Bayesian restricted likelihood and model diagnostics

Steve MacEachern (The Ohio State University)
along with Hang Joon Kim and Juhee Lee

motivated by a stream of work with many coauthors, including Xinyi Xu, Pingbo Lu, John Lewis, Yoonkyung Lee, and Xinyu Zhang

O'Bayes 2025
Athens, Greece
June 2025

Supported, in part, by the NSF under grant numbers SES-1921523 and DMS-2413823

Bayes and diagnostics

- Bayesian methods arise from a beautiful theory about how we *should* learn from data
  - the centerpiece of this theory is Bayes' Theorem
  - implicit in the development is a complete probability model for all observables
  - methods work best in a "high information" setting
- Sound data anlaysis? Must actively look for deficiencies in model
- This has given rise to a variety of techniques
  - prior predictive checks (Geisser, Johnson and coauthors; Box; many more)
  - posterior predictive checks (Rubin, Gelman and coauthors)
  - partial posterior predictive checks (Bayarri & Berger; many since; especially relevant to this talk, Blei, Moran & coauthors)
- Along with this, there is a large literature on robust Bayesian inference in a variety of forms

Three threads

- **Three ideas spring up when considering posterior predictive checks**
  - **partial updates**
    - **an update with a portion of the information in the sample, allowing us to pass from a low-information setting to mimic a high-information setting (e.g., the calibrated Bayes factor w/Xu,Lu,Xu)**
  - **a reliance on robust summaries**
    - **essential when we acknowledge imperfections in our model; applicable both to discrepancy measure for evaluation of model and also for partial update (e.g., Bayesian restricted likelihood w/Lewis,Lee)**
  - **intervention and "what if" questions**
    - **central to to understanding and to decision-making (e.g., the *intervention posterior* w/Hwang et al.)**

Close connection - model choice and model criticism

- Model diagnostics / criticism attempt to identify deficiencies in (Bayesian) model
  - generic misstatement of distributional form
  - problems with tail of distributiion / outliers
  - specified discrepancy between model and reality
- May focus on specified groups of observations
  - for this, requires "direction" or conceptual alternative
  - implicit in work that looks for local departures from model
  - often identify group of observations that departs from model
- Here, pursue this path for treatment vs control experiment
  - identified set of observations are those on treatment

Two-sample problem (easy version)

- **For clarity, we work with an unrealistically simple example**

$$Y_{ij} \mid \theta_i \overset{IID}{\sim} \mathbf{N}(\theta_i, \sigma^2), \qquad i = 1, 2;\ j = 1, \ldots, n_i$$

- **Perspective on Bayes for this talk**
  - **begin with a classical *family* of models**
  - **supplement with a (proper) prior distribution on unknown parameters (here the θ_i)**
  - **yields a joint distribution on $\mathbf{Y_1}$, $\mathbf{Y_2}$**
- **As in Wald's old book, Bayes corresponds to a point null**
  - **brings in language from testing**
  - **point null in the sense that there will be no supping over unknown parameters**
- **Following O'Bayes, we place a uniform prior on θ_1**

A purposive split

- **Sufficiency lets us work with only the sample means, $\bar{Y}_1$ and $\bar{Y}_2$**
  - **this limits our partial update to $\mathbf{Y_1}$ or, equivalently, $\bar{Y}_1$**
  - **aligns with our focus on θ_1 and θ_2**
- **Note that our model says nothing about the unknown θ_2**
  - **incomplete from a Bayesian perspective**
  - **we tie θ_2 to the discrepancy between control and treatment**
  - **we will handle θ_2 separately**
- **The partial update provides information about θ_1**
  - **moves us from low information to high information about θ_1**
  - **posterior for $\theta_1 \mid \mathbf{Y_1}$ is**

$$\theta_1 \mid \mathbf{Y_1} \sim \mathbf{N}(\bar{Y}_1, \sigma^2/n_1)$$

- **Under null, $\theta_2 = \theta_1$ and a Y_{2j} is just like a Y_{1j}**

The method

- **A summary of the method**
- **Begin with Bayesian model (point null) of "one big sample"**
  - **Step 1. Update model with $\mathbf{Y_1}$**
    - **result is a posterior for θ_1**
  - **Step 2. Invoke null hypothesis of no difference**
    - **provides a predictive distribution for $\mathbf{Y_2}$**
  - **Step 3. Derive the distribution for $\bar{Y}_2$**
    - **in many cases, we replace derivation with simulation**
  - **Step 4. Use the pdf to compute the tail area**
    - **the tail area is a valid p-value, uniform under the null**
- **Great flexibility in choices above**
  - **here, focus is on whether $\theta_2 = \theta_1$ or not**

Result!

