proportional to $\mathrm{H}_n^{-1}/\sqrt{n\omega^2}$.
- $\Pr\left\{\theta^\star \in C^{\mathrm{H}_n}_{1-\alpha}\right\} = 1-\alpha$?
- Correct width would be $\mathrm{H}_n^{-1}\mathrm{Var}\{\nabla_\theta\sqrt{n}D_n(\theta^\star)\}\mathrm{H}_n^{-1}$
- Any hope? Generalised information matrix identity: for some $\omega > 0$, $\mathrm{Var}\{\nabla_\theta\sqrt{n}D_n(\theta^\star)\}\mathrm{H}^{-1} = \omega \cdot \mathrm{I}$,
- Ryan Martin will talk about how we can find such an ω!
- May be a way around this for smooth losses: Frazier, et al. (2023) arXiv:2311.15485

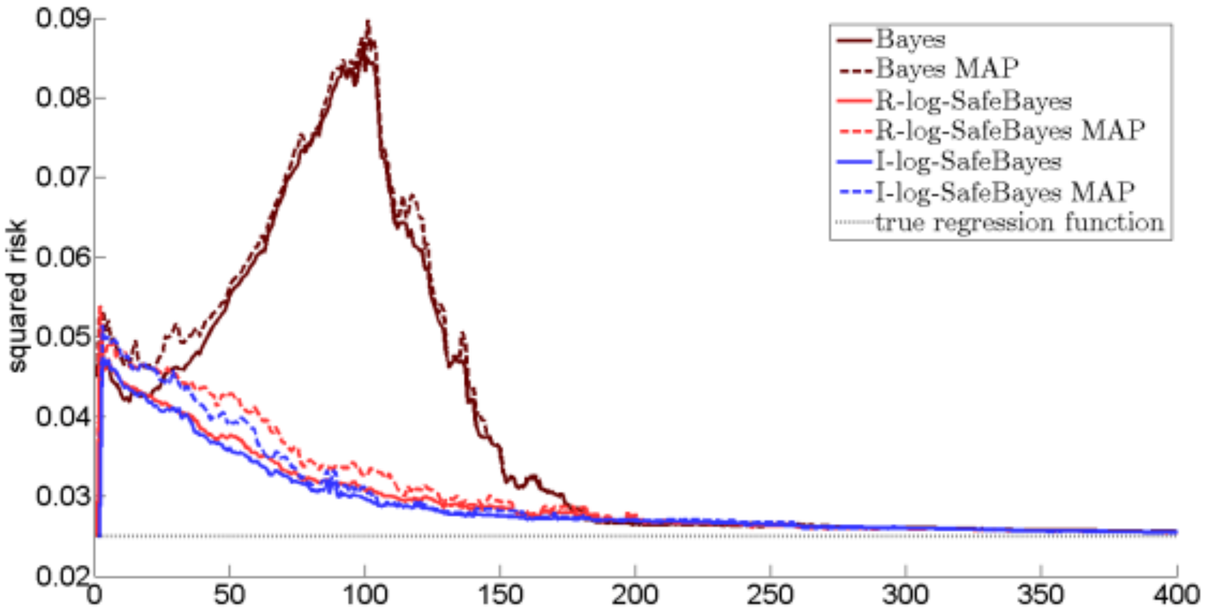
Why BvM? Brief History Lesson

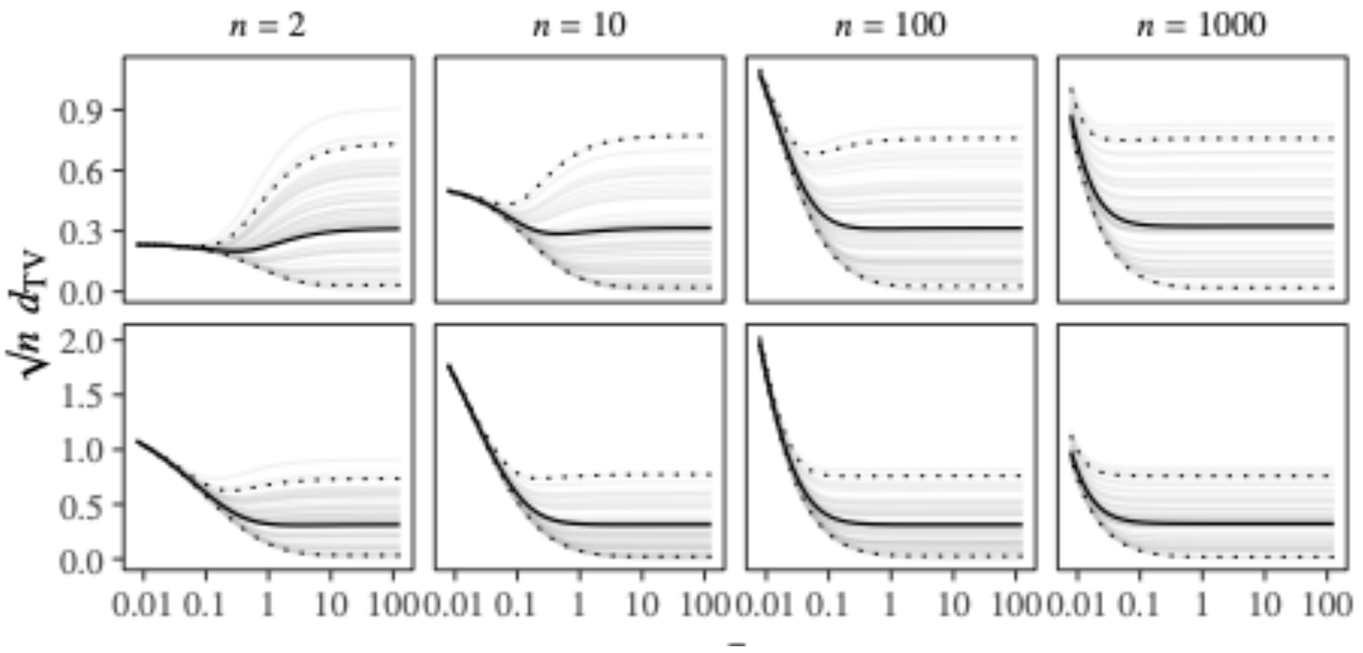
- Need some external notion of reliability with which to compare posteriors.
- Old view of BvMs…
- Newer view of BvMs…



