Bayesian Modeling via Goodness-of-Fit

Deep Mukhopadhyay and Doug Fletcher

Department of Statistical Science, Fox School of Business
Temple University
“The FDA (for example) doesn’t care about Pfizer’s prior opinion of how well its new drug will work, it wants objective proof. Pfizer, on the other hand may care very much about its own opinions in planning future drug development.”
Introduction
“The FDA (for example) doesn’t care about Pfizer’s prior opinion of how well it’s new drug will work, it wants objective proof. Pfizer, on the other hand may care very much about its own opinions in planning future drug development.”

- Frequentists view prior as a weakness that can hamper scientific objectivity and can corrupt the final statistical inference.

- whereas Bayesians view it as a strength to include relevant domain-knowledge into the data analysis.

WHO IS RIGHT?
• Frequentists view prior as a *weakness* that can hamper scientific objectivity and can corrupt the final statistical inference.

• whereas Bayesians view it as a *strength* to include relevant domain-knowledge into the data analysis.

**WHO IS RIGHT?**

In fact, Both Camps Are Absolutely Right!
Frequentists view prior as a weakness that can hamper scientific objectivity and can corrupt the final statistical inference.

whereas Bayesians view it as a strength to include relevant domain-knowledge into the data analysis.

Thus, probably a better question to ask is:

*How can we develop a ‘Bayes + Frequentist’ data analysis workflow that can incorporate relevant expert-knowledge without sacrificing the scientific objectivity?*

1. The answer lies in our ability to interrogate the credibility of an initial scientific prior in order to uncover its blind spots.

---

1 This question has a broader relevance for designing intelligent machine that can judiciously blend data and expert advice.
• This dataset includes $k = 70$ experiments;
• For each study, $y_i$ denotes the number of rats with tumors among $n_i$ rats: $y_i \mid \theta_i \stackrel{\text{ind}}{\sim} \text{Binomial}(n_i, \theta_i)$.

<table>
<thead>
<tr>
<th>$y_i$</th>
<th>$n_i$</th>
<th>Frequency</th>
<th>$y_i$</th>
<th>$n_i$</th>
<th>Frequency</th>
<th>$y_i$</th>
<th>$n_i$</th>
<th>Frequency</th>
<th>$y_i$</th>
<th>$n_i$</th>
<th>Frequency</th>
<th>$y_i$</th>
<th>$n_i$</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>20</td>
<td>7</td>
<td>2</td>
<td>25</td>
<td>1</td>
<td>2</td>
<td>17</td>
<td>1</td>
<td>10</td>
<td>48</td>
<td>1</td>
<td>6</td>
<td>22</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>19</td>
<td>4</td>
<td>2</td>
<td>24</td>
<td>1</td>
<td>7</td>
<td>49</td>
<td>1</td>
<td>4</td>
<td>19</td>
<td>3</td>
<td>6</td>
<td>20</td>
<td>3</td>
</tr>
<tr>
<td>0</td>
<td>18</td>
<td>2</td>
<td>2</td>
<td>23</td>
<td>1</td>
<td>7</td>
<td>47</td>
<td>1</td>
<td>5</td>
<td>22</td>
<td>1</td>
<td>16</td>
<td>52</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>17</td>
<td>1</td>
<td>2</td>
<td>20</td>
<td>6</td>
<td>3</td>
<td>20</td>
<td>2</td>
<td>11</td>
<td>46</td>
<td>1</td>
<td>15</td>
<td>47</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>20</td>
<td>4</td>
<td>1</td>
<td>10</td>
<td>1</td>
<td>2</td>
<td>13</td>
<td>1</td>
<td>12</td>
<td>49</td>
<td>1</td>
<td>15</td>
<td>46</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>19</td>
<td>2</td>
<td>5</td>
<td>49</td>
<td>1</td>
<td>9</td>
<td>48</td>
<td>1</td>
<td>5</td>
<td>20</td>
<td>2</td>
<td>9</td>
<td>24</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>18</td>
<td>2</td>
<td>2</td>
<td>19</td>
<td>1</td>
<td>10</td>
<td>50</td>
<td>1</td>
<td>6</td>
<td>23</td>
<td>1</td>
<td>5</td>
<td>19</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>27</td>
<td>1</td>
<td>5</td>
<td>46</td>
<td>1</td>
<td>4</td>
<td>20</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

• **MacroInference**: For drug development applications, one important goal is to estimate overall tumor probability $\theta$.