- **The p-value follows quickly in this case**

$$\begin{aligned} \theta_2 \mid \bar{Y}_1 &\sim \mathbf{N}(\bar{Y}_1, \sigma^2/n_1) \\ \bar{Y}_2 \mid \bar{Y}_1 &\sim \mathbf{N}(\bar{Y}_1, \sigma^2(n_1^{-1} + n_2^{-1})) \end{aligned}$$

- **The usual standardization to compute a p-value gives**

$$z = \frac{\bar{Y}_2 - \bar{Y}_1}{\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}}$$

- **We recognize the classical test statistic for the two-sample z-test**
  - **the split-sample p-value matches the classical p-value exactly**
- **If we wish, we have a test, motivated by Bayesian reasoning**
  - **on the Bayes side? no explicit model for the alternative**
  - **partial posterior, followed by a question about future data**

The intervention posterior

- **A classic question**
  - **if we change the world, what happens?**
  - **here, move an observation from the control to the treatment**
  - **elsewhere, change covariate, or tweak "parameter"**
- **Intervention posterior**
  - **replace a slice of a (posterior) distibution with a different slice**
  - **here, replace $\theta_2 \sim \mathbf{N}(\bar{Y}_1, \sigma^2/n_1)$ with different distribution**
  - **have chosen the degenerate $\theta_2 = \theta_1$**
- **In general, a change may affect a portion of prior / posterior**
  - **entirely a "what if" scenario**
  - **but scenario is guided by understanding of problem**
- **Choices that reduce stochasticity seem to work better**

From p-value to interval

- **The intervention posterior lets us do more**
  - **consider a different degenerate change:** $\theta_2 = \theta_1 + \delta$

$$\begin{aligned} \theta_2 \mid \mathbf{Y_1} &\sim \mathbf{N}(\bar{Y}_1 + \delta, \sigma^2/n_1) \\ \bar{Y}_2 \mid \bar{Y}_1 &\sim \mathbf{N}(\bar{Y}_1 + \delta, \sigma^2(n_1^{-1} + n_2^{-1})) \end{aligned}$$

- **From here, the familiar path**
  - **ask point null question repeatedly: Are these data consistent with specified intervention effect?**
    - **yes/no (fail to reject; reject) for each question**
    - **gather together all δ for which answer is 'yes'**
    - **result is an interval for $\delta = \theta_2 - \theta_1$**
  - **recognize interval as inverting family of classical tests**
- **Partial update predictive check produces interval**
  - **as with single check, Bayesian model for the alternative**

A few comments

- Bayes / classical match in this case is well known
  - labelling of treatment and control does not impact result
  - in predictive check literature, most easily follows from Bayarri & Berger, regression example with directed check
  - B&B's method updates with "all information not in check"
    - here, check based on $\bar{Y}_2$
    - with sufficiency, all remaining information is in $\bar{Y}_1$
- Split of data is not random split for update and holdout
  - purposive split, tied to potential model deficiency
- Approach is fully compatible with proper prior distributions
  - lose duality with classical intervals (not sure that I care)
- Strong contrast to methods based on Bayes factor
  - BF and predictive check methods ask different questions

Extensions

- **Result extends to case of common, uknown variance**

$$\pi(\theta, \sigma^2) \propto \frac{1}{\sigma^2}$$

- **Update with $\mathbf{Y_1}$ and S_2^2 to match classical results**
  - **instead of the z p-value and interval, it's a t with $n_1 + n_2 - 2$ df**
- **Dip into $\mathbf{Y_2}$ with use of S_2^2 for partial update is a choice**
  - **agrees with B&B's partial posterior predictive approach**
- **Rationale for choice**
  - **use all information not in "check" for partial update**
  - **wish to have parallel forms for analyses (next slide)**
- **Rationale for not using S_2^2 in partial update**
  - **robustness concerns (e.g., K-sample problem)**

Unknown, uncommon variances (Behrens-Fisher)

- **As a quick note, can handle different variances with same method**
  - **for improper priors, need update (S_2^2) to make prior proper**
  - **for vague priors, may be a good idea to update with S_2^2**
- **Leads to different solution than classical non-pooled t methods**
- **A few details**

$$\pi(\theta_1, \sigma_1^2, \sigma_2^2) \propto \frac{1}{\sigma_1^2}\frac{1}{\sigma_2^2}$$

  - **partial posterior for σ_i^2 (sample size and within sum of squares)**

$$\sigma_i^2 \mid RSS_i \sim \textbf{InvGamma}\left(\frac{n_i - 1}{2}, \frac{RSS_i}{2}\right)$$

  - **partial posterior for θ_1 gives partial (cond'l) predictive for $\bar{Y}_2$**

$$\bar{Y}_2 \mid \bar{Y}_1, RSS_1, RSS_2, \sigma_1^2, \sigma_2^2 \sim \mathbf{N}(\bar{Y}_1, \textstyle\sum \sigma_i^2/n_i)$$

- **Note irrelevance of labelling for what is to come**

Two examples

- **Method: Generate** σ_1^2, σ_2^2, $\bar{Y}_2$
  - **compute** $t = |\bar{Y}_2 - \bar{Y}_{1,obs}| / \sqrt{S_{1,obs}^2/n_1 + S_{2,obs}^2/n_2})$
  - **compare to** t_{obs}
  - **partial predictive p-value is fraction** $t \geq t_{obs}$
- **Ex 1. Data:** $n_1 = 9$, $n_2 = 4$, $S_1 = 1$, $S_2 = 5$ **(approx. df =** 3.11**)**
  - $\bar{Y}_2 - \bar{Y}_2$ **chosen so that p-value for** $t_{obs} = 0.05$
  - $1,000,000$ **replicates for simulation**
  - **partial predictive p-value is** 0.0515 ± 0.0002
- **Ex 2. Data:** $n_1 = 4$, $n_2 = 9$, $S_1 = 1$, $S_2 = 5$ **(approx. df =** 9.30**)**
  - $\bar{Y}_2 - \bar{Y}_2$ **chosen so that p-value for** $t_{obs} = 0.05$
  - $1,000,000$ **replicates for simulation**
  - **partial predictive p-value is** 0.0627 ± 0.0002
- **p-value from t approx differs from partial predictive p-value**

A limitation of these examples

- **Methods presented so far focus on a sharp feature of the model**
  - **well defined parameters - is $\theta_2 = \theta_1$?**
  - **different question (deficiency), different method is needed**
- **Methods are predicated on all data being "good"**
  - **fine for much of model selection, often in examples so far**
  - **but other situations are common**
  - **if tmt has an effect, lose equal variances, normality for $\bar{Y}_2$**
- **To effectively pick up deficiency in a Bayesian model**
  - **identify which inferences are of interest (e.g., centers of dist'ns)**
  - **data and context to assess which portions of model are solid**
  - **partial update based on good portion of model**
  - **diagnostics to assess whether to expand good portion**
- **Sacrifice some efficiency (power) for robustness**

The K-sample problem

- K **treatments**
  - **a random sample from each treatment**
  - **underlies ANOVA**
  - **fixed effects version and random effects version**
- **Two classical families of models**

$$Y_{ij} \mid \theta_i \overset{IND}{\sim} \mathbf{N}(\theta_i, \sigma^2)$$

  - **for random effects family, supplement with**

$$\theta_i \mid (\mu, \tau^2) \overset{IID}{\sim} \mathbf{N}(\mu, \tau^2)$$

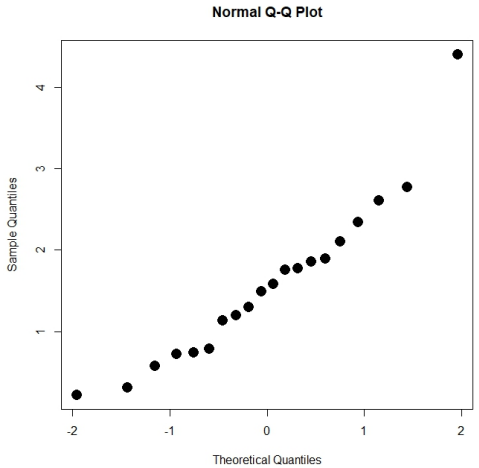
- **Fixed effects view**
  - **the set of θ_i are fixed but unknown constants**
- **Random effects view**
  - **τ^2 is a variance component, parameters of interest are μ and τ^2**

