Objective data-dependent distributions[1]

Ryan Martin
North Carolina State University
www4.stat.ncsu.edu/~rmartin

O'Bayes 2019
University of Warwick
June 30th, 2019

[1] Research supported by NSF DMS 1737929, 1737933, and 1811802

Introduction

- Classically, *objective Bayes* was about data-driven posteriors having good frequentist properties.
- Roughly, the construction proceeds by using Bayes's rule to combine likelihood with a "flat" prior.
- If the goal is to have a "good" data-dependent distribution, then why not consider other constructions?
- Extra flexibility might even be beneficial:
  - simpler prior specification
  - faster computation
  - robustness to model misspecification
  - ...

Intro, cont.

- Starting point is identification of the quantity of interest:
  - through a statistical model,
  - through some kind of loss function,
  - ...
- Assume we have relevant data $Y^n = (Y_1, \ldots, Y_n)$.
- Today, my goal is to construct a distribution Π^n, depending on Y^n, for that quantity of interest.
- *Not trying to be objective*, just trying to construct Π^n to achieve some objectives.
- My objectives:
  - fast asymptotic concentration rates
  - credible sets (approx) achieve nominal coverage

This talk

- Describe three different Triple-D constructions:
  1. *Bayes with empirical priors*
  2. *Gibbs posteriors*
  3. *Order-dependence and permutations*
- I'll mostly focus on specific examples.
- Theoretical and numerical results.
- Concluding remarks.

Triple-D(1): Bayes with empirical prior

- Consider the sparse normal means problem:
  - Independent $Y_1, \ldots, Y_n$, with $Y_i \sim \mathrm{N}(\theta_i, 1)$ and n large.
  - Only a few θ_i's are non-zero, but locations are unknown.
- Express the θ vector as a pair (S, θ_S):
  - $S \subseteq \{1, 2, \ldots, n\}$ denotes the location of non-zeros
  - θ_S the $|S|$-vector of non-zero values.
- Prior for $\theta = (S, \theta_S)$?
  - Sparsity suggests an informative marginal prior for S
  - No info about θ_S, why not let the data help?

Triple-D(1), cont.

- Marginal prior $\pi_n(S)$ for S:
  - $|S| \sim \text{Geo}(n^{-a})$, truncated to $\{0, 1, \ldots, n\}$.
  - Given the size, S is uniform over all configs of that size.
- Conditional prior for θ_S, given S:

$$\pi_n(\theta_S \mid S) = \mathrm{N}_{|S|}(Y_S, \gamma^{-1} I_{|S|}), \quad \gamma \in (0, 1).$$

- Conditional prior for θ_S, given S, times the marginal prior for S gives an *empirical prior* for θ, call it Π_n.
- Intuition:
  - informative prior on the thing we know about, S;
  - "non-informative" on the thing we don't know about, θ_S

Triple-D(1), cont.

- Combine prior and likelihood in *almost* the usual way:

$$\Pi^n(d\theta) \propto L_n(\theta)^\alpha \, \Pi_n(d\theta), \quad \alpha \in (0,1).$$

- Power $\alpha < 1$ might seem weird, but
  - it's not a restriction, just a feature of our construction
  - $\alpha = 1$ isn't "more Bayesian" or otherwise obviously "better"
- Relatively simple computations thanks to conjugacy.
- Good theoretical properties too:
  - exact minimax optimal ℓ_2 concentration rate[2]
  - model selection consistency[3]

[2] M. and Walker (`arXiv:1304.7366`, `1604.05734`)
[3] M. and Ning (`arXiv:1812.02150`)