Does it matter if we only care about prediction?

- Learning rates matter for inference…
- But what about prediction…
- Seems less sensitive, especially in large samples.
- More important is $\theta^\star$ and model!





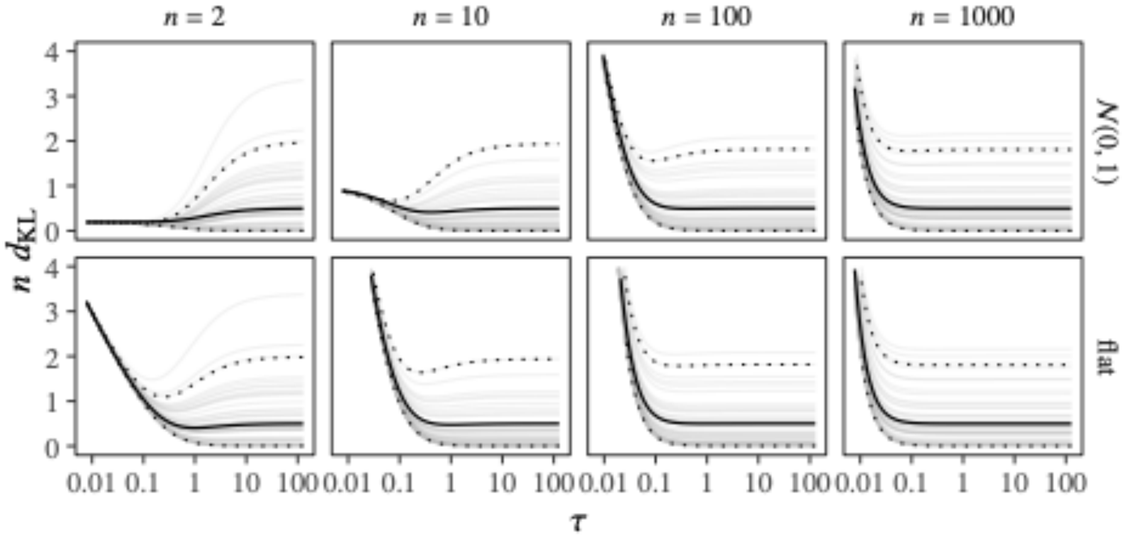


Figure B.1. Normal location example.

