Generalized Empirical Bayes Modeling *via* Frequentist Goodness of Fit

Subhadeep Mukhopadhyay, Douglas Fletcher

Temple University, Department of Statistical Science

Philadelphia, Pennsylvania, 19122, U.S.A.

Dedicated to 80th birthday anniversary of Brad Efron

Abstract

The two key issues of modern Bayesian statistics are: (i) establishing principled approach for *distilling* statistical prior that is *consistent* with the given data from an initial believable scientific prior; and (ii) development of a *consolidated* Bayes-frequentist data analysis workflow that is more effective than either of the two separately. In this paper, we propose the idea of "Bayes *via* goodness-of-fit" as a framework for exploring these fundamental questions, in a way that is general enough to embrace almost all of the familiar probability models. Several examples, spanning application areas such as clinical trials, metrology, insurance, medicine, and ecology show the unique benefit of this new point of view as a practical data science tool.

Keywords: Exploratory Bayes Modeling; Prior uncertainty modeling; Empirical Bayes.

1 Introduction

Bayesians and frequentists have long been ambivalent toward each other [1, 2, 3]. The concept of "prior" remains the center of this 250 years old tug-of-war: 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 incorporate relevant domain-knowledge into the data analysis. The question naturally arises: how can we develop a consolidated Bayes-frequentist data analysis workflow [4, 5, 6, 7] that enjoys the best of both worlds? The objective of this paper is to develop one such modeling framework.

We observe samples $y = (y_1, \ldots, y_k)$ from a known probability distribution $f(y|\theta)$, where the unobserved parameters $\theta = (\theta_1, \ldots, \theta_k)$ are independent realizations from unknown $\pi(\theta)$. Given such a model, Bayesian inference typically aims at answering the following two questions:

- MacroInference: How should we combine k model parameters to come up with an overall, macro-level aggregated statistical behavior of $\theta_1, \ldots, \theta_k$?

- MicroInference: Given the observables y_i, how should we simultaneously estimate individual micro-level parameters θ_i?

Thanks to Bayes' rule, answers to these questions are fairly straightforward and automatic once we have the observed data $\{y_i\}_{i=1}^k$ and a specific choice for $\pi(\theta)$. A common practice is to choose π as the parametric conjugate prior $g(\theta; \alpha, \beta)$, where the hyper-parameters are either selected based on an investigator's expert input or estimated from the data (current/historical) when little prior information is available .

Motivating Questions. However, an applied Bayesian statistician may find it unsatisfactory to work with an initial believable prior $g(\theta)$ at its face value, without being able to interrogate its credibility in the light of the observed data [8, 9] as this choice unavoidably shapes his or her final inferences and decisions. A good statistical practice thus demands greater transparency to address this trust-deficit. What is needed is a justifiable class of prior distributions to answer the following *pre*-inferential modeling questions: Why should I believe your prior? How to check its appropriateness (self-diagnosis)? How to quantify and characterize the uncertainty of the a priori selected g? Can we use that information to "refine" the starting prior (*auto*-correction), which is to be used for subsequent inference? In the end, the question remains: how can we develop a systematic and principled approach to go from a *scientific* prior to a *statistical* prior that is consistent with the current data? A resolution of these questions is necessary to develop a "dependable and defensible" Bayesian data analysis workflow, which is the goal of the "Bayes *via* goodness-of-fit" technology.

Summary of Contributions. This paper provides some practical strategies for addressing these questions by introducing a general modeling framework, along with concrete guidelines for applied users. The major practical advantages of our proposal are: (i) computational ease (it does not require Markov chain Monte Carlo (MCMC), variational methods, or any other sophisticated computational techniques); (ii) simplicity and interpretability of the underlying theoretical framework which is general enough to include *almost all* commonly encountered models; and (iii) easy integration with mainframe Bayesian analysis that makes it readily applicable to a wide range of problems. The next section introduces a new class of nonparametric priors $\mathrm{DS}(G, m)$ along with its role in exploratory graphical diagnostic and uncertainty quantification. The estimation theory, algorithm, and real data examples are discussed in Section 3. Consequences for inference are discussed in Section 4, which include methods of combining heterogeneous studies and a generalized nonparametric Stein-prediction formula that selectively borrows strength from 'similar' experiments in an automated manner. Section 4.2 describes a new theory of 'learning from uncertain data,' which is an important problem in many application fields including metrology, physics, and chemistry. Section 4.4 solves a long-standing puzzle of modern empirical Bayes, originally posed by Herbert Robbins [10]. We conclude the paper with some final remarks in Section 5. Connections with other Bayesian cultures are presented in the supplementary material to ensure the smooth flow of main ideas.

Real-data Applications. To demonstrate the versatility of the proposed "Bayes *via* goodness-of-fit" data analysis scheme, we selected examples from a wide range of models including normal, Poisson, and Binomial distributions. The full catalog of datasets is presented in Supplementary Table 6.

Notation. The notation g and G denote the density and distribution function of the starting prior, while π and Π denote the density and distribution function of the unknown oracle prior. We will denote the conjugate prior with hyperparameters α and β by $g(\theta; \alpha, \beta)$. Let $\mathscr{L}^2(\mu)$ be the space of square integrable functions with inner product $\int f(u)g(u)\, \mathrm{d}\mu(u)$. $\mathrm{Leg}_j(u)$ denotes jth shifted orthonormal Legendre polynomials on $[0, 1]$. They form a complete orthonormal basis for $\mathscr{L}^2(0, 1)$. Whereas $T_j(\theta; G) := \mathrm{Leg}_j[G(\theta)]$ is the modified shifted Legendre polynomials of rank-G transform $G(\theta)$, which are basis of the Hilbert space $\mathscr{L}^2(G)$. The composition of functions is denoted by the usual '$\circ$' sign.

2 The Model

Our model-building approach proceeds sequentially as follows: (i) it starts with a scientific (or empirical) parametric prior $g(\theta; \alpha, \beta)$, (ii) inspects the adequacy and the remaining uncertainty of the elicited prior using a graphical exploratory tool, (iii) estimates the necessary "correction" for assumed g by looking at the data, (iv) generates the final statistical estimate $\hat{\pi}(\theta)$, and (v) executes macro and micro-level inference. We seek a method that can yield answers to all five of the phases using only a *single* algorithm.

2.1 New Family of Prior Densities

This section serves two purposes: it provides a universal class of prior density models, followed by its Fourier non-parametric representation in a specialized orthonormal basis.

Definition 1. The Skew-G class of density models is given by

$$\pi(\theta) = g(\theta; \alpha, \beta)\, d[G(\theta); G, \Pi], \tag{2.1}$$

where $d(u; G, \Pi) = \pi(G^{-1}(u))/g(G^{-1}(u))$ for $0 < u < 1$ and consequently $\int_0^1 d(u; G, \Pi) = 1$.

A few notes on the model specification:

- It has a unique *two-component* structure that combines assumed parametric g with the d-function. The function d can be viewed as a "correction" density to counter the possible misspecification bias of g.
- The density function $d(u; G, \Pi)$ can also be viewed as describing the "excess" *uncertainty* of the assumed $g(\theta; \alpha, \beta)$. For that reason we call it the U-function.
- The motivation behind the representation (2.1) stems from the observation that $d[G(\theta); G, \Pi]$ is in fact the prior density-ratio $\pi(\theta)/g(\theta)$. Hence, it is straightforward to verify that the scheme (2.1) always yields a proper density, i.e., $\int_\theta g(\theta)\, d[G(\theta); G, \Pi] = 1$.

Since the square integrable $d[G(\theta); G, \Pi]$ lives in the Hilbert space $\mathscr{L}^2(G)$, we can approximate it by projecting into the orthonormal basis $\{T_j\}$ satisfying $\int T_i(\theta; G) T_j(\theta; G)\, \mathrm{d}G = \delta_{ij}$. We choose $T_j(\theta; G)$ to be $\mathrm{Leg}_j \circ G(\theta)$, a member of the LP-class of rank-polynomials [11]. The system $\{T_j\}$ possesses two attractive properties: they are polynomials of rank transform $G(\theta)$ thus constitutes a robust basis, and they are orthonormal with respect to $\mathscr{L}^2(G)$, for *any* arbitrary G (continuous). This is not to be confused with standard Legendre polynomials $\mathrm{Leg}_j(u), 0 < u < 1$, which are orthonormal with respect to Uniform$[0, 1]$ measure. For more details, see Supplementary Appendix B. The above discussion paves the way for the following definition.

Definition 2. $\Theta \sim \mathrm{DS}(G, m)$ distribution if it admits the following representation:

$$\pi(\theta) \;=\; g(\theta; \alpha, \beta)\, \Big[1 + \sum_{j=1}^{m} \mathrm{LP}[j; G, \Pi]\, T_j(\theta; G)\Big]. \tag{2.2}$$

The LP-Fourier coefficients $\mathrm{LP}[j; G, \Pi]$ are the key parameters that help us to express mathematically the "gap" between a priori anticipated G and the true prior Π. When all the expansion coefficients are zero, we automatically recover g.

We will now spend a few words on the LP-DS(G, m) class of prior models:

- When $\pi(\theta)$ is a member of DS(G, m) class of priors, the orthogonal LP-transform coefficients (2.2) satisfy

$$\mathrm{LP}[j; G, \Pi] \;=\; \langle d, T_j \circ G^{-1} \rangle_{\mathscr{L}^2(0,1)} \;=\; \mathbb{E}[T_j(\Theta; G); \Pi]. \tag{2.3}$$

  Thus, given a random sample $\theta_1, \ldots, \theta_k$ from $\pi(\theta)$, we could easily estimate the unknown LP-coefficients, and, thus, d and π, by computing the sample mean $k^{-1} \sum_{i=1}^{k} T_j(\theta_i; G)$.

But unfortunately, the θ_i's are unobserved. Section 3 describes an estimation strategy that can deal with the situation at hand. Before introducing this technique, however, we must acclimate the reader with the role played by the U-function $d(u; G, \Pi)$ for uncertainty quantification and characterization of the initial believable prior g. That's the objective of the next Section 2.2.

- Under definition 2, we have $\mathrm{DS}(G, m = 0) \equiv g(\theta; \alpha, \beta)$. The truncation point m in (2.2) reflects the *concentration* of permissible π around a known g. While this class of priors is rich enough to approximate any reasonable prior with the desired accuracy in the large-m limit, one can easily exclude absurdly rough densities and focus on a neighborhood around the domain-knowledge-based g by choosing m not "too big."

- The motivations behind the name 'DS-Prior' are twofold. First, our formulation operationalizes I. J. Good's '$\underline{\text{S}}$uccessive $\underline{\text{D}}$eepening' idea [12] 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 $\underline{\text{S}}$cientific g that encodes an expert's knowledge and a $\underline{\text{D}}$ata-driven d. That is to say that our framework embraces data and science, both, in a *testable* manner [13].

2.2 Exploratory Diagnostics and U-Function

Is your data compatible with the pre-selected $g(\theta)$? If yes, the job is done without getting into the arduous business of nonparametric estimation. If no, we can model the "gap" between the parametric g and the true unknown prior π, which is often *far easier* than modeling π from scratch (hence, one can learn from small number of cases)! If the observed $y_1, \ldots, y_k$ look very unexpected given $g(\theta; \alpha, \beta)$, it is completely reasonable to question the sanctity of such a self-selected prior. Here we provide a formal nonparametric exploratory procedure to describe comprehensively the uncertainty about the choice of g. Using the algorithm detailed in the next section, we estimate U-functions for four real data sets. Among them, the first three are binomial variate and the last one normal. The results are shown in Fig. 1.

- The rat tumor data [14] consists of observations of endometrial stromal polyp incidence in $k = 70$ groups of female rats. For each group, y_i is the number of rats with polyps and n_i is the total number of rats in the experiment.

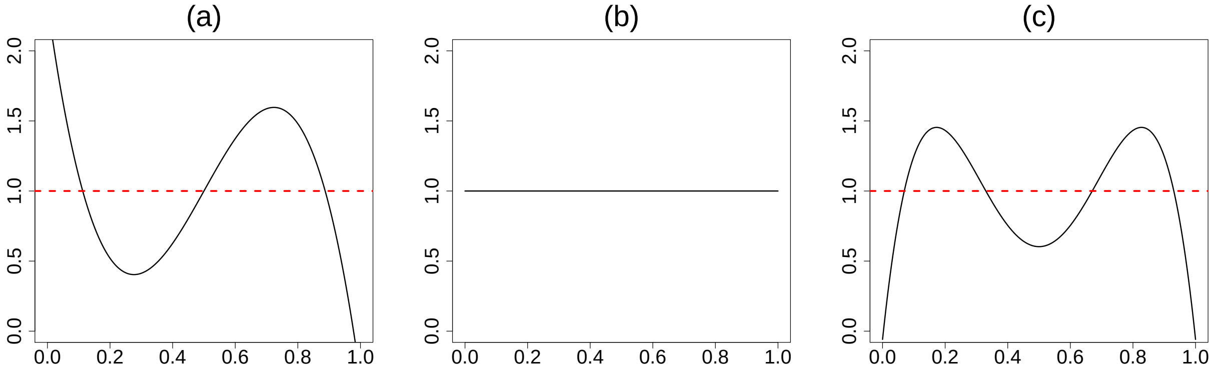

Figure 1: Graphical diagnostic tool: U-functions for (a) rat tumor data; (b) terbinafine and ulcer data; (c) rolling tacks data. The deviation from uniformity (red dotted line) indicates that the default prior contradicts the observed data. The flat shape of the U-function in panel (b) suggests Beta$(1.24, 34.7)$ and $\mathcal{N}(-1.17, 0.98)$ are consistent with the terbinafine and ulcer data, respectively.

- The terbinafine data [15] comprise $k = 41$ studies, which investigate the proportion of patients whose treatment terminated early due to some adverse effect of an oral anti-fungal agent: y_i is the number of terminated treatments and n_i is the total number of patients in the experiment.
- The rolling tacks [16] data involve flipping a common thumbtack 9 times. It consists of 320 pairs, $(9, y_i)$, where y_i represents the number of times the thumbtack landed point up.
- The ulcer data consist of forty randomized trials of a surgical treatment for stomach ulcers conducted between 1980 and 1989 [17, 18]. Each of the 40 trials has an estimated log-odds ratio $y_i|\theta_i \sim \mathcal{N}(\theta_i, s_i^2)$ that measures the rate of occurrence of recurrent bleeding given the surgical treatment.

Throughout, we have used the maximum likelihood estimates (MLE) for estimating the initial starting value of the hyperparameters. However, one can use any other reasonable choice, which may involve expert's judgment. What is important to note is the *shape* of the $\widehat{d}$; more specifically, its departure from uniformity, indicates the assumed conjugate prior $g(\theta; \alpha, \beta)$ needs a 'repair' to resolve the prior-data conflict. For example, the flat shape of the estimated $\widehat{d}$ in Fig. 1(b) indicates that our initial selection of $g(\theta; \alpha, \beta)$ is appropriate for the terbinafine and ulcer data. Therefore, one can proceed in turning the "Bayesian crank" with confidence using the parametric beta and normal prior respectively.

In contrast, Figs. 1(a,c) provide a strong warning in using $g = \text{Beta}(\alpha, \beta)$ for the rat tumor and the rolling tacks experiments. The smooth estimated U-functions expose the nature of the discrepancy that exists between g and the observed data by having an "extra" mode. Clearly, the answer does not lie in choosing a different (α, β) as this cannot rectify the missing bimodality. This brings us to an important point: the full Bayesian analysis, by assigning hyperprior distribution on α and β, is not always a fail-safe strategy and should be practiced with caution (not in a blind mechanical way). The bottom line is uncertainty in the prior probability model $\neq$ uncertainty in α, β. A foolproof prior uncertainty model, thus, has to allow ignorance in terms of the *functional shape* around g. The foregoing discussion motivates the following entropy-like measure of uncertainty.

Definition 3. The $q\,$LP statistic for uncertainty quantification is defined as follows:

$$\text{qLP}(G||\Pi) = \sum_j \big| \text{LP}[j; G, \Pi] \big|^2. \tag{2.4}$$

The motivation behind this definition comes from applying Parseval's identity in (2.2): $\int_0^1 d^2(u; G, \Pi) = 1 + \text{qLP}(G||\Pi)$. Thus, the proposed measure captures the departure of the U-function from uniformity. The following result connects our $q\,$LP statistic with relative entropy.

Theorem 1. *The $q\,$LP uncertainty quantification statistic satisfies the following relation:*

$$\text{qLP}(G||\Pi) \;\approx\; 2 \times \text{KL}(\Pi||G), \tag{2.5}$$

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

Proof. Express KL-information divergence using U-functions by substituting $G(\theta) = u$:

$$\text{KL}(\Pi||G) \;=\; \int \pi(\theta) \log \frac{\pi(\theta)}{g(\theta)} \, \mathrm{d}\theta \;=\; \int_0^1 d(u; G, \Pi) \log d(u; G, \Pi) \, \mathrm{d}u. \tag{2.6}$$

Complete the proof by approximating $d \log d$ in (2.6) via Taylor series $(d-1) + \frac{1}{2}(d-1)^2$. □

We conclude this section with a few additional remarks:

- Our exploratory uncertainty diagnostic tool encourages "interactive" data analysis that is similar in spirit to Gelman et al.[19]. Subject-matter experts can use this tool to "play" with different hyperparameter choices in order to filter out the reasonable ones. This functionality might be especially valuable when multiple expert opinions are available.

- When $\widehat{d}$ shows evidence of the prior-data conflict, the question remains: what to do next? It is not enough to check the adequacy without informing the user an explanation for the misfit or what is the "deeper" structure that is missing in the starting parametric prior. Fortunately, our DS(G, m) model suggests a simple, yet formal, guideline for upgrading: $\widehat{\pi}(\theta) = g(\theta; \hat{\alpha}, \hat{\beta}) \times \widehat{d}[G(\theta); G, \Pi]$, where the shape of $\widehat{d}(u; G, \Pi)$ captures the patterns which were not a priori anticipated. Hence our formalism *simultaneously* addresses the problem of uncertainty quantification and the subsequent model synthesis.

3 Estimation Method

3.1 Theory

In this Section, we lay out the key theoretical results that we use for designing our algorithm. Before deriving the general expressions under the LP-DS(G, m) model, it is helpful to start by recalling the results for the basic conjugate model, i.e., $\Theta \sim$ DS$(G, m = 0)$ and $y_i|\theta_i \overset{\text{ind}}{\sim} f(y_i|\theta_i)$ for $i = 1, \ldots, k$. Table 1 provides the marginal $f_G(y_i) = \int_{\theta_i} f(y_i|\theta_i) g(\theta_i) \, \mathrm{d}\theta_i$ and the posterior distribution $\pi_G(\theta_i|y_i) = \frac{f(y_i|\theta_i)g(\theta_i)}{f_G(y_i)}$ for four commonly encountered distributions, with the Bayes estimate of $h(\Theta_i)$ being denoted as $\mathbb{E}_G\big[h(\Theta_i)|y_i\big] = \int_{\theta_i} h(\theta_i)\pi_G(\theta_i|y_i) \, \mathrm{d}\theta_i$. The subscript '$G$' in these expressions underscores the fact that they are calculated for the conjugate g-model.

Table 1: Details on the distributions, their conjugate priors, and the resulting marginal and posterior distributions for four familiar distributions (two discrete and two continuous): Binomial, Poisson, Normal, and Exponential. For the normal-normal posterior $\lambda_i = \sigma_i^2/(\sigma_i^2 + \beta^2)$ and in the marginal of the Poisson-gamma $p = 1/(1+\beta)$. We use $\mathbf{B}(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ to denote the normalizing constant of beta distribution.

Family	Conjugate g-prior	Marginal $[f_G(y_i)]$	Posterior $[\pi_G(\theta_i \mid y_i)]$
Binomial(n_i, θ_i)	Beta(α, β)	$\binom{n_i}{y_i}\frac{\mathbf{B}(\alpha+y_i, \beta-y_i+n_i)}{\mathbf{B}(\alpha,\beta)}$	Beta$(\alpha + y_i, \beta - y_i + n_i)$
Poisson(θ_i)	Gamma(α, β)	$\binom{y_i+\alpha-1}{y_i} p^\alpha (1-p)^{y_i}$	Gamma$\big(\alpha + y_i, \frac{\beta}{1+\beta}\big)$
Normal(θ_i, σ_i^2)	Normal(α, β^2)	Normal$(\alpha, \sigma_i^2 + \beta^2)$	Normal$(\lambda_i \alpha + (1-\lambda_i) y_i, (1-\lambda_i)\sigma_i^2)$
Exp(λ)	Gamma(α, β)	$\frac{\alpha\beta}{(1+\beta y)^{\alpha+1}}$	Gamma$\big(\alpha + 1, \frac{\beta}{1+\beta y_i}\big)$

Next, we seek to extend these parametric results to LP-nonparametric setup in a systematic way. Especially, without deriving analytical expressions for each case separately, we want to establish a more general representation theory that is valid for all of the above and, in fact, extends to any conjugate pairs, explicating the underlying unity of our formulation.

Theorem 2. *Consider the following model:*

$$\begin{aligned} y_i|\theta_i &\overset{\text{ind}}{\sim} f(y_i|\theta_i), \qquad (i=1,\ldots,k)\\ \Theta_i &\overset{\text{ind}}{\sim} \pi(\theta), \end{aligned}$$

where $\pi(\theta)$ is a member of $\mathrm{DS}(G,m)$ *family (2.2), G being the associated conjugate prior. Under this framework, the following holds:*

(a) The marginal distribution of y_i is given by

$$f_{\mathrm{LP}}(y_i) \;=\; f_G(y_i)\left(1+\sum_j \mathrm{LP}[j;G,\Pi]\;\mathbb{E}_G[T_j(\Theta_i;G)|y_i]\right), \tag{3.1}$$

where $\mathbb{E}_G[T_j(\Theta_i;G)|y_i] = \int_{\theta_i} \mathrm{Leg}_j \circ G(\theta_i)\pi_G(\theta_i|y_i)\,\mathrm{d}\theta_i$.

(b) A closed-form expression for the posterior distribution of Θ_i given y_i is

$$\pi_{\mathrm{LP}}(\theta_i|y_i) = \frac{\pi_G(\theta_i|y_i)\big(1+\sum_j \mathrm{LP}[j;G,\Pi]\;T_j(\theta_i;G)\big)}{1+\sum_j \mathrm{LP}[j;G,\Pi]\;\mathbb{E}_G[T_j(\Theta_i;G)|y_i]} \tag{3.2}$$

(c) For any general random variable $h(\Theta_i)$, the Bayes conditional mean estimator can be expressed as follows:

$$\mathbb{E}_{\mathrm{LP}}[h(\Theta_i)|y_i] = \frac{\mathbb{E}_G[h(\Theta_i)|y_i] + \sum_j \mathrm{LP}[j;G,\Pi]\,\mathbb{E}_G[h(\Theta_i)T_j(\Theta_i;G)|y_i]}{1+\sum_j \mathrm{LP}[j;G,\Pi]\,\mathbb{E}_G[T_j(\Theta_i;G)|y_i]} \tag{3.3}$$

Proof. The marginal distribution for $\mathrm{DS}(G,m)$-nonparametric model can be represented as:

$$f_{\mathrm{LP}}(y_i) \;=\; \int f(y_i|\theta_i) \times \big\{g(\theta_i;\alpha,\beta)\,d[G(\theta_i);G,\Pi]\big\}\;\mathrm{d}\theta_i.$$

Expanding the U-function in the LP-bases (2.2) yields

$$f_{\mathrm{LP}}(y_i) \;=\; f_G(y_i) \;+\; \sum_j \mathrm{LP}[j;G,\Pi]\int T_j(\theta_i;G)f(y_i|\theta_i)g(\theta_i;\alpha,\beta)\;\mathrm{d}\theta_i. \tag{3.4}$$

The next step is to recognize that

$$f(y_i|\theta_i)\,g(\theta_i;\alpha,\beta) \;=\; f_G(y_i)\,\pi_G(\theta_i|y_i). \tag{3.5}$$

Substituting (3.5) in the second term of (3.4) leads to

$$\sum_j \mathrm{LP}[j;G,\Pi]\int T_j(\theta_i;G)f(y_i|\theta_i)g(\theta_i;\alpha,\beta)\;\mathrm{d}\theta_i \;=\; f_G(y_i)\sum_j \mathrm{LP}[j;G,\Pi]\,\mathbb{E}_G[T_j(\Theta_i;G)|y_i]. \tag{3.6}$$

Complete the proof of part (a) by replacing (3.6) into (3.4).

For part (b) of posterior distribution calculation we have

$$\pi_{\mathrm{LP}}(\theta_i|y_i) = \frac{f(y_i|\theta_i)\, g(\theta_i;\alpha,\beta)}{f_{\mathrm{LP}}(y_i)}\Big\{1+\sum_j \mathrm{LP}[j;G,\Pi]T_j(\theta_j;G)\Big\}. \tag{3.7}$$

Combine (3.1) and (3.5) to verify that

$$\frac{f(y_i|\theta_i)\, g(\theta_i;\alpha,\beta)}{f_{\mathrm{LP}}(y_i)} = \frac{\pi_G(\theta_i|y_i)}{1+\sum_j \mathrm{LP}[j;G,\Pi]\ \mathbb{E}_G[T_j(\Theta_i;G)|y_i]}. \tag{3.8}$$

Finish the proof of part (b) by replacing (3.8) into (3.7).

Part (c) is straightforward as

$$\mathbb{E}_{\mathrm{LP}}[h(\Theta_i)|y_i] = \int h(\theta_i)\pi_{\mathrm{LP}}(\theta_i|y_i)\ \mathrm{d}\theta_i,$$

which is same as

$$\frac{\int h(\theta_i)\pi_G(\theta_i|y_i)\{1+\sum_j \mathrm{LP}[j;G,\Pi]T_j(\theta_j;G)\}\ \mathrm{d}\theta_i}{1+\sum_j \mathrm{LP}[j;G,\Pi]\ \mathbb{E}_G[T_j(\Theta_i;G)|y_i]},$$

by (3.2). Hence, result (3.3) is immediate. □

Our LP-Bayes recipe (3.1)-(3.3), admits some interesting overall structure: The usual 'parametric' answer multiplied by a correction factor involving $\mathrm{LP}[j;G,\Pi]$'s. This decoupling pays dividends for theoretical interpretation as well as computation.

3.2 Algorithm

The critical parameters of our $\mathrm{DS}(G,m)$ model are the LP-Fourier coefficients, which, as is evident from (2.3), could be estimated simply by their empirical counterpart $\widehat{\mathrm{LP}}[j;G,\Pi] = k^{-1}\sum_{i=1}^k T_j(\theta_i;G)$. But as we pointed out earlier, $\theta_1,\ldots,\theta_k$ are unobservable. How can we then estimate those parameters? While the θ_i's are *unseen*, it is interesting to note that they have left their footprints in the observables $y_1,\ldots,y_k$ with distribution $f(y_i) = \int f(y_i|\theta_i)\pi(\theta_i)\ \mathrm{d}\theta_i$. Following the spirit of the EM-algorithm, an obvious proxy for $T_j(\theta_i;G)$ would be its posterior mean $\mathbb{E}_{\mathrm{LP}}[T_j(\Theta_i;G)|y_i]$, which also naturally arises in the expression (3.1). This leads to the following 'ghost' LP-estimates:

$$\widetilde{\mathrm{LP}}[j;G,\Pi]\ = k^{-1}\sum_{i=1}^k \mathbb{E}_{\mathrm{LP}}\big[T_j(\Theta_i;G)|y_i\big], \tag{3.9}$$

satisfying $\mathbb{E}\{\widetilde{\mathrm{LP}}[j;G,\Pi]\} = \widehat{\mathrm{LP}}[j;G,\Pi]$ $(j = 1\ldots,m)$, by virtue of the law of iterated expectations. These estimates can then be refined via iterations.

Type-II Method of Moments: Estimation of LP-Coefficients in DS(G, m)

`Step 0.` Input: Data $(y_1, \ldots, y_k)$ and m. Choice of α and β: based on expert's knowledge, otherwise, we use MLE empirical estimate as our default starting choice.

`Step 1.` Initialize: $\mathrm{LP}^{(0)}[j; G, \Pi] = 0$ for $j = 1, \ldots, m$. For iteration $\ell > 0$, perform steps (2-3) until convergence: $\sum_{j=1}^{m} \big|\widetilde{\mathrm{LP}}^{(\ell)}[j; G, \Pi] - \widetilde{\mathrm{LP}}^{(\ell-1)}[j; G, \Pi]\big|^2 \leqslant \epsilon$.

`Step 2.` Compute $\mathbb{E}_{\{\ell-1\}}[T_j(\Theta_i; G)|y_i]$ by plugging $\big\{\widetilde{\mathrm{LP}}^{(\ell-1)}[j; G, \Pi]\big\}_{j=1}^{m}$ into (3.3), where $h(\theta_i) = \mathrm{Leg}_j \circ G(\theta_i)$.

`Step 3.` Determine the 'ghost' LP-estimates:

$$\widetilde{\mathrm{LP}}^{(\ell)}[j; G, \Pi] \;=\; k^{-1} \textstyle\sum_{i=1}^{k} \mathbb{E}_{\{\ell-1\}}[T_j(\Theta_i; G)|y_i] \qquad (j = 1, \ldots, m).$$

`Step 4.` Return the final estimated LP-coefficients of $\mathrm{DS}(G, m)$ model together with $\widehat{d}(u; G, \Pi)$ and $\widehat{\pi}(\theta)$.

We conclude this section with a few remarks on the algorithm:

- Taking inspiration from I. J. Good's type II maximum likelihood nomenclature [20], we call our algorithm *Type-II* Method of Moments (MOM), whose computation is remarkably tractable and does not require *any* numerical optimization routine.
- To enhance the results, we smooth the output of MOM-II algorithm as follows: determine significantly non-zero LP-coefficients via Schwartz's BIC-based smoothing. Arrange $\widehat{\mathrm{LP}}[j; G, \Pi]$'s in a decreasing magnitude and choose m that maximizes

  $$\mathrm{BIC}(m) \;=\; \sum_{j=1}^{m} |\widehat{\mathrm{LP}}[j; G, \Pi]|^2 - \frac{m \log(k)}{k}.$$

  See Supplementary Appendix D for more details. Furthermore, Supplementary Appendix I discusses how MOM-II Bayes algorithm can be adapted to yield LP-maximum entropy prior density estimate [21].

3.3 Results

In addition to the rat tumor data (cf. Section 2.2), here we introduce and analyze three additional datasets: two binomial and one Poissonian example.

- The surgical node data [22] involves number of malignant lymph nodes removed during intestinal surgery. Each of the $k = 844$ patients underwent surgery for cancer, during which surgeons removed surrounding lymph nodes for testing. Each patient has a pair of data (n_i, y_i), where n_i represents the total nodes removed from patient i and $y_i \sim \text{Bin}(n_i, \theta_i)$ are the number of malignant nodes among them.
- The Navy shipyard data [23] consists of $k = 5$ samples of the number of defects y_i found in $n_i = 5$ lots of welding material.
- The insurance data [24], shown in Table 4, provides a single year of claims data for an automobile insurance company in Europe. The counts $y_i \sim \text{Poisson}(\theta_i)$ represent the total number of people who had i claims in a single year.

Figure 2 displays the estimated LP-DS(G, m) priors along with the default parametric (empirical Bayes) counterparts. The estimated LP-Fourier coefficients together with the choices of hyperparameters (α, β) are summarized below:

(a) Rat tumor data, g is the beta distribution with MLE $\alpha = 2.30$, $\beta = 14.08$:

$$\hat{\pi}(\theta) = g(\theta; \alpha, \beta)\big[1 - 0.50T_3(\theta; G)\big]. \tag{3.10}$$

(b) Surgical node data, g is the beta distribution with MLE $\alpha = 0.32$, $\beta = 1.00$:

$$\hat{\pi}(\theta) = g(\theta; \alpha, \beta)\big[1 - 0.07T_3(\theta; G) - 0.11T_4(\theta; G) + 0.09T_5(\theta; G) + 0.13T_7(\theta; G)\big]. \tag{3.11}$$

(c) Navy shipyard data, g is the Jeffreys prior with $\alpha = 0.5$, $\beta = 0.5$:

$$\hat{\pi}(\theta) = g(\theta; \alpha, \beta)\big[1 - 0.67T_1(\theta; G) + 0.90T_2(\theta; G)\big]. \tag{3.12}$$

(d) Insurance data, g is the gamma distribution with MLE $\alpha = 0.70$ and $\beta = 0.31$:

$$\hat{\pi}(\theta) = g(\theta; \alpha, \beta)\big[1 - 0.26T_2(\theta; G)\big]. \tag{3.13}$$

The rat tumor data shows a prominent bimodal shape, which should not come as a surprise in light of Fig. 1(a). For the surgical data, DS-prior puts excess mass around 0.4, which concurs with the findings of Efron [22, Sec 4.2]. In the case of the Navy shipyard data,

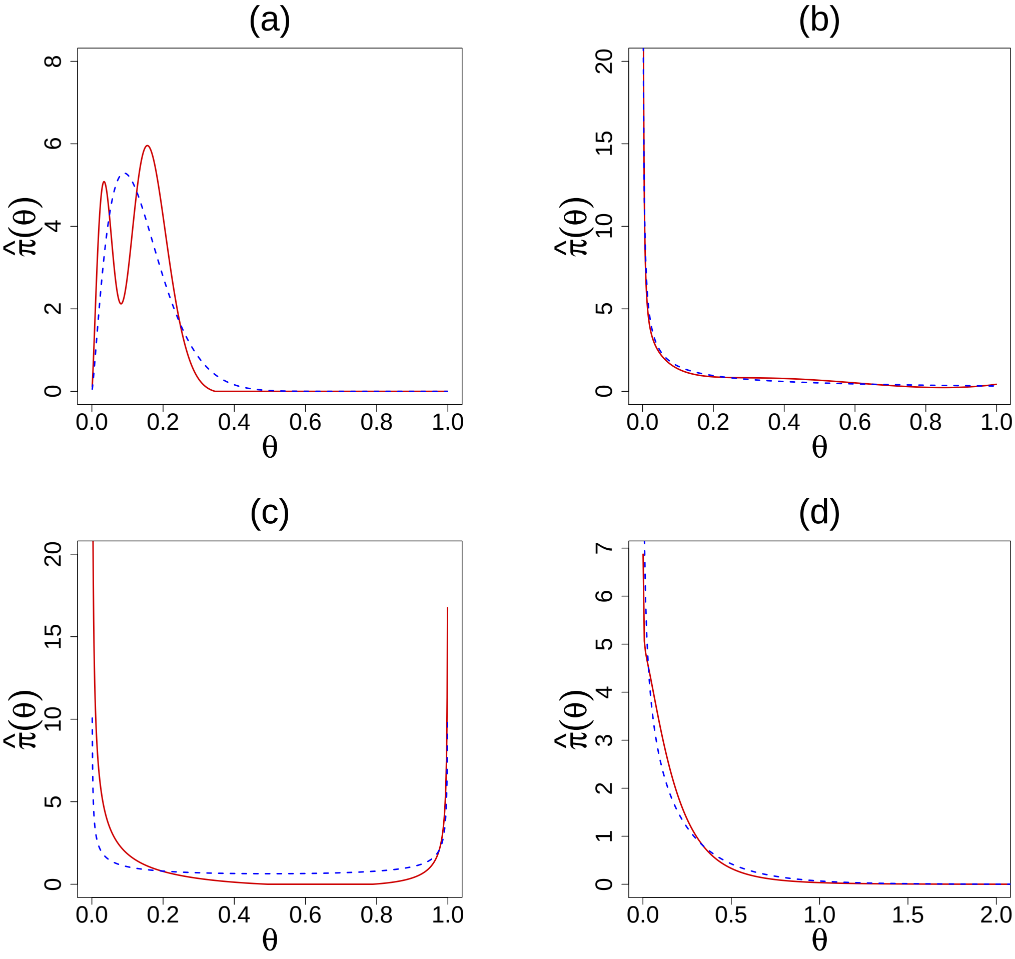

Figure 2: Comparisons of the DS(G, m) prior $\hat{\pi}(\theta)$ (solid red) with the respective parametric EB (PEB) priors $g(\theta; \alpha, \beta)$ (dashed blue) for the (a) rat tumor data, (b) surgical node data, (c) Navy shipyard data, and (d) insurance data.

our analysis corrects the starting "U" shaped Jeffreys prior to make it asymmetric with an extended peak at 0. This is quite justifiable looking at the proportions in the given data: $(0/5, 0/5, 0/5, 1/5, 5/5)$. Finally, for the insurance data, the starting gamma prior requires a second-order (dispersion parameter) correction to yield a bona-fide $\hat{\pi}$ (3.13), which makes it slightly wider in the middle with sharper peak and tail.

4 Inference

4.1 MacroInference

A single study hardly provides adequate evidence for a definitive conclusion due to the limited sample size. Thus, often the scientific interest lies in combining several *related but (possibly) heterogeneous* studies to come up with an overall macro-level inference that is more accurate and precise than the individual studies. This type of inference is a routine exercise in clinical trials and public policy research.

Terbinafine data analysis. For the terbinafine data, the aim is to combine $k = 41$ treatment arms with varying event rates and produce a pooled proportion of patients who withdrew from the study because of the adverse effects of oral anti-fungal agents. Recall that our U-function diagnostic in Fig. 1(b) indicated the parametric beta-binomial model with MLE estimates $\alpha = 1.24$ and $\beta = 34.7$ as a justifiable choice for this data. Thus the adverse event probabilities across $k = 41$ studies can be summarized by the prior mean $\frac{\alpha}{\alpha+\beta} = .034$. We apply parametric bootstrap using DS(G, m)-sampler (see Supplementary Appendix C) with $m = 0$ to compute the standard error (SE): 0.034 ± 0.006, highlighted in the Fig. 3(b). If one assumes a *single* binomial distribution for all the groups (i.e., under homogeneity), then the 'naive' average $\sum_{i=1}^{k} y_i / \sum_{i=1}^{k} n_i$ would lead to an overoptimistic biased estimate 0.037 ± 0.0034. In this example, heterogeneity arises due to overdispersion among the exchangeable studies. But there could be other ways too. An example is given in the following case study.

Rat tumor and rolling tacks data analysis. Can we always extract a "single" overall number to aptly describe k parallel studies? Not true, in general. In order to appreciate this, let us look at Figs. 3 (a,c), which depict the estimated DS-prior for the rat tumor and rolling tacks data. We highlight two key observations:

1. *Mixed population.* The bimodality indicates the existence of two distinct groups of θ_i's. We call this "*structured heterogeneity*," which is in between two extremes: homogeneity and complete heterogeneity (where there is no similarity between the θ_i's whatsoever). The presence of two clusters for the rolling tacks data was previously detected by Jun Liu [25]. The author further noted, "Clearly, this feature is unexpected and cannot be revealed by a regular parametric hierarchical analysis using the Beta-binomial priors." One plausible explanation for this two-group structure was attributed to the fact that the tack data were produced by two persons with some systematic difference in their flipping. On the other hand, the bimodal shape of the rat example was not previously anticipated [26, 27, 14]. The resulting two groups of rat tumor experiments are enumerated in the Table 2. Although we do not have the necessary biomedical background to scientifically justify this new discovery, we are aware that potentially numerous factors (e.g., experimental design, underlying conditions, selection of specific groups of female rats) may contribute to creating this systemic variation.

2. *From single mean to multiple modes.* An attempt to combine the two subpopulations using a single prior mean (as carried out for the terbinafine example) would result

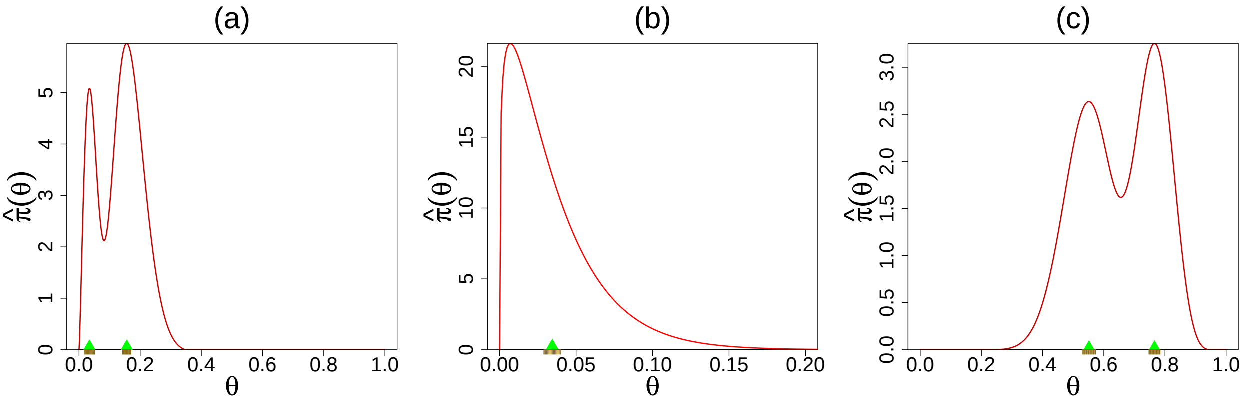

Figure 3: Estimated macro-inference summary along with standard errors (using smooth bootstrap) are shown. Panel (a) displays the rat tumor data modes located at 0.034 (±0.016) and 0.156 (±0.016). Panel (b) shows the estimated unimodal prior of the terbinafine data has a mean at 0.034 (±0.006). Panel (c) presents the modes of the rolling tacks data at 0.55 (±0.022) and 0.77 (±0.018).

Table 2: Two group partitions of the rat tumor studies based on K-means clustering on the posterior mode predictions (see Section 4.3 and Fig. 5(c)).

Group	Studies
1	(0,20), (0,20), (0,20), (0,20), (0,20), (0,20), (0,20), (0,19), (0,19), (0,19), (0,19) (0,18), (0,18), (0,17), (1,20), (1,20), (1,20), (1,20), (1,19), (1,19), (1,18), (1,18)
2	(3,27), (2,25), (2,24), (2,23), (2,20), (2,20), (2,20), (2,20), (2,20), (2,20), (1,10) (5,49), (2,19), (5,46), (2,17), (7,49), (7,47), (3,20), (3,20), (2,13), (9,48), (10,50) (4,20), (4,20), (4,20), (4,20), (4,20), (4,20), (4,20), (10,48), (4,19), (4,19), (4,19) (5,22), (11,46), (12,49), (5,20), (5,20), (6,23), (5,19), (6,22), (6,20), (6,20), (6,20) (16,52), (15,46), (15,47), (9,24)

in overestimating one group and underestimating another. We prefer *modes* of $\widehat{\pi}(\theta)$, along with their SEs, as a good representative summary, which can be easily computed by the nonparametric smooth bootstrap via DS(G, m) sampler.

Learning from big heterogeneous studies is one of the most important yet unsettled matters of modern macroinference [28, 18]. Our key insight is the realization that the 'science of combining' critically depends on the *shape* of the estimated prior. One interesting and commonly encountered case is multimodal structure of the learned prior. In such situations, instead of the prior-mean summary, we recommend group-specific modes. Our algorithm is also capable of finding data-driven clusters of the partially exchangeable studies in a fully automated manner.

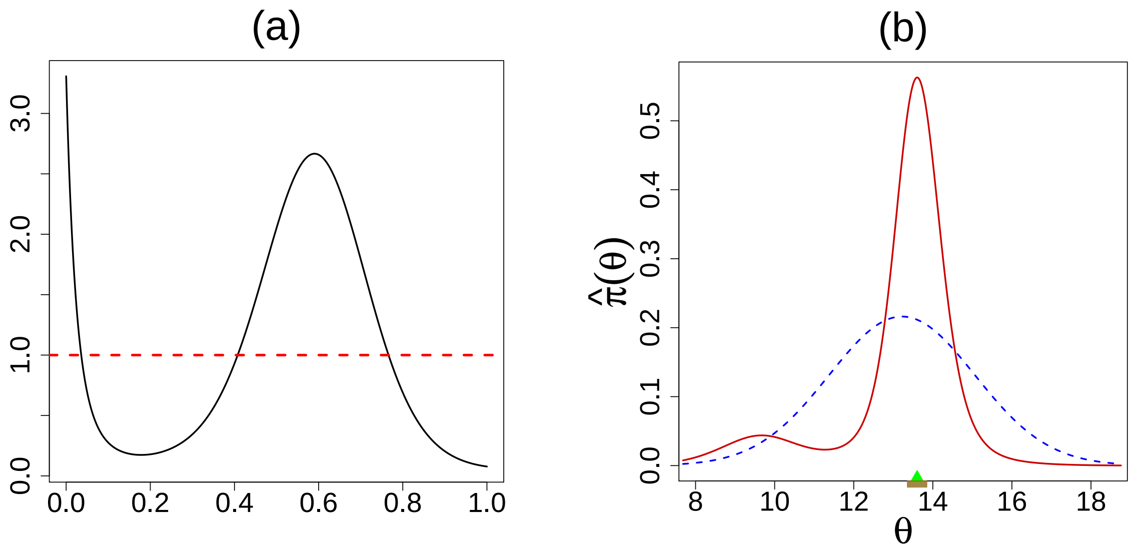

Figure 4: Panel (a) shows the U-function, while panel (b) compares the DS-prior $\hat{\pi}(\theta)$ (solid red) with the PEB prior $g(\theta; \alpha, \beta)$ (dashed blue) for the arsenic data. Based on the estimated macro-inference summary along with standard errors (using smooth bootstrap), the best consensus value is the mode 13.6 (± 0.242).

4.2 Learning From Uncertain Data

An important problem of measurement science that routinely appears in metrology, chemistry, physics, biology, and engineering can be stated as follows: measurements are made by k different laboratories in the form of $y_1, \ldots, y_k$ along with their estimated standard errors $s_1, \ldots, s_k$. Given this uncertain data, a fundamental problem of interest is inference concerning: (i) estimation of the consensus value of the measurand, and (ii) evaluation of the associated uncertainty. The data in Table 3 are an example of such an inter-laboratory study involving $k = 28$ measurements for the level of arsenic in oyster tissue. The study was part of the National Oceanic and Atmospheric Administration's National Status and Trends Program Ninth Round Intercomparison Exercise [29].

Table 3: Measurements (sorted) along with their uncertainty from different laboratories in arsenic data.

Laboratory	1	2	3	4	5	$\cdots$	25	26	27	28
Measurement (y_i)	9.78	10.18	10.35	11.60	12.01	$\cdots$	14.70	15.00	15.10	15.50
Uncertainty (s_i)	0.30	0.46	0.07	0.78	2.62	$\cdots$	0.30	1.00	0.20	1.60

Arsenic data analysis. We start with the DS-measurement model: $Y_i|\Theta_i = \theta_i \sim \mathcal{N}(\theta_i, s_i^2)$ and $\Theta_i \sim \mathrm{DS}(G, m)$ $(i = 1, \ldots, 28)$ with G being $\mathcal{N}(\mu, \tau^2)$. The shape of the estimated U-function in Fig. 4(a) indicates that the pre-selected prior $\mathcal{N}(\hat{\mu} = 13.22, \hat{\tau}^2 = 1.85^2)$ is clearly unacceptable for arsenic data, thereby disqualifying the classical Gaussian random effects model [30]. The DS-corrected $\widehat{\pi}$ shows some interesting asymmetric pattern with two-bumps. The left-mode represents measurements from three laboratories that are unlike the majority. The result of our macro-inference is shown in Fig. 4(b), which delivers the consensus value 13.6 ± 0.24. This is clearly far more resistant to fairly extreme low measurements and surprisingly, also more accurate when compared to the parametric EB estimate 13.22 ± 0.26. Most importantly, our scheme provides an automated solution to the fundamental problem of *which (as well as how)* measurements from the participating laboratories should be combined to form a best consensus value. Possolo [31] fits a Bayesian hierarchical model with prior as Student's t_ν, where the degrees of freedom was also treated as a random variable over some arbitrary range $\{3, \ldots, 118\}$. Although a heavy-tailed Student's t-distribution is a good choice to 'robustify' the analysis, it fails to capture the inherent asymmetry and the finer modal structure on the left. Distinguishing long-tail from bimodality is an important problem of applied statistics by itself.

To summarize, there are several attractive features of our general approach: (i) it adapts to the structure of the data, yet (ii) allows the use of expert opinion to go from knowledge-based prior to statistical prior; (iii) if multiple expert opinions are available, one can also use the U-diagnostic for reconciliation–exploratory uncertainty assessment; (iv) it avoids the questionable exercise of detecting and discarding apparently unusual measurements [32], and finally (v) our theory is still applicable for very small number of parallel cases (cf. Fig. 2(c)), a situation which is not uncommon in inter-laboratory studies.

4.3 MicroInference

The objective of microinference is to estimate a specific microlevel θ_i given y_i. Consider the rat tumor example where, along with earlier $k = 70$ studies, we have an additional current experimental data, that shows $y_{71} = 4$ out of $n_{71} = 14$ rats developed tumors. How can we estimate the probability of a tumor for this new clinical study? There could be at least three ways to answer this question:

- Frequentist MLE estimate: An obvious estimate would be the sample proportion $\widetilde{\theta}_i$: $y_{71}/n_{71} = 0.286$. This operates in an isolated manner, completely ignoring the additional historical information of $k = 70$ studies.
- Parametric empirical Bayes estimate: It is reasonable to expect that the historical data from earlier studies may be related to the current 71st study, thus borrowing information can result in improved estimator of θ_{71}. Bayes posterior mean estimate

$\breve{\theta}_i = \mathbb{E}_G[\Theta_i|y_i]$ operationalizes this heuristic, which in the Binomial case takes the following form:

$$\breve{\theta}_i = \frac{n_i}{\alpha+\beta+n_i}\,\widetilde{\theta}_i + \frac{\alpha+\beta}{\alpha+\beta+n_i}\mathbb{E}_G[\Theta]. \tag{4.1}$$

This is famously known as Stein's shrinkage formula [33, 34], as it pulls the sample proportions toward the *overall* mean of the prior $\frac{\alpha}{\alpha+\beta}$. For smaller ($n_i$) studies, shrinkage intensity is higher, which allows them to learn from other experiments.

- Nonparametric Elastic-Bayes estimate: Is it a wise strategy to shrink all $\widetilde{\theta}_i$'s toward the grand mean 0.14? Interestingly, this shrinking point is near the valley between the twin-peaks of the rat tumor prior density estimate (verify from Fig. 3(a)) and therefore may not represent a preferred location. Then, *where to shrink?* Ideally, we want to learn only from the *relevant* subset of the full dataset–*selective shrinkage*, e.g., for rat data, it would be the group 2 of Table 2. This brings us to the question: how to rectify the parametric empirical Bayes estimate $\breve{\theta}_i$? The formula (3.3) gives us the required (nonlinear) adjusting factor:

$$\widehat{\theta}_i = \frac{\breve{\theta}_i + \sum_j \widehat{\mathrm{LP}}[j;G,\Pi]\,\mathbb{E}_G[\Theta_i T_j(\Theta_i;G)|y_i]}{1 + \sum_j \widehat{\mathrm{LP}}[j;G,\Pi]\,\mathbb{E}_G[T_j(\Theta_i;G)|y_i]}, \tag{4.2}$$

  dictating the magnitude and direction of shrinkage in a completely data-driven manner via LP-Fourier coefficients. Note that when $d \equiv 1$, i.e., all the $\mathrm{LP}[j;G,\Pi]$ are zero, (4.2) reproduces the parametric $\breve{\theta}_i$. Due to its flexibility and adaptability, we call this the Elastic-Bayes estimate. This can be considered as a nonparametric class of shrinkage estimators that starts with the classical Stein's formula and rectifies it by looking at the data.

Rat tumor example. Figure 5 compares Stein's empirical Bayes estimate with our Elastic-Bayes estimate for the all $k = 70$ tumor rates. Posterior mean, median, and mode of θ_j's are shown side by side in three plots. The departure from the 45° reference line is a consequence of "adaptive shrinkage." Elastic-Bayes automatically shrinks the empirical $\widetilde{\theta}_i$ towards the representative modes (0.034 and 0.156), whereas the Stein's PEB estimate uses the grand mean (≈ 0.14) as the shrinking target for *all* the tumor rates. This is particularly prominent in Fig. 5 (c) for maximum a posteriori (MAP) estimates. As before, for heterogeneous population, we prescribe posterior mode as the final prediction.

The Pharma-example. Our DS Elastic-Bayes estimate is especially powerful in the presence of prior-data conflict. To illustrate this point, we report a small simulation study. The goal is to compare MSE for frequentist MLE, parametric empirical Bayes, and nonparamet-

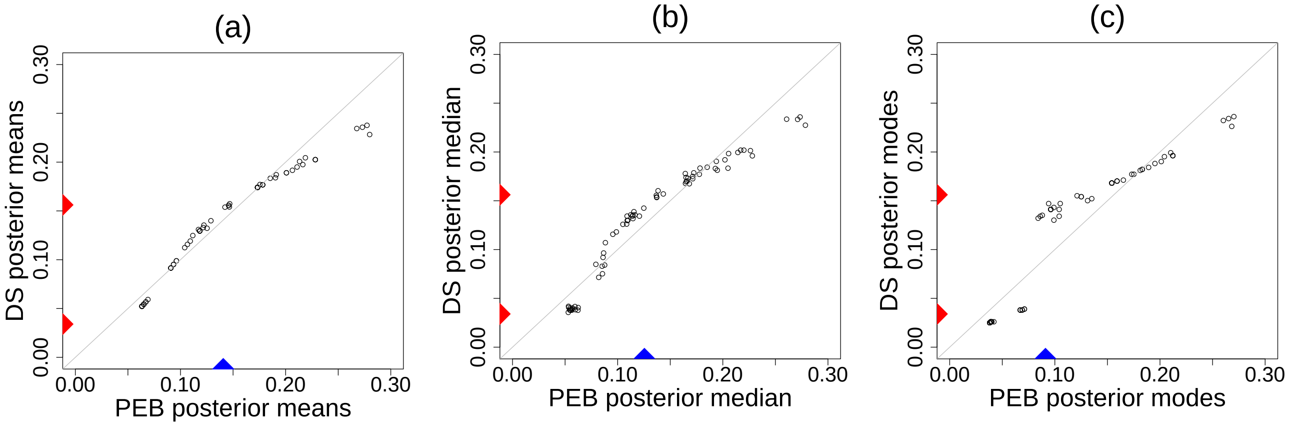

Figure 5: Comparisons of DS Elastic-Bayes and PEB posterior predictions of the rat tumor data: (a) posterior means, (b) posterior medians, and (c) posterior modes. The vertical red triangles indicate the location of the modes on the DS prior; the blue triangles respectively denote the mean, median, and mode of the parametric Beta($\hat{\alpha} = 2.3, \hat{\beta} = 14.08$).

ric Elastic-Bayes estimates for a new study y_{new} in various levels of prior-data conflict. To capture the prior-data conflict, we consider the following model for $\pi(\theta)$ and y_{new}:

$$\pi(\theta) = \eta \text{Beta}(5, 45) + (1 - \eta)\text{Beta}(30, 70)$$
$$y_{\text{new}} \sim \text{Bin}(50, 0.3).$$

The parameter η varies from 0 to 0.50 in increments of 0.05; as η increases we introduce more heterogeneity into the true prior distribution and exacerbate the prior-data conflict between $\pi(\theta)$ and y_{new}; see Fig. 6(a). We simulated $k = 100$ θ_i from $\pi(\theta)$, with which we generate $y_i|\theta_i \sim \text{Bin}(60, \theta_i)$. Using the Type-II MoM algorithm on the simulated data set, we found $\hat{\pi}$. After generating y_{new}, we then determined the frequentist MLE, parametric EB (PEB), and the nonparametric elastic Bayes estimates of the mode. For each value of η, we repeated this process 250 times and found the mean squared error (MSE) for each estimate. To better illustrate the impact of prior-data conflicts, we used ratio of PEB MSE to frequentist MSE and PEB MSE to DS MSE. The results are shown in Fig. 6 (b).

The Elastic-Bayes estimate outperforms the Stein's estimate for all η. More importantly the efficiency of our estimate continues to increase with the heterogeneity. This is happening because elastic Bayes performs *selective* shrinkage of sample proportion towards the appropriate mode (near 0.3) and thus gains "strength" by combining information from 'similar' studies even when the contamination in the study population increases. An interesting observation is the performance of the frequentist MLE estimate; as the data becomes more heterogeneous, the frequentist MLE shows improvement with respect to the Stein's PEB estimate. Our simulation depicts a scenario that is very common in historic-controlled clinical

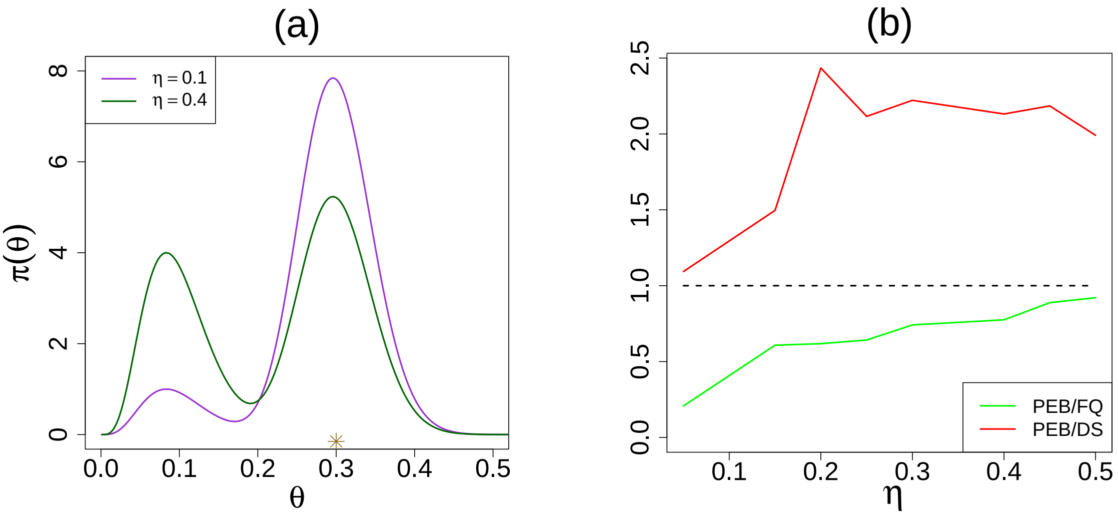

Figure 6: Panel (a) illustrates the prior-data conflict for $\eta = 0.1$ versus $\eta = 0.4$; '*' denotes 0.3, the true mean of y_{new}. Panel (b) shows the MSE ratios for PEB to Frequentist MLE (PEB/FQ; green) and PEB to DS (PEB/DS; red) with respect to η. Notice that as more prior-data conflict is introduced, DS outperforms PEB while frequentist MLE performance improves.

trials, where the heterogeneity arises due to changing conditions. Additional comparisons with other empirical Bayes procedures can be found in Supplementary Appendix G.

Three additional real examples. Figure 7 shows the posterior plots for specific studies in four of our data sets: surgical node, rat tumor, Navy shipyard, and rolling tacks. In studies like the surgical node data, personalized predictions are typically valuable. Figure 7(a) shows posterior distributions for three selected patients, which are indistinguishable from Efron's deconvolution answer [35, Fig. 4]; the patient with $n_i = 32$ and $y_i = 7$ shows almost certainly $\theta_i > 0.5$, i.e., he or she is highly prone to positive lymph nodes, and thus should be referred to follow-up therapy. With regard to the rat tumor data, Fig. 7(b) depicts the DS-posterior distribution of θ_{71} along with its parametric counterpart $\pi_G(\theta_{71}|y_{71}, n_{71})$. Interestingly, the DS nonparametric posterior shows less variability; this possibly has to do with the selective learning ability of our method, which learns from similar studies (e.g. group 2), rather than the whole heterogeneous mix of studies. We see similar phenomena in the rolling tacks data, where panel (d): $y_i = 3$, is more reflective of the first mode and panel (f): $y_i = 8$, of the second. Panel (e) shows the bimodal posterior for $y_i = 6$ case. Finally, the Navy shipyard data (Fig. 7 (c)) exhibits another advantage of DS priors: it works equally well for small k. The DS-posterior mean estimate for $y_6 = 0$ is 0.0471, which is consistent with the findings of Sivaganesan and Berger [36, p. 117].

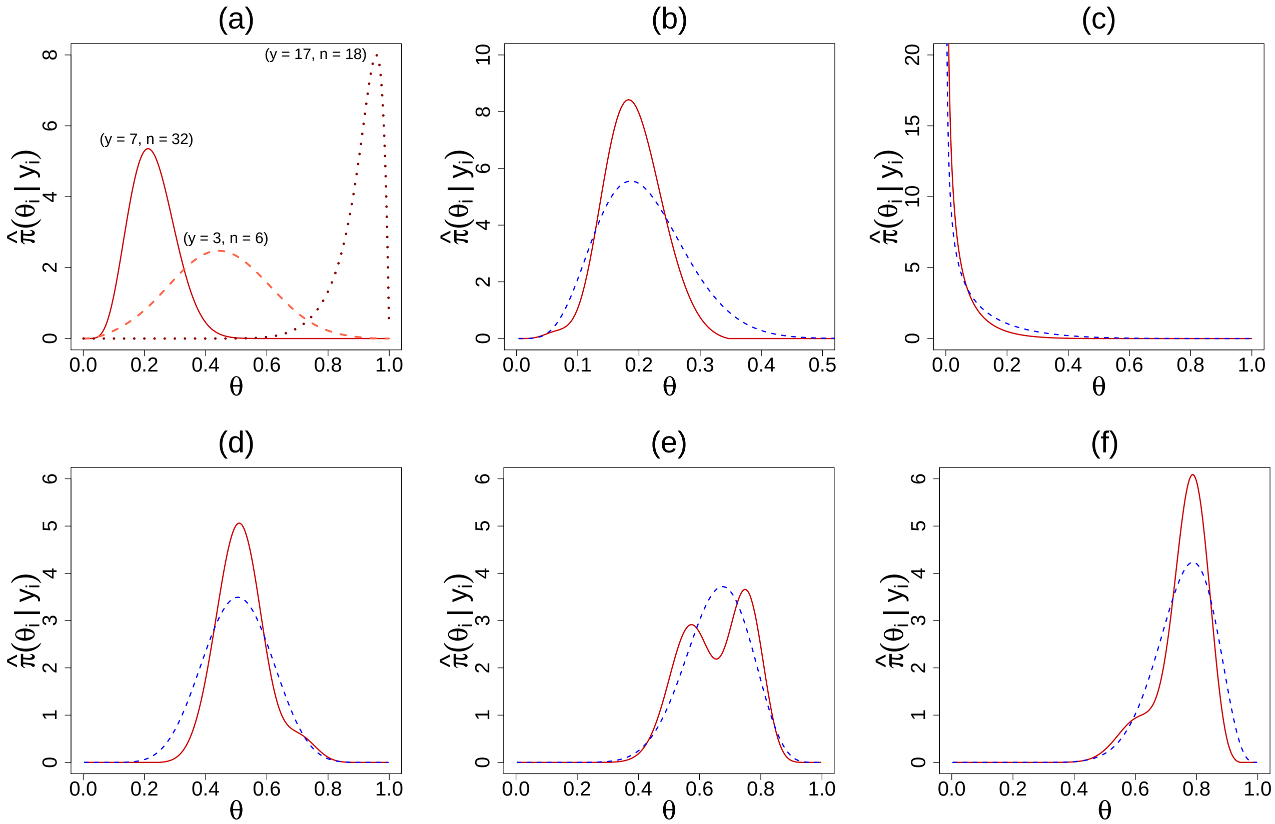

Figure 7: Panel (a) shows DS posterior plots of three observations from the surgical node data: $(y = 7, n = 32)$, $(y = 3, n = 6)$, and $(y = 17, n = 18)$. For panels (b) through (f), red denotes the DS posterior and blue dashed is the PEB posterior. Panel (b) is $\breve{\pi}(\theta_{71}|y_{71} = 4)$ for the rat tumor data. Panel (c) displays $\breve{\pi}(\theta_6|y_6 = 0)$ for the Navy shipyard data. The second row shows the posterior distributions of (d) $y_i = 3$, (e) $y_i = 6$, and (f) $y_i = 8$ from the rolling tacks data.

4.4 Poisson Smoothing: The Two Cultures

We consider the problem of estimating a vector of Poisson intensity parameters $\theta = (\theta_1, \ldots, \theta_k)$ from a sample of $Y_i|\theta_i \sim \text{Poisson}(\theta_i)$, where the Bayes estimate is given by:

$$\mathbb{E}[\Theta|Y = y] \;=\; \frac{\int_0^\infty \theta\big[e^{-\theta}\theta^y/y!\big]\pi(\theta)\; d\theta}{\int_0^\infty \big[e^{-\theta}\theta^y/y!\big]\pi(\theta)\; d\theta}; \quad y = 0, 1, 2, \ldots. \tag{4.3}$$

Two primary approaches for estimating (4.3):

- Parametric Culture [37, 38]: If one assumes $\pi(\theta)$ to be the parametric conjugate Gamma distribution $g(\theta; \alpha, \beta) = \frac{1}{\beta^\alpha \Gamma(\alpha)}\theta^{\alpha-1}\, e^{-\theta/\beta}$, then it is straightforward to show that Stein's estimate takes the following analytical form $\check{\theta}_i = \frac{y_i+\alpha}{\beta^{-1}+1}$, weighted average of the MLE y_i and the prior mean $\alpha\beta$.
- Nonparametric Culture [4, 7, 39]: This was born out of Herbert Robbins' ingenious observation that (4.3) can alternatively be written in terms of marginal distribution

$(y+1)\frac{f(y+1)}{f(y)}$, and thus can be estimated non-parametrically by substituting empirical frequencies. This remarkable "prior-free" representation, however, does not hold in general for other distributions. As a result, there is a need to develop methods that can bite the bullet and estimate the prior π from the data. Two such promising methods are Bayes deconvolution [7] and the Kiefer-Wolfowitz non-parametric MLE (NPMLE) [40, 39]. Efron's technique can be viewed as *smooth* nonparametric approach, whereas NPMLE generates a discrete (atomic) probability measure. For more discussion, see Supplementary Appendix A2.

The Third Culture. Each EB modeling culture has its own strengths and shortcomings. For example, PEB methods are extremely efficient when the true prior is Gamma. On the other hand, the NEB methods possess extraordinary robustness in the face of a misspecified prior yet they are inefficient when in fact $\pi \equiv \text{Gamma}(\alpha, \beta)$. Noticing this trade-off, Robbins raised the following intriguing question [10]: *how can this efficiency-robustness dilemma be resolved in a logical manner?* To address this issue, we must design a data analysis protocol that offers a mechanism to answer the following *intermediate* modeling questions (before jumping to estimate $\widehat{\pi}$): Can we assess whether or not a Gamma-prior is adequate in light of the sample-information? In the event of a prior-data conflict, how can we estimate the 'missing shape' in a completely data-driven manner? All of these questions are at the heart of our 'Bayes *via* goodness-of-fit' formulation, whose goal is to develop a third culture of generalized empirical Bayes (gEB) modeling by uniting the parametric and non-parametric philosophies. Compute the DS Elastic-Bayes estimate by substituting $\breve{\theta}_i = \frac{y_i+\alpha}{\beta^{-1}+1}$ in the Eq. (4.2), which reduces to the PEB answer when $d(u; G, \Pi) \equiv 1$ (i.e, the true prior is a Gamma) and modifies non-parametrically, only when needed; thereby turning Robbins' vision into action (see Supplementary Appendices A and G for more discussions on this point).

Table 4: For the insurance data set, estimates for the number of claims expected in the following year by an individual who made y claims during the present year, $\hat{\mathbb{E}}(\theta|Y=y)$, by five different methods.

Claims y	0	1	2	3	4	5	6	7
Counts	7840	1317	239	42	14	4	4	1
Gamma PEB	0.164	0.398	0.633	0.87	1.10	1.34	1.57	1.80
Robbins' EB	0.168	0.363	0.527	1.33	1.43	6.00	1.75	—
Deconvolve	0.164	0.377	0.642	1.14	2.13	3.45	4.47	5.08
NPMLE	0.168	0.362	0.534	1.24	2.21	2.53	2.58	2.58
DS Elastic-Bayes	0.156	0.322	0.517	0.744	1.02	1.56	3.01	5.24

The insurance data. Table 4 reports the Bayes estimates $\mathbb{E}[\theta|Y=y]$ for the insurance data. We compare five methods: parametric Gamma, classical Robbins' EB, Efron's Deconvolve, Koenker's NPMLE, and our procedure. The raw-nonparametric Robbins' estimator is clearly erratic at the tail due to data-sparsity. The PEB estimate overcomes this limitation and produces a stable estimate; but *is it dependable?* Should we stop here and report this as our final result? Our exploratory U-diagnostic tells that (consult Sec 3.3) the PEB estimate needs a second-order correction to resolve the discrepancy between the Gamma prior and data. The improved LP-Stein estimates are shown in the last row of Table 4.

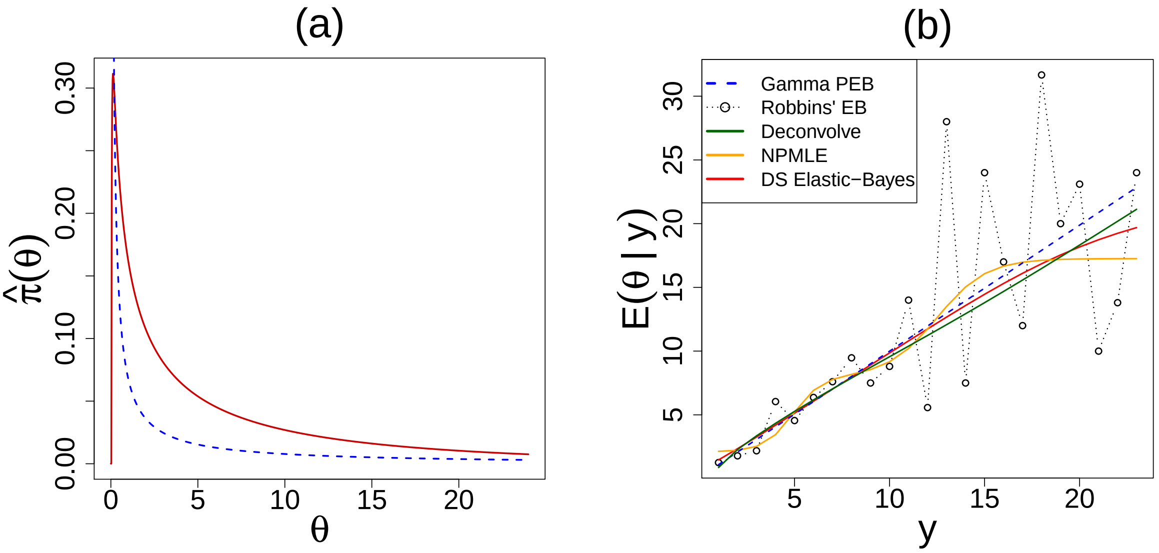

Figure 8: Panel (a) displays the estimated DS$(G, m = 4)$ prior (solid red) with the PEB Gamma prior $g(\theta; \alpha, \beta)$ (dashed blue) for the butterfly data; these results indicate that Fisher's Gamma-prior guess required some correction. Panel (b) shows estimates for the number of butterfly species caught in the following year $\hat{\mathbb{E}}(\theta \mid x)$ by the Gamma PEB, Robbins' formula, Bayesian deconvolution, NPMLE, and our Elastic-Bayes estimate.

The butterfly data. The next example is Corbet's Butterfly data [37]– one of the earliest examples of empirical Bayes. Alexander Corbet, a British naturalist, spent two years in Malaysia trapping butterflies in the 1940s. The data consist of the number of species trapped exactly y times in those two years for $y = 1, \ldots, 24$. Figure 8(b) plots different Bayes estimates. The Robbins' procedure suffers from similar 'jumpiness.' The blue dotted line represents the linear PEB estimate with $\alpha = 0.104$ and $\beta = 89.79$ (same as of Efron and Hastie [24, Eq. 6.24]) estimated from the zero-truncated negative binomial marginals. Our DS-estimate is almost sandwiched between the PEB and Deconvolve answer. The NPMLE method (the orange curve) yields some strange looking sinusoidal pattern, probably due to overfitting. In conclusion, we must say that the triumph of our procedure as compared to the other Bayes estimators lies in its automatic adaptability that Robbins alluded in his 1980 article [10].

5 Discussions

We laid out a new mechanics of data modeling that effectively consolidates Bayes and frequentist, parametric and nonparametric, subjective and objective, quantile and information-theoretic philosophies. However, at a practical level, the main attractions of our "Bayes *via* goodness-of-fit" framework lie in its (i) ability to quantify and protect against prior-data conflict using exploratory graphical diagnostics; (ii) theoretical simplicity that lends itself to analytic closed-form solutions, avoiding computationally intensive techniques such as MCMC or variational methods.

We have developed the concepts and principles progressively through a range of examples, spanning application areas such as clinical trials, metrology, insurance, medicine, and ecology, highlighting the core of our approach that gracefully combines Bayesian way of thinking (parameter probability where prior knowledge can be encoded) with a frequentist way of computing via goodness-of-fit (evaluation and synthesis of the prior distribution). If our efforts can help to make Bayesian modeling more attractive and transparent for practicing statisticians (especially non-Bayesians) by even a tiny fraction, we will consider it a success.

Data availability

All datasets and the computing codes are available via free and open source `R`-software package `BayesGOF`. The online link: https://CRAN.R-project.org/package=BayesGOF

References

[1] Efron, B. Why isn't everyone a Bayesian? *The Am. Stat.* **40**, 1–5 (1986).

[2] Sims, C. Understanding non-Bayesians. *Unpubl. chapter, Dep. Econ. Princet. Univ.* (2010).

[3] Stigler, S. M. Thomas Bayes's Bayesian inference. *J. Royal Stat. Soc. Ser. A (General)* **125**, 250–258 (1982).

[4] Robbins, H. An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, 157–164 (1956).

[5] Good, I. The Bayes/non-Bayes compromise: A brief review. *J. Am. Stat. Assoc.* **87**, 597–606 (1992).

[6] Rubin, D. B. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals Stat.* **12**, 1151–1172 (1984).

[7] Efron, B. Robbins, empirical Bayes and microarrays. *The Annals Stat.* **31**, 366–378 (2003).

[8] Dempster, A. P. A subjectivist look at robustness. *Bull. Intern. Stat. Inst* **46**, 349–374 (1975).

[9] Berger, J. O. An overview of robust Bayesian analysis (with discussion). *Test* **3**, 5–124 (1994). DOI 10.1007/BF02562676.

[10] Robbins, H. An empirical Bayes estimation problem. *Proc. Natl. Acad. Sci.* **77**, 6988–6989 (1980).

[11] Mukhopadhyay, S. & Parzen, E. LP approach to statistical modeling. *arXiv preprint arXiv:1405.2601* (2014).

[12] Good, I. J. The philosophy of exploratory data analysis. *Philos. science* **50**, 283–295 (1983).

[13] Gelman, A., Simpson, D. & Betancourt, M. The prior can often only be understood in the context of the likelihood. *Entropy* **19**, 555 (2017).

[14] Gelman, A. *et al. Bayesian Data Analysis, Third Edition*. Chapman & Hall/CRC Texts in Statistical Science (Taylor & Francis, 2013).

[15] Young-Xu, Y. & Chan, K. A. Pooling overdispersed binomial data to estimate event rate. *BMC Med. Res. Methodol.* **8**, 58 (2008).

[16] Beckett, L. & Diaconis, P. Spectral analysis for discrete longitudinal data. *Adv. Math.* **103**, 107–128 (1994).

[17] Sacks, H. S., Chalmers, T. C., Blum, A. L., Berrier, J. & Pagano, D. Endoscopic hemostasis: an effective therapy for bleeding peptic ulcers. *J. Am. Med. Assoc.* **264**, 494–499 (1990).

[18] Efron, B. Empirical Bayes methods for combining likelihoods. *J. Am. Stat. Assoc.* **91**, 538–550 (1996).

[19] Gelman, A., Meng, X.-L. & Stern, H. Posterior predictive assessment of model fitness via realized discrepancies. *Stat. Sinica* 733–760 (1996).

[20] Good, I. J. *Good thinking: The foundations of probability and its applications* (Univ. Minnesota Press, Minneapolis, 1983).

[21] Mukhopadhyay, S. Large-scale mode identification and data-driven sciences. *Electron. J. Stat.* **11**, 215–240 (2017).

[22] Efron, B. Empirical Bayes deconvolution estimates. *Biom.* **103**, 1–20 (2016).

[23] Martz, H. & Lian, M. Empirical bayes estimation of the binomial parameter. *Biom.* **61**, 517–523 (1974).

[24] Efron, B. & Hastie, T. *Computer Age Statistical Inference*, vol. 5 (Cambridge University Press, 2016).

[25] Liu, J. S. Nonparametric hierarchical Bayes via sequential imputations. *The Annals Stat.* 911–930 (1996).

[26] Tarone, R. E. The use of historical control information in testing for a trend in proportions. *Biom.* **38**, 215–220 (1982).

[27] Dempster, A. P., Selwyn, M. R. & Weeks, B. J. Combining historical and randomized controls for assessing trends in proportions. *J. Am. Stat. Assoc.* **78**, 221–227 (1983).

[28] Cox, D. R. Comment: The 1988 Wald Memorial Lectures: The present position in Bayesian statistics. *Stat. Sci.* **5**, 76–78 (1990).

[29] Willie, S. & Berman, S. Ninth round intercomparison for trace metals in marine sediments and biological tissues. *NRC/NOAA* (1995).

[30] Rukhin, A. L. & Vangel, M. G. Estimation of a common mean and weighted means statistics. *J. Am. Stat. Assoc.* **93**, 303–308 (1998).

[31] Possolo, A. Five examples of assessment and expression of measurement uncertainty. *Appl. Stoch. Model. Bus. Ind.* **29**, 1–18 (2013).

[32] Toman, B. & Possolo, A. Laboratory effects models for interlaboratory comparisons. *Accreditation Qual. Assur.* **14**, 553–563 (2009).

[33] Stein, C. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proc. Third Berkeley Symp. on Math. Stat. Probab.* **1**, 197–206 (1955).

[34] Efron, B. & Morris, C. Data analysis using Stein's estimator and its generalizations. *J. Am. Stat. Assoc.* **70**, 311–319 (1975).

[35] Cox, D. & Efron, B. Statistical thinking for 21st century scientists. *Sci. Adv.* **3**, e1700768 (2017).

[36] Sivaganesan, S. & Berger, J. Robust Bayesian analysis of the binomial empirical Bayes problem. *Can. J. Stat.* **21**, 107–119 (1993).

[37] Fisher, R. A., Corbet, A. S. & Williams, C. B. The relation between the number of species and the number of individuals in a random sample of an animal population. *The J. Animal Ecol.* 42–58 (1943).

[38] Maritz, J. Empirical Bayes estimation for the poisson distribution. *Biom.* **56**, 349–359 (1969).

[39] Gu, J. & Koenker, R. On a problem of Robbins. *Int. Stat. Rev.* **84**, 224–244 (2016).

[40] Kiefer, J. & Wolfowitz, J. Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals Math. Stat.* 887–906 (1956).

Additional information

Supplementary information: It includes (i) connection with other major Bayesian modeling cultures, (ii) Details of `BayesGOF` R-Software together with additional numerical illustrations, (iii) important extensions to examples with covariates and (iv) Maximum-entropy DS(G, m) modeling.

Competing Interests: The authors declare no competing interests.

Supplementary Material for “Bayesian Modeling via Goodness-of-fit”

Subhadeep Mukhopadhyay*, Douglas Fletcher
Temple University, Department of Statistical Science
Philadelphia, Pennsylvania, 19122, U.S.A.

* To whom correspondence should be addressed; E-mail: deep@temple.edu

This supplementary document contains nine Appendices, organized as follows:

- Appendix A: Connections with other Bayesian modeling cultures.
- Appendix B: More insights into the LP-basis functions.
- Appendix C: The DS(G, m) sampler.
- Appendix D: Other practical considerations.
- Appendix E: Software.
- Appendix F: Data Catalogue.
- Appendix G: The Robbins’ puzzle.
- Appendix H: Example with covariates.
- Appendix I: Maximum-Entropy enhancement.

A. CONNECTIONS WITH OTHER BAYESIAN MODELING CULTURES

In this section, we explore the relationship of our approach with other existing Bayesian data modeling cultures from philosophical and computational perspective. We will show that our formulation can be interpreted from surprisingly diverse perspectives.

A1. Robust Bayesian Methods

Our view of going from a unique prior assumption to a class of priors for robust Bayesian modeling was shaped by the Jim Berger’s outstanding article [1]. In the same spirit of the ϵ-contamination class [2], our U-function $d(u; G, \Pi)$ can be thought of as an automatic robustifier for standard (conjugate) priors. Thus, our approach may attain similar goals in a more computationally friendly way. Finally, we completely agree with Berger [1] that ‘The major objection of non-Bayesians to Bayesian analysis is uncertainty in the prior, so eliminating this concern can make Bayesian methods considerably more appealing.’

A2. Empirical Bayes Methods

Empirical Bayes approaches use data to determine the prior. While parametric empirical Bayes [PEB] [3] fixes the hyperparameters based on the data, nonparametric empirical Bayes [NEB] [4] makes no assumptions on the prior's form and develops it based solely on the data. In particular, Brad Efron [5, 6] advocate a *smooth* nonparametric exponential family model: $\log \pi(\theta) = \sum_{j=0}^{m} \beta_j \theta^j$ for the prior distribution where $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_m)$ is estimated by maximizing the marginal log-likelihood function.

Example 1. The dotted line in Figure 9(a) denotes the non-parametrically estimated Efron's $\widehat{\pi}$ based on two-dimensional sufficient vector $S = (\theta, \theta^2)$ for the ulcer data [5]. At a first glance, it appears strikingly close to the conjugate normal prior $\mathcal{N}(-1.17, 0.98)$, marked as the bold red line. Perhaps the reader may be curious to know whether '$\pi(\theta) \equiv$ PEB Normal' here? This is indeed the case, as already shown in Figure 1(b) of the main paper. Our generalized empirical Bayes (gEB) framework automatically reduces to PEB when the data is consistent with the assumed parametric prior and modifies it non-parametrically otherwise. The output of the combined inference from $k = 40$ clinical trials is shown as a green triangle -1.17 ± 0.197, which is quite close[†] to the Efron's nonparametric answer [5] -1.22 ± 0.26. The negative macro-estimate of the log-odds ratio parameters suggests that the new surgical treatment for stomach ulcers is overall more effective than the existing one.

Another attractive NEB technique is based on non-parametric maximum likelihood estimate (NPMLE): maximize the log-likelihood $\sum_{i=1}^{k} \log\big\{ \int f(y_i|\theta)\ \mathrm{d}\Pi(\theta) \big\}$ over the set of all $\pi(\theta)$ on $\mathbb{R}$, which is known to be a notoriously difficult problem. Thanks to Gu and Koenker [8], an approximate NPMLE can be estimated via convex optimization technique (interior point method) instead of classical EM (Expectation-Maximization) algorithm [9], thereby making it a computationally feasible alternative.

Example 2. NPMLE imposes no structural constraint and produces an estimated prior as discrete measure supported on at most k points within the data range. Figure 9(b) shows its application to the child illness data [10], which comes from a study that followed $k = 602$ pre-school children in north-east Thailand from June 1982 through September 1985. Researchers recorded the number of times (y) a child became sick during every 2-week period. Using the DS-Bayes method, we have $\widehat{\pi}(\theta)$ where $g(\theta)$ is a gamma distribution

[†] The slight gain in accuracy for our method lies in the style of estimation that proceeds *via* goodness-of-fit. Constructing prior by validating its credibility (using frequentist criterion) may also strengthen the Bayesian objectivity that Brad Efron [7] alluded to his article "Why isn't everyone a Bayesian?"

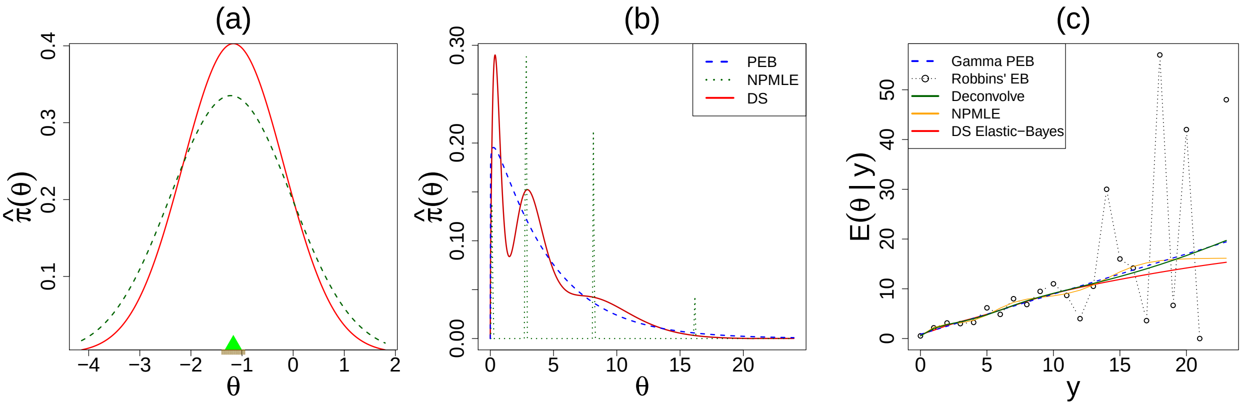

Figure 9: Comparisons of DS(G, m) (red) with other empirical Bayes modeling cultures (green): (a) The DS-estimated prior is compared with Efron's exponential prior model [5]; (b) The DS distribution for the child illness data compared to NPMLE (the dotted line); (c) Estimates for the number of illnesses in the following year $\hat{\mathbb{E}}(\theta \mid x)$ by Gamma PEB, Robbins' formula, Bayesian deconvolution, NPMLE, and our elastic-Bayes estimate.

Table 5: Run-time comparisons between DS-Bayes and two other BNP methods: Dirichlet prior (DP), and Bernstein-Dirichlet (BDP) model. All methods were run using an Intel®Core$^{\text{TM}}$ i5-7200 CPU @ 2.50GHz. `DPpackage` uses `C++` complier to speed-up, while ours is a prototype version implemented in `R`.

Data Set	# Studies (k)	DS Time	DP Time	Ratio DP to DS	BDP Time	Ratio BDP to DS
Rat Tumor	70	1.83	10.42	5.69	3457.75	1889.5
Surgical Node	844	30.95	189.3	6.12	45292.15	1463.4
Terbinafine	41	1.7	5.46	3.2	1883.18	1107.8
Rolling Tacks	320	8.27	59.16	7.15	16569.78	2003.6
Arsenic	28	0.47	13.09	27.8	433.29	254.9

with $\hat{\alpha} = 1.06$ and $\hat{\beta} = 4.19$ as

$$\hat{\pi}(\theta) = \text{Gamma}(\theta; \alpha, \beta)\big[1 - 0.13T_3(\theta; G) - 0.28T_6(\theta; G)\big]. \tag{5.1}$$

Our method produces a smooth, grid-free $\widehat{\pi}$ that accurately captures the overall shape. Figure 9(c) plots the Bayes estimates $\mathbb{E}[\Theta_i|Y = y]$ for all competing methods. For Efron's `Deconvolve` we have used c0 $= 2$ and pDegree $= 25$, which seems to produce a reasonable prior density estimate for this example. A careful look at the plot reveals an 'oscillating' NPMLE Bayes estimates (orange curve), which many not be particularly desirable.

A3. Dirichlet-Process-based Approaches

Bayesian nonparametric [BNP] technique assigns prior distribution on infinite-dimensional spaces of probability models. The majority of work on Bayesian nonparametrics utilizes a Dirichlet process prior [11]. The computational cost of BNP is severe and produces prior on a set of discrete probability measures that demands an additional layer of smoothing. Figure 10 contrasts Dirichlet-process based Beta-Binomial models [12] with our DS-Bayes model. There are few remarks warranted here:

- BNP method requires careful tuning of several hyper-priors values, which from our experience can be quite sensitive (see Figure 10). Without practical guidance, this "fishing expedition" can potentially overwhelm one who seeks to confidently use it in practice. On the contrary, our method finds practically the same answer without adjusting multiple hyper-prior values.
- The posterior inferences of BNP are highly complex and require computationally expensive MCMC. In contrast, the beauty of our approach is that it provides compact analytical expressions that make the computation much more amicable.
- The flexibility of BNP comes with the heavy task of estimating a massive number of parameters–"massively parametric Bayes." Contrast this with DS(G, m) model, which provides a reduced-dimensional characterization of the prior distribution with a closed form solution that is computationally efficient (see Table 5) and produces smooth estimates in one-shot. For additional comments see the 'Critical Appraisal' section.

A4. Weakly Informative Priors

A weakly informative prior [WIP] is a proper prior that intentionally provides less information than available prior knowledge. This lies somewhere between a fully subjective and a fully objective prior [13, 14].

One can also view our approach from a WIP-angle where $d(u; G, \Pi)$ acts as a "spreading/weakening function" of the subjective prior $g(\theta)$, which we *learn from the data.* In the DS(G, m) language: m is the radius of spread; the larger the m, the greater possibility you allow for changing the shape (the process of weakening) of the presumed scientific prior distribution $g(\theta)$. These analogies suggest that our concepts and notations might provide a systematic way to formulate the WIP philosophy by addressing the debates around "WIP is a subjective prior with ad hoc large but bounded support." This reformulation can also bring some tangible computational gain.

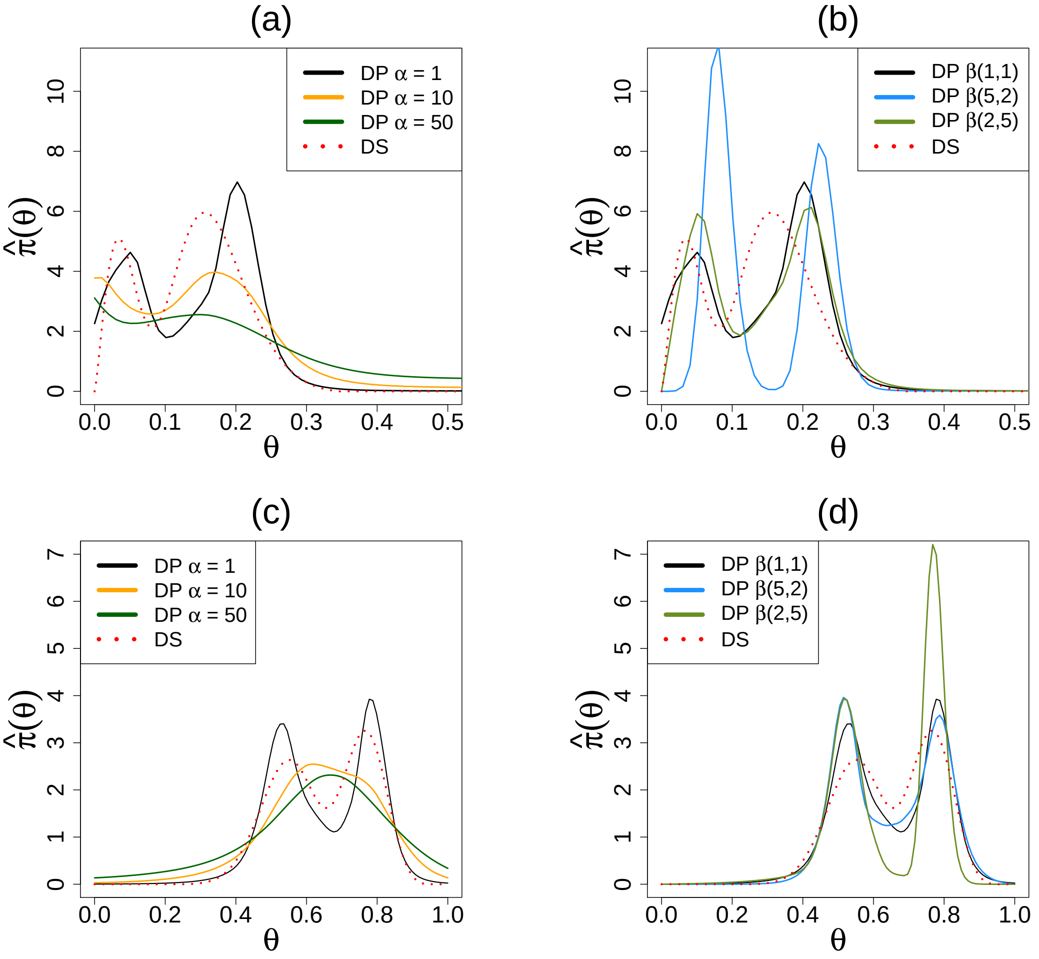

Figure 10: Illustrations of the different settings for BNP modeling with a Dirichlet process prior. Panel (a) displays results for the rat tumor data using uniform base prior while varying α. Panel (b), also for the rat tumor data, fixes $\alpha = 1$ and varies the base prior between uniform, Beta$(5, 2)$ and Beta$(2, 5)$. Panels (c) and (d) use the same settings as (a) and (b), but applied to the rolling tacks data.

A Critical Appraisal

We close this section by highlighting some of the unique aspects and practical advantages of our technique:

- *Clarifying the Motivation*: Let's start by reminding ourselves that the core motivation behind the 'Bayes *via* goodness-of-fit' is more than just another recipe for estimating the prior from data. To understand the mysterious prior in a transparent and definitive way, it is critical to ask: How can we provide automatic protection from unqualified specifications of prior distribution? How do we assess the prior-uncertainty using exploratory graphical tools? How can we prescribe a revised statistical-prior starting from the user-specified scientific-prior? As it stands, these fundamental questions are usually left unanswered in traditional Bayes framework and create a major obstacle for non-Bayesian practitioners to confidently use Bayesian tools. Consequently, there is a need to address these issues in a formal manner to bring much-needed transparency. This paper has taken some solid steps toward this goal with a methodology that is readily usable for wide-range of applied problems. We believe that our technology can become an integral part of applied Bayesian modeling.

- *Theoretical Novelty*: Our proposed theory, which is general enough to include almost all commonly-used models, yields analytic closed-form solutions for posterior modeling. This is noteworthy for the simple reason that none of the nonparametric methods mentioned above can stand by this claim.
- *Theoretical Simplicity*: The whole 'Bayes *via* Goodness-of-fit' framework can be developed starting from a few basic principles, without requiring any exotic theoretical treatment. This could add invaluable transparency to the theory and practice of (empirical) Bayesian statistics.
- *Exploratory Side*: Our approach brings a distinct exploratory flavor into the empirical-Bayes modeling. It encourages interactive data analysis rather than blindly 'turning the crank.' Through numerous examples, we demonstrated how this mode of operation often leads to more insights into the data that are typically infeasible under a business-as-usual Bayesian modus operandi.
- *Computational Side*: Simplicity of implementation and computational ease are the two hallmarks of our method. No expensive MCMC or even sophisticated optimization routines are required! We made a sincere effort to design a practical Bayesian data analysis tool that is both simpler to comprehend and easy to implement.
- *A Third Empirical Bayes Culture.* Our empirical Bayes approach is neither parametric nor nonparametric. As argued in Section 4.4 (of the main paper), our algorithmic approach blends conventional PEB and Robbins-style full-fledged NEB. Our goal is to combine the best of both worlds, in the sense that it reduces to PEB (ulcer data example) when in fact the default parametric g is appropriate, while in the event of prior-data conflict (rat tumor or child illness data), it automatically produces reliable nonparametric procedures. And in this whole story, the U-function $d(u; G, \Pi)$ acts as the "connector" between these two extreme philosophies. Overall, we are hopeful that our Generalized EB (gEB) modeling framework might expedites the development of a new *genre* of 'unified' Bayesian algorithms [15] by leveraging the rich interplay between two extreme EB philosophies.

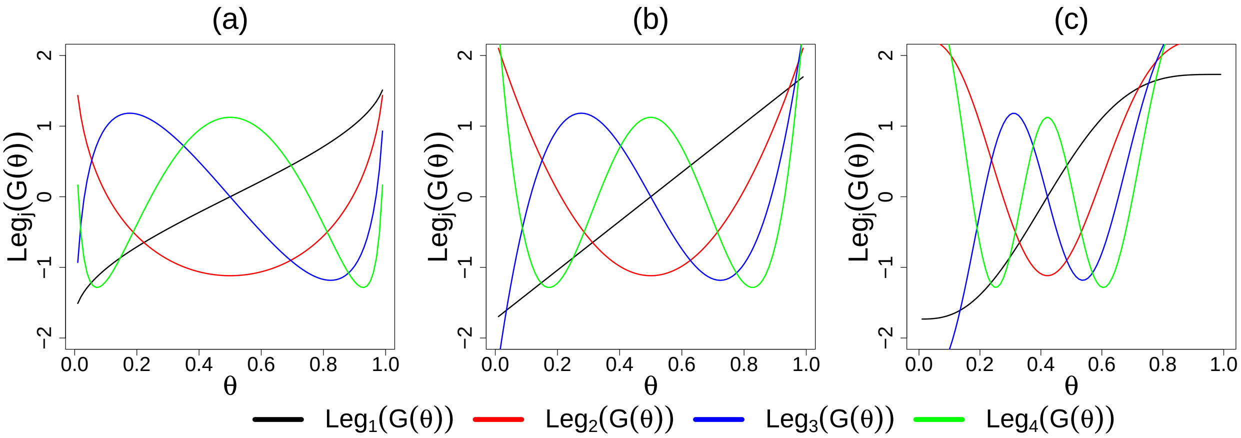

Figure 11: LP-polynomials $T_j(\theta; G_{\alpha,\beta})$ for `family= "beta"` with the following (α, β) choices: (a) Jeffrey's prior ($\alpha = \beta = 0.5$), (b) uniform prior ($\alpha = \beta = 1$), and for (c) Beta($\alpha = 3, \beta = 4$). Note that for $U[0, 1]$ (the middle panel): $T_j \equiv \text{Leg}_j$, as $G(\theta)$ is simply θ in this case.

B. MORE INSIGHTS INTO THE LP-BASIS FUNCTIONS

Here we will show the shapes of the LP-polynomials, focusing only the Binomial case. It works similarly for other families.

The $\{T_j(\theta; G_{\alpha,\beta})\}_{j\geqslant 1}$ denotes the class of orthonormal polynomials of the beta distribution with parameters α and β. Let $T_j(\theta; G_{\alpha,\beta}) = \text{Leg}_j\{G_{\alpha,\beta}(\theta)\}$ and $G_{\alpha,\beta}(\theta) = \frac{1}{\mathbf{B}(\alpha,\beta)} \int_0^\theta \phi^{\alpha-1}(1-\phi)^{\beta-1} d\phi$. Figure 11 displays the shapes of top four LP polynomials for three different sets of parameters. We generate these polynomials with the following R code:

```
LP.basis.beta <- function(y, g.par, m){
#######################################
##  g.par: parameters for the beta distribution
#######################################
require(orthopolynom)
u <- pbeta(y, g.par[1], g.par[2]) # computes G(y)
poly <-  slegendre.polynomials(m,normalized=TRUE)
TY <- matrix(NA,length(u),m)
for(j in 1:m) TY[,j] <- predict(poly[[j+1]],u)
return(TY)}
```

C. 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

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

otherwise, discard $\Theta$ and return to Step 1.

Step 3. Repeat until simulated sample of size $k$, $\{\theta_1^*, \theta_2^*, \cdots, \theta_k^*\}$.
```

Note that when $\widehat{d} \equiv 1$ then the DS(G,m) automatically samples from parametric G.

D. OTHER PRACTICAL CONSIDERATIONS

In the event that *no* prior knowledge is available, selecting the parametric conjugate prior G with empirically estimated α, β in conjunction with our Type-II Method of Moments algorithm (sec. 3.2) will provide a quick estimate of the oracle π. The algorithm finds the 'best' approximating prior model given `m.max`: the maximum complexity that the subject-matter experts want to entertain. From our experience with DS(G,m) model, we found `m.max` $= 8$ works satisfactorily well in practice (in fact in all our examples 8 was our default choice), which encompasses the space of reasonable priors around G. Given this maximum radius, our method generates a deviance plot, where the "elbow" shape (see Figure 12 (a)) denotes the most likely model dimension. This procedure is fully incorporated into our algorithm so that practitioners can use it in a completely automatic manner.

Illustration. Consider the model: $y_i|\theta_i \sim$ Binomial$(50, \theta_i)$ with $i = 1, \ldots, k = 90$ and the true prior distribution $\pi(\theta) = .3\text{Beta}(4,6) + .7\text{Beta}(20,10)$. Our goal is to see how well we can approximate the unknown π without any prior knowledge of its shape. The following R code can be used to reproduce our findings reported in Figure 12.

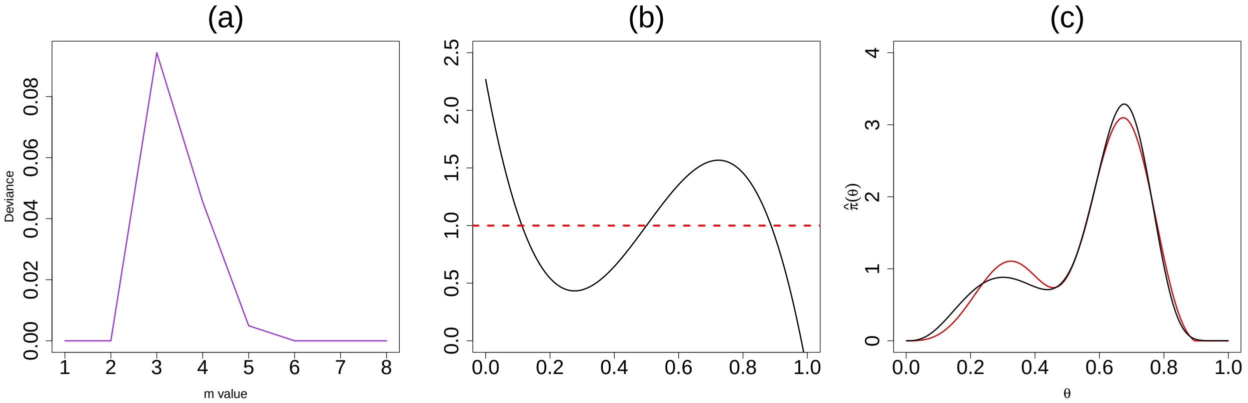

Figure 12: Analysis for simulated data based on Type-II Method of Moments algorithm. The first panel (a) finds the "elbow" in the BIC(m) deviance plot at $m = 3$; (b) shows the U-function, while (c) plots the true $\pi(\theta)$ (black) along with the estimated DS prior (red) $\hat{\pi}(\theta) = g(\theta; \hat{\alpha}, \hat{\beta})\big[1 - 0.48T_3(\theta; G)\big]$ with MLE $\hat{\alpha} = 4.16$ and $\hat{\beta} = 3.04$.

```
set.seed(8701)
k <- 90
n.i <- 50
n.vec <- rep(n.i,k)
k1 <- ceiling(.7*k)
#Test Simulation: Mixed beta Distribution
theta.sim <- c(rbeta(k1,20,10), rbeta( (k-k1),4,8))
y.sim <- sapply(theta.sim, rbinom, size = n.i, n = 1)
sim.df <- data.frame(y = y.sim, N = n.vec)
##Run Type II MoM Algorithm
sim.start <- gMLE.bb(sim.df$y,sim.df$N)$estimate
sim.LP.par <- DS.prior(sim.df, g.par = sim.start, family = "Binomial")
```

The `sim.start` object holds the MLE estimate for the initial parameters for G. From the `sim.LP.par` object, we generate diagnostic and analysis plots for appropriate m, U-function, and the DS(G, m) estimate, as shown in Figure 12.

E. SOFTWARE

We provide an `R` package, `BayesGOF` [16] to perform all the tasks outlined in the paper. We now summarize the main functions and their usage for the Rat binomial data example:

```
 #Phase I: Modeling
library("BayesGOF")
data(rat)
rat.start <- gMLE.bb(rat$y, rat$n)$estimate
rat.ds <- DS.prior(rat, g.par = rat.start, family = "Binomial")
plot(rat.ds, plot.type = "Ufunc") # Figure 1(a)
plot(rat.ds, plot.type = "DSg") # Figure 2(a)
```

The package also provide functionalities for Macro and MicroInference:

```
 #Phase II: Inference
rat.ds.macro <- DS.macro.inf(rat.ds, num.modes = 2, method = "mode")
plot(rat.ds.macro) # Figure 3(a)
rat.ds.pos <- DS.micro.inf(rat.ds, y.0 = 4, n.0 = 14)
plot(rat.ds.pos)  # Figure 5(b)
```

We hope this software will encourage applied data scientists to apply our method for their real problems.

F. DATA CATALOGUE

Table 6: List of datasets by distribution family and sources. They are sorted first by family, then according to k: from large to small-scale studies.

Dataset	# Studies (k)	Family	Sources
Surgical Node	844	Binomial	Efron (2016) [6]
Rolling Tacks	320	Binomial	Beckett and Diaconis (1994) [17]
Rat Tumor	70	Binomial	Gelman et al. (2013, Ch. 5) [14]
Terbinafine	41	Binomial	Young-Xu and Chan (2008) [18]
Naval Shipyard	5	Binomial	Martz et al. (1974) [19]
Galaxy	324	Gaussian	De Blok et al.(2001) [20]
Ulcer	40	Gaussian	Sacks et al.(1990) [21]
Arsenic	28	Gaussian	Willie and Berman (1995) [22]
Insurance	9461	Poisson	Efron and Hastie (2016) [23]
Child Illness	602	Poisson	Wang (2007) [10]
Butterfly	501	Poisson	Fisher et al. (1943) [24]
Norberg	72	Poisson	Norberg(1989) [25]

G. THE ROBBINS' PUZZLE

In Section 4.3 of the main paper, we presented a simulated scenario (Pharma-example) that demonstrated the power of the DS Elastic-Bayes estimate when there is significant prior-data conflict. Here we include further comparisons with two recent methods: Efron's Bayesian deconvolution (implemented in the `deconvolveR` package), and Koenker's NPMLE (implemented in the `REBayes` package).

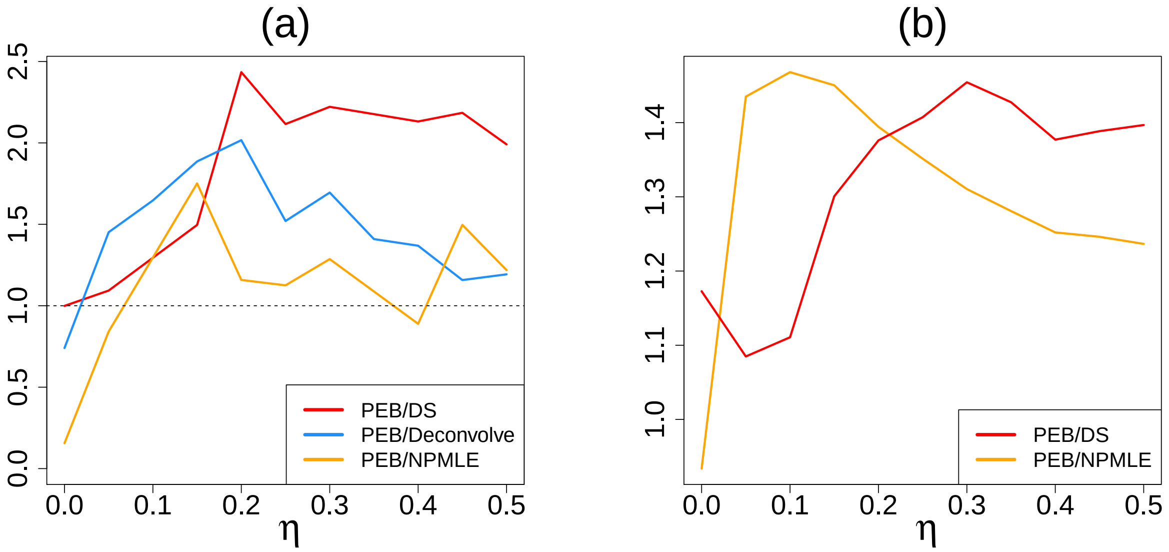

Figure 13: Results of two separate simulations comparing DS with other methods. In (a), the MSE ratios for PEB to empirical Bayes deconvolution (PEB/Dec; blue), PEB to Kiefer-Wolfowitz NPMLE using REBayes `Bmix` (PEB/NPMLE; orange) and PEB to DS (PEB/DS; red) with respect to η. Panel (b) shows the ratio of empirical risks after applying both DS and NPMLE methods to Robbins' 'compound decision' problem.

Example 1. Here we will operate under the exact settings presented in Section 4.3. Figure 13(a) shows that as η increases, DS tends to outperform the other two methods, although Deconvolve performs superbly for η smaller than 0.15. Two specially interesting extreme cases are $\eta = 0$ and $\eta = 0.5$. The first scenario describes the situation when the underlying parametric *beta* distribution is the right choice for the prior where, as expected, the Stein's parametric shrinkage estimator dominates other nonparametric approaches. On the other hand, the $\eta = 0.5$ is a complicated situation where $\pi(\theta) = \frac{1}{2}\text{Beta}(5, 45) + \frac{1}{2}\text{Beta}(30, 70)$, and consequently, the parametric EB [PEB] is less efficient compared to the nonparametric ones. The most interesting and surprising result, however, comes from DS Elastic-Bayes, which acts like the Stein prediction formula when underlying parametric assumption is correct (the null $\eta = 0$ case) but adapts itself non-parametrically in a completely automated manner when the true $\pi(\theta)$ deviates from the assumed g, thereby elegantly addressing the robustness-efficiency puzzle of Robbins [26].

Example 2. Next, we investigate the prediction problem originally introduced by Robbins [27] and discussed in Gu and Koenker [8]. We observe $Y_i = \theta_i + \epsilon_i$, $i = 1 \cdots k$, where $\epsilon_i \overset{\text{ind}}{\sim}$ Normal$(0, 1)$, and $\theta_i = \pm 1$ with probability η and $1 - \eta$ respectively. Our goal is to estimate the k-vector $\theta \in \{-1, 1\}^k$ under the loss $\mathrm{L}(\hat{\theta}, \theta) = k^{-1} \sum_{i=1}^{k} |\hat{\theta}_i - \theta_i|$. For comparison purpose, we computed the ratio of PEB empirical risk† to the the DS method (EB/DS) and to the NPMLE estimator (EB/KW) for $k = 1000$. Figure 13(b) shows a very interesting result: Kiefer-Wolfowitz NPMLE method performs significantly better than the DS-elastic Bayes when $0 < \eta < 0.2$. While for other values of η, including η equals to zero point, our micro-estimation procedure demonstrates tremendous promise. This further validates the flexibility and adaptability of our technique even in the discrete settings.

H. EXAMPLE WITH COVARIATES

The 'Bayes via goodness-of-fit' methodology can easily accommodate additional covariates. We demonstrate this capability using the following example.

The Norberg Example. The Norberg insurance dataset [25] consists of $k = 72$ Norwegian occupational categories, where y_i denotes the number of claims made against a policy. Additionally, we have the total number of years each group was exposed to risk E_i; when normalized by a factor of 344, E_i gives the expected number of claims during a contract period. Similar to Norberg [25], we assume $Y_i \sim$ Poisson$(\theta_i E_i)$. Given the normalized E_i, we interpret θ_i as the occupational-specific rate of risk.

DS-Bayes analysis yields the following estimated prior, where g is the conjugate gamma prior with MLE $\alpha = 6.02$ and $\beta = 0.20$:

$$\hat{\pi}(\theta) \;=\; g(\theta; \alpha, \beta)\big[1 - 0.70T_1(\theta; G) + 0.83T_2(\theta; G) - 0.53T_3(\theta; G)\big]. \tag{5.2}$$

In Figure 14(a), the U-function clearly indicates potential prior-data conflict when using $\pi(\theta) = $ Gamma$(6.02, 0.20)$. Figure 14(b) displays the DS prior (red) along with the parametric EB (blue) and the Kiefer-Wolfowitz NPMLE estimate (green). We see a definite bimodality for $\hat{\pi}(\theta)$, indicating that there are two distinct groups of risk profiles. The macroinference plot in Figure 14(c) reinforces the structured heterogeneity of the data. In terms of risk-profile, we consider the mode at 0.59 as occupational categories with comparatively lower risk; these are occupations less likely to make a claim based on their risk exposure. The mode at 1.46 represents those occupations at a higher risk, thus more likely

† Mean loss is computed over 500 replications.

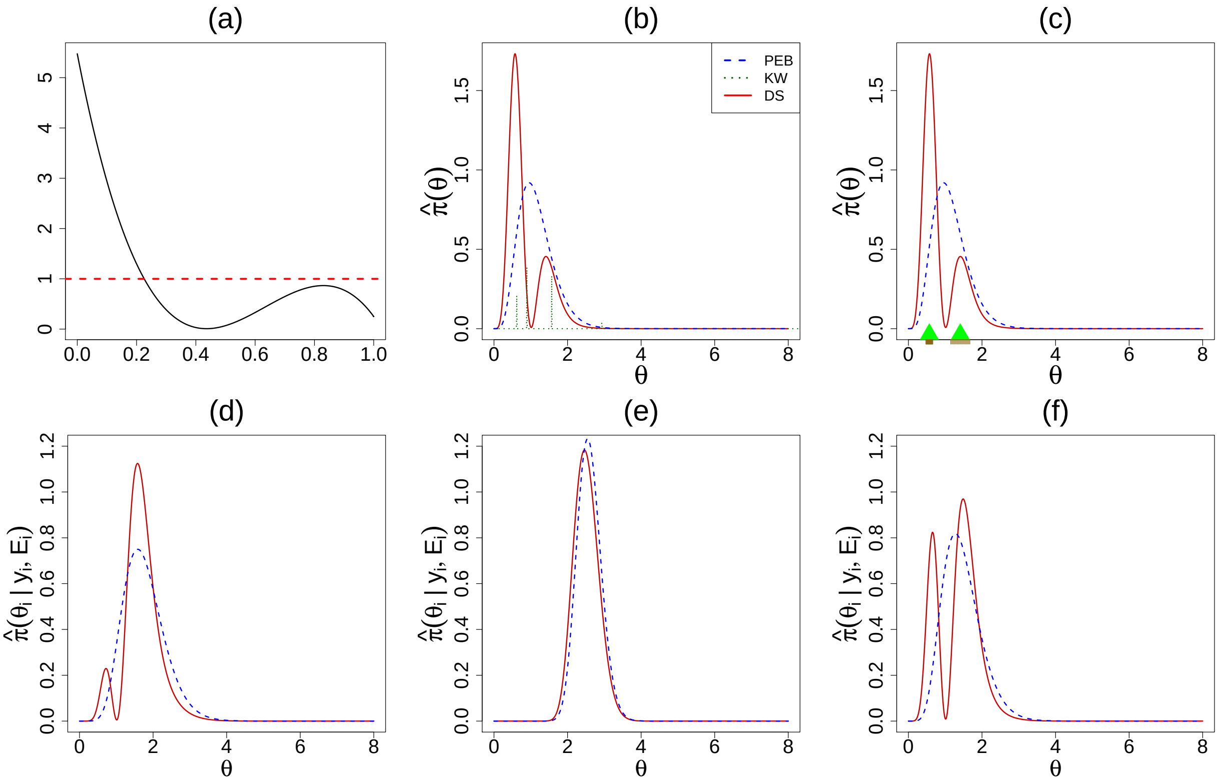

Figure 14: Demonstration of DS-Bayes with covariates on the Norberg insurance dataset. In (a), we display the U-function. Panel (b) shows the DS-prior (red), the PEB prior (blue) and the Kiefer-Wolfowitz NPMLE prior (green). Panel (c) shows the macroinference with standard errors (using smooth bootstrap): two modes located at 0.57(±0.094) and 1.41(±0.261). Panels (d) through (f) show microinference for occupational groups 13, 22, and 53 (respectively).

to make a claim based on their exposure. Of particular interest are panels (d), (e), and (f). These panels show the microinference for three specific occupational groups: group 13 ($Y_{13} = 4$, $E_{13} = 0.45$), group 22 ($Y_{22} = 57$, $E_{22} = 19.1$), and group 53 ($Y_{53} = 2$, $E_{53} = 0.25$). In Figure 14(d), we have an occupational category that identifies as higher risk with a small lower risk component. The unimodality in Figure 14(e) clearly indicates that category is a higher risk of claim based on exposure. Finally, the occupational category in Figure 14(f) is tricky. Here, we have bimodality with an almost equal probability of being a high or low-risk occupation. While the other two groups provide clear alternatives for an insurance company, the occupational group 53 needs the company's judgment in assigning the policy.

I. MAXIMUM-ENTROPY ENHANCEMENT

For more enhanced result, we offer an extension to maximum entropy DS(G, m) model, which assumes the following representation of the prior distribution:

$$\breve{\pi}(\theta) \;=\; g(\theta;\alpha,\beta)\,\exp\Big[c_0 + \sum_j c_j\,T_j(\theta;G)\Big], \tag{5.3}$$

where c_0 is some normalizing constant and the c_j's are the LP-maximum entropy coefficients. The following algorithm outlines the process to solve for the unknown c_j's starting from the $\mathcal{L}^2$ estimate.

Orthogonal Series to Maximum Entropy Estimator

`Step 0.` Input: BIC-smoothed LP-Fourier ($\mathcal{L}^2$) coefficients $\widehat{\mathrm{LP}}[j;G,\Pi]$, $j = 1,\ldots,m$.

`Step 1.` Define the set $\mathcal{J} = \Big\{j : |\widehat{\mathrm{LP}}[j;G,\Pi]| > 0\Big\}$, collection of j's for which we have significant non-zero $\mathcal{L}^2$ orthogonal coefficients.

`Step 2.` To estimate the maximum entropy coefficients c_j in $\breve{\pi}(\theta)$ of (5.3), solve the following sets of moment equality constraints:

$$\widehat{\mathrm{LP}}[j;G,\Pi] = \int T_j(\theta;G)\breve{\pi}(\theta)d\theta, \quad \text{for } j \in \mathcal{J}. \tag{5.4}$$

`Step 3.` Output: $\big(\hat{c}_0, \{\hat{c}_j\}_{j\in\mathcal{J}}\big)$; accordingly the estimated maximum entropy $\breve{d}$ and $\breve{\pi}$.

Two Data Examples. Here we carry out the maximum entropy analysis for `rat` (binomial variate) and `galaxy` data (normal variate). The `galaxy` data consists of $k = 324$ observed rotation velocities y_i and their uncertainties of Low Surface Brightness (LSB) galaxies [20].

(a) Rat Tumor data, g is beta distribution with MLE $\alpha = 2.30$, $\beta = 14.08$:

$$\breve{\pi}(\theta) = g(\theta;\alpha,\beta)\exp\big[-0.13 - 0.52T_3(\theta;G)\big]. \tag{5.5}$$

(b) Galaxy data, g is normal distribution with MLE $\mu = 85.5$, $\tau^2 = 3304$:

$$\breve{\pi}(\theta) = g(\theta;\mu,\tau^2)\exp\big[-0.15 + 0.26T_3(\theta;G) - 0.28T_4(\theta;G) + 0.46T_5(\theta;G)\big]. \tag{5.6}$$

The resulting LP-maximum-entropy DS(G,m) priors are shown in Figure 15. In both examples, we see the maximum entropy estimates (green dashed lines) are very similar to the $\mathcal{L}^2$ with some adjustments to the modal shapes.

The `BayesGOF` package in R implements this algorithm as an option for the `DS.prior` function. The following code demonstrates how to generate both the $\mathcal{L}^2$ and maximum entropy representations of the LP coefficients for both the rat tumor and galaxy data sets.

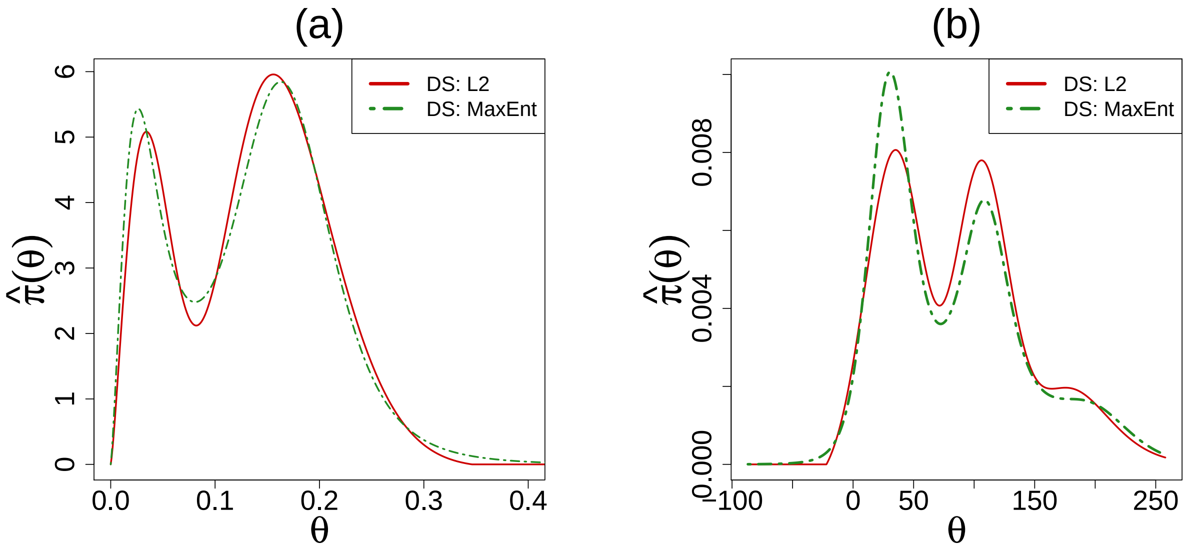

Figure 15: Comparison of $\mathcal{L}^2$ (solid red line) and maximum entropy (two-dash green line) estimates of DS prior. Panel (a) shows the comparison for the `rat` tumor data, while panel (b) illustrates the difference (in modal shapes) for the `galaxy` data.

```
library(BayesGOF)
#---Rat Tumor Data
data(rat)
rat.start <- gMLE.bb(rat$y, rat$n)$estimate
rat.ds.L2 <- DS.prior(rat, max.m = 4, g.par = rat.start,
                      family = "Binomial", LP.type = "L2")
## Shown in Figure 15(a) as solid red line
rat.ds.ME <- DS.prior(rat, max.m = 4, g.par = rat.start,
                      family = "Binomial", LP.type = "MaxEnt")
## Shown in Figure 15(a) as two-dashed green line
#---Galaxy Data
data(galaxy)
gal.start <- gMLE.nn(galaxy$y, galaxy$se)$estimate
gal.ds.L2 <- DS.prior(galaxy, max.m = 5, g.par = gal.start,
                      family = "Normal", LP.type = "L2")
## Shown in Figure 15(b) as solid red line
gal.ds.ME <- DS.prior(galaxy, max.m = 5, g.par = gal.start,
                      family = "Normal", LP.type = "MaxEnt")
## Shown in Figure 15(b) as two-dashed green line
```

References

[1] Berger, J. O. An overview of robust Bayesian analysis (with discussion). *Test* **3**, 5–124 (1994). DOI 10.1007/BF02562676.

[2] Berger, J. & Berliner, L. M. Robust Bayes and empirical Bayes analysis with ε-contaminated priors. *The Annals Stat.* 461–486 (1986).

[3] Morris, C. N. Parametric empirical Bayes inference: theory and applications. *J. Am. Stat. Assoc.* **78**, 47–55 (1983).

[4] Efron, B. Robbins, empirical Bayes and microarrays. *The Annals Stat.* **31**, 366–378 (2003).

[5] Efron, B. Empirical Bayes methods for combining likelihoods. *J. Am. Stat. Assoc.* **91**, 538–550 (1996).

[6] Efron, B. Empirical Bayes deconvolution estimates. *Biom.* **103**, 1–20 (2016).

[7] Efron, B. Why isn't everyone a Bayesian? *The Am. Stat.* **40**, 1–5 (1986).

[8] Gu, J. & Koenker, R. On a problem of Robbins. *Int. Stat. Rev.* **84**, 224–244 (2016).

[9] Laird, N. Nonparametric maximum likelihood estimation of a mixing distribution. *J. Am. Stat. Assoc.* **73**, 805–811 (1978).

[10] Wang, Y. On fast computation of the non-parametric maximum likelihood estimate of a mixing distribution. *J. Royal Stat. Soc. Ser. B (Statistical Methodol.* **69**, 185–198 (2007).

[11] Ferguson, T. S. A Bayesian analysis of some nonparametric problems. *The Annals Stat.* 209–230 (1973).

[12] Liu, J. S. Nonparametric hierarchical Bayes via sequential imputations. *The Annals Stat.* 911–930 (1996).

[13] Gelman, A., Jakulin, A., Pittau, M. G. & Su, Y.-S. A weakly informative default prior distribution for logistic and other regression models. *The Annals Appl. Stat.* 1360–1383 (2008).

[14] Gelman, A. *et al. Bayesian Data Analysis, Third Edition.* Chapman & Hall/CRC Texts in Statistical Science (Taylor & Francis, 2013).

[15] Berger, J. O. Bayesian analysis: A look at today and thoughts of tomorrow. *J. Am. Stat. Assoc.* **95**, 1269–1276 (2000).

[16] Mukhopadhyay, S. & Fletcher, D. *BayesGOF: Bayesian Modeling via Goodness of Fit* (2018). R package version 3.1.

[17] Beckett, L. & Diaconis, P. Spectral analysis for discrete longitudinal data. *Adv. Math.* **103**, 107–128 (1994).

[18] Young-Xu, Y. & Chan, K. A. Pooling overdispersed binomial data to estimate event rate. *BMC Med. Res. Methodol.* **8**, 58 (2008).

[19] Martz, H. & Lian, M. Empirical Bayes estimation of the binomial parameter. *Biom.* **61**, 517–523 (1974).

[20] De Blok, W., McGaugh, S. S. & Rubin, V. C. High-resolution rotation curves of low surface brightness galaxies II. Mass models. *The Astron. J.* **122**, 2396 (2001).

[21] Sacks, H. S., Chalmers, T. C., Blum, A. L., Berrier, J. & Pagano, D. Endoscopic hemostasis: an effective therapy for bleeding peptic ulcers. *J. Am. Med. Assoc.* **264**, 494–499 (1990).

[22] Willie, S. & Berman, S. Ninth round intercomparison for trace metals in marine sediments and biological tissues. *NRC/NOAA* (1995).

[23] Efron, B. & Hastie, T. *Computer Age Statistical Inference*, vol. 5 (Cambridge University Press, 2016).

[24] Fisher, R. A., Corbet, A. S. & Williams, C. B. The relation between the number of species and the number of individuals in a random sample of an animal population. *The J. Animal Ecol.* 42–58 (1943).

[25] Norberg, R. Experience rating in group life insurance. *Scand. Actuar. J.* **1989**, 194–224 (1989).

[26] Robbins, H. An empirical Bayes estimation problem. *Proc. Natl. Acad. Sci.* **77**, 6988–6989 (1980).

[27] Robbins, H. Asymptotically subminimax solutions of compound statistical decision problems. In *Proceedings of the Second Berkley Symposium on Mathematical Statistics and Probability*, vol. I, 131–149 (Berkeley: University of California Press, 1951).