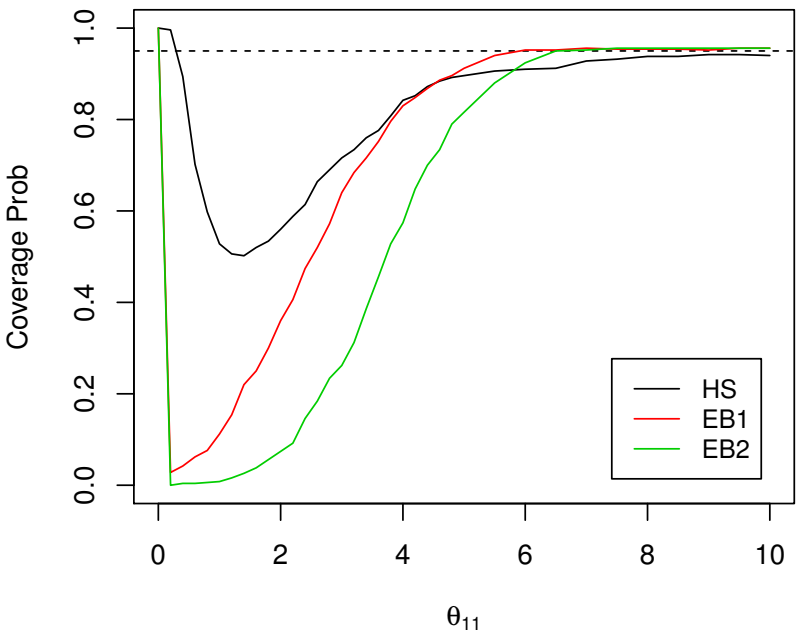
Triple-D(1), cont.

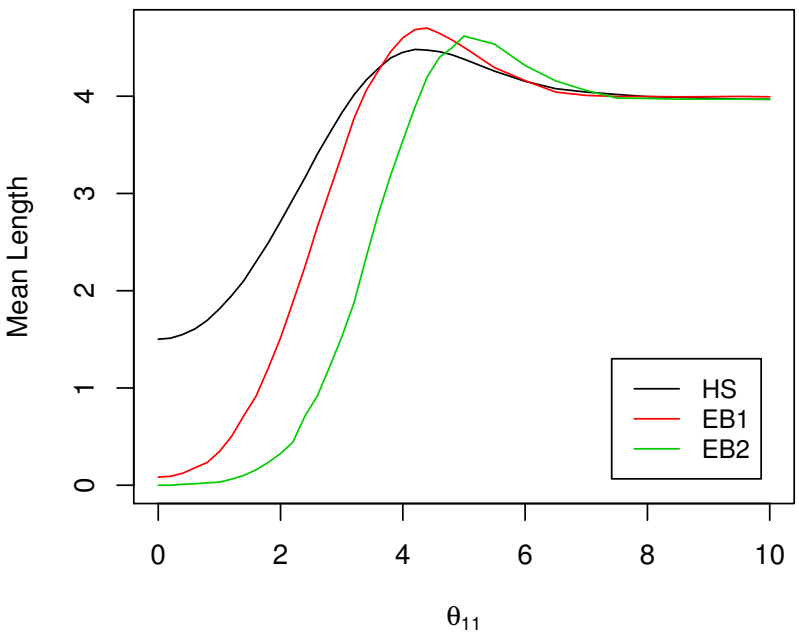
- Under certain conditions, we have also established a particular Bernstein–von Mises theorem.[4]
- This leads to claims about uncertainty quantification.
- For example, consider the marginal credible intervals:
  - Suppose each non-zero $|\theta_i^\star|$ is $\gtrsim (\log n)^{1/2}$, except $\theta_k^\star$.
  - How large does $|\theta_k^\star|$ have to be for its marginal credible interval to hit the target coverage?
  - Turns out $|\theta_k^\star|$ a little less than $(\log n)^{1/2}$ is large enough.
- Compares favorably to results for the horseshoe prior.

[4] M. and Ning (`arXiv:1812.02150`)

Triple-D(1), cont.

- $n = 200$, $\theta^\star =$ (five 7's, five 1.5's, $\theta^\star_{11}$, rest equal 0).
- Coverage and length of 95% credible interval for θ_{11}.
  - HS: $\sigma^2 = 1$ fixed, τ via MMLE, `horseshoe` package
  - EB1: $\alpha = 0.99$, $\gamma = 0.01$, beta-binomial version
  - EB2: $\alpha = 0.99$, $\gamma = 0.01$, as above





Triple-D(1), cont.

- Extensions to the regression setting
  - estimation, etc (M., et al, `arXiv:1406.7718`)
  - prediction (M. and Tang, `arXiv:1903.00961`)
- Piecewise constant (M. and Shen, `arXiv:1712.03848`)
- Monotone density estimation (M. `arXiv:1706.08567`)
- General theory (M. and Walker, `arXiv:1604.05734`)
- Others in the works...

Triple-D(2): Gibbs posterior

- Parameters aren't always defined through a statistical model
  - quantiles, quantile regression, VaR, etc
  - area under the ROC curve
  - ...
- Of course, they CAN be defined through a model but then there's sure to be
  - nuisance parameters, hence marginalization
  - extra priors and extra computation
  - risk of misspecification bias
- Either way, it's advantageous to get a direct posterior for the quantity of interest.
- That's the benefit of a *Gibbs posterior*.

Triple-D(2), cont.

- Example: minimum clinically important difference (MCID).
- Data $Y_i = (X_i, Z_i)$, $i = 1, \ldots, n$, iid:
  - $X_i \in \mathbb{R}$ is a diagnostic measure on patient i;
  - $Z_i \in \{-1, +1\}$, where "$Z_i = \pm 1$" means patient i found the treatment to be effective/ineffective.
- Quantity of interest is the MCID:

$$\theta = \theta(\mathsf{P}) = \arg\min_{\vartheta} \underbrace{\mathsf{P}\{Z \neq \mathrm{sign}(X - \vartheta)\}}_{R(\vartheta)},$$

  where P is the distribution of $Y = (X, Z)$.
- *MCID is not naturally a model parameter*, so assuming a model (e.g., logistic regression) may introduce bias.

Triple-D(2), cont.

- To avoid bias, M-estimation replaces the risk R by its empirical version, R_n, and proceeds with minimization.
- Similarly, the Gibbs posterior is given by

$$\Pi_\omega^n(d\theta) \propto e^{-\omega\, nR_n(\theta)}\, \Pi(d\theta),$$

  where Π is a prior for θ and $\omega > 0$ is the learning rate.
- A direct posterior for the interest parameter θ
  - no nuisance parameters to marginalize out
  - since there's no model, no risk of misspecification bias.
- Concentration rates come from features of R_n.[5][6][7]

[5] MCID case (Syring and M., `arXiv:1501.01840`)
[6] image boundary curve (Syring and M., `arXiv:1606.08400`)
[7] area under ROC curve (Wang and M., `arXiv:1906.08296`)

Triple-D(2), cont.

- Good concentration rates are important, but we also care about uncertainty quantification.
- Learning rate ω is important for this.
- By construction, *for any* $\omega > 0$,
  - Π_ω^n is roughly centered around the M-estimator,
  - which is close to the true $\theta^\star$ asymptotically.
- Then ω only really affects the spread.
- *Idea:* choose ω so that credible regions from Π_ω^n have nominal coverage probability.

Triple-D(2), cont.

- We[8] developed a calibration algorithm.
- Written below in the "oracle" case where P is *known*; if unknown, then replace P by the empirical distribution $\mathbb{P}_n$.

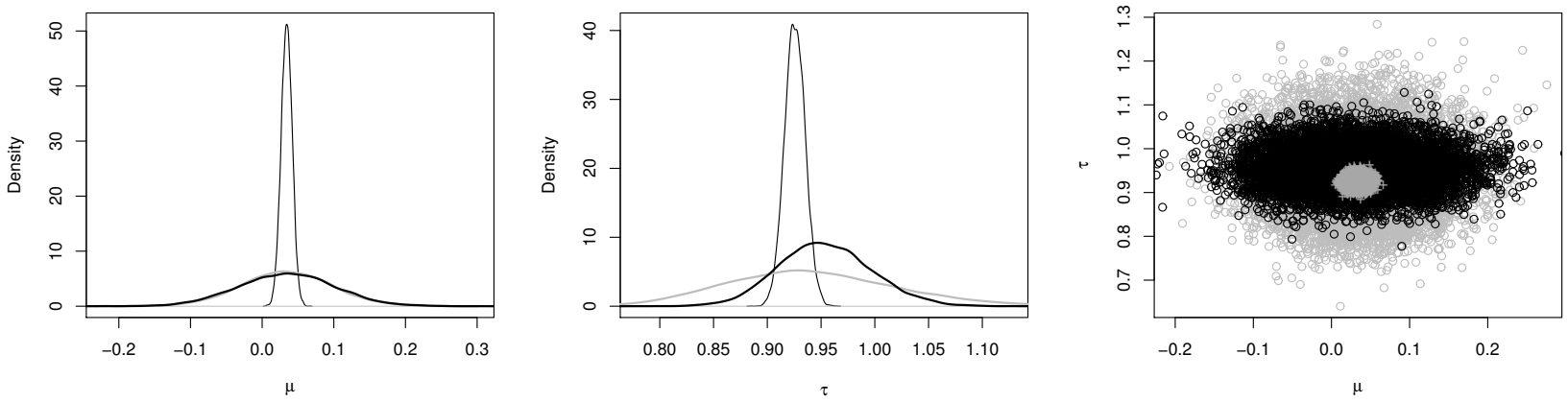
General Posterior Calibration Algorithm:

0. Initialize ω_0 and set $t = 0$
1. Sample data sets from P to approximate the coverage prob of $100(1-\alpha)\%$ Gibbs credible region $C_{\omega_t,\alpha}$.
   - compute $C_{\omega_t,\alpha}$ for each data set;
   - check coverage using $\theta(\mathsf{P})$
2. If coverage prob is close to $1-\alpha$, then stop; otherwise, update $\omega_t \to \omega_{t+1}$ via stochastic approximation, increment t, and go back to Step 1.

[8] Syring and M. (arXiv:1509.00922)

Triple-D(2), cont.

- Simple Gaussian process model with two parameters (μ, τ).
- Compare several posteriors:[9]
  - full Bayes posterior
  - composite Bayes posterior
  - our calibrated composite posterior



(a) μ marginal (b) τ marginal (c) (μ, τ) joint

[9] Syring and M. (`arXiv:1509.00922`)

Triple-D(2), cont.

- Two-dim quantile regression example[10]
- 95% intervals based on several methods:

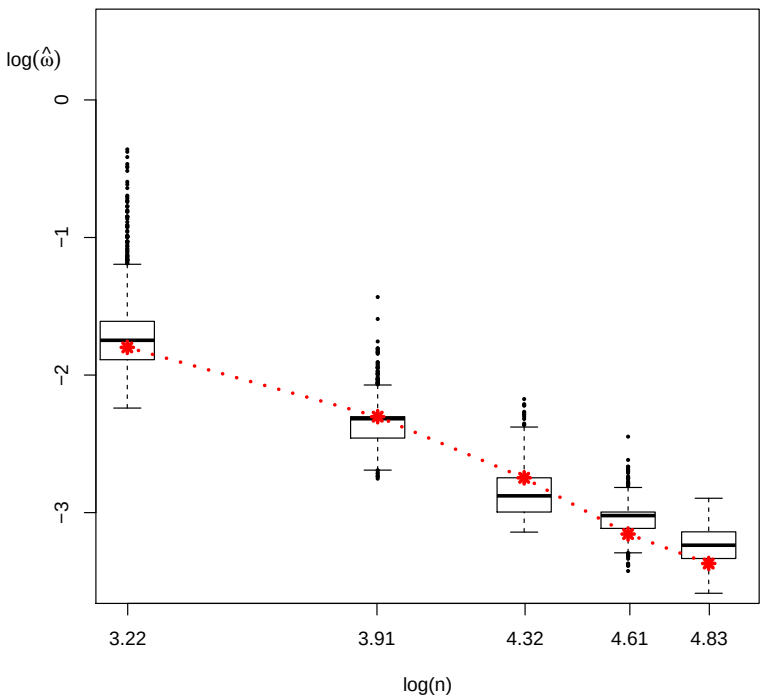
BEL.s Bayesian empirical likelihood (Yang and He 2012)
Normal asymptotic normality
GPC Gibbs with GPC

		Coverage Prob $\times 100$			Avg Length $\times 100$		
n		BEL.s	Normal	GPC	BEL.s	Normal	GPC
100	θ_0	97	95	95	106	100	91
	θ_1	98	98	95	58	55	47
400	θ_0	95	95	95	50	50	46
	θ_1	97	97	95	26	25	23
1600	θ_0	96	96	95	25	25	23
	θ_1	96	96	95	13	12	11

[10] Syring and M. (`arXiv:1509.00922`)

Triple-D(2), cont.

- The *oracle version* of the GPC algorithm returns $\omega_n^\star$, say.
- Let's compare with $\hat{\omega}_n$ based on the *empirical version*.
- Plot of $\omega_n^\star$ and boxplot of $\hat{\omega}_n$ versus n (on log scale).[11].



[11] Wang and M. (`arXiv:1906.08296`)

Triple-D(2), cont.

- Key selling point: "model-free" posterior, with
  - automatic robustness
  - good convergence rates
  - calibration tools available for uncertainty quantification
- Other benefits of a direct Gibbs posterior:
  - only needs a prior for "real" parameters
  - avoids marginalization, saves on computation
- When a risk function is given, a Gibbs approach is natural.
- How to reverse-engineer risk functions?

Triple-D(3): permutations

- Consider estimation of the density f based on iid data Y^n.
- A version of Newton's algorithm for estimating a density[12]
  - Start with an initial guess $\hat{f}_0$ and weights $(w_i) \subset (0,1)$.
  - Then for $i = 1, \ldots, n$, update:

$$\hat{f}_i(y) = \hat{f}_{i-1}(y)\big[1 + w_i\big\{c_\rho\big(\widehat{F}_{i-1}(y), \widehat{F}_{i-1}(Y_i)\big) - 1\big\}\big].$$

  - Take $\hat{f}_n$ as the final estimate based on Y^n.
- Properties:
  - very fast to compute
  - asymptotic L_1 consistency
- But uncertainty quantification seems difficult

[12] Hahn et al (`arXiv:1508.07448`)

Triple-D(3), cont.

- Where is the Triple-D?
- Key observation: $\hat{f}_n$ depends on the data ordering.
- Initially seems problematic, one might consider averaging over permutations to reduce/eliminate this dependence.
- But maybe the order-dependence is beneficial...
- Create Π^n for f by randomly permuting the data sequence:
  - $S \sim \text{Unif}(\text{permutations})$
  - $\hat{f}_n^S :=$ estimate based on data $Y_{S(1)}, \ldots, Y_{S(n)}$.
- Read off quantiles for uncertainty quantification about f.

Triple-D(3), cont.

- Simple idea, but why would this work?
- Define functionals $\Psi_n = \int \psi \, \hat{f}_n \, dy$ and $\Psi_n^S = \int \psi \, \hat{f}_n^S \, dy$.
- Since data are iid, distribution doesn't depend on order, so

$$\begin{aligned} \mathrm{V}(\Psi_n) &= \mathrm{V}(\Psi_n^S) \\ &= \mathrm{E}\{\mathrm{V}(\Psi_n^S \mid Y^n)\} + \mathrm{V}\{\mathrm{E}(\Psi_n^S \mid Y^n)\}. \end{aligned}$$

- Consistency implies $\mathrm{V}\{\mathrm{E}(\Psi_n^S \mid Y^n)\} \approx 0$, hence

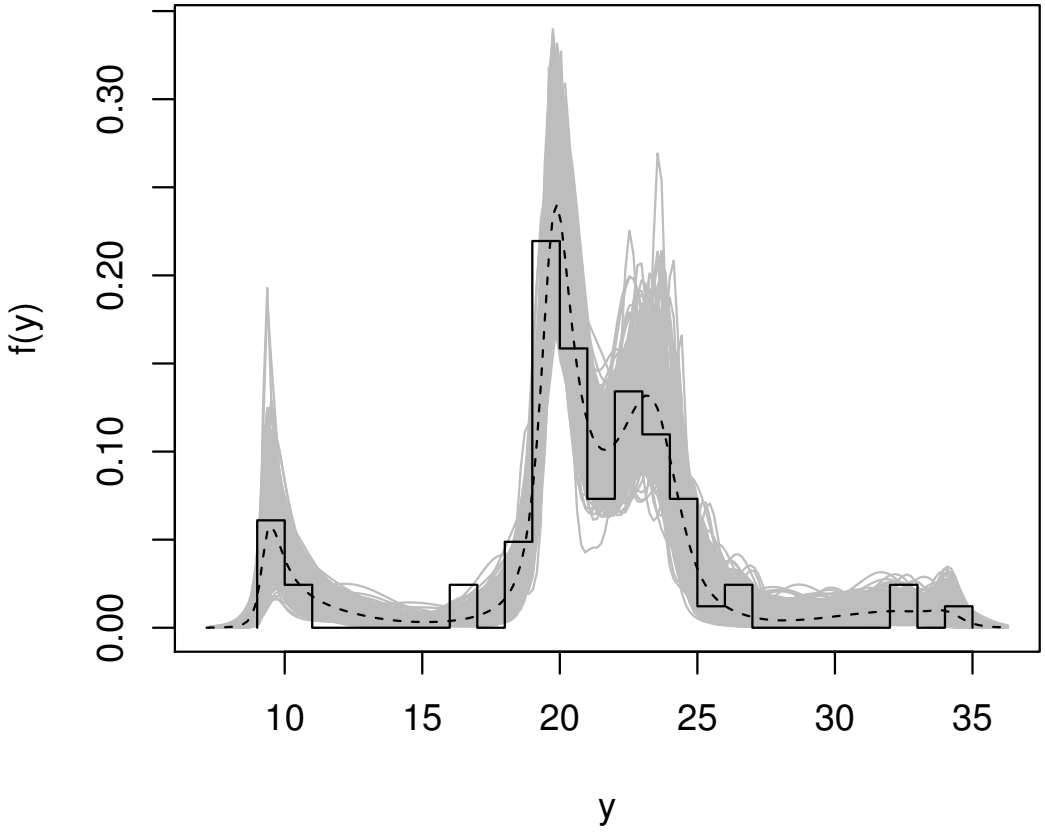
$$\mathrm{E}\{\mathrm{V}(\Psi_n^S \mid Y^n)\} \approx \mathrm{V}(\Psi_n).$$

- Permutation dist "gets the spread right," i.e., its variance is an approx unbiased estimate of sampling dist variance.[13]

[13] Dixit and M. (`arXiv:1906.05349`)

Triple-D(3), cont.

- For illustration, consider the famous galaxy data.
- $n = 82$, $\rho = 0.95$, and $w_i = (i+1)^{-0.75}$.
- Results based on 500 random permutations.



Triple-D(3), cont.

- Can be applied to any order-dependent estimator.[14]
- Limited scope? Not too many estimators are order-dependent.
- *New idea:* take an arbitrary estimator and artificially create order- dependence...
- How to do it? Use Cesáro averaging!

- Aside: a usual prior on f induces a prior on data orderings, so the choice of a uniform prior is "objective"

[14] Mixing densities in Dixit and M. (`arXiv:1906.05349`)

Conclusions

- We can construct "posteriors" without dealing with the difficulties of classical Bayes:
  - statistical models
  - priors being "objective"
  - marginalizing nuisance parameters
- Good properties can still be achieved:
  - fast convergence rates
  - valid uncertainty quantification
- That is, we can construct posteriors to *achieve an objective*.
- Modern take on "objective Bayes"...?

The end... except for a commercial

Thank you!

rgmarti3@ncsu.edu
www4.stat.ncsu.edu/~rmartin

An open-access publication platform is now available, featuring an *author-driven* peer-review process.

RESEARCHERS.ONE

For more details, check out

www.researchers.one
www.twitter.com/@ResearchersOne