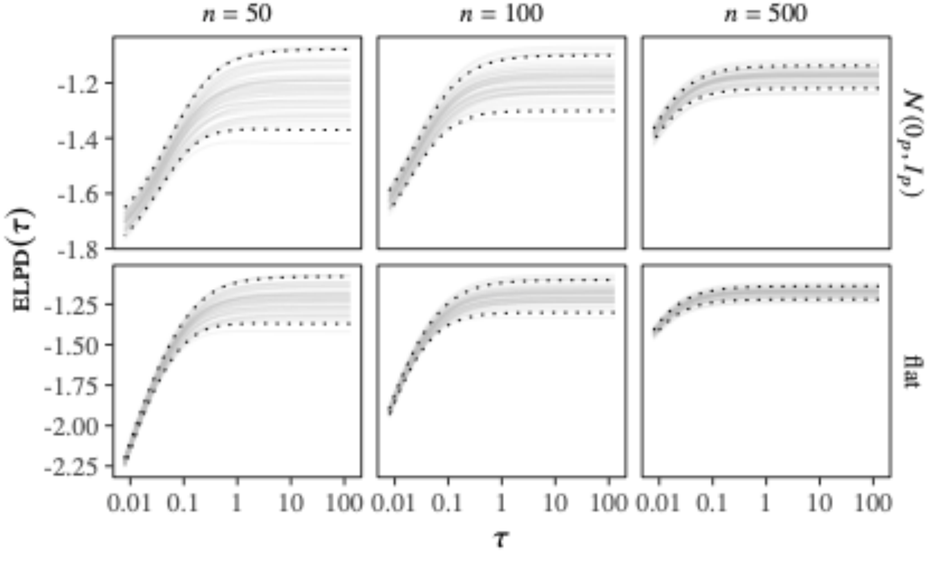
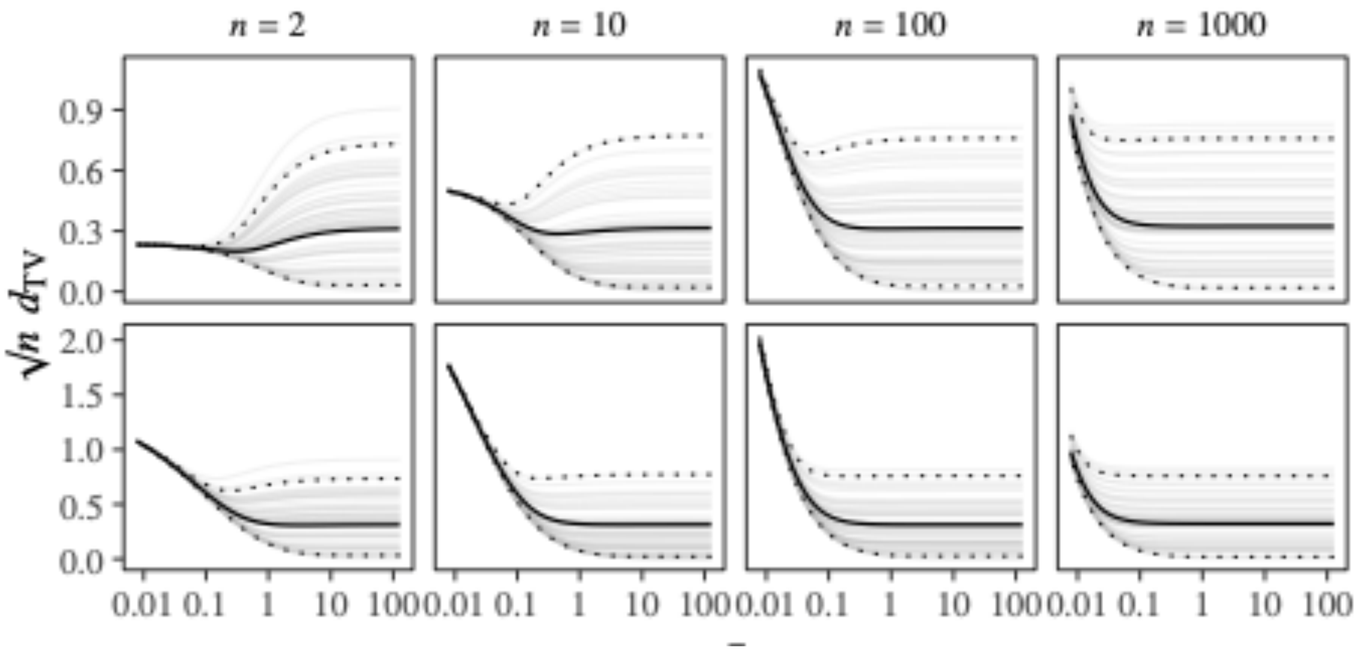


Figure B.2. Misspecified linear regression example.



Important in smaller samples, definitely!

