Quantifying Observed Prior Impact

David Jones

Texas A&M University

Joint work with Robert Trangucci and Yang Chen (UMich)

November 17, 2020

Introduction

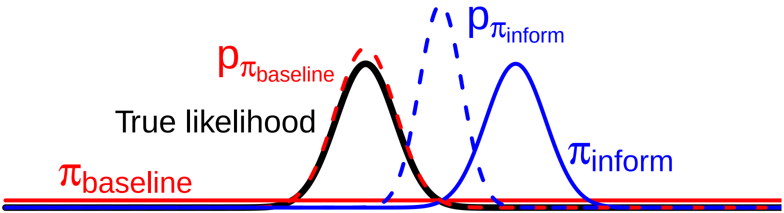
$$p(\theta|y) \propto f(y|\theta)\pi(\theta)$$

Questions:

1) How much information does the prior contain?
2) What is the effect of the prior?
   (a) Average effect
   (b) Effect for data at hand

Appealing approach: effective prior sample size

Data Dependent Prior Impact



Multiple Instrument Motivating Example

Multiple Instrument Example

Goal: combine flux estimates / calibrate instruments

Chen et al. (2019) considered the model:

$$y_{ij} = -0.5\ \sigma_{ij}^2 + B_i + G_j + e_{ij}, \qquad e_{ij} \overset{\text{indep}}{\sim} N(0, \sigma_{ij}^2),$$

where

- i indexes instruments
- j indexes sources
- y_{ij} = log photon counts for instrument i and source j
- B_i = log Effective Area of instrument i
- G_j = log Flux of source j

Multiple Instrument Example

Goal: combine flux estimates / calibrate instruments

Chen et al. (2019) considered the model:

$$y_{ij} = -0.5\ \sigma_{ij}^2 + B_i + G_j + e_{ij}, \qquad e_{ij} \overset{\text{indep}}{\sim} N(0, \sigma_{ij}^2),$$

where

- i indexes instruments
- j indexes sources
- y_{ij} = log photon counts for instrument i and source j
- B_i = log Effective Area of instrument i
- G_j = log Flux of source j

Problem: B_i and G_j are not initially identifiable

Multiple Instrument Example: Instrument Specific Priors

Identifiable due to prior information:

$$B_i \sim \mathcal{N}(b_i, \tau_i^2), \quad G_j \sim \text{flat on real line}$$

- Weighting of data from instrument i depends on τ_i^2, in a joint analysis to estimate the G_j

Question: are some priors driving the final result much more than others?

Existing Ways to Measure Prior Impact

Approach 1: Match Prior to Hypothetical Previous Posterior

Approach 1: Match Prior to Hypothetical Previous Posterior

Baseline prior: $\pi_{\text{base}}(\mu) \propto 1$

Model: $y_1, \ldots, y_n \overset{iid}{\sim} N(\mu, \sigma^2)$

Baseline posterior: $\mu \sim N\left(\bar{y}, \frac{\sigma^2}{n}\right)$

Approach 1: Match Prior to Hypothetical Previous Posterior

Baseline prior: $\pi_{\text{base}}(\mu) \propto 1$

Model: $y_1, \ldots, y_n \overset{iid}{\sim} N(\mu, \sigma^2)$

Baseline posterior: $\mu \sim N\left(\bar{y}, \frac{\sigma^2}{n}\right)$

- Suppose informative prior is π_{inform}: $\mu \sim N(\mu_0, \tau^2)$

Approach 1: Match Prior to Hypothetical Previous Posterior

Baseline prior: $\pi_{\text{base}}(\mu) \propto 1$

Model: $y_1, \ldots, y_n \overset{iid}{\sim} N(\mu, \sigma^2)$

Baseline posterior: $\mu \sim N\left(\bar{y}, \frac{\sigma^2}{n}\right)$

- Suppose informative prior is π_{inform}: $\mu \sim N(\mu_0, \tau^2)$
- Choose hypothetical previous data so $\bar{y} \approx \mu_0$ and $\sigma^2/n \approx \tau^2$

