Flexible model assessment via approximate Calibrated Posterior Predictive p-values

Sally Paganin, joint work with Perry de Valpine

spaganin@hsph.harvard.edu @sampling_sally https://salleuska.github.io/

Paper out soon, with software

NIMBLE

Model assessment in theory

Posterior predictive p-values (ppp)

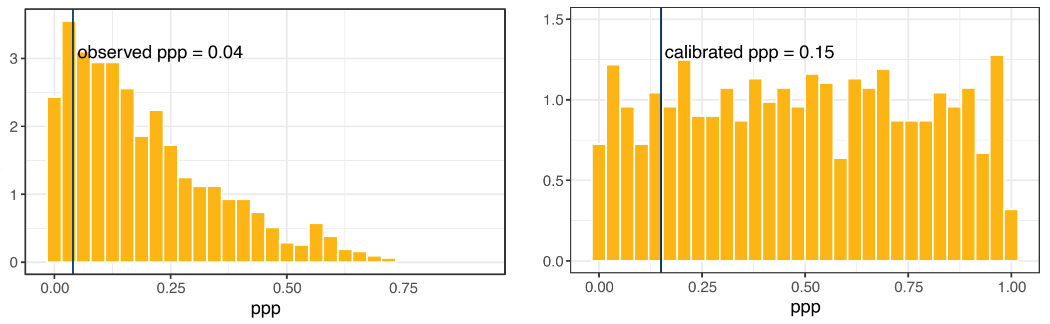
Tool for Bayesian model assessment

$$\textbf{model} = \underbrace{\text{data distribution}}_{p(y|\theta)} + \underbrace{\text{prior distribution}}_{\pi(\theta)}$$

1. **discrepancy measure** $D(y, \theta)$
   reflects aspects of the data we want the model to capture
2. **posterior predictive** $p(y^*|y) = \int p(y^*|\theta)\pi(\theta|y)d\theta$

$$\text{ppp} = \Pr\{D(Y^*, \theta) \geq D(y, \theta)\}$$

Hard to interpret!



Calibrated ppp (cppp)

Interpret the ppp as a **statistic** - $\text{ppp}(Y) \sim F$ **unknown**

$$\text{cppp} = \Pr\{\text{ppp}(\tilde{Y}) \leq \text{ppp}(y)\} = \mathbb{E}_{\tilde{Y}}\left[\mathbb{I}\left\{\text{ppp}(\tilde{Y}) \leq \text{ppp}(y)\right\}\right]$$

can be interpret with respect a Uniform distribution

[Hjort et al. (2006)] - bootstrap-like procedure

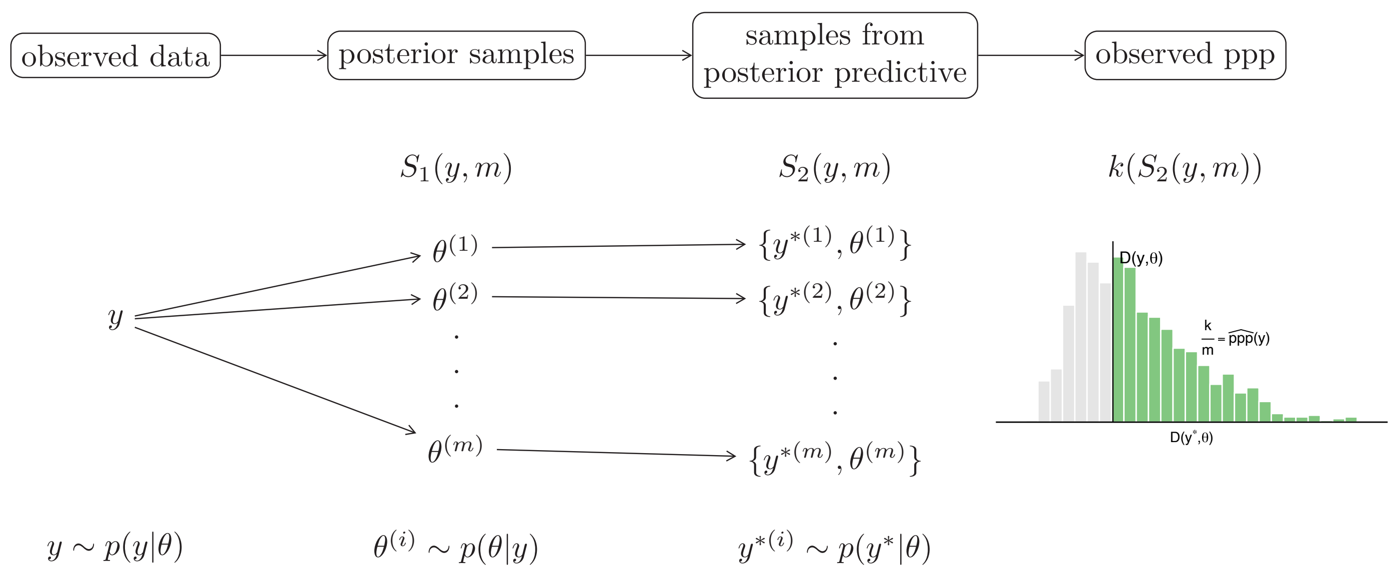
1. Simulate new data $\tilde{y} \sim g(\tilde{y}|y)$ multiple times
   $g(\tilde{y}|y)$ **calibration density** (e.g., prior, posterior predictive, ...)
2. For each new data repeat MCMC estimation & ppp computation