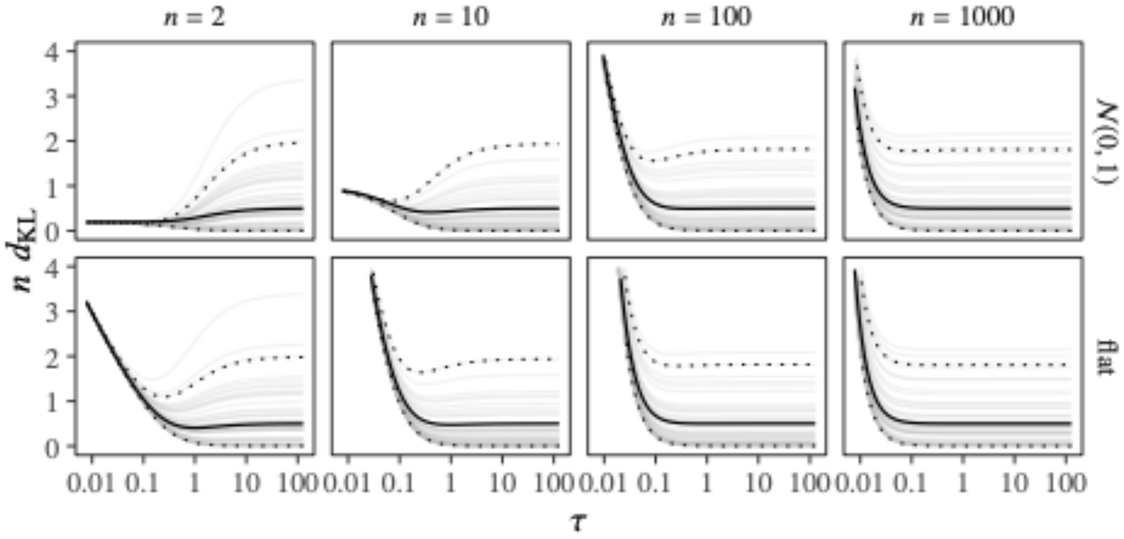


Figure B.1. Normal location example.

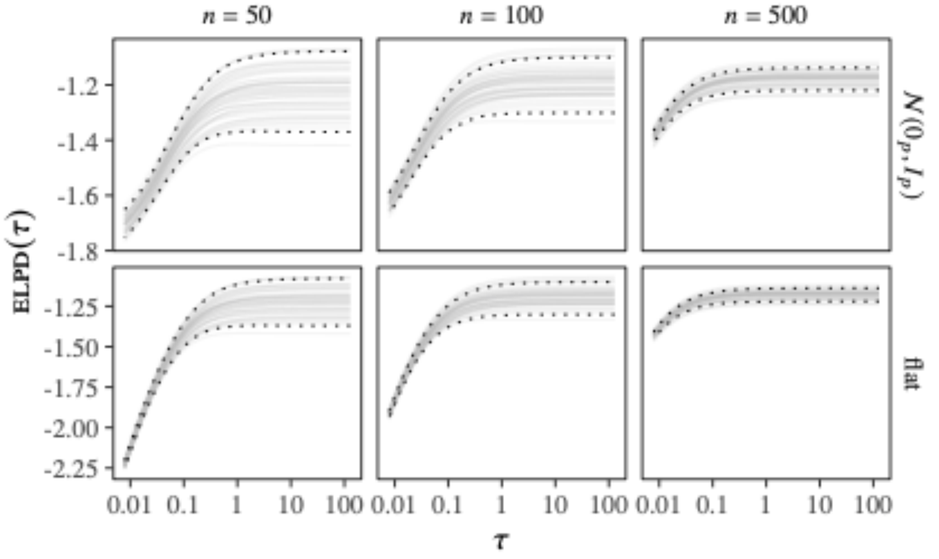
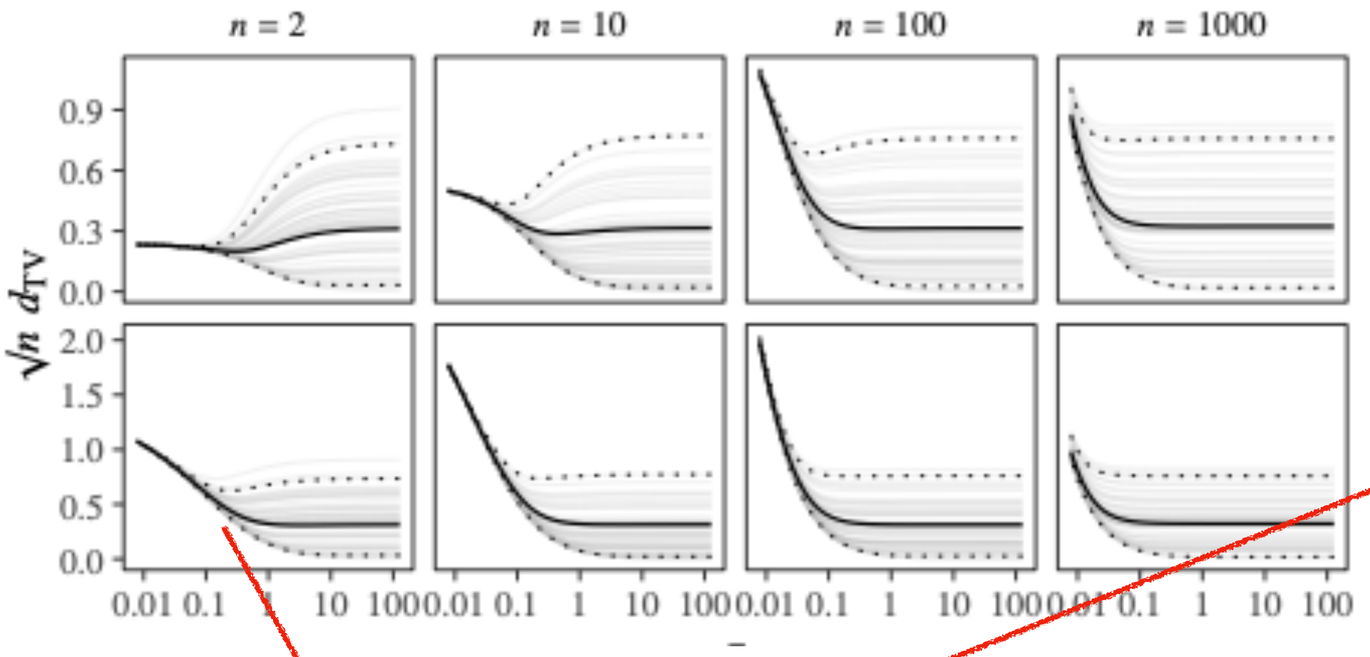