Approach 1: Match Prior to Hypothetical Previous Posterior

Baseline prior: $\pi_{\text{base}}(\mu) \propto 1$

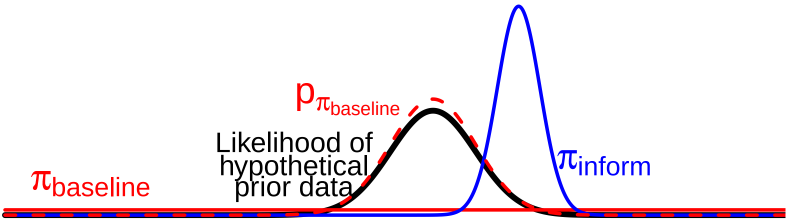
Model: $y_1, \ldots, y_n \overset{iid}{\sim} N(\mu, \sigma^2)$

Baseline posterior: $\mu \sim N\left(\bar{y}, \frac{\sigma^2}{n}\right)$

- Suppose informative prior is π_{inform}: $\mu \sim N(\mu_0, \tau^2)$
- Choose hypothetical previous data so $\bar{y} \approx \mu_0$ and $\sigma^2/n \approx \tau^2$

Problem: does not tell us anything about the actual analysis

Approach 1: Match Prior to Hypothetical Previous Posterior



- **Clarke (1996):** choose **specific hypothetical dataset** to minimize KL divergence between **hypothetical posterior** and our informative prior
- **Morita et al. (2008):** similar but chooses the sample size and **averages over the hypothetical data**

Effective Prior Sample Size

Informative prior: $\mu \sim N\left(\mu_0, \frac{\sigma^2}{k}\right)$

Model: $y_1, \ldots, y_n \overset{iid}{\sim} N(\mu, \sigma^2)$

Informative prior posterior: $\mu \sim N\left(\mu_1 = \alpha \bar{y}_{1:n} + (1-\alpha)\mu_0, \frac{\sigma^2}{n+k}\right),$

where $\alpha = \frac{n}{n+k}$

Effective Prior Sample Size

$$\textbf{Informative prior: } \mu \sim N\left(\mu_0, \frac{\sigma^2}{k}\right)$$

$$\textbf{Model: } y_1, \ldots, y_n \overset{iid}{\sim} N(\mu, \sigma^2)$$

$$\textbf{Informative prior posterior: } \mu \sim N\left(\mu_1 = \alpha \bar{y}_{1:n} + (1-\alpha)\mu_0, \frac{\sigma^2}{n+k}\right),$$

$$\text{where } \alpha = \frac{n}{n+k}$$

Interpretation:

- EPSS = k

Effective Prior Sample Size

Informative prior: $\mu \sim N\left(\mu_0, \frac{\sigma^2}{k}\right)$

Model: $y_1, \ldots, y_n \overset{iid}{\sim} N(\mu, \sigma^2)$

Informative prior posterior: $\mu \sim N\left(\mu_1 = \alpha \bar{y}_{1:n} + (1-\alpha)\mu_0, \frac{\sigma^2}{n+k}\right),$

where $\alpha = \frac{n}{n+k}$

Interpretation:

- EPSS = k

Problems:

- What if the model is not conjugate?
- If μ_0 is far from $\bar{y}$ then the actual impact of the prior on the posterior distribution can be arbitrarily large

Approach 2: Extending Effective Prior Sample Size

Baseline prior: $\pi_{\text{base}}(\mu) \propto 1$

Informative prior: $\mu \sim N\left(\mu_0, \frac{\sigma^2}{k}\right)$

Model: $y_1, \ldots, y_n \overset{iid}{\sim} N(\mu, \sigma^2)$

Baseline prior posterior: $\mu \sim N\left(\bar{y}_{1:m}, \frac{\sigma^2}{m}\right)$

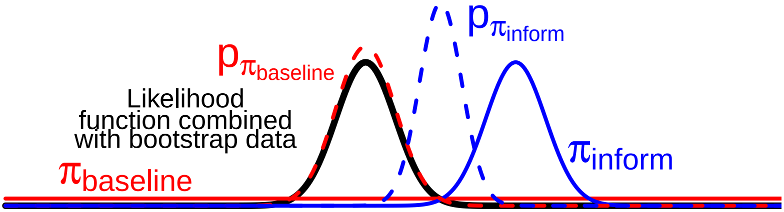
Informative prior posterior: $\mu \sim N\left(\mu_1 = \alpha\bar{y}_{1:n} + (1-\alpha)\mu_0, \frac{\sigma^2}{n+k}\right),$

where $\alpha = \dfrac{n}{n+k}$

Basic idea:

- Set $m = n + k$
- Then variance of two posteriors agree
- $\Rightarrow$ EPSS of π_{inform} is k

Approach 2: Extending Effective Prior Sample Size

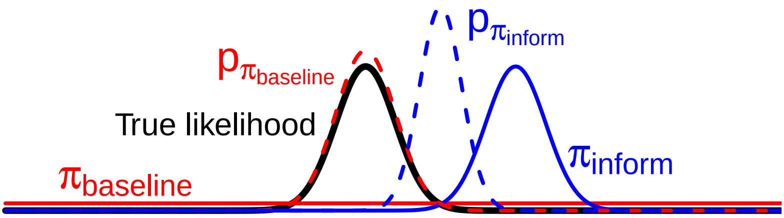


- **Reimherr et al. (2014):** how many extra samples needed to minimize the separation between posteriors?
- Not real posteriors – based on ($< n$) "bootstrap" samples
- Captures importance of prior location, but only on average

What about the prior impact for my specific dataset?

Data Dependent Prior Impact

Intuition: how many extra samples do I need to minimize the separation between the **baseline prior posterior** and the **informative prior posterior**?



General formulation

For $j = 1, \ldots, N$:

- **Generate new samples:** $y_{n+1}, \ldots, y_{M_{max}}$
- **Compute distances:** for $m = M_{min}, \ldots, n, \ldots, M_{max}$ compute

$$D_m = \text{Dist}(p(\theta | y_{1:n}, \pi_{\text{inform}}), p(\theta | y_{1:m}, \pi_{\text{base}}))$$

- **Simulation specific PSS:** $\text{PSS}_j = \underset{m}{\text{argmin}}\ D_m - n$

End for loop

Report final PSS = $\frac{1}{N} \sum \text{PSS}_j$

General formulation

For $j = 1, \ldots, N$:

- **Generate new samples:** $y_{n+1}, \ldots, y_{M_{max}}$
- **Compute distances:** for $m = M_{min}, \ldots, n, \ldots, M_{max}$ compute

$$D_m = \mathrm{Dist}(p(\theta | y_{1:n}, \pi_{\text{inform}}), p(\theta | y_{1:m}, \pi_{\text{base}}))$$

- **Simulation specific PSS:** $\mathrm{PSS}_j = \underset{m}{\mathrm{argmin}}\ D_m - n$

End for loop

Report final PSS = $\frac{1}{N} \sum \mathrm{PSS}_j$

MOPESS: Mean Observed Prior Effective Sample Size

Key components

1. **How to generate extra samples?**
   - **Our approach:** posterior predictive simulation

$$\theta \sim p_{\pi_{\text{informative}}}$$
$$(y_{n+1}, \ldots, y_m) \sim f_\theta$$

$\Longrightarrow$ Bayes estimator of PSS

Key components

1. **How to generate extra samples?**
   - **Our approach:** posterior predictive simulation

$$\theta \sim p_{\pi_{\text{informative}}}$$
$$(y_{n+1}, \ldots, y_m) \sim f_\theta$$

$\Longrightarrow$ Bayes estimator of PSS

2. **What is the distance?**
   - e.g. Wasserstein distance

Key components

1. **How to generate extra samples?**
   - **Our approach:** posterior predictive simulation

$$\theta \sim p_{\pi_{\text{informative}}}$$
$$(y_{n+1}, \ldots, y_m) \sim f_\theta$$

$\Longrightarrow$ Bayes estimator of PSS

2. **What is the distance?**
   - e.g. Wasserstein distance

3. **How to set the weights w_j in PSS $= \sum w_j$PSS$_j$?**
   - Distance never exactly zero
   - Not discussed in previous work i.e. $w_j = \frac{1}{N}$