• **MicroInference**: Given an additional new study: $y_{71} = 4$ out of $n_{71} = 14$ rats developed tumor; *How can we estimate $\theta_{71}$?*
Step 1. Construct Scientific Beta prior $g(\theta; \alpha, \beta)$

Let’s say we are given some additional information: the probability of tumor is expected to be around 0.14 with sd 0.084; solve for $\hat{\alpha} = 2.3$ and $\hat{\beta} = 14.08$; construct the starting Beta($\theta; \alpha, \beta$):

- Expert clinical knowledge: It comes from the medical officers’ knowledge on the disease and the treatment.

- External clinical evidences:
  - Database search: based on aggregating results from similar studies from electronic databases PubMed, ScienceDirect, Google Scholar etc.
  - Use of pilot/historical datasets [i.e, $k=70$ studies in our context] to quickly estimate a meaningful $\hat{\alpha}$ and $\hat{\beta}$. 
Step 2. Bayesian Inference

The Model:

\[ y_i | \theta_i \overset{\text{ind}}{\sim} \text{Binomial}(n_i, \theta_i), \quad (i = 1, \ldots, k) \]
\[ \theta_i \sim \text{Beta}(2.3, 14.08). \]

- **MacroInference**: The probabilities of tumor across \( k = 70 \) studies can be summarized by the prior mean:

\[
\frac{\alpha}{\alpha + \beta} = \frac{2.3}{2.3 + 14.1} = 0.141
\]

- **MicroInference**: Given \( k = 70 \) historical studies, the probability of a tumor \( \theta_{71} \) for the new clinical study:

\[
\pi_G(\theta_{71} | y_{71}) = \text{Beta}(\alpha + y_{71}, \beta - y_{71} + n_{71})
\]
\[
\mathbb{E}_G[\Theta_{71} | y_{71} = 4] = \hat{\theta}_{71} = \frac{\alpha + y_{71}}{\alpha + \beta + n_{71}} = \frac{2.3 + 4}{2.3 + 14.1 + 14} = 0.207
\]
Bayesian learning is completely automatic (Thanks to Bayes’ rule) once we pick a $\pi(\theta)$.

The Achilles’ heel: Why a regulator should believe your handpicked prior $g(\theta)$ at its face value?

**Million Dollar Question: How can we defend the pre-selected $g(\theta)$?**

- How to check the appropriateness of the $g(\theta)$?
- Beyond Yes/No answer, can we quantify and characterize the uncertainty of $g$ to better understand the nature of misfit?
- Finally, we would like to provide a simple, yet formal guideline for upgrading (repairing) the starting $g(\theta)$?
Bayesian Learning as “one coherent whole”

Step 1
Begin with a scientific (or empirical) parametric prior $g$.

Step 2
Inspect the ‘goodness’ of the elicited prior.

Step 3
Estimate the required ‘correction’ for $g$.

Step 4
Generate the final statistical estimate $\hat{\pi}(\theta)$

Step 5
Execute Macro and MicroInference
Pre-Inferential Modeling
Step 2: Why should I believe your prior?

```
library("BayesGOF")
rat.ds <- DS.prior(rat, g.par = c(2.3,14.08), family = "Binomial")
plot(rat.ds, plot.type = "Ufunc")
```

- The **U-function** allows us to visualize the compatibility of \( g \equiv \text{Beta}(2.3, 14.08) \) with the observed data.
- If **U-function** \( \equiv 1 \rightarrow \text{No conflict.} \)
- Shape of **U-function** \( \rightarrow \text{Insight into unexpected deeper structure.} \)
- There is a misfit between \( \text{Beta}(2.3, 14.08) \) and the observed data by having an “extra” mode.