Figure B.2. Misspecified linear regression example.



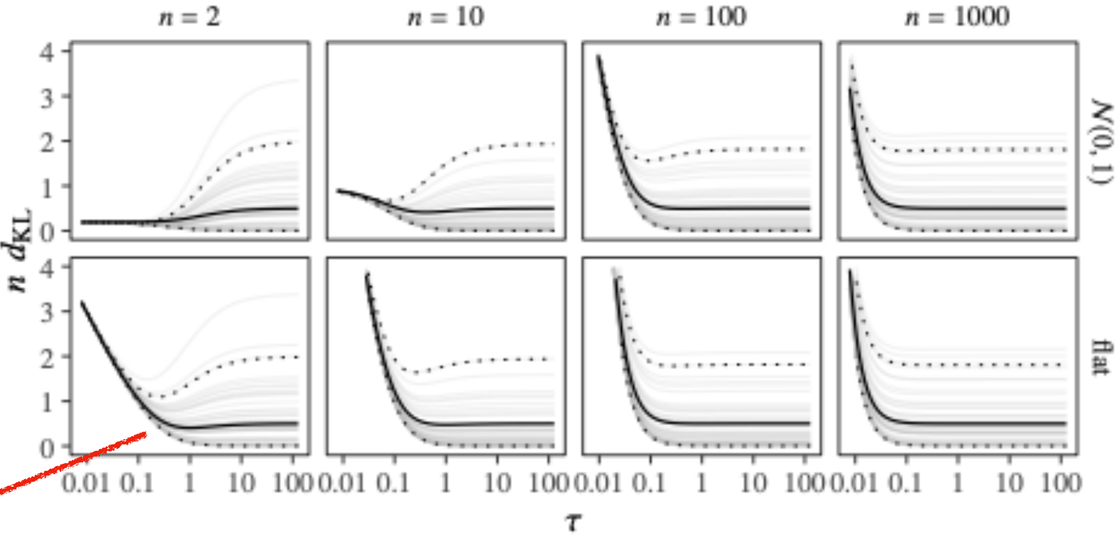


Figure B.1. Normal location example.

Important in smaller samples, definitely!