High computational cost!

- r = number of calibration replicates ($\approx 10^2$)
- m = number of posterior samples ($\approx 10^4$)
- naive computational cost $c = r \times m$ ($\approx 10^8$)

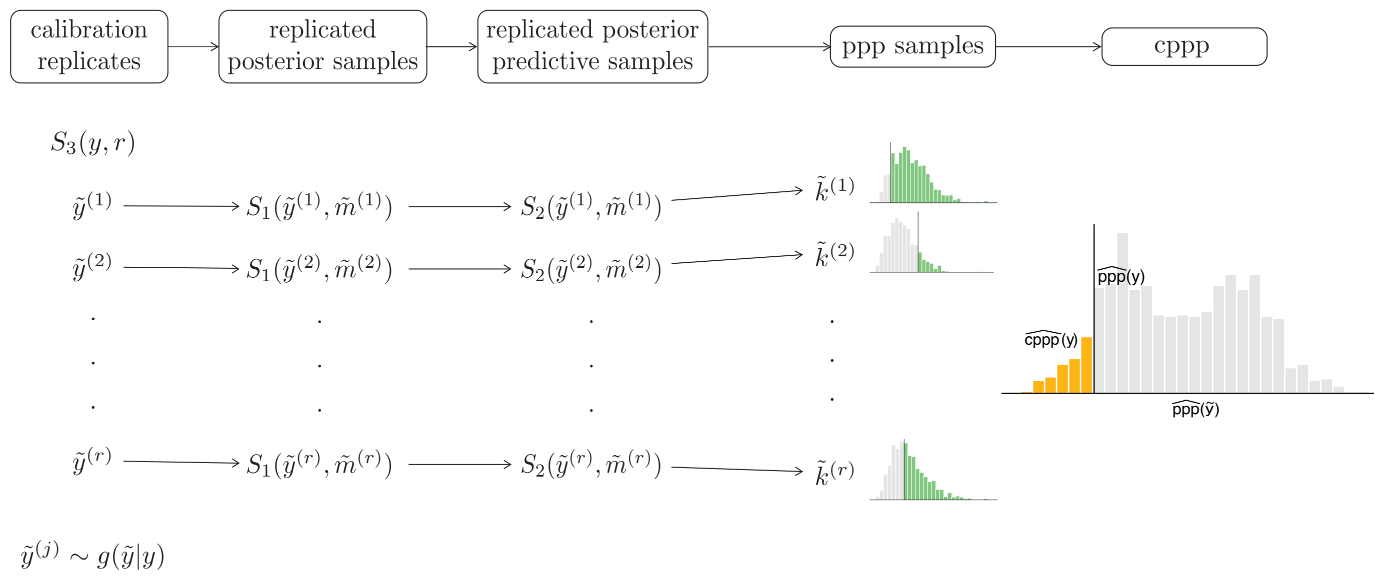
Model assessment in practice

Monte Carlo estimation of ppp



$$\widehat{\text{ppp}}(y) = \frac{1}{m}\sum_{i=1}^{m}\mathbb{I}\left\{D(y^{*(i)}, \theta^{(i)}) \geq D(y, \theta^{(i)})\right\} = \frac{k}{m}$$