**Figure 1:** U-function: Rat Tumor Data. Informs users on the “nature” of misfit.
Certifying Business-as-usual Bayesian Modeling

- **Terbinafine** data comprise $k = 41$: $y_i$ is the number of patients whose treatment terminated early due to some adverse effect

  $$g(\theta) = \text{Beta}(1.24, 34.7)$$

- The **ulcer** data consists of $k = 40$ studies; each trial has a log-odds ratio $y_i|\theta_i \sim \mathcal{N}(\theta_i, s_i^2)$ measures the rate of recurrent bleeding given the surgical treatment.

  $$g(\theta) = \mathcal{N}(-1.17, 0.98)$$

*Figure 2:* A ‘flat’ U-function indicates no adjustment required. One can *safely* proceed in turning the Bayesian crank.

Our approach: neither Parametric nor Nonparametric, it *includes both*. The U-function “connects” the two philosophies.
The Model

- A universal class of prior density models:
  \[ \pi(\theta) = g(\theta) \times d[G(\theta); G, \Pi] \]

where
\[
d[u; G, \Pi] = \frac{\pi(G^{-1}(u))}{g(G^{-1}(u))}, \quad 0 < u < 1
\]
satisfying \( \int_0^1 d[u; G, \Pi] \, du = 1. \)

- It has a unique two-component structure that combines assumed parametric \( g \) with the \( d \)-function.

- \( d(u; G, \Pi) \) refines the initial guess \( g \).

- It also describes the excess uncertainty of the assumed \( g(\theta; \alpha, \beta) \). For that reason we call it the U-function.
Generalized Empirical Bayes Prior

\[ \pi(\theta) = g(\theta) \times d[G(\theta); G, \Pi] \]

- Parametric or Scientific Prior
- Nonparametric or Empirical rectifier

1. Something in between: PEB \( \subseteq \) gEB \( \subseteq \) NEB.
2. Combines parametric stability with nonparametric flexibility.
3. Works for small as well as large number of parallel cases.
The DS$(G, m)$ prior

- The square integrable $d[G(\theta); G, \Pi] \in L^2(G)$ can be expanded as:

$$\text{DS}(G, m) : \quad \pi(\theta) = g(\theta)[1 + \sum_{j=1}^{m} \text{LP}[j; G, \Pi] T_j(\theta; G)]$$

- where the $\{T_j\}$ are orthonormal basis with respect to measure $G$:

$$\int T_i(\theta; G)T_j(\theta; G)dG = \delta_{ij}$$

- We choose $T_j(\theta; G)$ to be $\text{Leg}_j[G(\theta)]$, a member of LP-rank polynomials. Robust + Automatic for arbitrary $G$ continuous.

- $\text{DS}(G, m = 0) \equiv g(\theta; \alpha, \beta)$ The truncation point $m$ reflects the concentration of permissible $\pi$ around a known $g$. 
Step 3: How Can I ‘Quantify’ Prior Uncertainty?

- Prior uncertainty quantification:

\[
q_{\text{LP}}(G||\Pi) = \sum_j \left| \text{LP}[j; G, \Pi] \right|^2 = \int_0^1 d^2(u; G, \Pi) \, du - 1.
\]

- It captures the departure of the U-function from uniformity.

- Some interesting connection with relative entropy:

\[
q_{\text{LP}}(G||\Pi) \approx 2 \times \text{KL}(\Pi||G).
\]

where \( \text{KL}(\Pi||G) \) is the Kullback–Leibler (KL) divergence between the true prior \( \pi \) and its parametric approximate \( g \).

- One can use this tool to “play” with multiple expert opinions [hyperparameters], in order to filter out the reasonable ones.
Step 4. How Can I ‘Repair’ My Starting $g(\theta)$?

- If $g$ is inconsistent with the data: what to do next?
- DS($G, m$) model: A simple, yet formal, guideline for upgrading:

$$\hat{\pi}(\theta) = g(\theta; \hat{\alpha}, \hat{\beta}) \times \hat{d}[G(\theta); G, \Pi].$$

- Our formalism addresses (in one-shot): (1) Quantification (What); (2) Characterization (Why); (3) Synthesis (How)

Modeling the “gap” between the parametric $g$ and the true $\pi$ often far easier than modeling $\pi$ from scratch.
Estimation & Algorithm
The Basic Idea

• If $\theta_i$ were observable, we could estimate the LP-Fourier coeffs $LP[j; G, \Pi] = \langle d, T_j \circ G^{-1} \rangle_{\mathcal{L}^2(0,1)}$ by their empirical counterpart:

$$\hat{LP}[j; G, \Pi] = \widetilde{E}_{LP} [T_j(\Theta_i; G)] = k^{-1} \sum_{i=1}^k T_j(\theta_i; G).$$

• But $\theta_i$’s are unobserved. An obvious proxy for $T_j(\theta_i; G)$ would be its posterior mean $E_{LP}[T_j(\Theta_i; G)|y_i]$, leads to ‘ghost’ LP-estimates:

$$\widetilde{LP}[j; G, \Pi] = k^{-1} \sum_{i=1}^k E_{LP}[T_j(\Theta_i; G)|y_i]$$

Simple Estimation Strategy

Step 1. Initialize: $LP^{(0)}[j; G, \Pi] = 0$ for $j = 1, \ldots, m$.

Step 2. Compute ‘ghost’ LP-estimates $\{\widetilde{LP}^{(\ell-1)}[j; G, \Pi]\}_{j=1}^m$

Step 3. Repeat until convergence: $\sum_{j=1}^m |\widetilde{LP}^{(\ell)}[j; G, \Pi] - \widetilde{LP}^{(\ell-1)}[j; G, \Pi]|^2 \leq \epsilon$
Closed-form Posterior Modeling

\[ y_i|\theta_i \overset{\text{ind}}{\sim} f(y_i|\theta_i), \quad (i = 1, \ldots, k) \]  

\[ \theta_i \overset{\text{ind}}{\sim} \pi(\theta) \]

where \( \pi(\theta) \sim DS(G, m) \) model with conjugate \( G \).

- The posterior distribution of \( \Theta_i \) given \( y_i \):

\[
\pi_{LP}(\theta_i|y_i) = \frac{\pi_G(\theta_i|y_i)(1 + \sum_j \text{LP}[j; G, \Pi] T_j(\theta_i; G))}{1 + \sum_j \text{LP}[j; G, \Pi] \mathbb{E}_G[T_j(\Theta_i; G)|y_i]}
\]

- For any general random variable \( h(\Theta_i) \), the Bayes estimate:

\[
\mathbb{E}_{LP}[h(\Theta_i)|y_i] = \frac{\mathbb{E}_G[h(\Theta_i)|y_i] + \sum_j \text{LP}[j; G, \Pi] \mathbb{E}_G[h(\Theta_i)T_j(\Theta_i; G)|y_i]}{1 + \sum_j \text{LP}[j; G, \Pi] \mathbb{E}_G[T_j(\Theta_i; G)|y_i]}
\]
The derived analytical expressions are valid for any conjugate pairs—Towards a general representation theory.

| Family         | Conjugate g-prior | Marginal \( f_G(y_i) \) | Posterior \( \pi_G(\theta_i | y_i) \) |
|----------------|-------------------|------------------------|-----------------------------------|
| Binomial \((n, \theta)\) | Beta(\(\alpha, \beta\)) | \( \binom{n_i}{y_i} \frac{(\alpha+y_i, \beta-y_i+n_i)}{(\alpha, \beta)} \) | Beta(\(\alpha+y_i, \beta-y_i+n_i\)) |
| Poisson \((\theta)\)       | Gamma(\(\alpha, \beta\)) | \( (y_i+\alpha^{-1})p^\alpha (1-p)^{y_i} \) | Gamma(\(\alpha+y_i, \frac{\beta}{1+\beta} \)) |
| Normal \((\theta, \sigma^2)\) | Normal(\(\alpha, \beta^2\)) | Normal(\(\alpha, \sigma_i^2 + \beta^2\)) | Normal(\(\lambda_i \alpha + (1-\lambda_i)y_i, (1-\lambda_i)\sigma_i^2 \)) |
| Exp(\(\lambda\))     | Gamma(\(\alpha, \beta\)) | \( \frac{\alpha \beta}{(1+\beta y)^{\alpha+1}} \) | Gamma(\(\alpha + 1, \frac{\beta}{1+\beta y_i} \)) |

**Table 1:** The marginal and posterior distributions for four familiar distributions (two discrete and two continuous): Binomial, Poisson, Normal, and Exponential.
Bayesian Inference
`rat.macro <- DS.macro.inf(rat.ds, method = "mode")`

`plot(rat.macro)`

- Bimodality implies two distinct groups of $\theta_i$, a case which is in between two extremes: homogeneity and complete heterogeneity.
- A single mean would overestimate one group and underestimate the other.
- Modes are better representative:

  \[
  \hat{\pi}(\theta) = g(\theta; \alpha, \beta) \left[ 1 - 0.5 T_3(\theta; G) \right]
  \]

  \(\triangle\) Mode 1: 0.034 ± 0.014
  \(\triangle\) Mode 2: 0.156 ± 0.012

**Figure 4:** Estimated $\hat{\pi}$ with mode (red triangles) ± SDs.
**MacroInference: Structured Heterogeneity**

```r
> rat.macro <- DS.macro.inf(rat.ds, method = "mode")
> plot(rat.macro)
```

**Figure 4:** Estimated $\hat{\pi}$ with mode (red triangles) ± SDs.

- Bimodality implies **two distinct groups** of $\theta_i$, a case which is in between two extremes: homogeneity and complete heterogeneity.
- A **single** mean would **overestimate** one group and **underestimate** the other.
- **Modes** are better representative:

  $$\hat{\pi}(\theta) = g(\theta; \alpha, \beta) \left[1 - 0.5T_3(\theta; G)\right]$$

  △ Mode 1: 0.034 ± 0.014
  △ Mode 2: 0.156 ± 0.012

The ‘science of combining’ critically depends on the **shape** of $\hat{\pi}$. 

---
MicroInference: Balancing Robustness & Efficiency

What’s your estimate for $\theta_{71}$ (prob of a tumor for the new study)?

• Stein’s formula: shrinks $\tilde{\theta}_i = y_i/n_i$ towards prior mean $\approx .14$

$$\tilde{\theta}_i = \frac{n_i}{\alpha + \beta + n_i} \hat{\theta}_i + \frac{\alpha + \beta}{\alpha + \beta + n_i} \mathbb{E}_G[\Theta]$$

• Where to shrink? How can we rectify parametric Stein’s formula?

$$\hat{\theta}_i = \frac{\tilde{\theta}_i + \sum_j \text{LP}[j; G, \Pi] \mathbb{E}_G[\Theta_i T_j(\Theta_i; G)|y_i, n_i]}{1 + \sum_j \text{LP}[j; G, \Pi] \mathbb{E}_G[T_j(\Theta_i; G)|y_i, n_i]}$$

• When all $\text{LP}[j; G, \pi] = 0$, it reduces to Stein’s formula \([\text{Efficiency}]\)

• LP-coeffs determine the magnitude and direction of shrinkage in a completely data-driven manner, when needed. \([\text{Robustness}]\)

Robbins (1980): Can we resolve this efficiency-robustness dilemma?
> rat.micro.y71 <- DS.micro.inf(rat.ds, y.θ = 4, n.θ = 14)
> plot(rat.micro.y71, xlim = c(0,0.5))

• Interestingly, $\hat{\pi}(\theta_{71}|y_{71} = 4)$ (red curve) shows less variability than PEB (blue dotted). Possibly due to the selective shrinkage ability of our method, which learns from similar studies (e.g. group 2), rather than the whole heterogeneous mix of studies.

• Adaptively shrinks empirical $\tilde{\theta}_i = y_i/n_i$ towards the respective mode; PEB uses the grand mean ($\approx 0.14$) for ALL estimates.
Table 2: List of datasets along by distribution family and sources. They are sorted by family and according to $k$: from large to small-scale studies.

<table>
<thead>
<tr>
<th>Dataset</th>
<th># Studies ($k$)</th>
<th>Family</th>
<th>Sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>Surgical Node</td>
<td>844</td>
<td>Binomial</td>
<td>Efron (2016)</td>
</tr>
<tr>
<td>Rolling Tacks</td>
<td>320</td>
<td>Binomial</td>
<td>Beckett and Diaconis (1994)</td>
</tr>
<tr>
<td>Rat Tumor</td>
<td>70</td>
<td>Binomial</td>
<td>Gelman et al. (2013, Ch. 5)</td>
</tr>
<tr>
<td>Terbinafine</td>
<td>41</td>
<td>Binomial</td>
<td>Young-Xu and Chan (2008)</td>
</tr>
<tr>
<td>Naval Shipyard</td>
<td>5</td>
<td>Binomial</td>
<td>Martz et al. (1974)</td>
</tr>
<tr>
<td>Galaxy</td>
<td>324</td>
<td>Gaussian</td>
<td>De Blok et al. (2001)</td>
</tr>
<tr>
<td>Ulcer</td>
<td>40</td>
<td>Gaussian</td>
<td>Sacks et al. (1990); Efron (1996)</td>
</tr>
<tr>
<td>Arsenic</td>
<td>28</td>
<td>Gaussian</td>
<td>Willie and Berman (1995)</td>
</tr>
<tr>
<td>Insurance</td>
<td>9461</td>
<td>Poisson</td>
<td>Efron and Hastie (2016)</td>
</tr>
<tr>
<td>Butterfly</td>
<td>501</td>
<td>Poisson</td>
<td>Efron and Hastie (2016)</td>
</tr>
<tr>
<td>Norberg</td>
<td>72</td>
<td>Poisson</td>
<td>Norberg (1989)</td>
</tr>
</tbody>
</table>
BayesGOF R-Package

BayesGOF: Bayesian Modeling via Goodness of Fit

A Bayesian data modeling scheme that performs four interconnected tasks: (i) characterizes the uncertainty of the elicited parametric prior; (ii) provides exploratory diagnostic for checking prior-data conflict; (iii) computes the final statistical prior density estimate; and (iv) executes macro- and micro-inference. Primary reference is Mukhopadhyay, S. and Fletcher, D. (2018, Technical Report, 

Version: 2.1
Depends: 
  orthopoly, VGAM
Suggests: 
  knitr, rmarkdown
Published: 2018-02-08
Author: Subhadeep Mukhopadhyay, Douglas Fletcher
Maintainer: Doug Fletcher <tug25070 at temple.edu>
License: GPL-2
NeedsCompilation: no
CRAN checks: BayesGOF results

Downloads:

Reference manual: BayesGOF.pdf
Vignettes: Bayes via Goodness of Fit
Package source: BayesGOF_2.1.tar.gz
Windows binaries: r-devel: BayesGOF_2.1.zip, r-release: BayesGOF_2.1.zip, r-oldrel: BayesGOF_2.1.zip
OS X El Capitan binaries: r-release: BayesGOF_2.1.tar.gz
OS X Mavericks binaries: r-oldrel: BayesGOF_1.4.tar.gz
Old sources: BayesGOF archive

Linking:

Please use the canonical form https://CRAN.R-project.org/package=BayesGOF to link to this page.

It has been downloaded > 3500 times
Conclusion
The High-Order Bits

The main attractions of the “Bayes via goodness of fit” framework:

(1) A systematic strategy to go from a scientific prior to a statistical prior by examining the credibility of a self-selected $g$.

(2) It has a distinct exploratory flavor that encourages interactive Bayesian learning rather than blindly “turning the crank.”

(3) The theory is general enough to include almost all commonly used models + yields closed-form analytic solutions for posterior modeling.

(4) Most importantly, No expensive MCMC or variational methods are required. Easy to implement + Computationally fast.
A Personal Story...

On what lead me to this research:

• It may seem that I had the noble intention to declutter Bayesian statistics. But in reality, that was not the case.
On what lead me to this research:

• It may seem that I had the noble intention to declutter Bayesian statistics. But in reality, that was not the case.

• **Tuesday, Aug 2nd, 2016**: I met Brad at the JSM to discuss some ideas, which had nothing to do with Empirical Bayes.
On what lead me to this research:

- It may seem that I had the noble intention to declutter Bayesian statistics. But in reality, that was not the case.

- **Tuesday, Aug 2nd, 2016**: I met Brad at the JSM to discuss some ideas, which had nothing to do with Empirical Bayes.

- Halfway through our conversation, I told him: “I enjoyed reading the last chapter of the CASI book.” Brad promptly replied: “g-modeling is close to my heart.”

I interpreted it: ‘a problem that really matters’ and devoted my next one year to figure out the right question to ask. The rest were details.
If “Statistics learns from experience” then Statisticians learn from Brad Efron, and It will continue.
If “Statistics learns from experience” then Statisticians learn from Brad Efron, and It will continue.

Thank You, and Happy 80th Birthday Brad.
Some Related References


APPENDIX: OTHER PRACTICAL CONSIDERATIONS
The motivations behind the name ‘DS-Prior’ are twofold. First, our formulation operationalizes I. J. Good’s ‘Successive Deepening’ idea for Bayesian data analysis:

“A hypothesis is formulated, and, if it explains enough, it is judged to be probably approximately correct. The next stage is to try to improve it. The form that this approach often takes in EDA is to examine residuals for patterns, or to treat them as if they were original data” (I. J. Good, 1983, p. 289).

Secondly, our prior has two components: A Scientific $g$ that encodes an expert’s knowledge and a Data-driven $d$. That is to say that our framework embraces data and science, both, in a testable manner.
A1. Determining an appropriate $m$

'Elbow' plot for determining an appropriate $m$. The plot shows the BIC deviance for the LP coefficients for each $m$ value.

\[
\text{BIC}(m) = \sum_{j=1}^{m} |\hat{LP}[j; G, \Pi]|^2 - \frac{m \log(k)}{k}.
\]
A2. The DS\((G, m)\) Sampler

The following algorithm generates samples from the DS\((G, m)\) model via accept/reject scheme.

**DS\((G, m)\) Sampling Algorithm**

**Step 1.** Generate \(\Theta\) from \(g\); independent of \(\Theta\), generate \(U\) from Uniform\([0, 1]\).

**Step 2.** Accept and set \(\Theta^* = \Theta\) if

\[
\hat{d}[G(\theta); G, \Pi] > U \max_u \hat{d}(u; G, \Pi);
\]

otherwise, discard \(\Theta\) and return to Step 1.

**Step 3.** Repeat until simulated sample of size \(k\), \(\{\theta_1^*, \theta_2^*, \ldots, \theta_k^*\}\).

Note when \(\hat{d} \equiv 1\), DS\((G, m)\) automatically samples from parametric \(G\).
A3. When No Prior Knowledge is Available

Model: \( y_i|\theta_i \sim \text{Binomial}(50, \theta_i) \) with \( i = 1, \ldots, k = 90 \) and the true prior \( \pi(\theta) = .3\text{Beta}(4, 6) + .7\text{Beta}(20, 10) \). How well we approximate the unknown \( \pi \) without any prior knowledge of its shape?

**Figure 6:** The first panel (a) finds the “elbow” in the BIC\((m)\) deviance plot at \( m = 3 \); (b) shows the U-function, while (c) plots the true \( \pi(\theta) \) (black) along with the estimated DS prior (red) \( \hat{\pi}(\theta) = g(\theta; \hat{\alpha}, \hat{\beta})[1 - 0.48T_3(\theta; G)] \) with MLE \( \hat{\alpha} = 4.16 \) and \( \hat{\beta} = 3.04 \).
Setting: We observe \( Y_i = \theta_i + \epsilon_i, \ i = 1 \cdots k \), where \( \epsilon_i \overset{\text{ind}}{\sim} \text{Normal}(0, 1) \), and \( \theta_i = \pm 1 \) with probability \( \eta \) and \( 1 - \eta \) respectively.

Goal: Estimate \( k \)-vector \( \theta \in \{ -1, 1 \}^k \) under \( L(\hat{\theta}, \theta) = k^{-1} \sum_{i=1}^{k} |\hat{\theta}_i - \theta_i| \).

**Figure 7:** The ratio of empirical risks: DS and NPMLE methods to Robbins’ ‘compound decision’ problem.
A5. The Expansion Basis: Shapes

- Robust basis: Polynomial of rank transform $G(\theta)$, not $\theta$.
- Orthonormal with respect to $L^2(G)$, for arbitrary $G$ (continuous).
- This is not to be confused with standard Legendre polynomials $\text{Leg}_j(u), 0 < u < 1$, which are orthonormal with respect to $U[0, 1]$.

![Figure 8: LP-polynomials $T_j(\theta; G_{\alpha, \beta})$ for family= "beta" for (a) Jeffrey’s prior ($\alpha = \beta = 0.5$), (b) Uniform ($\alpha = \beta = 1$), and for (c) ($\alpha = 3, \beta = 4$).]
A6. The Pharma-Example

The following example depicts a scenario that is very common in historic-controlled clinical trials:

\[
\pi(\theta) = \eta \text{ Beta}(5, 45) + (1 - \eta) \text{ Beta}(30, 70)
\]

\[y_{\text{new}} \sim \text{ Bin}(50, 0.3)\]

- 0 \leq \eta \leq 0.5: larger values indicate more heterogeneity in the historical studies.

- Generate 100 \(\theta_i\) from \(\pi(\theta)\), and then \(y \leftarrow \text{rbinom}(100, 60, \theta)\).

- \(y_{\text{new}} \leftarrow \text{rbinom}(1, 50, 0.3)\)

- How accurately we can estimate \(\theta_{\text{new}}\) under various levels of contamination?

- Repeat process 250 times for each value of \(\eta\) and find MSE for each estimate.

**Figure 9:** Prior-data conflict for \(\eta = 0.1\) versus \(\eta = 0.4\); and ‘*’ denotes .3, the true mean of \(y_{\text{new}}\).
Effect of Selective Shrinkage

Figure 10: Comparing MSE of different methods.

- Interesting pattern of freq. MLE as heterogeneity increases.
- For all $\eta$: $\text{MSE(DS-Bayes)} \leq \text{MSE(PEB)}$ The efficiency continues to increase with $\eta$ due to selective shrinkage ability –“borrowing strength” from similar studies only (near .3).
- Efron’s Bayes deconvolution and Koenker’s NPMLE are also promising, specially for $0 < \eta \leq 0.15$. 
A7. Finite Bayes Correction (Efron 2018): Rat Data $\theta_{71}$

- Finite Bayes: The “inflated” (green) posterior dist. $\tilde{\pi}(\theta_{71} | y_{71})$.
- 90% gEB credible intervals: $(0.1904 - 0.092, 0.1904 + 0.132)$. 

![Graph showing posterior distributions and credible intervals](graph.png)