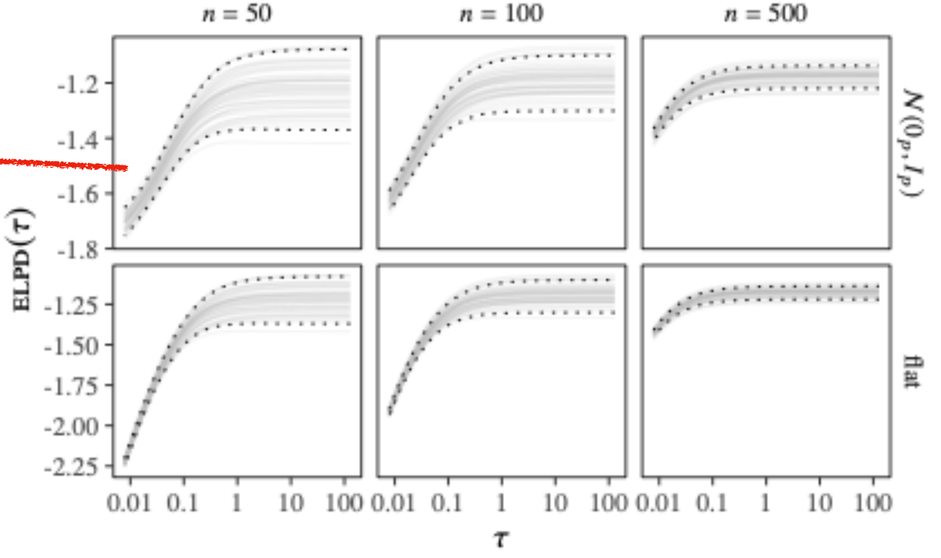
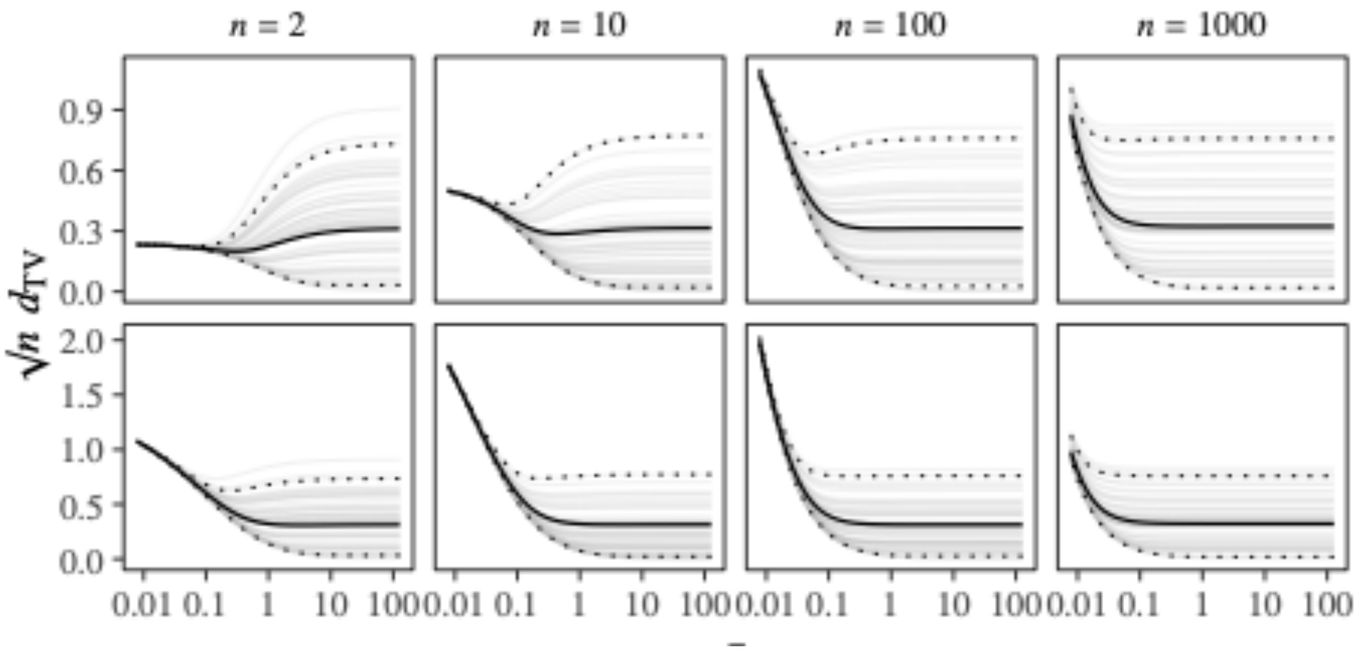


Figure B.2. Misspecified linear regression example.



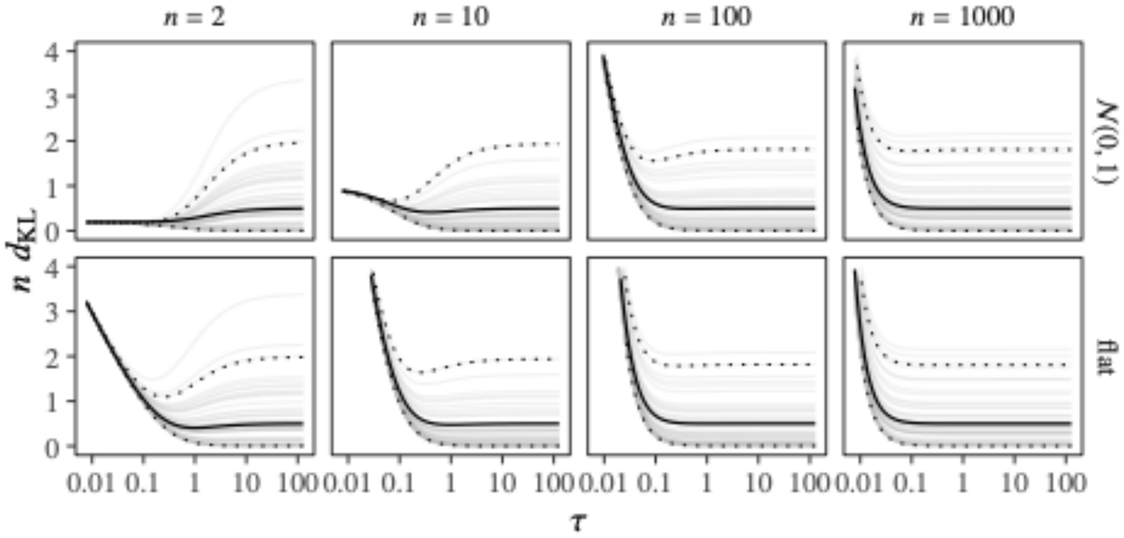


Figure B.1. Normal location example.

Important in smaller samples, definitely!

But not really in large samples!