Illustrations

Simple numerical example

Baseline prior: $\pi_{\text{base}}(\mu) \propto 1$

Informative prior: $\mu \sim N\left(\mu_0, \frac{\sigma^2}{k}\right)$

Model: $y_1, \ldots, y_n \overset{iid}{\sim} N(\mu = 0, \sigma^2)$

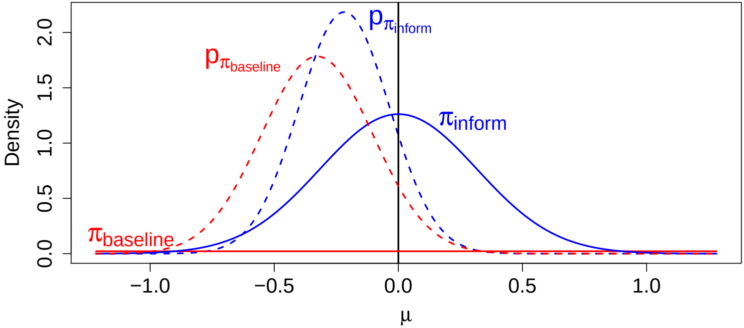
Baseline posterior: $\mu \sim N\left(\bar{y}_{1:m}, \frac{\sigma^2}{m}\right)$

Informative posterior: $\mu \sim N\left(\mu_1 = \alpha\bar{y}_{1:n} + (1-\alpha)\mu_0, \frac{\sigma^2}{n+k}\right),$

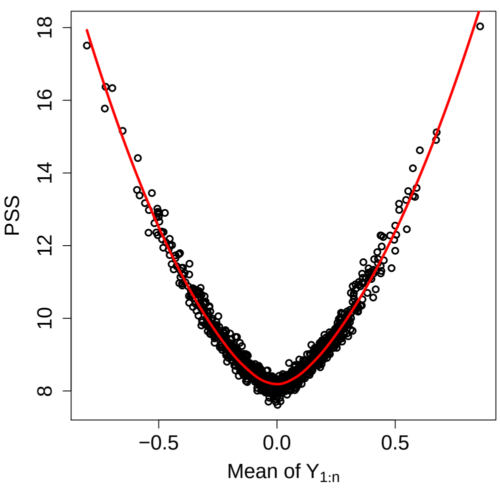
where $\alpha = \frac{n}{n+k}$

Simple numerical example: agreeing prior

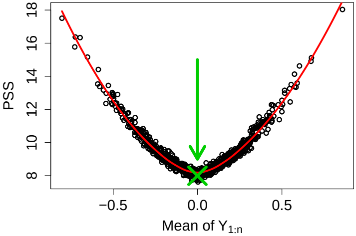
- $\mu_0 = 0$, $k = 10$
- $n = 20$, $\sigma^2 = 1$
- 1000 simulated datasets $y_{1:n}^{(1)}, \ldots, y_{1:n}^{(1000)}$

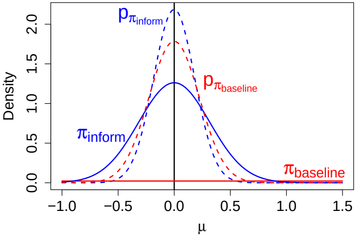


PSS as a function of data mean



Low PSS example case

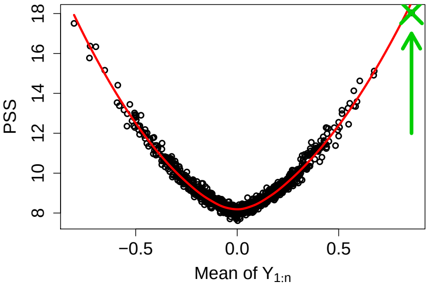


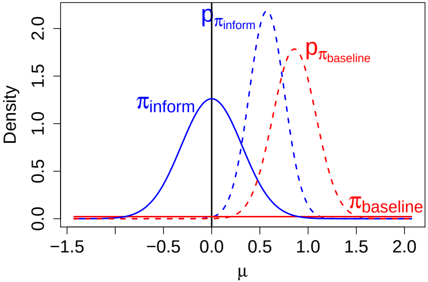


Reimherr et al. (2014): when we are "lucky" and the prior lines up exactly with the truth this corresponds to "super-information"

Different framing: high concordance vs. little impact

High PSS example case





Regression example

Model:

$$Y_i|\boldsymbol{\beta}, X_i = x_i \sim \mathcal{N}(\beta_1 + \beta_2 x_i, \sigma^2)$$

Priors:

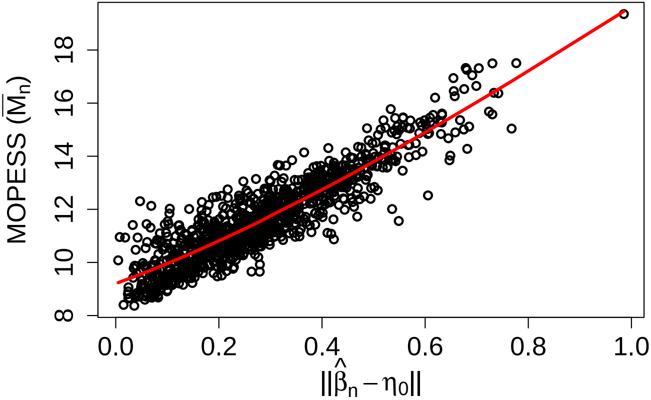
$$\pi_{\text{inform}}: \quad \boldsymbol{\beta} \sim \mathcal{N}\left(\boldsymbol{\eta}_0, \Sigma_0\right), \quad \text{where} \quad \Sigma_0 = \begin{bmatrix} \tau_1^2 & 0 \\ 0 & \tau_2^2 \end{bmatrix},$$

$$\pi_{\text{base}}(\boldsymbol{\beta}) \propto 1$$

Setup:

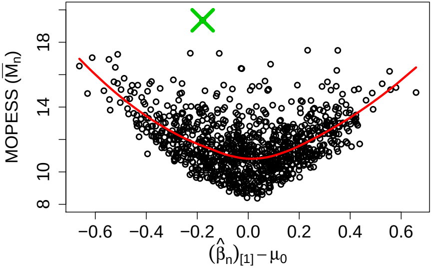
- For simplicity assume known: $\boldsymbol{\eta}_0 = (\mu_0, \gamma_0) = (0, 0)$, $\tau_1^2, \tau_2^2 = 0.1$, and $\sigma^2 = 1$
- Nominal EPSS for β_i under $\pi_{\text{inform}} = \sigma^2/\tau_i^2 = 10$, for $i \in [1, 2]$

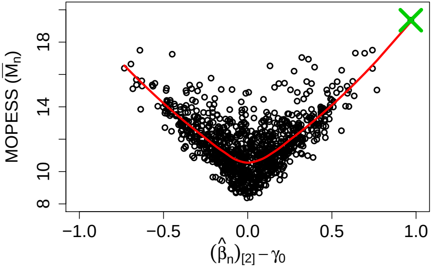
Regression example: MOPESS



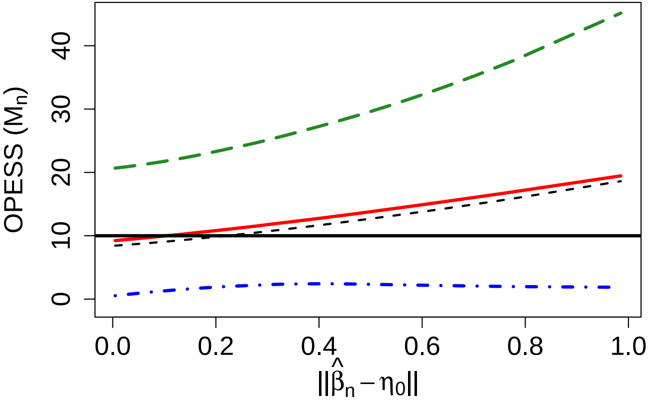
- Simulation based on $X_i \sim \mathcal{N}(0, 1)$ (more generally we can resample from the empirical distribution)

Regression example: MOPESS





Regression example: OPESS



Conceptual developments and future work

Conceptual developments:

- Prior impact depends on the data
- Directly compare the posterior distributions under different priors
- Future data: posterior predictive distribution

Conceptual developments and future work

Conceptual developments:

- Prior impact depends on the data
- Directly compare the posterior distributions under different priors
- Future data: posterior predictive distribution

Future statistical work:

- Distance almost never exactly zero
- Connections with sensitivity analysis

Conceptual developments and future work

Conceptual developments:

- Prior impact depends on the data
- Directly compare the posterior distributions under different priors
- Future data: posterior predictive distribution

Future statistical work:

- Distance almost never exactly zero
- Connections with sensitivity analysis

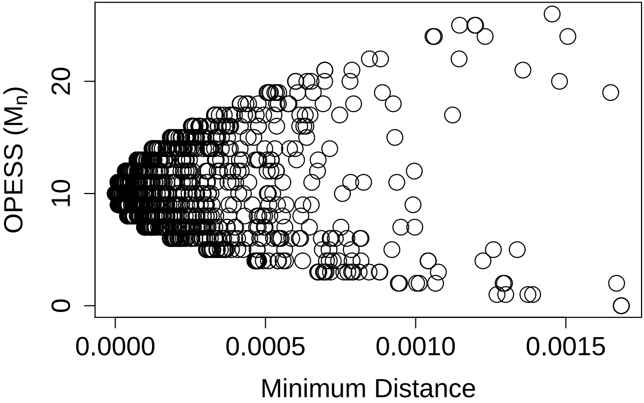
Future astrostatical work:

- Multiple instrument application: what is the impact of priors from different telescope teams?
- Gravitational waves application? :)

References

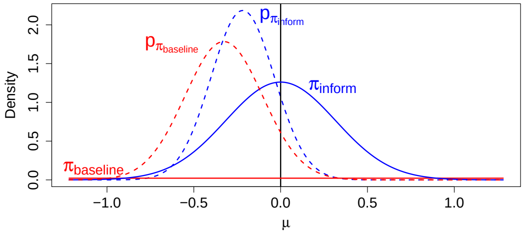
1. **Jones DE, Trangucci RN, and Chen Y. "Quantifying Observed Prior Impact." arXiv preprint arXiv:2001.10664 (2020).**
2. Chen Y, Meng XL, Wang X et al. "Calibration concordance for astronomical instruments via multiplicative shrinkage." Journal of the American Statistical Association 114.527 (2019): 1018-1037.
3. Reimherr M, Meng XL, and Nicolae DL. "Being an informed Bayesian: Assessing prior informativeness and prior likelihood conflict." arXiv preprint arXiv:1406.5958 (2014).
4. Morita S, Thall PF, and Müller P. "Determining the effective sample size of a parametric prior." Biometrics 64.2 (2008): 595-602.
5. Clarke B. "Implications of reference priors for prior information and for sample size." Journal of the American Statistical Association 91.433 (1996): 173-184.

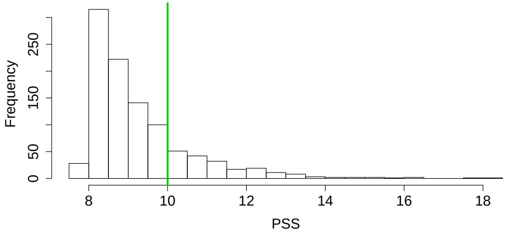
Minimum distance



Simple numerical example: agreeing prior

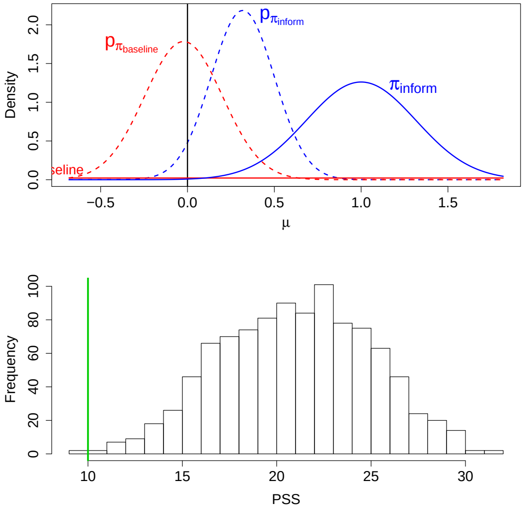
- $\mu_0 = 0$, $k = 10$
- $n = 20$, $\sigma^2 = 1$
- 1000 simulated datasets $y_{1:n}^{(1)}, \ldots, y_{1:n}^{(1000)}$





Simple numerical example: disagreeing prior

- $\mu_0 = 1$, $k = 10$
- $n = 20$, $\sigma^2 = 1$
- 1000 simulated datasets $y_{1:n}^{(1)}, \ldots, y_{1:n}^{(1000)}$



Strong impact