A Bayesian model

- **The Bayesian model supplements the classical model with a prior distribution**
  - **– isn't clear how to move from classical family to Bayesian model**
- **Fixed effects version**
  - **– consider the θ_i to be exchangeable**
  - **– place prior on set of K unknown numbers**
  - **– perhaps θ_i a random sample from $\mathrm{N}(\mu, \tau^2)$ dist'n**
- **Random effects version**
  - **– place prior on dist'n from which random effects are drawn**
  - **– follow principle of full support**
  - **– use Bayesian nonparametric model**
- **Sadly, muddy interface between O'Bayes and BNP**
  - **– we use fixed effects version with O'Bayes prior on (μ, τ^2, σ^2)**

An example

- $K = 20$ **treatments,** $n_i = 6$ **observations per treatment**
  - **parameter values are** $\mu = 1.5$, $\sigma^2 = 0.5$, $\tau^2 = 0.25$
  - 20^{th} **treatment has** 3.5 **added to its mean**
- **Sample (tmt) means are normally distributed with one outlier**



- **A Shapiro-Wilk test for normality gives a p-value of** 0.1339
- **Many features of interest in this model**
  - **we focus on largest sample mean - is it too big for the model?**

The robust update

- We pursue a partial update, followed by a predictive check
  - sufficiency implies that we need only consider the $(\bar{Y}_i, S_i^2)$
  - independence of sample mean and variance allows us to split these pairs
- For the update
  - condition on a subset of the sample means
  - condition on all of the sample variances
  - specifics of conditioning will be varied
  - interest in conditioning on "middle means"
- This conditioning differs from the usual partial update
  - uses less information than Bayarri & Berger
  - held out data is chosen with model's deficiency in mind
  - decompose likelihood into trustworthy/untrustworthy portions

The predictive distribution and computation

- **Conditioning statistics chosen with an eye to easy updating**
  - **basic Gibbs sampler for conjugate model**
  - **complete data updates for μ, τ^2, σ^2, θ_i standard**
- **Not conditioning on $\bar{Y}_i$ means we treat it as missing data**
  - **novel step is to fill in "missing" values – extreme $\bar{Y}_i$s**
  - **begin with unrestricted full condition for $\bar{Y}_i$ – it's normal**
  - **impose restriction to "large" or "small" – truncated normal**
- **With samples from partial posterior, can compute anything**
  - **here, statistic for discrepancy is $\max_i \bar{Y}_i$**
- **Conditioning of two forms**
  - **all but largest sample mean (strongly focused test)**
  - **middle 4, 6, 8, 10, 12, 14, 16, or 18 sample means**

Results

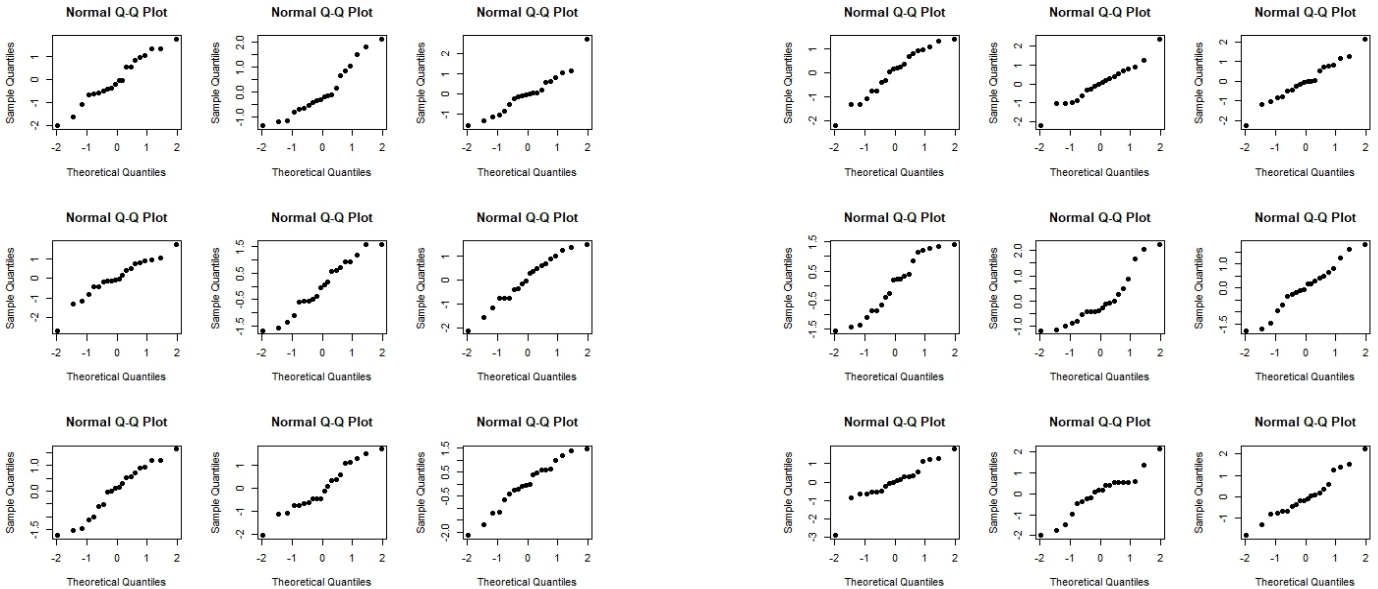
- **Predictive p-values from the partial update**
  - **recall that the Shapiro-Wilk p-value is** 0.134
  - **this misses the outlier, not much evidence against normality**

	19	18	16	14	12	10	8	6	4	PP
Ex 1	0.014	0.018	0.036	0.045	0.067	0.102	0.098	0.250	0.646	0.206
Ex 2	0.014	0.017	0.036	0.044	0.068	0.099	0.097	0.244	0.568	NA
Ex1R	0.858	0.820	0.747	0.773	0.858	0.899	0.682	0.809	0.929	0.790

- **Based on** $1,000,000$ **Gibbs iterates with** $100,000$ **dropped for burn-in**
  - **Ex 1 - as described, large mean is really big**
  - **Ex 2 - condition on means and variances for middle means**
  - **Ex1R - reverse data; outlying mean is on the bottom**
- **The partial predictive p-value works!**
  - **picks up outlier on high side, does not if no outlier on that side**

Graphics as discrepancy measures

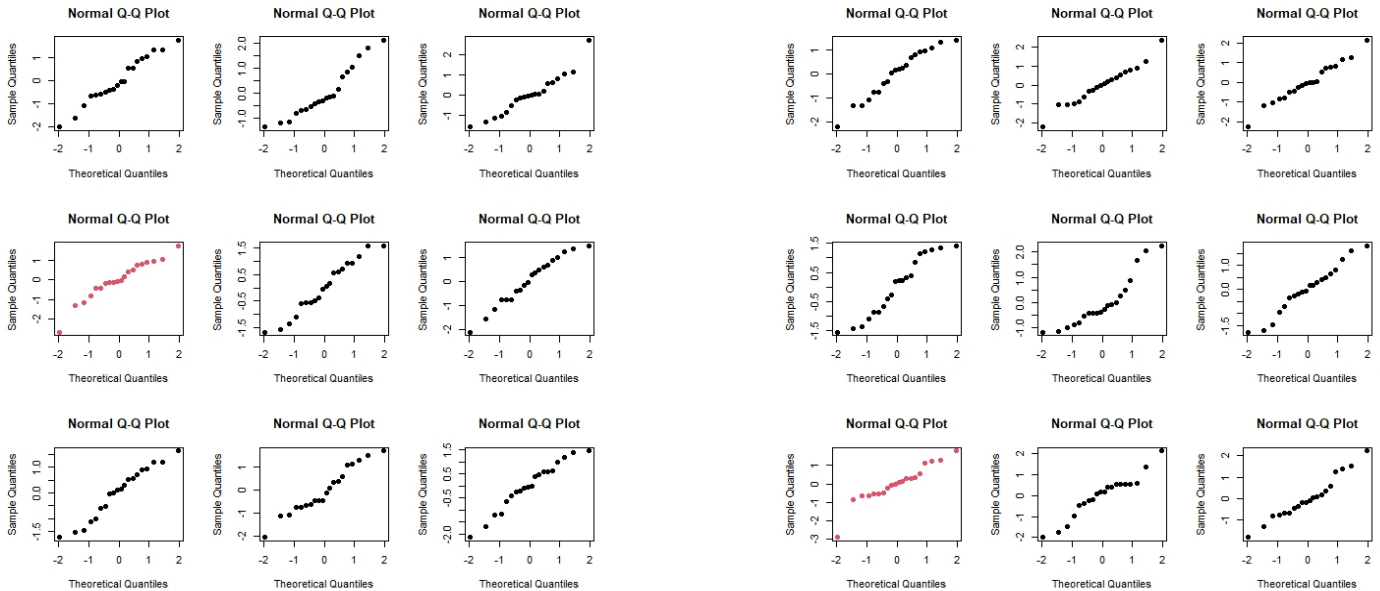
- **How do we identify a deficiency in a model?**
  - **train our eye on graphics, then try to extract a rule**



- **Here, eight are normals, one is t w/**4 **df**
  - **random samples of size** 20

Training one's eye

- **Write a bit of R code**
  - **repeat many times, revealing "bad sample" after each replicate**



- **Simulation, so we control all – n, departure, distance from null**
  - **begin with easy identifications, then move to more subtle deficiencies. same approach we use when teaching residual analysis**

Parting words

- Splitting data is needed for predictive work in low-info setting
  - low end - minimal update before computing predictive
    - echoes old work on Bayes factor with (min training sample)
  - high end - maximal update before computing predictive
    - as in old work on predictive marginal likelihood (last case)
    - leave one out cross-validation; B&B
  - advocate middle ground, pass from low-info to high-info
    - akin to calibrated Bayes factor (arises in BNP)
- Splits need not be "by case"
  - decompose likelihood (B&B; Bayesian restricted likelihood)
  - we're comfortable sacrificing some information
    - for computational simplicity
    - for robustness – here, searching for deficiencies
    - for freedom to investigate many departures

Parting words

- Purposive splits
  - recover known results as split partial update, followed with predictive check
    - may be on cases / may be decomposition of likelihood
      - framework allows us to target parameters
      - allows us to target tails / outliers
      - fully compatible with graphical as well as numerical checks
- Fits within framework of Bayesian restricted likelihood
  - predictive checks identify deficiencies but don't fix them
  - if you know how to fix defects, do so
  - if not, may wish to consider living with imperfection
    - partial update from decomposition is okay!

A very few references

- **Big ones for predictive checks - much more from these and others**
  - **Geisser, S. (1975). The predictive sample reuse method with applications. JASA 70, 320-328.**
  - **Box, G.E.P. (1980). Sampling and Bayes' inference in scientific modelling and robustness (w/disc). JRSSB 143, 383-430.**
  - **Bayarri, M.J., and Berger, J.O. (2000). P values for composite null models. JASA 95, 1127-1142.**
  - **Moran, G.E., Cunningham, J.P., and Blei, D.M. (2023). The posterior predictive null. BA 18, 1071-1097.**
- **A few I'm connected to with relevance to this talk**
  - **Lewis, J., Lee, Y., and MacEachern, S.N. (2012). Robust inference via the blended paradigm.**
  - **Xu, X., Lu, P., MacEachern, S.N., and Xu, R. (2019). Calibrated Bayes factors for model comparison. JSCS 89, 591-614.**
  - **Lewis, J., MacEachern, S.N., Lee, Y. (2021). Bayesian restricted likelihood methods: conditioning on insufficient statistics in Bayesian regression (w/disc). BA 16, 1393-1462.**
  - **Hwang, Y., Kim, H.J., Chang, W., Hong, C., and MacEachern, S.N. (2025). Bayesian model calibration and sensitivity analysis for oscillating biological experiments. Technom 67, 333-343.**