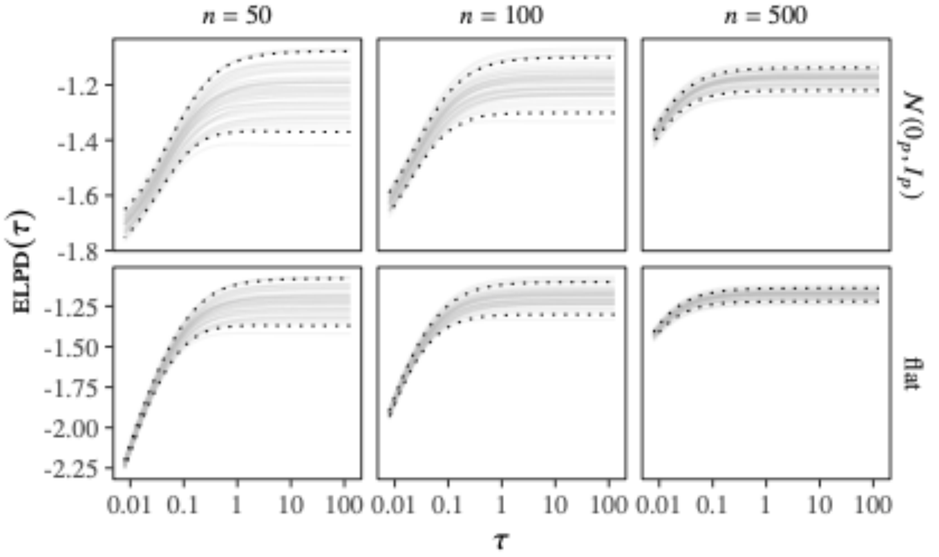
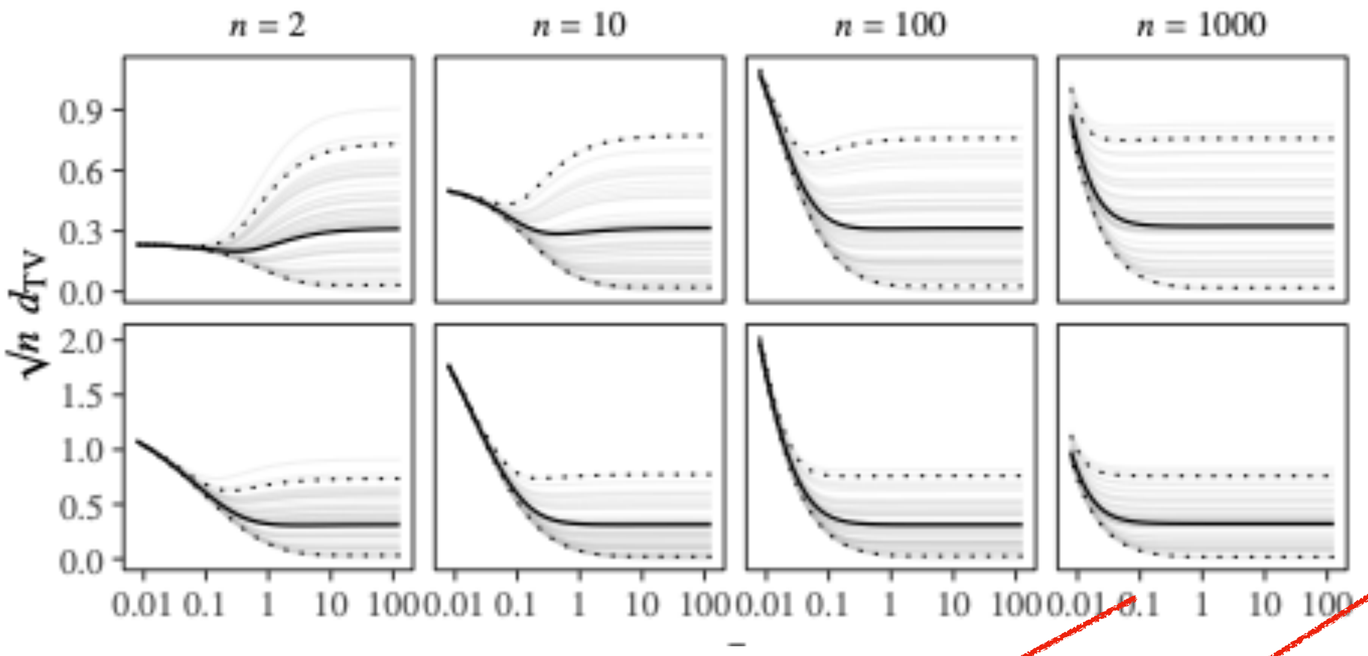


Figure B.2. Misspecified linear regression example.