Monte Carlo estimation of cppp



$$\widehat{\text{cppp}}(y) = \frac{1}{r}\sum_{j=1}^{r}\mathbb{I}\left\{\widehat{\text{ppp}}(\tilde{y}^{(j)}) \leq \text{ppp}(y)\right\} = \frac{1}{r}\sum_{j=1}^{r}\mathbb{I}\left\{\tilde{k}^{(j)} \leq \tilde{m}\text{ppp}(y)\right\}$$

Model assessment: can we do better?

Some considerations

$\widehat{\text{cppp}}(y)$ is an estimate of the true $\text{cppp}(y)$

- **bias**: $\tilde{m} \to \infty$, $\widehat{\text{ppp}}(\tilde{y}) \to \text{ppp}(\tilde{y}) \to$ **bias** $\to 0$
- **variance** $= \frac{1}{r}\left\{\underbrace{\mathbb{E}_{\tilde{Y}}\left[\mathbb{V}_{\tilde{K}|\tilde{Y}}\right]}_{1} + \underbrace{\mathbb{V}_{\tilde{Y}}\left[\mathbb{E}_{\tilde{K}|\tilde{Y}}\right]}_{2}\right\}$

1. average variance of determining $\text{ppp}(\tilde{y}) \leq \text{ppp}(y)$
   small as $\tilde{m}$ increases
2. variance across calibration replicates of $\Pr(\widehat{\text{ppp}}(\tilde{y}) \leq \text{ppp}(y))$
   as $\tilde{m}$ increases, this is the variance of a $Bernoulli(\text{cppp}(y))$

Question: how to choose $\tilde{m}$ and r?

Results

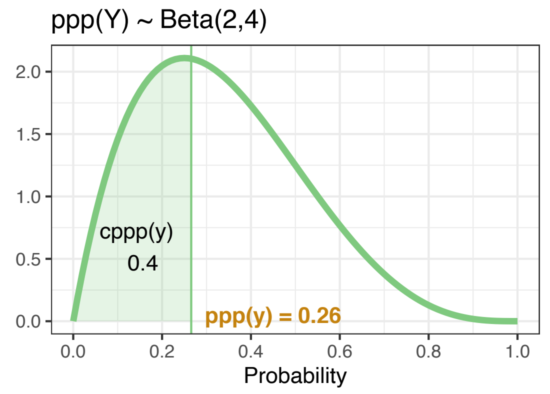
Good approximation
→ **bet on r, small $\tilde{m}$**
if $\tilde{m}$ is large enough, RMSE is dominated by variance

Faster
→"large enough " $\tilde{m}$ is small compared to
usual number of MCMC iterations ($\approx 10^4, 10^5$)

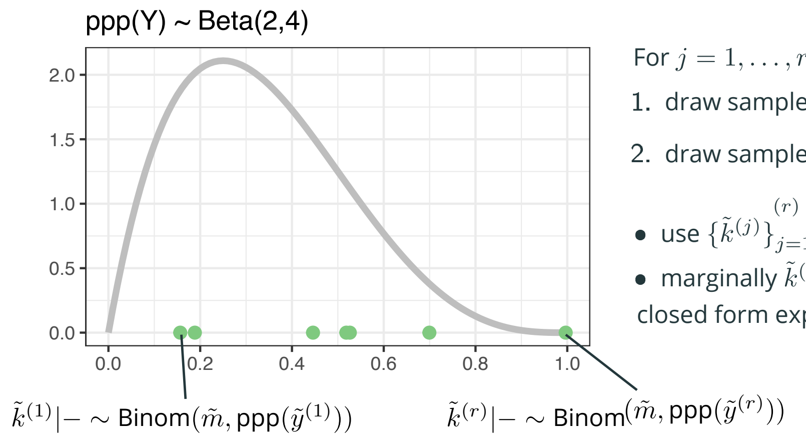
Error quantification: variance estimation
With MCMC samples, $\tilde{m}$ = ESS (Effective Sample Size)
hard to estimate with small $\tilde{m}$ (short chains)

- **Plug-in method**: uses original MCMC samples (from the data) to inform on ESS for the short ones
- **Bootstrap method**: uses two levels
  calibration replicates & MCMC samples

A simple example



1. $\text{ppp}(Y) \sim F(\cdot)$ **known**
2. Fix $\text{cppp}(y)$ value
3. $\text{ppp}(y) = F^{-1}(\text{cppp}(y))$

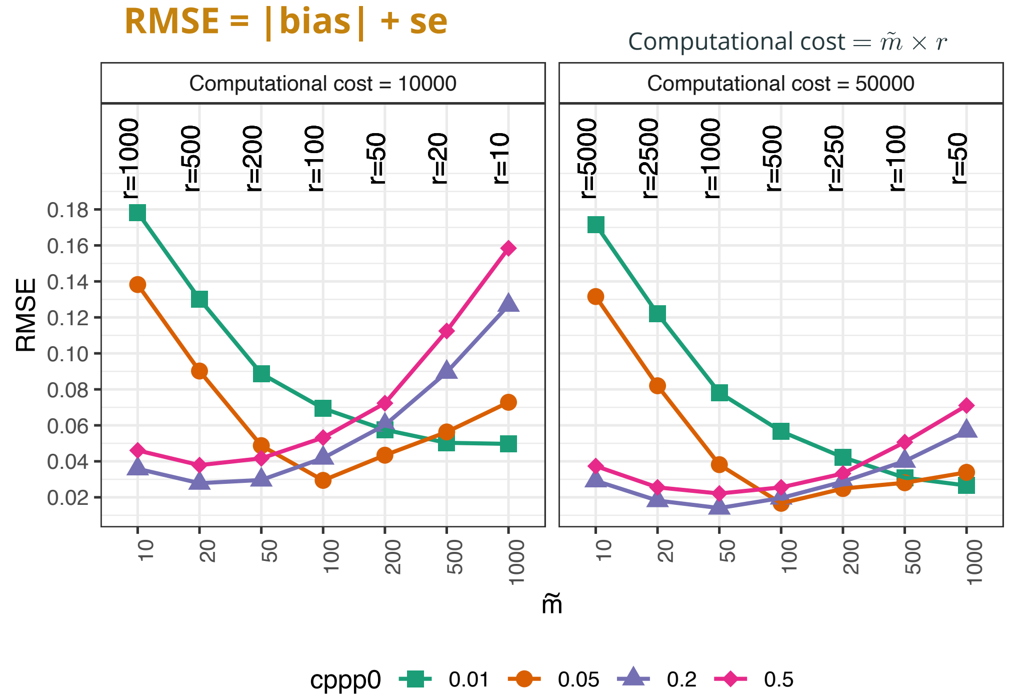


For $j = 1, \ldots, r$

1. draw samples $\text{ppp}(\tilde{y}^{(j)})$
2. draw samples $\tilde{k}^{(j)}|\text{ppp}(\tilde{y}^{(j)})$

- use $\{\tilde{k}^{(j)}\}_{j=1}^{(r)}$ to calculate $\text{cppp}(y)$
- marginally $\tilde{k}^{(j)} \sim \text{BetaBinomial}(\tilde{m}, 2, 4)$

closed form expressions for bias and variance



References

Gelman, A., Meng, X.L., & Hal, S. (1996). **Posterior predictive assessment of model fitness via realized discrepancies** *Statistica sinica*, 733–760.

Meng, X.L. (1994). **Posterior predictive p-values** *The Annals of Statistics*, **22**(3), 1142–1160.

Hjort, N.L., Dahl, F.A., & Steinbakk, G.H. (2006). **Post-processing posterior predictive p-values** *Journal of the American Statistical Association*, **101**(475), 1157–1174.

Robins, J. M., van der Vaart, A., & Ventura, V. (2000). **Asymptotic distribution of p values in composite null models** *Journal of the American Statistical Association*, **95**(452), 1143–1156.