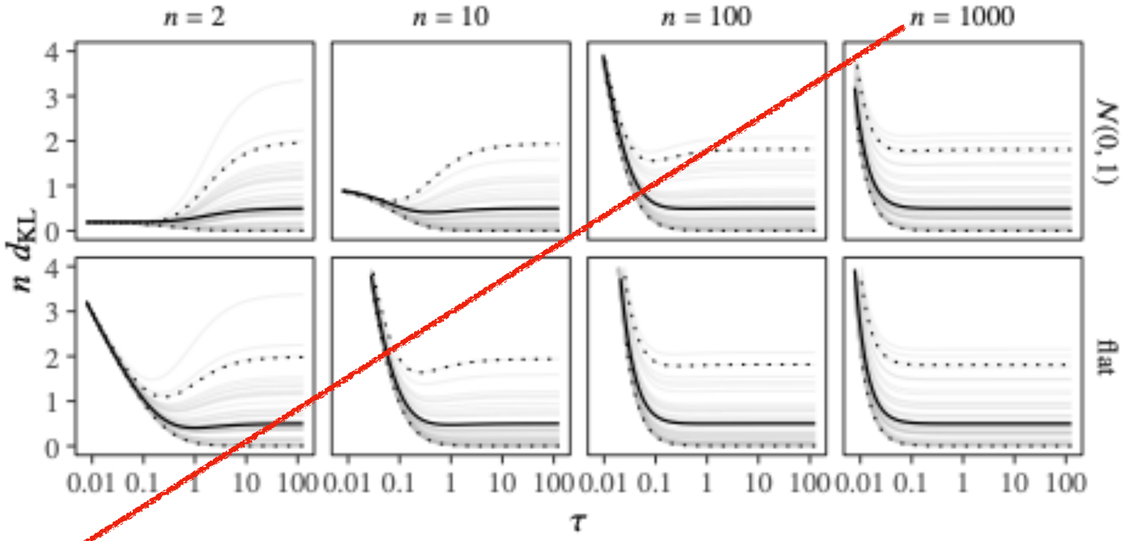


Figure B.1. Normal location example.

Important in smaller samples, definitely!

But not really in large samples!

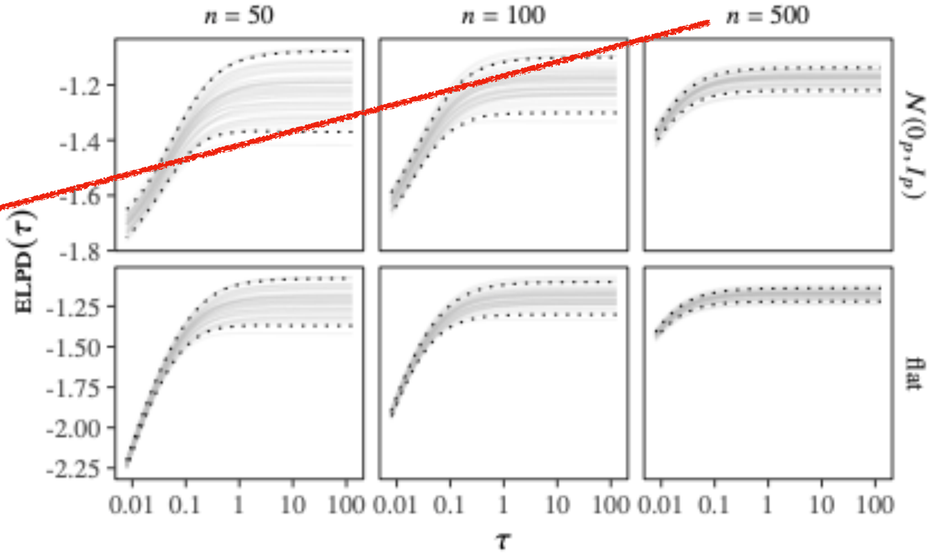


Figure B.2. Misspecified linear regression example.

What's behind this behaviour?

Posterior Concentration!

Result. For $\omega_n \geq 0$, even $\omega_n \to 0$, such that , $n\omega_n\epsilon_n^2 \to \infty$, $\omega_n \gg \log(n)/n$, the Gibbs posterior predictive

$$p_n^{(\omega)}(\,\cdot \mid y_{1:n}, \mathrm{D}_n) = \int_\Theta f_\theta(\,\cdot \mid y_{1:n})\pi^{(\omega_n)}(\theta \mid \mathsf{D}_n)d\theta,$$

satisfies

$$\mathbb{E}\,\left[d_{\mathrm{TV}}\left\{f_{\theta^\star}\left(\,\cdot \mid y_{1:n}\right), p_n^{(\omega)}\left(\,\cdot \mid y_{1:n}, \mathrm{D}_n\right)\right\}\right] \leq \varepsilon_n + o(1)\,.$$

Similar result achievable in KL divergence as well.

We use asymptotics not to obtain direct knowledge of the phenomenon we are observing, but to save ourselves the embarrassment of believing in easy and ultimately verifiably false truths.

- Inspired by Thomas Carlyle, Chartism, Chapter II, Statistics: